



Lothar Schmidt-Atzert · Stefan Krumm
Manfred Amelang *Hrsg.*

Psychologische Diagnostik

6. Auflage

Psychologische Diagnostik

Lothar Schmidt-Atzert · Stefan Krumm · Manfred Amelang
(Hrsg.)

Psychologische Diagnostik

Mitbegründet von Prof. Dr. Werner Zielinski

6., vollständig überarbeitete Auflage

Hrsg.

Lothar Schmidt-Atzert
Fachbereich Psychologie
Philipps-Universität Marburg
Marburg, Hessen, Deutschland

Manfred Amelang
Psychologisches Institut
Universität Heidelberg
Heidelberg, Baden-Württemberg
Deutschland

Stefan Krumm
Arbeitsbereich Psychologische
Diagnostik, Differentielle und
Persönlichkeitspsychologie
Freie Universität Berlin
Berlin, Deutschland

Zusätzliches Material zu diesem Buch finden Sie auf <http://www.lehrbuch-psychologie.springer.com>.

ISBN 978-3-662-61642-0 ISBN 978-3-662-61643-7 (eBook)
<https://doi.org/10.1007/978-3-662-61643-7>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über ► <http://dnb.d-nb.de> abrufbar.

© Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 1994, 1997, 2002, 2006,
2012, 2021

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag, noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Einbandabbildung: © LIGHTFIELD STUDIOS/stock.adobe.com

Planung/Lektorat: Joachim Coch, Judith Danziger, Stefanie Teichert
Springer ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort zur 6. Auflage

Zwischen der Fertigstellung des Manuskripts der 6. Auflage und dem der 5. Auflage liegen fast 10 Jahre. Das ist eine lange Zeit, in der sich im Fach Psychologische Diagnostik viel getan hat. Neue Tests kamen auf den Markt, einige bewährte wurden aktualisiert. Die Digitalisierung und mit ihr der Zugriff auf sehr große Datenmengen, ausgefieilte Algorithmen bis hin zu künstlicher Intelligenz haben Einzug in die Diagnostik gehalten. In einigen Bereichen, in denen Standards relevant sind, gab es Neuerungen: In der Klinischen Psychologie sind die für die Diagnostik maßgeblichen Diagnosesysteme DSM und ICD durch neue Versionen abgelöst worden. Für die Eignungsdiagnostik ist es relevant, dass die DIN 33430 im Jahr 2016 in revidierter Form erschienen ist. Das Pendant zur DIN 33430 sind in der Verkehrspychologie die „Begutachtungsleitlinien zur Kraftfahreignung“, zu der 2018 neue Kommentare erschienen sind. Qualitätsprobleme bei familienrechtspychologischen Gutachten waren viele Jahre lang in Fachkreisen ein „heiße“ Thema und wurden auch in einer breiten Öffentlichkeit diskutiert. Eine Arbeitsgruppe, in der nicht nur Psychologenverbände, sondern u. a. auch Psychiater- und Juristenverbände mitwirkten, legte 2015 die „Mindestanforderungen an die Qualität von Sachverständigengutachten im Kindschaftsrecht“ vor (2019 folgte bereits eine überarbeitete Version). Zur Diagnostik (und Behandlung) diverser psychischer Störungen wurden unter Federführung der „Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften“ mehrere evidenz- und konsensbasierte Leitlinien vorgelegt, an denen auch psychologische Fachverbände mitgewirkt haben. Im ► Kap. 7 zur Diagnostik in der Pädagogischen Psychologie stellen wir die 2018 veröffentlichte Leitlinie „Diagnostik und Behandlung der Rechenstörung“ vor. Der Föderation Deutscher Psychologenvereinigungen gelang es 2017 nach langer Vorarbeit die neuen „Qualitätsstandards für psychologische Gutachten“ zu verabschieden. Die genannten Initiativen dokumentieren starke Bestrebungen, die Qualität psychologischer Diagnostik zu sichern und zu verbessern.

Hier soll jedoch nicht der Eindruck entstehen, die psychologische Diagnostik agiere nur auf der Basis von Leitlinien. Leitlinien und Qualitätsstandards sind wichtig, aber viele für die psychologische Diagnostik relevante Fragen können nur adäquat beantwortet werden, wenn einschlägige Forschungsergebnisse mit grundlegendem psychologisch-diagnostischem Know-how kombiniert und gemeinsam angewendet werden. Das vorliegende Buch soll dazu einen Beitrag leisten.

Wie bereits bei der 5. Auflage standen wir vor der Herausforderung, Bewährtes beizubehalten, aber auch Platz für Neues zu schaffen. Konkret stellt sich in der Zusammenarbeit der Autoren die manchmal schwer beantwortbare Frage: „Was ist bewährt, heute noch gültig und relevant oder gar zeitlos – und wie viel Neues verträgt ein Buch, ohne die in 5 Auflagen gewachsene Tradition zu verletzen?“ Das ohnehin schon umfangreiche Werk immer dicker werden zu lassen, war nicht unser Ziel (dennoch waren es am Ende mehr Seiten als geplant). Wir hoffen, dass es uns letztendlich gelungen ist, eine optimale Balance zwischen Breite und Tiefe hergestellt sowie Tradition und Aktualität auf eine Weise verknüpft zu haben, die bei den Leserinnen und Lesern Gefallen findet. Nicht alles ist neu. In allen Kapiteln wurden auch bewährte Formulierungen und Textpassagen aus der vorherigen Auflage übernommen, auch wenn ein neuer Erstautor hinzugekommen ist. Zu den Änderungen gehört auch eine gendergerechte Schreibweise. Wir haben uns daher auf geschlechtsneutrale Formulierungen wie „Testpersonen“ sowie Geschlechtsbezeichnung wie „Psychologinnen und Psychologen“ verständigt. Ähnlich wie andere sprachliche Formen (bspw. Gendersternchen, Binnen-I) sollen die von

uns gewählten Bezeichnungen auch Personen diversen Geschlechts inkludieren. Wir möchten diese Menschen bitten, sich auch mit der von uns gewählten Schreibweise angesprochen zu fühlen.

■ Zielgruppe und Verwendung des Lehrbuchs im Studium

Das Fach Psychologische Diagnostik ist im Psychologiestudium fest verankert. Auch mit der Novellierung des Psychotherapiegesetzes und der Einführung des „Direktstudiums Psychotherapie“ gehört die Diagnostik weiter zu den wichtigen Grundlagen unseres Fachs, die vornehmlich im Bachelorstudium vermittelt werden. Die Anwendungsfächer profitieren weiter von diesen Grundlagen. Die Anwendungsfächer, also neben der Psychotherapieausbildung insbesondere die Pädagogische Psychologie, die Arbeits-, Organisations- und Wirtschaftspsychologie sowie die Neuro-, Rechts- und Verkehrspychologie, finden auch in der 6. Auflage dieses Buchs wieder ihren Raum. Wir meinen, dass Grundlagen alleine wenig nützen, wenn nicht auch erkennbar ist, wofür sie gut sind. Deshalb sind wir weiter dem Konzept treu geblieben, Grundlagen (also insbesondere Testtheorie und diagnostische Verfahren) und Anwendungen angemessen zu berücksichtigen. Darüber hinaus dürfen zum mindest ausgewählte Kapitel für Studierende anderer Fächer von Interesse sein. Und wir sind überzeugt, dass das vorliegende Lehrbuch sich auch weiter in der Praxis als nützlich erweisen wird.

■ Dank für vielfältige Unterstützung

Ein wichtiges Anliegen war uns eine weiterhin hohe Benutzerfreundlichkeit. Wir haben uns um eine klare und verständliche Sprache bemüht. Unsere studentischen Hilfskräfte übernahmen nicht nur Recherchen, sondern haben auch große Teile des Manuskripts mit kritischem Blick gelesen und uns Hinweise zur Optimierung der Verständlichkeit gegeben. Unser Dank gilt Mareike Breda, Alexandra Göbel, Luca Kröger, Juliane Maurer und Laura Restrepo. Ein besonderer Dank gebührt Nico Remmert und Nomi Reznik, die sich mit viel Mühe und Geduld u. a. der Gestaltung von Grafiken sowie der Pflege des Literaturverzeichnisses gewidmet und wertvolles Feedback zu einigen Kapiteln des Buchs gegeben haben. Darüber hinaus danken wir allen Kolleginnen und Kollegen herzlich, die mit wertvollen Hinweisen und im ständigen fachlichen Austausch zum Gelingen dieses Buchs beigetragen haben: Prof. Dr. Michael Eid, Dr. Karin Funsch, Dr. Jan-Philipp Freudenstein, Katharina Schmidt und Dr. Julian Schulze.

Über Rückmeldungen zur 5. Auflage haben wir uns gefreut. Neben Lob und Anerkennung erhielten wir von aufmerksamen Leserinnen und Lesern auch Hinweise auf Verbesserungsmöglichkeiten und (meist kleine) Fehler. Wir danken Ersin Cetin, Matthias Fligge, Gunnar Jürgens, Prof. Dr. Martin Kersting, Juliane Müller, Melanie Nath, Dr. Hansjörg Plieninger, Ingo Seifert, Manuel Siegert, Sandra Weber und Prof. Dr. Werner Wittmann für ihre Hinweise auf Fehler. Wir sind auch weiter dankbar für alle Hinweise auf „neue“ Fehler, die uns bei der Überarbeitung möglicherweise unterlaufen sein könnten.

Herrn Joachim Coch (Senior Editor beim Springer-Verlag) und Frau Judith Danziger (Project Manager beim Springer-Verlag) danken wir für ihre große Geduld bei der sich immer wieder hinziehenden Fertigstellung des Manuskripts und für die sehr gute Zusammenarbeit in allen Fragen. Bei Frau Stefanie Teichert (Lektorat) möchten wir uns für die sehr gute und effiziente Zusammenarbeit bedanken.

Lothar Schmidt-Atzert

Würzburg

Stefan Krumm

Berlin

Manfred Amelang

Heidelberg

im Juni 2020

Vorwort zur 1. Auflage

Die neue Rahmenprüfungsordnung für das Fach Psychologie sowie die daran ansetzenden hochschulspezifischen Prüfungsordnungen und Studienpläne sehen eine Verklammerung von Psychologischer Diagnostik und Intervention vor. Damit soll deutlich gemacht werden, daß sich Psychologische Diagnostik nicht in der Beschreibung bestimmter Gegebenheiten erschöpfen darf, sondern stets im Hinblick auf konkrete Fragestellungen erfolgt und deshalb starke Handlungs- oder Interventionsimplikationen aufweist.

Für diese Verknüpfung von Psychologischer Diagnostik und Intervention fehlt es unseres Erachtens an kompakten Darstellungen – ungeachtet der zahlreichen und z. T. qualitativ exzeptionellen Behandlungen von jedem einzelnen der beiden Teilgebiete in der Literatur.

Der hiermit vorgelegte Text richtet sich ausdrücklich und primär an Studierende des Faches Psychologie. Unsere Konzeption ging dahin, den Umfang auf das für ein Prüfungsfach Zentrale und wirklich unabdingbar Notwendige zu beschränken.

Inhaltlich sollte der Stoff eine nach Möglichkeit optimale Mischung aus methodischen Prinzipien, instrumentellen Fakten und Informationen über Anwendungen bzw. Interventionsbereiche darstellen. Die Menge des Stoffes sollte so bemessen sein, daß sie im Zuge der Vorbereitung auf eine Prüfung auch wirklich bewältigt und die Materie angemessen verarbeitet werden kann.

Das bedeutete in didaktischer Hinsicht unter anderem, daß die Darstellung nicht durch mögliche „Ziselierungen“, also Quer- und Tiefenverweise sowie Belege jeder einzelnen Feststellung mit Zitaten anderer Autoren usw., belastet werden durfte. Solche Zusatzinformationen sind zwar für wissenschaftliches Arbeiten unerlässlich, würden jedoch den eher linearen Duktus eines Lehrbuches etwas beeinträchtigen und damit die Lektüre erschweren.

Um die Rezeption weiter zu erleichtern, haben wir Merksätze, Randbemerkungen und Übungsfragen vorgesehen.

Obwohl die Planungen für das Buchprojekt längere Zeit zurückreichen, erfolgte seine Realisierung dann doch für einen von uns (M. A.) unter erheblichem Zeitdruck und erschwert durch den Umstand, simultan auch anderweitigen Dienstverpflichtungen entsprechen zu müssen.

Um so dankbarer sind wir deshalb für die tatkräftige und umsichtige Unterstützung, die wir von selten unserer Mitarbeiterinnen und Mitarbeiter sowie Hilfskräfte, insbesondere in der Endphase der Fertigstellung, auf ganz verschiedene Weise erfahren haben: Karin Holthausen und – mehr noch – Dorothea Benz besorgten die Schreibarbeiten, Heiner Rindermann, Jörg Müller und Nicole Petrow setzten die Formeln, Abbildungen und Tabellen, Claudia Schmidt-Rathjens und Jochen Czemmel arbeiteten die Rechenbeispiele für die Gütekriterien aus, Margarete Edelmann und Gerhard Rothmann halfen mit Literaturexzerpten und Übersichten aus dem ABO-Bereich, Viktor Oubaid erstellte einen großen Teil der Randbemerkungen und Übungsfragen, Sabine Pöhltz war für das Literaturverzeichnis und dessen Kongruenz zum laufenden Text verantwortlich, Claudia Müller für die Grundstruktur des Stichwortverzeichnisses – und die Koordination für all das sowie dessen Endredaktion lag in den Händen von Claudia Krüger.

Ihnen allen danken wir auch an dieser Stelle ganz herzlich und fügen hinzu, daß für verbleibende Unzulänglichkeiten selbstverständlich wir allein die Verantwortung tragen.

Ohne das nachhaltige Interesse des Verlages in Gestalt von Heike Berger und den von ihr ausgehenden Anregungen und zeitlichen Vorstellungen wäre das Projekt weder in der nun vorliegenden Form noch zum jetzigen Zeitpunkt erschienen. Auch ihr danken wir sowie der Lektorin Dr. Regine Körkel-Hinkfoth für ihre vorzügliche Korrekturarbeit.

Manfred Amelang

Werner Zielinski

Heidelberg

im September 1994

Inhaltsverzeichnis

1	Einleitung	1
	<i>Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang</i>	
1.1	Definition	2
1.2	Anwendungsgebiete und Fragestellungen	4
1.2.1	Anwendungsgebiete	4
1.2.2	Beispiele für diagnostische Fragestellungen	7
1.3	Verhältnis zu anderen Disziplinen der Psychologie	10
1.4	Ziele der Psychologischen Diagnostik	11
1.5	Der diagnostische Prozess	15
1.6	Meilensteine in der Geschichte der Psychologischen Diagnostik	22
1.7	Gesetzliche Rahmenbedingungen und ethische Richtlinien	27
1.7.1	Menschenwürde und Privatsphäre	28
1.7.2	Geheimnisse, Schweigepflicht und Datenschutz	29
1.7.3	Offenbarungspflicht	31
1.7.4	Rechtliche Regelungen für spezifische Anwendungsfelder der Psychologischen Diagnostik	32
1.7.5	Ethische Richtlinien	33
1.8	Zusammenfassung	36
	Literatur	37
2	Grundlagen diagnostischer Verfahren	39
	<i>Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang</i>	
2.1	Allgemeines zu psychologischen Tests	41
2.1.1	Was versteht man unter einem psychologischen Test?	41
2.1.2	Bandbreite psychologischer Tests	43
2.1.3	Rückschluss auf ein latentes Merkmal	46
2.2	Die Klassische Testtheorie	49
2.2.1	Zentrale Annahmen	49
2.2.2	Reliabilität von Messungen	52
2.2.3	Grenzen der Klassischen Testtheorie	54
2.3	Item-Response-Theorien	55
2.3.1	Item-Response-Theorien für dichotome Antwortformate	57
2.3.2	Item-Response-Theorien für ordinale Antwortformate	76
2.3.3	Item-Response-Theorien zur Klassifikation von Personen	80
2.4	Konstruktionsprinzipien psychologischer Tests	84
2.4.1	Ziel der Messung und Messgegenstand	84
2.4.2	Generieren von Testitems	87
2.5	Grundzüge von Itemanalysen	109
2.5.1	Itemschwierigkeit (nach der Klassischen Testtheorie)	109
2.5.2	Itemstreuung bzw. Itemvarianz	112
2.5.3	Trennschärfe	113
2.5.4	Itemladungen auf Faktoren	118
2.5.5	Itemvalidität	126
2.5.6	Itemanalysen nach Probabilistischen Testtheorien	126
2.6	Testgütekriterien	132
2.6.1	Objektivität	133
2.6.2	Reliabilität	138
2.6.3	Validität	157
2.6.4	Nebengütekriterium: Normierung	182
2.6.5	Weitere Nebengütekriterien	190
2.7	Zusammenfassung	198
	Literatur	201

3	Diagnostische Verfahren	209
	<i>Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang</i>	
3.1	Einleitung	211
3.2	Leistungstests	216
3.2.1	Allgemeines zu Leistungstests.....	216
3.2.2	Aufmerksamkeits- und Konzentrationstests	225
3.2.3	Intelligenztests	253
3.2.4	Spezielle Fähigkeitstests	284
3.2.5	Entwicklungstests	288
3.2.6	Schultests	301
3.3	Persönlichkeitsfragebögen	303
3.3.1	Persönlichkeitsmerkmale und ihre Messung.....	303
3.3.2	Allgemeine Vor- und Nachteile von Persönlichkeitsfragebögen.....	312
3.3.3	Persönlichkeitstestsysteme.....	322
3.3.4	Verfahren zur Erfassung aktueller Zustände	361
3.4	Objektive Persönlichkeitstests	367
3.4.1	Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität	369
3.4.2	Objektiver Leistungsmotivations-Test (OLMT)	371
3.4.3	Implizite Assoziationstests (IAT)	375
3.4.4	Weitere Forschung zu objektiven Persönlichkeitstests und impliziten Assoziationstests	376
3.4.5	Weitere digitale Ansätze – Machine Learning und künstliche Intelligenz	377
3.4.6	Objektive sprachbasierte Eignungsdiagnostik	381
3.5	Projektive Verfahren	382
3.5.1	Klassische projektive Tests	384
3.5.2	Abgeleitete Testprinzipien und semiprojektive Tests.....	393
3.5.3	Zeichnerische und Gestaltungsverfahren.....	398
3.6	Verhaltensbeobachtung und -beurteilung	400
3.6.1	Arten der Verhaltensbeobachtung.....	401
3.6.2	Systematische Verhaltensbeobachtung	407
3.6.3	Verhaltensbeurteilung	415
3.6.4	Gütekriterien von Beobachtungs- und Beurteilungsverfahren	419
3.7	Diagnostisches Interview	426
3.7.1	Standardisierte strukturierte Interviews	432
3.7.2	Interviews selbst konstruieren	443
3.7.3	Techniken der Gesprächsführung	451
3.8	Zusammenfassung	460
	Literatur	461
4	Durchführung einer diagnostischen Untersuchung und Gutachtenerstellung	477
	<i>Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang</i>	
4.1	Persönliche Voraussetzungen und ethisch verantwortliches Vorgehen	481
4.2	Auftragsannahme und Fragestellung	483
4.3	Ableitung von psychologischen Fragen	486
4.3.1	Psychologische Fragen finden	487
4.3.2	Darstellung der psychologischen Fragen.....	489
4.4	Auswahl der Verfahren und Untersuchungsplanung	489
4.4.1	Geeignete diagnostische Verfahren finden	490
4.4.2	Untersuchungsplanung	494
4.5	Durchführung und Auswertung diagnostischer Verfahren	497
4.5.1	Aufklärung	497
4.5.2	Gute Arbeitsbedingungen.....	502
4.5.3	Standardisierung der Untersuchungsbedingung	503
4.5.4	Testauswertung	504

Inhaltsverzeichnis

4.5.5	Darstellung und Interpretation der Ergebnisse	505
4.6	Das psychologische Gutachten	512
4.6.1	Der Befund	512
4.6.2	Stellungnahme	514
4.6.3	Wenn der Begutachtungsprozess nicht erfolgreich verläuft	515
4.6.4	Formale Gestaltung des Gutachtens	516
4.6.5	Beurteilung der Qualität eines Gutachtens	520
4.7	Zusammenfassung	522
	Literatur	524
5	Diagnostische Strategien und Evaluation des Vorgehens	527
	<i>Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang</i>	
5.1	Diagnostische Strategien	528
5.1.1	Status- vs. Veränderungsdiagnostik.....	528
5.1.2	Selektion vs. Modifikation	530
5.1.3	Strategien der Integration von Daten zu einer diagnostischen Entscheidung.....	532
5.1.4	Einstufige vs. mehrstufige diagnostische Entscheidungen.....	553
5.2	Evaluation des Vorgehens	556
5.2.1	Prozessevaluation der Psychologischen Diagnostik.....	556
5.2.2	Ergebnisevaluation der Psychologischen Diagnostik.....	558
5.2.3	Schätzung des Nutzens Psychologischer Diagnostik	559
5.3	Zusammenfassung	563
	Literatur	564
6	Diagnostik in der Arbeits-, Organisations- und Wirtschaftspsychologie	567
	<i>Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang</i>	
6.1	Organisationsdiagnostik	570
6.1.1	Fragebögen zur Beschreibung der Arbeit und des Klimas in Teams bzw. Organisationen	571
6.1.2	Arbeits- und Anforderungsanalyse.....	573
6.2	Diagnostik von Personenmerkmalen	582
6.2.1	Selektion von Personen: Personalauswahl.....	584
6.2.2	Selektion von Bedingungen: Berufs- und Ausbildungswahl.....	609
6.2.3	Modifikation von Personen: Personalentwicklung	613
6.2.4	Modifikation von Bedingungen.....	617
6.3	Evaluation der Psychologischen Diagnostik in der Arbeits-, Organisations- und Wirtschaftspsychologie	618
6.3.1	Evaluation von Arbeits- und Anforderungsanalysen	619
6.3.2	Evaluation von Personalauswahlverfahren	621
6.3.3	Evaluation von diagnostischen Verfahren zur Berufs- und Ausbildungswahl	624
6.3.4	Evaluation von diagnostischen Verfahren zur Feststellung des Personalentwicklungsbedarfs	625
6.3.5	Evaluation der Diagnostik von individuellen Merkmalen zum Zwecke der Modifikation von Arbeitskontexten/Arbeitsgestaltung	626
6.3.6	Messung der Auswirkung von „Passung“	627
6.4	Ein Qualitätsstandard für berufsbezogene Eignungsbeurteilungen – die DIN 33430	632
6.5	Zusammenfassung	636
	Literatur	637

7	Diagnostik in der Pädagogischen Psychologie	643
	<i>Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang</i>	
7.1	Diagnostik zur Schullaufbahnberatung	644
7.1.1	Schuleingangsdagnostik	645
7.1.2	Diagnostik zur Feststellung von sonderpädagogischem Förderbedarf	648
7.1.3	Diagnostik beim Übertritt in den tertiären Bildungsbereich	654
7.2	Diagnostik bei Schulschwierigkeiten.....	660
7.2.1	Diagnostik bei Lernschwierigkeiten.....	660
7.2.2	Diagnostik von Teilleistungsstörungen.....	661
7.3	Hochbegabungsdagnostik	667
7.4	Tests im Bildungsbereich.....	673
7.4.1	Schultests	673
7.4.2	Tests zur Evaluierung des Bildungssystems.....	677
7.5	Zusammenfassung	682
	Literatur	684
8	Diagnostik in der Klinischen Psychologie und Psychotherapie ...	689
	<i>Thomas Fydrich</i>	
8.1	Aufgaben der klinisch-psychologischen Diagnostik	690
8.1.1	Rahmenbedingungen für klinisch-psychologische Diagnostik und Intervention.....	692
8.1.2	Das diagnostische Interview.....	694
8.2	Problem-, Verhaltens- und Plananalyse als Ansatz der kognitiv-verhaltenstherapeutischen Diagnostik	696
8.3	Psychische Störungen und ihre Klassifikation.....	700
8.3.1	Klassifikation psychischer Störungen.....	701
8.3.2	Diagnostische Verfahren zur Klassifikation psychischer Störungen.....	706
8.4	Psychometrische Verfahren	708
8.4.1	Verhaltenstheoretisch und kognitiv orientierte Fragebogenverfahren	708
8.4.2	Beobachtungsmethoden.....	712
8.4.3	Persönlichkeitstests in der Klinischen Psychologie und Psychotherapie	713
8.4.4	Verfahren und Ansätze auf klientenzentrierter, psychodynamischer, systemischer und interpersoneller Grundlage	714
8.5	Verbindung von Diagnostik und Intervention: Die Indikation	717
8.6	Erfolgskontrolle als Teil der Qualitätssicherung.....	719
8.6.1	Zieldefinition, Therapieverlaufs- und Veränderungsdiagnostik	720
8.6.2	Kriterium der klinisch bedeutsamen Verbesserung.....	721
8.7	Zusammenfassung	724
	Literatur	726
9	Diagnostik in weiteren Anwendungsfeldern	731
	<i>Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang</i>	
9.1	Neuropsychologische Diagnostik	732
9.1.1	Neuropsychologische Untersuchung	735
9.1.2	Spezielle Probleme der neuropsychologischen Diagnostik	738
9.2	Rechtspsychologische Diagnostik.....	746
9.2.1	Glaubhaftigkeit von Zeuginnen und Zeugen	749
9.2.2	Schuldunfähigkeit und verminderte Schuldfähigkeit.....	750
9.2.3	Kriminalprognose.....	751
9.2.4	Familiengericht: Sorgerechtsentscheidungen	758

Inhaltsverzeichnis

9.3	Verkehrpsychologische Diagnostik	762
9.3.1	Begutachtung der Fahreignung für den Straßenverkehr	763
9.3.2	Spezielle Probleme der verkehrpsychologischen Diagnostik.....	771
9.4	Zusammenfassung	774
	Literatur	777
	Serviceteil	
	Stichwortverzeichnis.....	785

Schmidt-Atzert, Krumm, Amelang: Psychologische Diagnostik, 5. Auflage

Der Wegweiser zu diesem Lehrbuch

The diagram illustrates the structure of the book. A central vertical column contains the main text and sections. Marginal notes are shown in boxes along the left edge, connected by lines to specific parts of the text. At the bottom, there is a summary box.

Leitsystem: zur schnellen Orientierung

Marginalien: Stichworte für die Orientierung

Merksätze: besonders wichtig

Beispiel: So wird das Gelernte anschaulich

Übersichten: z.B. Empfehlungen und Formelableitungen

Zentrale Annahme der Klassischen Testtheorie

2

40 S. Krumm et al.

2.1 Allgemeines zu psychologischen Tests

Dieses Kapitel befasst sich mit grundlegenden Anforderungen an diagnostische Instrumente. Der Schwerpunkt liegt dabei auf psychologischen Tests. Diese stellen zwar nur eine von mehreren Möglichkeiten dar, diagnostisch relevante Informationen über Personen zu erheben. Im Vergleich zu anderen Möglichkeiten, beispielsweise Interviews oder Verhaltensbeobachtungen, sind die grundlegenden Anforderungen an psychologische Tests deutlich genauer spezifiziert und formalisiert. Diese Anforderungen können jedoch – mit wenigen Einschränkungen – auf diagnostische Interviews und Verhaltensbeobachtungen übertragen werden.

! Die zentrale Annahme der Klassischen Testtheorie ist, dass Messungen fehlerbehaftet sind. Sie nimmt an, dass eine einzelne Messung aufgrund von unsystematischen Einflussfaktoren ein höheres oder niedrigeres Ergebnis liefert als aufgrund der tatsächlichen Merkmalsausprägung zu erwarten wäre.

► Beispiel

Die Fragestellung in einem Gutachten lautet, ob Marco S. (Schüler, 8 Jahre) geistig behindert ist. Daraus wurde u. a. die psychologische Frage abgeleitet, ob Marcos IQ unter 70 liegt (eine etablierte Definition von geistiger Behinderung). In dem eingesetzten Intelligenztest erreichte Marco einen IQ von 64. Dieser wurde im Ergebnisbericht so interpretiert: „Der IQ von 64 spricht dafür, dass Marco im Vergleich zu gleichaltrigen Kindern eine sehr niedrige Intelligenz hat. Unter Berücksichtigung der Messgenauigkeit des Tests ($\alpha=.97, p=.05$, einseitige Fragestellung) kann sein IQ auch maximal 69 betragen, womit er weiter im sehr niedrigen Bereich liegt.“

Im Befund wird dieses Ergebnis unter Berücksichtigung weiterer Ergebnisse dahingehend interpretiert, dass mit dem Testergebnis seine Intelligenz unterschätzt wird. Das Ergebnis spricht zwar für eine sehr niedrige Intelligenz mit einem IQ unter 70, ist aber nicht als Beleg für eine geistige Behinderung zu werten. Dafür sprechen die Verhaltensbeobachtung („wirkte unmotiviert“) und das diagnostische Interview („ühlte sich mit dem Test sehr an schulische Aufgaben erinnert und hatte deshalb ‚keine Lust‘; „fassste Fragen gut auf und beantwortete sie für sein Alter angemessen differenziert“). ◀

Einteilungen der Testverfahren

- Nach Messanspruch (z. B. Persönlichkeitsfragebögen, Intelligenztests)
- In Leistungs- vs. Persönlichkeitstests bzw. Persönlichkeitsfragebögen (► Kap. 3)
- Nach psychologischer Disziplin (z. B. neuropsychologische Tests)
- Nach Zielgruppe (z. B. Tests für Kinder)
- Nach Administrationsform (z. B. Onlinetest, Paper-Pencil-Test)

Qualitätsstandards für psychologische Gutachten

Die „Qualitätsstandards für psychologische Gutachten“ (hier kurz: Qualitätsstandards) wurden am 18. Oktober 2017 vom Vorstand der Föderation Deutscher Psychologenvereinigungen verabschiedet. Sie ersetzen die „Richtlinien für die Erstellung psychologischer Gutachten“ der Föderation Deutscher Psychologenvereinigungen von 1988. Sie sind in einem langen Prozess entstanden.

Die Deutsche Gesellschaft für Psychologie (DGPs) beauftragte im August 2009 eine Arbeitsgruppe, Qualitätsstandards für psychodiagnostische Gutachten auszuarbeiten und deren Konsequenzen für die Lehre an den Hochschulen

Aufgabenbeispiele, vertiefende und wichtige Studien, Hintergrundinformationen

Definition

Psychologische Fragen sind ein wesentliches Element des Begutachtungsprozesses und auch des schriftlichen Gutachtens. Sie sind Hypothesen oder selbst gesetzte Arbeitsaufträge, die zur Beantwortung der globalen Fragestellung benötigt werden.

Definition:
erläutert wichtige
Fachbegriffe

Um bei der Menge an verfügbaren Testverfahren den Überblick zu behalten, ist eine Systematik der Tests hilfreich. Das wichtigste Kriterium für eine Einteilung von Tests ist der Messgegenstand (welches Merkmal soll erfasst werden?). Aber auch andere Unterteilungen sind geläufig.

Fazit Wie man sieht, sind die Erläuterungen zu Probabilistischen Testtheorien deutlich umfangreicher als die zur Klassischen Testtheorie. Das liegt daran, dass Probabilistische Testtheorien sehr viel elaboriertere Annahmen darüber machen, wie Testantworten und -ergebnisse in Abhängigkeit von der infrage stehenden Merkmalsausprägung zustandekommen. Hingegen begnügt sich die Klassische Testtheorie mit wenigen Grundannahmen (Axiomen) und daraus folgenden Ableitungen. Dies scheint jedoch gleichzeitig ihr „Erfolgsprinzip“ zu sein: Sie ist immer noch die weitaus häufiger genutzte Basis für die Testkonstruktion. Aus diesem Grund werden in ▶ Abschn. 2.5 und 2.6.2.3 die auf der Klassischen Testtheorie beruhenden Analysen (zur Konstruktion und Evaluation von Tests) ausführlicher besprochen als analoge, auf Probabilistischen Testtheorien fußende Analysen.

Fazit:
Rekapitulieren Sie
das Gelernte!

Freie Literatursuche

Ein freier Zugriff auf Testinformationen ist über die Suchmaschine PubPsych möglich, die man über das ZPID erreicht (▶ <https://pubpsych.zpid.de/pubpsych/>). Auf ein Stichwort hin findet man aber nicht nur Tests, sondern auch viel Fachliteratur zu dem Thema. Die Suche kann auf Tests eingrenzt werden, indem bei „Publikationstyp“ die Kategorie „Tests/Questionnaires“ angeklickt wird. Bei den aufgeführten Verfahren kommt man durch Anklicken von „PSYNDEX Tests Zusatzinformationen“ (soweit vorhanden) zu den oben beschriebenen Testinformationen.

**Weiterführende
Literatur und
Internetressourcen:**
Tipps für die weitere
Lektüre

Übungsfragen

- ▶ Abschn. 4.1–4.5:
 - Nennen Sie die 5 zentralen Schritte des diagnostischen Prozesses!
 - Nennen Sie 2 zentrale Anforderungen der International Test Commission (ITC) bezüglich der Darstellung und Interpretation der Ergebnisse!

Weiterführende Literatur und Internetressourcen

Zur weiteren Vertiefung in das Thema „Rechtsfragen psychologischer Diagnostik“ eignen sich besonders die Bücher von Joussen (2004) und Zier (2002).

Hilfreiche Internetressourcen:

- Zur Relevanz von Psychologischer Diagnostik in psychologischen Berufsbildern:
 - ▶ <https://www.onetonline.org/>
- Leitlinien zur verkehrspychologischen Diagnostik: ▶ <https://www.bast.de/>
- Kompetenzen zur Testanwendung in den Richtlinien der International Test Commission: ▶ https://www.intestcom.org/files/guideline_test_use.pdf sowie ▶ https://www.psyndex.de/pub/tests/itc_richtlinien.pdf (deutsche Fassung)
- Ethische Richtlinien der APA und der DGPs: ▶ <https://www.apa.org/ethics/code/index> und ▶ <https://www.dgps.de/>

Übungsfragen:
Fit für die Prüfung!

Lernmaterialien zum Lehrbuch *Psychologische Diagnostik* im Internet – ► www.lehrbuch-psychologie.springer.com



Lothar Schmidt-Atzert · Stefan Krumm
Manfred Amelang Hrsg.

Psychologische Diagnostik

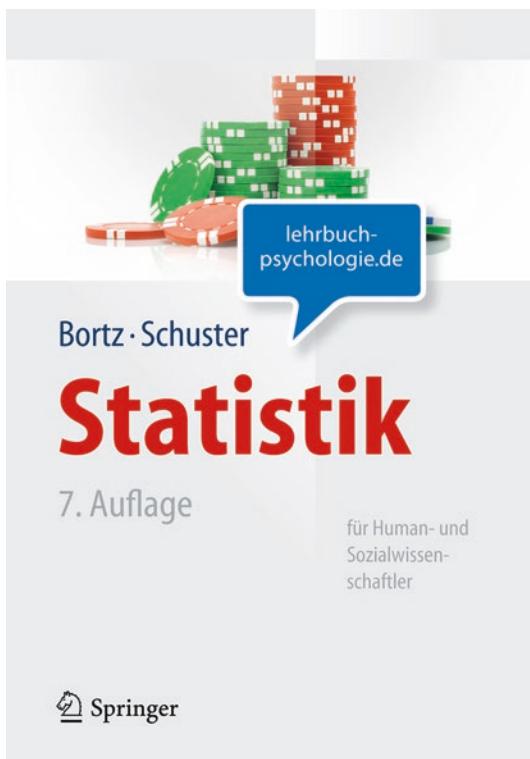
6. Auflage

MOREMEDIA

Springer

- Kapitelzusammenfassungen: Das steckt drin im Lehrbuch
- Verständnisfragen: Üben Sie für die Prüfung
- Karteikarten: Überprüfen Sie Ihr Wissen
- Glossar mit zahlreichen Fachbegriffen
- Umfangreiche kommentierte Linkssammlung
- Foliensätze, Abbildungen und Tabellen für Dozentinnen und Dozenten zum Download

Weitere Websites unter ► www.lehrbuch-psychologie.springer.com



lehrbuch-psychologie.de

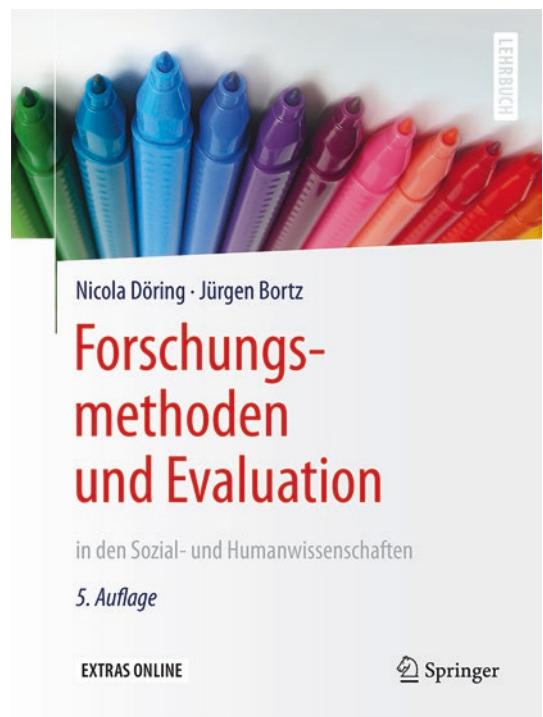
Bortz · Schuster

Statistik

7. Auflage

für Human- und Sozialwissenschaftler

Springer



LEHRBUCH

Nicola Döring · Jürgen Bortz

Forschungs- methoden und Evaluation

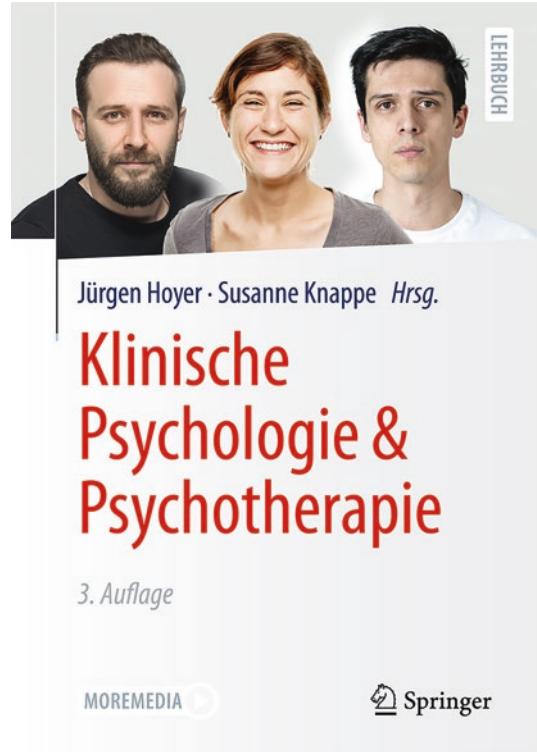
in den Sozial- und Humanwissenschaften

5. Auflage

EXTRAS ONLINE

Springer

- Rechnen mit SPSS und R: Syntax-Kommentare zur Berechnung der Software
- Glossar der wichtigsten Fachbegriffe
- Zusammenfassungen der 28 Buchkapitel
- Karteikarten: Überprüfen Sie Ihr Wissen
- Vorlesungsfolien, Tabellen und Abbildungen für Dozentinnen und Dozenten zum Download
- Verständnisfragen und Antworten
- Glossar mit zahlreichen Fachbegriffen und englischer Übersetzung
- Karteikarten: Fachbegriffe pauken
- Lösungen zu den Lernquiz des Buchs zum Download
- Foliensätze sowie Tabellen und Abbildungen für Dozentinnen und Dozenten zum Download



- Antworten auf die Kontrollfragen des Buchs
- Karteikarten: Überprüfen Sie Ihr Wissen
- Glossar mit zahlreichen Fachbegriffen Zusammenfassungen aller Buchkapitel
- Prüfungsfragen, Tabellen und Abbildungen für Dozentinnen und Dozenten zum Download

- Verständnisfragen: Üben Sie für die Prüfung
- Karteikarten: Überprüfen Sie Ihr Wissen
- Glossar mit zahlreichen Fachbegriffen
- Umfangreiche kommentierte Linkssammlung
- Abbildungen und Tabellen für Dozentinnen und Dozenten zum Download



Einleitung

Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang

Inhaltsverzeichnis

- 1.1 **Definition – 2**
- 1.2 **Anwendungsgebiete und Fragestellungen – 4**
 - 1.2.1 Anwendungsgebiete – 4
 - 1.2.2 Beispiele für diagnostische Fragestellungen – 7
- 1.3 **Verhältnis zu anderen Disziplinen der Psychologie – 10**
- 1.4 **Ziele der Psychologischen Diagnostik – 11**
- 1.5 **Der diagnostische Prozess – 15**
- 1.6 **Meilensteine in der Geschichte der Psychologischen Diagnostik – 22**
- 1.7 **Gesetzliche Rahmenbedingungen und ethische Richtlinien – 27**
 - 1.7.1 Menschenwürde und Privatsphäre – 28
 - 1.7.2 Geheimnisse, Schweigepflicht und Datenschutz – 29
 - 1.7.3 Offenbarungspflicht – 31
 - 1.7.4 Rechtliche Regelungen für spezifische Anwendungsfelder der Psychologischen Diagnostik – 32
 - 1.7.5 Ethische Richtlinien – 33
- 1.8 **Zusammenfassung – 36**
- Literatur – 37**

Begriffsbestimmung

1.1 Definition

Wie viele andere Begriffe in der Psychologie haben auch diejenigen von Diagnose und Diagnostik ihre Wurzeln im Griechischen, wo das Verb „diagnoskein“ eine kognitive Funktion mit den Bedeutungen „gründlich kennenlernen“, „entscheiden“ und „beschließen“ bezeichnet. Das Erleben und Verhalten einzelner oder mehrerer Menschen gründlich kennenzulernen, mit dem Ziel, eine möglichst gute Entscheidung zu treffen (z. B. welche Therapieform angemessen ist, welcher Studiengang zu den eigenen Interessen passt), beschreibt den Begriff „Psychologische Diagnostik“ in der Tat ganz gut.

Definition

Definition

Psychologische Diagnostik ist eine Teildisziplin der Psychologie. Sie dient der Beantwortung von Fragestellungen, die sich auf die Beschreibung, Klassifikation, Erklärung oder Vorhersage menschlichen Verhaltens und Erlebens beziehen. Sie schließt die gezielte Erhebung von Informationen über das Verhalten und Erleben eines oder mehrerer Menschen sowie deren relevanter Bedingungen ein. Die erhobenen Informationen werden für die Beantwortung der Fragestellung interpretiert. Das diagnostische Handeln wird von psychologischem Wissen geleitet. Zur Erhebung von Informationen werden Methoden verwendet, die wissenschaftlichen Standards genügen.

Erläuterung der Definitionselemente

Einige Definitionselemente sollen kurz erläutert werden:

Beantwortung von Fragestellungen Psychologische Diagnostik erfolgt nicht zum Selbstzweck, sondern wird durch einen Auftrag (Übernahme einer Fragestellung) in Gang gesetzt. Die Fragestellung kann eine Beschreibung oder Klassifikation („Erfüllt Person A die diagnostischen Kriterien einer Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung?“), eine Erklärung („Durch welche kognitiven Beeinträchtigungen sind die Leistungsschwankungen von Person B erklären?“) oder eine Vorhersage betreffen („Wird Person C mit hoher Wahrscheinlichkeit ein Studium erfolgreich abschließen können?“). Kurz: Diagnostik wird nicht etwa um ihrer selbst willen betrieben, sondern um eine konkrete Fragestellung zu beantworten. Das Beantworten einer Fragestellung setzt einen diagnostischen Prozess (► Abschn. 1.5) in Gang, der in der Regel mit einem konkreten Ergebnis – einer Antwort auf die Fragestellung – endet.

Gezielte Erhebung von Informationen Informationen werden nicht schematisch oder wahllos erhoben. Aus den vielen, prinzipiell ermittelbaren Informationen werden nur diejenigen tatsächlich – mittels Tests, Interviews etc. – erhoben, die zur Beantwortung der Fragestellung beitragen.

Verhalten und Erleben eines oder mehrerer Menschen Gegenstand der Psychologischen Diagnostik sind Menschen, und zwar sowohl einzelne als auch mehrere Personen (denkbar sind etwa Paare, Familien, Teams). Dies inkludiert auch die Erfassung, wie sich Menschen in bestimmten Umwelten oder Situationen verhalten bzw. wie sie diese erleben; oder wie sie ein bestimmtes Objekt wahrnehmen bzw. bewerten (z. B. ein Produkt, eine Werbeanzeige).

Relevante Bedingungen Wenn es für die Beantwortung der Fragestellung nützlich ist, können auch Informationen über relevante Bedingungen, denen die untersuchte(n) Person(en) ausgesetzt ist (sind), erhoben werden. Dies kann allgemeine Lebensumstände (Beziehungs- oder Beschäftigungsstatus), aber auch konkrete Situationen betreffen, denn menschliches Verhalten und Erleben ist nicht nur durch Eigenschaften der Person zu erklären, sondern auch durch situative Randbedingungen (s. dazu auch ► Abschn. 1.4).

Informationen werden für die Beantwortung der Fragestellung interpretiert Damit wird die Trennung von Informationen und deren Bewertung betont. Manchmal können Informationen unterschiedlich interpretiert werden. Die Interpretation erfolgt mit dem Ziel, die Fragestellung zu beantworten.

Von psychologischem Wissen geleitet Das diagnostische Handeln muss von psychologischem Wissen geleitet sein und unter Verwendung wissenschaftlicher Methoden erfolgen, wie auch Fissen (2004) sowie Eid und Petermann (2006) betonen. Würde beim diagnostischen Prozess nur das konkrete Handeln thematisiert, gelänge die Abgrenzung zur Laiendiagnostik nicht. Beim psychologischen Wissen kann es sich z. B. um die Kenntnis von Symptomen handeln, die charakteristisch für eine bestimmte psychische Störung sind. Gleichsam erfordert die Interpretation von vorliegenden Informationen, z. B. aus verschiedenen diagnostischen Instrumenten, entsprechendes Fachwissen.

Methoden, die wissenschaftlichen Standards genügen Zur Beschreibung menschlichen Verhaltens und Erlebens stehen Methoden wie Leistungstests, Fragebögen, Beobachtungsmethoden und diagnostische Interviews zur Verfügung. Diese sind dann als Werkzeuge der Psychologischen Diagnostik zu verstehen, wenn sie wissenschaftlichen Standards – sog. „Gütekriterien“ (► Abschn. 2.6) – genügen. Manche Methoden genügen wissenschaftlichen Standards nicht und sollten daher weder eingesetzt noch als Bestandteil Psychologischer Diagnostik verstanden werden. Zu letzteren Methoden gehören u. a. die Vermessung der Schädelform, die Deutung von Gesichtszügen oder der Handschrift.

Abgrenzung

Psychologische Diagnostik muss von reinem Testen, von medizinischer Diagnostik und von Evaluation abgegrenzt werden:

- **Testen:** Einfach einen Test durchzuführen ist noch keine Diagnostik. Der Begriff „Test“ bezieht sich nur auf eine von mehreren möglichen Methoden der Datenerhebung. Im Rahmen von Psychologischer Diagnostik werden auch andere Methoden wie etwa Interviews oder Verhaltensbeobachtungen eingesetzt. Hinzu kommt, dass selbst bei reiner Anwendung von Tests die Aufgaben der Psychologischen Diagnostik darüber hinausreichen. So erfordert die Anwendung verschiedener Tests sowohl eine Interpretation der im Rahmen des Testens entstandenen Informationen als auch eine Integration verschiedener Ergebnisse zu einem diagnostischen Urteil.
- **Medizinische Diagnostik:** Der Mensch ist auch Gegenstand der medizinischen Diagnostik. Hier stehen allerdings körperliche Merkmale im Fokus und nicht – oder zumindest seltener – Verhalten und Erleben. Es gibt allerdings Bereiche, in denen Diagnostik sowohl medizinisch als auch psychologisch sein kann. Dies gilt z. B. für die Diagnostik psychischer Störungen oder des Krankheitsverhaltens (z. B. der Medikamentencompliance).
- **Evaluation:** „Evaluation ist die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes. Solche Evaluationsgegenstände können z. B. Programme, Projekte, Produkte, Maßnahmen, Leistungen, Organisationen, Politik, Technologie oder Forschung sein“ (Gesellschaft für Evaluation 2008, S. 15). Das heißt, man benötigt zur Evaluation unter Umständen keine psychologisch-diagnostischen Verfahren. Dienen die zu evaluierenden Maßnahmen dazu, psychische Merkmale von Menschen (Beispiel: Depressivität) oder deren Verhalten (Beispiel: Zwangsvorhaben) zu verändern, können diese Veränderungen mithilfe psychologisch-diagnostischer Methoden (Tests, Fragebögen, Interviews etc.) erfasst werden. Psychologische Diagnostik ist dann ein Mittel zum Zweck der Evaluation.

1.2 Anwendungsgebiete und Fragestellungen

1.2.1 Anwendungsgebiete

Psychologische Maßnahmen ohne Psychologische Diagnostik?

In der Medizin gilt das geflügelte Wort: „Vor die Behandlung hat der liebe Gott die Diagnose gestellt“. Es betont, dass eine wirksame Behandlung nur dann erfolgen kann, wenn die Art der Erkrankung bekannt ist. Analog ließe sich für die Psychologie formulieren: „Vor jede **psychologische Entscheidung oder Maßnahme** hat der liebe Gott die **Psychologische Diagnostik** gestellt.“ Die Sinnhaftigkeit dieser Aussage ist offenkundig, sobald man sich psychologische Entscheidungen und Maßnahmen in verschiedenen Anwendungsgebieten der Psychologie *ohne* Psychologische Diagnostik vorstellt.

Geht es ohne Psychologische Diagnostik?

Was wäre ...

- eine Psychotherapie ohne Diagnostik der psychischen Störung?
- eine Hochbegabtenförderung ohne Diagnostik, ob bei teilnehmenden Kindern eine Hochbegabung vorliegt?
- eine eignungsdiagnostische Entscheidung ohne Eignungsdiagnose?

Die Antwort auf diese Fragen muss stets lauten: unseriös. Die jeweiligen Maßnahmen (Therapie, Training, Einstellung in einem Unternehmen) würden ohne Grundlage erfolgen. Personen würden in den meisten Fällen an den eigentlich für sie angemessenen Maßnahmen nicht teilnehmen, dafür aber andere, für sie unangemessene Maßnahmen durchlaufen.

Es geht also nicht ohne Psychologische Diagnostik. Aber natürlich reicht es nicht aus, dass Psychologische Diagnostik stattfindet. Sie muss auch so durchgeführt werden, dass *zutreffende* Entscheidungen hinsichtlich der bestmöglichen Maßnahmen getroffen werden. Fehler bei der Psychologischen Diagnostik wiegen meist schwer, da sie sich – meist irreversibel – negativ auf die danach folgenden Schritte auswirken. Es ist ein zentrales Anliegen dieses Buchs, Leserinnen und Lesern nahezubringen, wie Psychologische Diagnostik möglichst zutreffend und so fehlerfrei wie möglich gelingen kann.

Entsprechend groß ist in vielen psychologischen Berufen der Anteil diagnostischer Tätigkeiten an der gesamten Arbeitszeit. Roth und Herzberg (2008) haben rund 400 praktisch tätige Psychologinnen und Psychologen postalisch befragt. Im Durchschnitt gaben die Befragten an, dass 27 % ihrer Arbeitszeit auf Psychologische Diagnostik entfällt. Dabei traten deutliche Unterschiede zwischen den einzelnen Anwendungsfeldern hervor (► Tab. 1.1).

Demnach ist der Anteil der Psychologischen Diagnostik an der Gesamtaktivität in der Forensischen und der Verkehrspychologie am größten und in der Klinischen Psychologie am kleinsten. Die Streuung ist in allen Bereichen groß. Das bedeutet, dass es in jedem der genannten Bereiche Psychologinnen und Psychologen gibt, die deutlich mehr oder deutlich weniger diagnostizieren, als es die Mittelwerte vermuten lassen.

Die Bedeutung der Psychologischen Diagnostik wird auch durch Aufgabenbeschreibungen aus Berufen, in denen Psychologinnen und Psychologen arbeiten, unterstrichen. Die Datenbank O*Net (Occupational Information Network, ► <https://www.onetonline.org/>) bietet umfangreiche

Psychologische Diagnostik als Teil der praktischen Tätigkeit

Große Streuung der praktischen Anteile Psychologischer Diagnostik

O*Net betont Relevanz Psychologischer Diagnostik

► **Tab. 1.1** Durchschnittlicher Anteil von Psychologischer Diagnostik in einzelnen Arbeitsfeldern

Arbeitsfeld	Anteil an der Gesamtaktivität (%)
Klinische Psychologie	24
Gesundheitspsychologie	26
Pädagogische Psychologie	29
Arbeits- und Organisationspsychologie	30
Forensische Psychologie	44
Verkehrspychologie	44
Andere Bereiche	31

Quelle: Nach Roth und Herzberg (2008), N = 398

Aufgabenbeschreibungen solcher Tätigkeiten. Betrachtet man die Beschreibungen für klinische, Schul- oder Arbeits- und Organisationspsychologinnen und -psychologen, so wird deutlich, dass ca. 20–30% der darin aufgeführten Tätigkeiten der Psychologischen Diagnostik zuzuordnen sind.

Tätigkeitsbeschreibung „Klinische Psychologin bzw. klinischer Psychologe“ aus O*Net (2019; Übersetzung der Autoren)

- Mit Klientinnen und Klienten interagieren, um sie dabei zu unterstützen, Einsicht zu erlangen, Ziele zu definieren und Handlungen zu planen, um erfolgreich eine persönliche, soziale, berufliche oder Bildungsentwicklung und -anpassung zu erreichen.
- *Psychologische, emotionale oder Verhaltensprobleme identifizieren und Störungen diagnostizieren, indem Informationen aus Interviews, Tests, Aufnahmen oder anderen Bezugsquellen verwendet werden.*
- Eine Vielfalt von Behandlungsmethoden wie Psychotherapie, Hypnose, Verhaltensmodifikation, Stressreduktionstherapie, Psychodrama oder Spieltherapie nutzen.
- Individuen und Gruppen im Hinblick auf Probleme wie Stress, Substanzmissbrauch oder Familiensituationen beraten, um Verhalten zu verändern und persönliche, soziale oder berufliche Anpassung zu verbessern.
- Den Umgang mit Problemen mit Klientinnen und Klienten diskutieren.
- *Berichte über Klientinnen und Klienten verfassen und notwendige Dokumentationen erstellen.*
- Sich mit anderen Ärztinnen und Ärzten sowie Therapeutinnen und Therapeuten besprechen oder Beratung anbieten.
- *Medizinische, psychologische, soziale und familiengeschichtliche Informationen durch Interviews mit Individuen, Paaren oder Familien erheben oder durch die Sichtung von Unterlagen erlangen und analysieren.*
- *Die Effektivität von Beratung und Behandlungen sowie die Genauigkeit und Vollständigkeit von Diagnosen beurteilen, und dabei Pläne und Diagnosen, sofern notwendig, modifizieren.*
- *Psychologische Test auswählen, anwenden, auswerten und interpretieren, um Informationen über Intelligenz, Leistungen, Interessen und Persönlichkeit von Individuen zu erhalten.*
- Individuelle Behandlungspläne entwickeln und einsetzen, indem Art, Häufigkeit, Intensität und Dauer der Therapie festgelegt werden.
- Klientinnen und Klienten an andere Spezialistinnen und Spezialisten, Institutionen oder, wenn nötig, unterstützende Dienstleisterinnen und Dienstleister verweisen.
- Wissen über relevante Forschung aufrechterhalten.
- *Bezugsquellen wie Bücher, Manuale oder Journals heranziehen, um Symptome zu identifizieren, Diagnosen zu stellen oder Behandlungsansätze zu entwickeln.*
- *Individuen beim Spielen, in Gruppeninteraktionen oder in anderen Kontexten beobachten, um Hinweise auf kognitive Störungen, abweichendes Verhalten oder Verhaltensauffälligkeiten aufzudecken.*
- Informationen für Individuen zur Verfügung stellen, sodass sie schulische oder berufliche Pläne machen können.
- Psychologische Unterstützungsprogramme in psychiatrischen Zentren oder Kliniken in Zusammenarbeit mit Psychiaterinnen und Psychiatern und anderem Fachpersonal planen und entwickeln.
- Personal sowie Praktikantinnen und Praktikanten, die in die Beurteilung und Behandlung von Patienten eingebunden sind, leiten, koordinieren und deren Handlungen bewerten.

- Trainingsprogramme für Personal und Studierende entwickeln, sie leiten und an ihnen teilnehmen.
- Psychologische oder administrative Dienstleistungen und Empfehlungen für private Unternehmen und Dienststellen hinsichtlich psychologischer Gesundheitsprogramme und individueller Fälle zur Verfügung stellen.

Anmerkung: Psychologisch-diagnostische Tätigkeiten sind *kursiv* hervorgehoben.
(Abdruck mit freundlicher Genehmigung von O*Net Online)

1.2.2 Beispiele für diagnostische Fragestellungen

Für viele Anwendungsbereiche lassen sich typische Fragestellungen identifizieren. Im Folgenden werden Beispiele aufgeführt, die auch deutlich machen, dass die einzelnen Fragestellungen den Einsatz unterschiedlicher diagnostischer Verfahren verlangen. Ausführliche Informationen zur Diagnostik in Anwendungsfeldern der Psychologie finden sich in ► Kap. 6 bis 9.

In der **Klinischen Psychologie** stellt sich oft die Frage, ob eine Person, die über bestimmte Symptome wie Antriebslosigkeit oder Konzentrationsstörungen klagt, eine **psychische Störung** aufweist oder **nicht**. Liegt eine psychische Störung vor, so wird diese qualitativ näher bestimmt. Diesen Vorgang bezeichnet man als **kategoriale Diagnostik**, da es gilt, aufgrund der vorhandenen Symptome die dazu passende Störungskategorie zu finden. Gebräuchliche **Kategoriensysteme** zur Einordnung von psychischen Störungen sind die „Internationale Klassifikation psychischer Störungen“ in der 10. Revision“ (**ICD-10**; Dilling et al. 2010), die in Deutschland voraussichtlich ab 2022 von der 11. Revision (ICD-11; WHO 2018) abgelöst werden soll, und das „**Diagnostiche und Statistische Manual Psychiatrischer Störungen**“ in der 5. Revision (**DSM-5**) der American Psychiatric Association (2018). Entscheidend für die Diagnose einer psychischen Störung ist, dass eine bestimmte Anzahl und/oder Kombination genau definierter Symptome vorliegt. Die notwendigen Informationen über das Vorliegen solcher Symptome werden meist mit einem diagnostischen Interview sowie zusätzlich eingesetzten Fragebögen gewonnen.

Beispiele für typische
Fragestellungen

Psychische Störungen
diagnostizieren

Symptome der Posttraumatischen Belastungsstörung nach ICD-11 (WHO 2018; verkürzt und übersetzt durch die Autoren)

1. Wiedererleben des traumatischen Erlebnisses in Form von intrusiven Gedanken, Flashbacks, Albträumen, verbunden mit starken emotionalen und physiologischen Reaktionen
2. Vermeidung von Gedanken und Erinnerungen an das traumatische Erlebnis oder Vermeidung von Aktivitäten, Situationen und Menschen, die an das traumatische Erlebnis erinnern
3. Anhaltende Wahrnehmung erhöhter akuter Gefahr.

Diese Symptome sollen über mehrere Wochen bestehen und bedeutsame Beeinträchtigungen in wesentlichen Lebensbereichen zur Folge haben.

Die **Gesundheitspsychologie** befasst sich hauptsächlich mit der Förderung und Erhaltung von Gesundheit sowie der Vermeidung von Krankheit (Matarazzo 1980). Dabei kann es um das Gesundheitsverhalten und die Vermeidung von Krankheit im Allgemeinen gehen – also um alltägliche Dinge wie Ess-, Bewegungs- oder Erholungsverhalten. Es können aber auch spezifische Kontexte adressiert werden, z. B. das berufsbezogene

Psychologische Diagnostik zur
Förderung und Erhaltung von
Gesundheit

Psychologische Diagnostik in Erziehung, Bildung und Weiterbildung

Stresserleben und dessen Bewältigung. Wenn die subjektive Wahrnehmung der Zielpersonen im Vordergrund steht, eignen sich Fragebögen als diagnostischer Zugang sehr gut. Aber auch diagnostische Interviews, Verhaltensbeobachtungen und objektive Datenquellen (Krankheitstage, zu Fuß zurückgelegte Strecke pro Tag) liefern wichtige Informationen.

Die ***Pädagogische Psychologie*** befasst sich mit Erziehung, Bildung und Weiterbildung. Psychologische Diagnostik betrifft häufig den Leistungsstand und die Leistungsfähigkeit, was den Einsatz von Leistungstests nahelegt. Beispielsweise stellt sich manchmal bei der Einschulung die Frage, ob ein Kind den Anforderungen der Schule schon gewachsen ist. Für diese Fragestellung stehen spezielle Schulreifetests zur Verfügung. Bei der Einschulung oder im Laufe der Schulzeit, kann sich die Frage stellen, ob spezielle Fördermaßnahmen einschließlich einer Beschulung in einer speziellen Förderschule angemessen sind. Während der Schulzeit treten manchmal Leistungsprobleme auf, deren Ursache es zu erkennen gilt. In diesem Fall können Intelligenz- und Konzentrationstests helfen, die kognitive Leistungsfähigkeit bzw. die Konzentrationsfähigkeit einzuschätzen. Beim Verdacht auf eine Teilleistungsstörung sind Leistungstests indiziert, die Auskunft über die Lese-, Rechen- und Rechtschreibfertigkeit geben. Mit Schulleistungstests kann der Leistungsstand in einem Schulfach so ermittelt werden, dass ein objektiver Vergleich mit Schülerinnen und Schülern der gleichen Klassenstufe möglich ist. Ein prominentes Beispiel hierfür sind die weltweit und regelmäßig durchgeführten PISA-Testungen (PISA = Programme for International Student Assessment). Da schulische Leistungen auch von der Motivation, den Interessen, Schulangst, der Förderung durch Eltern und weiteren Bedingungen abhängen, wird man häufig auch entsprechende Fragebögen einsetzen und diagnostische Interviews mit Schülerinnen und Schülern sowie deren Eltern und Lehrkräften führen. Bei Verhaltensproblemen liegt es nahe, auch das Verhalten im Unterricht zu beobachten.

Die ***Arbeits- und Organisationspsychologie*** befasst sich mit dem Erleben und Verhalten von Menschen in beruflichen Kontexten. Psychologische Diagnostik kommt u.a. im Rahmen der Personalauswahl, der Personalentwicklung und in der Laufbahnberatung zum Einsatz. Bei der Personalauswahl variieren die eingesetzten diagnostischen Verfahren stark mit dem Untersuchungsanlass: Intelligenz- und andere Leistungstests finden trotz ihrer hohen Validität überwiegend nur bei der Auswahl von Auszubildenden Verwendung; fast immer wird ein mehr oder weniger stark strukturiertes Interview durchgeführt (Schuler et al. 2007). Bei der Personalentwicklung ist die Zielsetzung eine andere. Das Unternehmen will die Potenziale und Defizite seiner Mitarbeiterinnen und Mitarbeiter erkennen und auf Basis dieser Erkenntnisse die Weiterentwicklung der Mitarbeiterinnen und Mitarbeiter systematisch fördern. Für die Personalentwicklung werden häufig Assessment-Center konzipiert, die auf die speziellen Bedürfnisse des Unternehmens abgestimmt sind. Auch der Einsatz von Persönlichkeitsfragebögen kann sinnvoll sein. Diagnostik im Rahmen der Laufbahnberatung nehmen beispielsweise Jugendliche in Anspruch, die Hilfe bei der Berufswahl benötigen. Hier können z.B. berufsbezogene Interessentests wertvolle Erkenntnisse liefern.

Die ***Forensische Psychologie*** ist ein weites Feld mit vielen Fragestellungen, die sich durch die unterschiedlichen Zielsetzungen in Gerichtsverfahren oder im Strafvollzug ergeben. Beispielsweise kann die **Glaubwürdigkeit von Zeuginnen und Zeugen** bzw. deren Aussagen zu beurteilen sein. Je nach Grund für Zweifel an der Glaubwürdigkeit wird die **kognitive Leistungsfähigkeit** einer Zeugin oder eines Zeugen beispielsweise mit einem Intelligenztest untersucht oder die mögliche **Motivation** für eine

Psychologische Diagnostik in beruflichen Kontexten

Psychologische Diagnostik im Kontext von Gerichtsverfahren und Strafvollzug

Einleitung

Falschaussage durch ein **diagnostisches Interview** erkundet. Die Aussagen selbst werden inhaltsanalytisch auf **Anzeichen für Echtheit** analysiert. Bei Täterinnen und **Tätern** kann sich die Frage ergeben, ob sie strafrechtlich verantwortlich sind. Im Strafvollzug kann eine vorzeitige Entlassung zur Diskussion stehen und daher eine **Kriminalprognose** erstellt werden. Es gilt, das **Rückfallrisiko** einzuschätzen und eventuell die Notwendigkeit von Maßnahmen festzustellen, die das Rückfallrisiko bei der Entlassung mindern. Dabei kommen neben **Aktenanalysen** auch **diagnostische Interviews** mit Täterinnen und Tätern sowie mit deren Bezugspersonen, **Persönlichkeitstests**, **Checklisten zu Risikofaktoren** etc. zum Einsatz.

In der *Verkehrpsychologie* stellt die Beurteilung der Fahreignung einen häufigen Untersuchungsanlass dar. Eine Beurteilung der Fahreignung wird beispielsweise vorgenommen, wenn Bedenken gegen die Eignung zum Führen eines Kraftfahrzeuges bestehen. Alkoholisiertes Fahren, Verkehrsaufälligkeit aufgrund einer Suchtproblematik und wiederholte Verkehrsdelikte sind häufige Gründe für diese Bedenken. Betroffen sind in Deutschland jährlich etwa 90.000 Personen (Bundesanstalt für Straßenwesen 2019). Die Behörde, die die Fahrerlaubnis wiedererteilt, verlangt ein medizinisch-psychologisches Gutachten. Diese Gutachten dürfen nur von amtlich anerkannten medizinisch-psychologischen Untersuchungsstellen angefertigt werden. Dabei kommt dem diagnostischen Interview eine herausragende Bedeutung zu. Auch Kraftfahrerinnen und Kraftfahrer, die eine Fahrerlaubnis zur Beförderung von Fahrgästen erhalten wollen (Bus- und Taxifahrer/-innen), müssen sich einer Fahreignungsbegutachtung unterziehen. In diesem Fall steht die psychische Leistungsfähigkeit im Vordergrund. Daher werden bei diesem Untersuchungsanlass Leistungstests zur Messung der Konzentrations- und Reaktionsfähigkeit, Aufmerksamkeit, Belastbarkeit und Orientierungsleistung eingesetzt.

Als weitere Bereiche, in denen Psychologinnen und Psychologen diagnostisch tätig sind, seien exemplarisch die Entwicklungs-, die Gerontopsychologie und die Neuropsychologie genannt. In Erziehungsberatungsstellen, sozialpädiatrischen oder sozialpädagogischen Zentren und beispielsweise Frühförderstellen wird Diagnostik bei Kindern und Jugendlichen durchgeführt, die in Bezug auf ihre psychische oder geistige Entwicklung auffällig geworden sind. Dieses Arbeitsgebiet tangiert die Pädagogische und die Klinische Psychologie. Je nach Fragestellung wird mit Entwicklungstests ein „breites“ Bild der Entwicklung erstellt oder es wird gezielt in einzelnen Funktionsbereichen wie der Motorik, der Sprache, der Intelligenz oder im Bereich des Sozialverhaltens mit Leistungstests und Verhaltensbeobachtung geprüft, ob eine altersgerechte Entwicklung vorliegt. Weitere Instrumente sind das diagnostische Interview und manchmal auch Fragebögen. Mit dem anderen „Ende“ der Entwicklungsspanne befasst sich die Gerontopsychologie. Oftmals stellt sich die Frage, ob vermeintliche Leistungsdefizite objektivierbar sind. Dazu dienen Leistungstests zur Prüfung des Gedächtnisses, der Intelligenz oder der Konzentrationsfähigkeit. Die Neuropsychologie betrifft die gesamte Entwicklungsspanne – aber nur, wenn eine hirnorganische Ursache oder Beteiligung an einem Störungsbild vermutet wird. Vor allem in Neurologischen Kliniken und in Rehabilitationseinrichtungen werden Menschen behandelt, die durch einen Hirntumor, eine Schädel-Hirn-Verletzung, einen Schlaganfall oder bestimmte organische Erkrankungen hauptsächlich im Leistungsbereich Defizite aufweisen. Einzelne Funktionsbereiche können oftmals durch gezieltes Training wieder stark verbessert werden. In anderen Fällen wird nach Stärken gesucht, die zur Kompensation von Defiziten nutzbar sind. Dementsprechend kommt den Leistungstests in der neuropsychologischen Diagnostik eine besondere Bedeutung zu.

Psychologische Diagnostik zur Überprüfung der Fahreignung

Weitere Anwendungsgebiete
Psychologischer Diagnostik

Wissensaustausch mit anderen Fachgebieten

1.3 Verhältnis zu anderen Disziplinen der Psychologie

Psychologische Diagnostik ist ohne bestimmte Konstrukte, Theorien und Forschungsergebnisse aus anderen Bereichen der Psychologie – sowohl aus Grundlagenfächern wie auch aus Anwendungsfeldern – undenkbar. Der Wissenstransfer ist keineswegs einseitig; Erkenntnisse aus der Diagnostik bereichern umgekehrt auch andere Fachgebiete.

Am Beispiel der Intelligenzdiagnostik lässt sich dieses Zusammenspiel gut erläutern. Zur Messung der Intelligenz werden Intelligenztests verwendet. Nachfolgend sind exemplarisch 2 Intelligenztestaufgaben aufgeführt.

Intelligenztestaufgaben

Aufgabe 1: Ordnen Sie den fehlenden Begriff zu.

Dach : Haus = Deckel : ?

- (a) Herd (b) Topf (c) Henkel (d) Dampf (e) Küche

Erläuterung: Das im 1. Wortpaar (Dach : Haus) implizierte Verhältnis soll auf das 2. Wortpaar angewandt werden. Die richtige Lösung ist (b) Topf.

Aufgabe 2: Vervollständigen Sie die folgende Zahlenreihe.

4 8 3 9 2 ?

Erläuterung: Die Regel, nach der die Zahlenreihe aufgebaut ist, soll erkannt und die Reihe nach dieser Logik fortgesetzt werden. In diesem Beispiel werden jeweils um 1 ansteigende Zahlen wechselweise addiert und subtrahiert (+4, -5, +6, -7). Daher ist die richtige Lösung 10.

Testentwicklung profitiert von Grundlagenforschung – und umgekehrt

Die Diagnostik der Intelligenz ist in vielen Anwendungsbereichen relevant

Solche Aufgaben fallen jedoch nicht „vom Himmel“ oder sind zufällig erdacht. Um einen Intelligenztest entwickeln zu können, muss man vielmehr wissen, „was Intelligenz ist“, und die entsprechenden Aufgaben so entwickeln, dass sie genau dieses Verständnis von Intelligenz reflektieren. Die Differentielle Psychologie befasst sich ausführlich mit Theorien und Modellen der Intelligenz. Mit diesem Wissen lässt sich zu Beginn einer Testentwicklung festlegen, ob man einen Test zur Allgemeinen Intelligenz, zu einer Intelligenzkomponente (wie etwa dem räumlichen oder dem schlussfolgernden Denken) oder einen Intelligenzstrukturtest – also einen Intelligenztest, mit dem die angenommene Struktur der Intelligenz in Gänze abbildbar ist – konstruieren will. Die Forschung zu Intelligenzmodellen zeigt, welche Aufgabentypen für das Vorhaben geeignet sind. Auch Anwenderinnen und Anwender eines Intelligenztests sind auf dieses Grundlagenwissen angewiesen. Bereits bei der Auswahl eines Tests sollte ihnen bewusst sein, dass es *die* Intelligenz nicht gibt, sondern nur bestimmte Modelle oder „Arten“ der Intelligenz. Schließlich können sie ein Testergebnis nur angemessen interpretieren, wenn sie den von ihnen eingesetzten Test konzeptionell richtig einordnen.

Bei der Entscheidung, Intelligenz zu messen, sollte zuvor bedacht werden, ob und wie die Kenntnis des Testergebnisses zur Klärung der Fragestellung beitragen kann. Je nach Anwendungsbereich kann der Zusammenhang zwischen Intelligenz und Schulerfolg, Ausbildungserfolg oder Berufserfolg relevant sein. Wie eng dieser Zusammenhang ist, wird in der Pädagogischen Psychologie bzw. der Arbeits- und Organisationspsychologie erforscht. Unter Umständen muss man wissen, welche Rolle die Intelligenz bei der Definition von Lernbehinderung oder

Tab. 1.2 Bedeutung von Erkenntnissen aus verschiedenen Fachgebieten für die Diagnostik

Fach	Theorie/Modell/Forschungsergebnis	Relevanz für Diagnostik
Differentielle Psychologie	Fünf-Faktoren-Modell der Persönlichkeit (mit folgenden 5 Persönlichkeitseigenschaften: Offenheit für Erfahrungen, Gewissenhaftigkeit, Extraversion, Verträglichkeit, emotionale Stabilität)	Strukturmodell als Grundlage für Fragebogenentwicklungen; 5 breite Faktoren der Persönlichkeit gut zur Validierung von anderen Fragebögen geeignet
Entwicklungspsychologie	Piagets Stadienmodell der kognitiven Entwicklung (sensorisches Stadium, präoperationales Stadium, konkret-operacionales Stadium, formal operationales Stadium)	Das Stadienmodell kann als Grundlage für Entwicklungstests verwendet werden
Sozialpsychologie	Impression Management, d. h. der Versuch, sich so darzustellen, wie man gerne gesehen werden möchte	Dieser Einfluss auf Ergebnisse diagnostischer Instrumente sollte beachtet werden.
Biologische Psychologie	Physiologische Stressreaktionen (z. B. Kortisolaußschüttung)	Validierung von Stressfragebögen an stresskontingenten physiologischen Maßen (z. B. kann ein Fragebogen zur subjektiv empfundenen Beanspruchung anhand des Haarkortisol validiert werden)
Methodenlehre	Unterschiedliche Korrelationskoeffizienten je nach Skalenniveau (z. B. Produkt-Moment-Korrelationen für metrische Variablen oder Rangkorrelationen für ordinalskalierte Variablen)	Relevant für zutreffende Berechnung des Zusammenhangs zwischen 2 diagnostischen Instrumenten

Hochbegabung spielt oder bei welchen Arten der Demenz die Intelligenz bzw. bestimmte Intelligenzkomponenten betroffen sind – dies fällt in die Bereiche der Klinischen Psychologie und der Neuropsychologie.

Die meisten Forschungsergebnisse zur Intelligenz wären nicht zustande gekommen, wenn zur Messung der Intelligenz keine guten Tests vorhanden wären. Mit der Entwicklung von Intelligenztests werden sowohl der Grundlagenlagenforschung wie auch der angewandten Forschung wichtige Forschungsinstrumente zur Verfügung gestellt. Bei der Validierung der Tests fallen oftmals Forschungsergebnisse an, die für andere Disziplinen relevant sind.

Tests sind wichtig in der Forschung

Psychologische Diagnostik wird zur Beantwortung konkreter Fragestellungen in unterschiedlichen Anwendungsfeldern durchgeführt. Deshalb ist es unerlässlich, auch deren Forschungsergebnisse zu beachten und zu nutzen.

■ Tab. 1.2 zeigt einige Beispiele für solche Forschungsergebnisse und deren Relevanz für die Diagnostik.

Fazit Insgesamt wird deutlich, dass die Psychologische Diagnostik fest mit vielen anderen psychologischen Teildisziplinen verbunden ist. Theorien, Modelle und Forschungsergebnisse aus anderen Teildisziplinen fließen in die Entwicklung diagnostischer Verfahren ein und werden bei der Beantwortung konkreter Fragestellungen herangezogen.

1.4 Ziele der Psychologischen Diagnostik

Psychologische Diagnostik wird nicht zum Selbstzweck betrieben, sondern dient immer der Beantwortung von vorgegebenen Fragestellungen. Diese Fragestellungen können nach Inhaltsbereichen (z. B. klinische, forensische, eignungsdiagnostische Fragestellungen), aber auch nach abstrakten Zielen unterteilt werden, die unabhängig von den inhaltlichen Themen verfolgt werden. In ■ Tab. 1.3 sind exemplarisch verschiedene Ziele der Psychologischen Diagnostik in ihren Anwendungsfeldern sowie der gesellschaftliche Nutzen der Diagnostik dargestellt.

Beantwortung konkreter Fragestellungen

Tab. 1.3 Zweck und Nutzen Psychologischer Diagnostik anhand von Beispielen aus Anwendungsbereichen der Psychologie

Bereich	Zweck Psychologischer Diagnostik	Gesellschaftlicher Nutzen
Pädagogische Psychologie	Schullaufbahnberatung (Eruieren der Schule, Schulform oder Klasse, in der Schülerinnen und Schüler mit ihren Fähigkeiten, Interessen und Persönlichkeitsmerkmalen wahrscheinlich einen guten Abschluss erreichen werden)	Höhere Lebenszufriedenheit der richtig platzierten Schülerinnen und Schüler, eventuell später bessere Berufschancen, effizienter Einsatz der Ressource Schule
Klinische Psychologie	Erkennen und genaue Bestimmung von psychischen Störungen	Patientinnen und Patienten werden dadurch einer Therapie zugeführt, die die bestmögliche Heilungschance verspricht
Forensische Psychologie	Erkennen von Straftäterinnen bzw. Straftätern, die ein hohes Risiko aufweisen, nach ihrer Entlassung wieder Straftaten zu begehen	Gesellschaft wird vor Straftaten geschützt; Straftäterinnen und Straftäter erfahren eventuell weitere Behandlung, die ihnen später ein straffreies Leben ermöglicht.
Personalpsychologie	Potenzialanalyse (zur Erkennung der Stärken und Schwächen von Mitarbeiterinnen und Mitarbeitern)	Einsatz von Mitarbeiterinnen und Mitarbeitern, der ihren Fähigkeiten gerecht wird; Unternehmen und Behörden „funktionieren“ dadurch besser
Verkehrspychologie	Überprüfung der Fahreignung von Personen, die wegen Trunkenheit am Steuer oder anderer Delikte ihren Führerschein verloren haben	Die Gesellschaft wird vor Gefährdung durch andere Verkehrsteilnehmerinnen und -teilnehmer geschützt; Betroffenen wird eventuell ein Weg aufgezeigt, wie sie an sich arbeiten können, um wieder eine Fahrerlaubnis zu erhalten

Im Folgenden werden die im Rahmen der Definition der Psychologischen Diagnostik formulierten Ziele (Beschreibung, Klassifikation, Erklärung und Vorhersage) thematisiert (vgl. ► Abschn. 1.1).

Beschreiben

■ Beschreibung, Klassifikation, Erklärung und Vorhersage

Nehmen wir an, eine Schülerin oder ein Schüler wird diagnostisch untersucht, weil die Eltern berichten, sie bzw. er habe große Angst vor der Schule. Bevor Ursachen erkundet und Interventionen geplant werden, ist eine exakte Beschreibung der Schulangst nötig. Eine Verhaltensbeobachtung im Unterricht und auf dem Pausenhof sowie ein diagnostisches Interview mit den Eltern und den Lehrkräften helfen, das Problemverhalten zu beschreiben. Das Verhalten kann qualitativ und quantitativ näher bestimmt werden: Welche einschlägigen Verhaltensweisen (Vermeidung von Kontakt zu Mitschülerinnen und Mitschülern, Nichtbeteiligung am Unterricht etc.) treten auf? Wie häufig kommen diese Verhaltensweisen vor und wie ausgeprägt sind sie? Beschreibungen sind jedoch keineswegs auf Verhaltensbeobachtungen beschränkt. Im Prinzip ist jede Form der Diagnostik im 1. Schritt eine Beschreibung, die dann interpretiert und zum Zwecke der Klassifikation, Erklärung oder Vorhersage genutzt wird.

Klassifizieren

Ein Spezialfall der Beschreibung ist die Klassifikation. In der Klinischen Psychologie und der Psychiatrie sind Klassifikationssysteme (DSM-5, ICD-10/ICD-11; ► Kap. 7) gebräuchlich. Sie dienen dazu, psychische Störungen anhand von Symptomen zu diagnostizieren. In unserem Beispiel könnte die Frage auftreten: Hat die Schülerin bzw. der Schüler eine psychische Störung, etwa eine soziale Phobie (6B04)? Im Kindesalter könnte auch eine Trennungsangst (6B05) diagnostiziert werden. Die Nummer in Klammern

Einleitung

dient der exakten Klassifikation; es geht nicht um eine irgendwie geartete Form der Angst vor Menschen, sondern um eine Störung, die nach ICD-11 durch ganz bestimmte Symptome definiert ist. Eine Klassifikation setzt also immer genau definierte und voneinander abgegrenzte Kategorien voraus. Die Klassifikationssysteme DSM-5 (American Psychiatric Association 2018) und ICD-11 (WHO 2018) erfüllen diese Voraussetzung. Aber auch in anderen Bereichen wird klassifiziert. Von einer Klassifikation wird man auch sprechen, wenn Bewerberinnen und Bewerber nach einer Eignungsuntersuchung in die Kategorien „ungeeignet“, „bedingt geeignet“ oder „geeignet“ eingeteilt werden. Kritische Leserinnen und Leser werden vielleicht einwenden, der Grad der Eignung sei doch eine Dimension. Dieser Einwand ist berechtigt. Man kann jedoch eine Dimension in Kategorien einteilen. Dies geschieht meist aus pragmatischen Gründen (im Beispiel würde man den „Ungeeigneten“ eine Absage erteilen, den „Geeigneten“ ein Stellenangebot machen und die „bedingt Geeigneten“ als Reserve betrachten für den Fall, dass nicht alle Geeigneten das Stellenangebot annehmen). Auch bei der Diagnostik einer intellektuellen Hochbegabung wird aus einer dimensionalen Beschreibung, den Intelligenztestergebnissen, eine Einteilung in Kategorien vorgenommen. Personen, die in einem oder mehreren Intelligenztests einen Intelligenzquotienten ($IQ \geq 130$) erreichen, werden der Klasse „hochbegabt“ zugeordnet, alle anderen werden nicht als hochbegabt klassifiziert (Rost 2009). Spätestens am Beispiel der Hochbegabung wird deutlich, dass solche Klassen künstlich sind – also von Menschen für diagnostische Zwecke gebildet wurden. Dazu gehören auch die psychischen Störungen. Dies mag verwundern, denkt man doch, dass Krankheiten zur Natur gehören wie die Statur oder die Haarfarbe eines Menschen. Dass es sich um künstliche Klassen handelt, erkennt man daran, dass sich selbst die verbreiteten Klassifikationssysteme DSM-5 und ICD-11 bei einer Reihe von Störungen unterscheiden.

Sucht man Erklärungen für eine herausragende Leistung oder ein Problemverhalten, so liegt es auf der Hand, dass dem zu erklärenden Phänomen eine Ursache vorausgegangen sein muss. Die Schulangst des Kindes hat sich wahrscheinlich langsam entwickelt, wobei bestimmte Ereignisse eine Rolle gespielt haben. Bei einigen Störungen wird sogar per Definition angenommen, dass ein früheres Ereignis zu den aktuell vorhandenen Symptomen geführt hat. So wird bei der Posttraumatischen Belastungsstörung (ICD-11 Code: 6B40) festgelegt, dass ein Trauma („an extremely threatening or horrific event or series of events“, beispielsweise ein schwerer Unfall, Vergewaltigung, Kriegsergebnisse) vorliegen muss. Diagnostik, die zum Zweck der Erklärung durchgeführt wird, wird sich daher stark auf die Vorgeschichte einer Störung beziehen. Ein diagnostisches Interview oder die Analyse von vorhandenen Akten sind in solchen Fällen geeignete Erhebungsinstrumente.

Oftmals bestehen die Ursachen fort, und die Diagnostik kann sich auf die Gegenwart beziehen. So wird etwa nach Bedingungen gesucht, die ein Fehlverhalten aufrechterhalten. In unserem Schulangstbeispiel könnte eine ausgeprägte Ängstlichkeit diagnostiziert werden. Man darf annehmen, dass diese Ängstlichkeit schon länger besteht und zum Teil für die akuten Probleme verantwortlich ist. Es bleibt in aller Regel bei mehr oder weniger plausiblen Erklärungen. Tatsächliche „Beweise“, dass ein Ereignis die Ursache für ein Problemverhalten ist, sind nicht möglich.

Erklären

Identifikation von
aufrechterhaltenden Bedingungen

Erklärung von Verhalten durch Eigenschaften und Situationen

Fragt man Laien, warum Menschen sich so verhalten, wie sie es tun, lautet eine gängige Antwort: Das liegt an ihrem Charakter. Damit ist auch eine wissenschaftliche Position treffend beschrieben. Die grundlegende Annahme eigenschaftstheoretischer Konzepte besteht darin, dass sich das Erleben und Verhalten von Menschen in Form von Eigenschaften (Traits) beschreiben lässt. Diese werden aufgefasst als „relativ breite und zeitlich stabile Dispositionen zu bestimmten Verhaltensweisen, die konsistent in verschiedenen Situationen auftreten“ (Stemmler et al. 2010, S. 51).

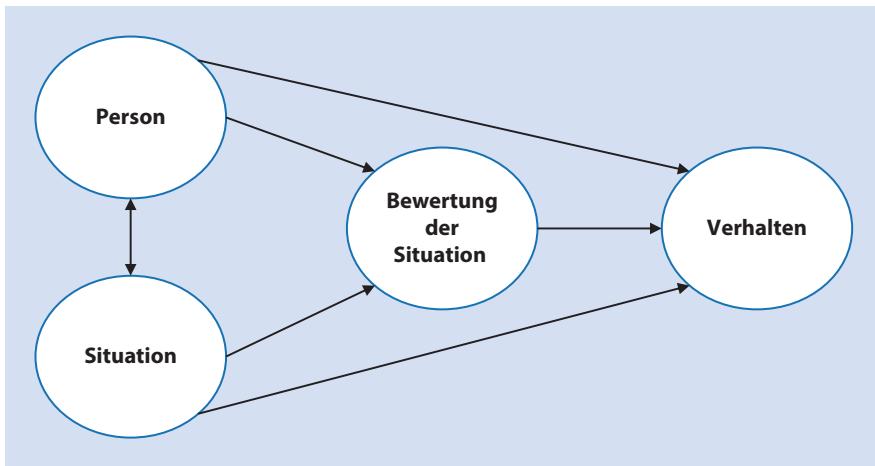
Nach der eigenschaftstheoretischen Konzeption von Persönlichkeit besteht der zweckmäßigste Weg zur Ergründung der Ursachen des Verhaltens (und auch zur Vorhersage des zukünftigen Verhaltens) von Personen darin, deren Eigenschaften mit geeigneten diagnostischen Verfahren genau zu erfassen. Dabei wird angenommen, dass sich die Person auch in anderen Situationen gemäß ihrer Eigenschaftsausprägung verhält.

Eine solche Sichtweise lässt natürlich außer Acht, dass situative Gegebenheiten ebenfalls das Verhalten von Menschen beeinflussen können. Eine rote Ampel, eine um Hilfe rufende Person oder die Teilnahme an einer Vorlesung sind situative Gegebenheiten, in denen viele Menschen ähnliche Verhaltensweisen zeigen (anhalten, sich der hilfesuchenden Person zuwenden, ruhig zuhören) – und zwar relativ unabhängig von ihren Eigenschaftsausprägungen. Das Ausmaß, in dem eine Situation bestimmte Verhaltensweisen nahelegt, wird als *Situationsstärke* bezeichnet (Meyer et al. 2010). Starke Situationen rufen über viele Personen hinweg ähnliche Verhaltensweisen hervor. Bei der Erklärung von menschlichem Verhalten müssen also neben Eigenschaftsausprägungen auch Charakteristika der Situation berücksichtigt werden.

Mittlerweile liegen mehrere Taxonomien vor, die inhaltlich unterscheidbare Wahrnehmungen von Situationen auflisten. Beispielsweise identifizierten Rauthmann et al. (2014) 8 Dimensionen der Situationswahrnehmung, die sie zu dem Akronym DIAMONDS zusammenfassen. DIAMONDS steht für **D**uty, **I**ntellect, **A**dversity, **M**ating, **P**ositivity, **N**egativity, **D**eception und **S**ociality. So könnte in einer unangenehmen Prüfungssituation vor allem die Wahrnehmungen vorherrschen, seine Pflicht erfüllen zu müssen (Duty), eine intellektuelle Leistung erbringen zu müssen (Intellect) und eine eher unangenehme Erfahrung zu machen (Negativity). Mit dem „Situational Eight DIAMONDS (S8*)“ (Rauthmann und Sherman 2015, 2017) liegt zudem ein auf den DIAMONDS basierendes Messinstrument vor, mit dem die Situationswahrnehmung von Personen erfasst werden kann.

Ob menschliches Verhalten nun stärker durch situative Gegebenheit oder eher durch Eigenschaften der Person beeinflusst wird, lässt sich anhand einer umfangreichen Metaanalyse von Richard et al. (2003) beantworten. Diese Autorinnen und Autoren nahmen 322 zuvor publizierte Metaanalysen zu sozialpsychologischen Effekten auf und berechneten mittlere Effekte der (sozialen) Situation und der Person. Die Ergebnisse dieser Analysen zeigen, dass beide mittleren Effekte in etwa gleich groß waren (mittleres r der Situation = .22, mittleres r der Person = .19).

In interaktionistischen Ansätzen zur Erklärung (und Prognose) von Verhalten geht man davon aus, dass neben Merkmalen der Person (also vor allem Eigenschaften) und der Situation zusätzlich auch die Interaktion zwischen beidem – also Person und Situation – relevant ist (vgl. Endler und Magnusson 1976). Ein Beispiel für eine solche interaktionistische Annahme findet sich in aktuellen Modellen der Situationswahrnehmung und -bewertung (Funder 2016; Rauthmann et al. 2014), denen zufolge Personeneigenschaften und Situationscharakteristika in Interaktion miteinander zu einer Bewertung der Situation führen und diese wiederum – zusammen mit direkten Einflüssen von Eigenschaften und Situationsmerkmalen – in bestimmten Verhaltensweisen resultieren (Abb. 1.1).



■ Abb. 1.1 Situations-Bewertungs-Modell. (Nach Funder 2016, S. 205, copyright © 2016, reprinted by Permission of SAGE Publications, Inc.)

In vielen Fällen wird von Diagnostikerinnen und Diagnostikern erwartet, dass sie eine Prognose abgeben. Vorhersagen können sich auf den Schul- oder den Berufserfolg oder etwa auf den Verlauf einer psychischen Störung beziehen. Die Antwort kann immer nur eine Wahrscheinlichkeitsaussage sein (z. B. „sehr wahrscheinlich wird Frau Schmitt den Anforderungen des Berufs gewachsen sein“). So zeigt die Forschung zwar, dass zum Teil enge Zusammenhänge zwischen Prädiktoren wie Intelligenz und Kriterien wie Berufserfolg bestehen (z.B. Schmidt und Hunter 1998), dass aber die Variation des Kriteriums nie vollständig aufgeklärt werden kann. Zudem zeigt uns die Erfahrung, aber auch die einschlägige Forschung, dass der Erfolg in der Schule, im Studium oder im Beruf von vielen Faktoren abhängt. Einige dieser Faktoren, etwa die Motivation oder die Gewissenhaftigkeit, kann man messen und bei der Vorhersage berücksichtigen. Die Vorhersage verbessert sich dadurch. Von einer perfekten Vorhersage sind wir aber zumeist weit entfernt. Einige Faktoren, von denen wir begründet annehmen, dass sie relevant sind, können überhaupt nicht oder nur sehr unzuverlässig erfasst werden. Viele Ereignisse, beispielsweise der Verlust eines Angehörigen durch Krankheit, Unfall oder ein Verbrechen, treten ohne Vorankündigung auf. Sie erschweren jede Vorhersage.

Prognostizieren

Fazit Prognosen sind in vielen Anwendungsfeldern der Diagnostik bedingt möglich. Forschungsergebnisse belegen einen Zusammenhang zwischen den Eigenschaften, die wir momentan messen können (z. B. Intelligenz, Motivation, Berufserfahrung), und dem vorherzusagenden Kriterium (z. B. dem Berufserfolg). Dieses Wissen erlaubt Prognosen, die aber nie zu 100 % exakt sein können und daher nur eine Wahrscheinlichkeitsaussage darstellen. Welche Alternativen gibt es zu den Wahrscheinlichkeitsaussagen? Alternativen wie „es dem Zufall überlassen“ oder „andere fragen“, die noch weniger Expertise haben, führen wahrscheinlich noch häufiger zu falschen Vorhersagen.

1.5 Der diagnostische Prozess

Um Fragestellungen, die sich auf die Beschreibung, Klassifikation, Erklärung oder Vorhersage menschlichen Verhaltens und Erlebens beziehen (s. Definition „Psychologische Diagnostik“, ▶ Abschn. 1.1), zutreffend zu beantworten, ist es wichtig, dass die dazu notwendigen Teilschritte

Sinnvolle Abfolge diagnostischer Teilschritte



Abb. 1.2 Schrittweises Vorgehen. (© MAK/► stock.adobe.com)

in sinnvoller Abfolge vollzogen werden. Die Abfolge der Teilschritte wird auch als diagnostischer Prozess bezeichnet. Dieser wird hier aufgrund seiner zentralen Bedeutung bereits vorab kurz skizziert und dann nochmals in ▶ Kap. 4 ausführlicher behandelt.

Definition

Als **diagnostischer Prozess** wird die Abfolge von Maßnahmen zur Gewinnung diagnostisch relevanter Informationen und deren Integration zur Beantwortung einer Fragestellung bezeichnet.

Diagnostischer Prozess mehr als Abfolge der diagnostischen Instrumente

Der diagnostische Prozess umfasst also weit mehr als nur die Abfolge der diagnostischen Instrumente. Vielmehr ist damit der gesamte diagnostische Entscheidungsprozess von der Formulierung der Fragestellung einer Klientin oder eines Klienten bis zu deren Beantwortung gemeint (Abb. 1.2).

Schritte des diagnostischen Prozesses

- Schritt 1: Formulierung der globalen Fragestellung
- Schritt 2: Differenzierung der globalen Fragestellung in dafür infrage kommende Teilfragen (sog. „psychologische Fragen“)
- Schritt 3: Auswahl der zur Beantwortung der Teilfragen bestmöglichen diagnostischen Instrumente
- Schritt 4: Durchführung und Auswertung der diagnostischen Instrumente
- Schritt 5: Integration der Ergebnisse zur Beantwortung der Teilfragen und der globalen Fragestellung

Idealisierter diagnostischer Prozess muss nicht der realen Praxis entsprechen

Zur konkreten Beschreibung des diagnostischen Prozesses wurden auch andere Modellvorstellungen entwickelt (z. B. Jäger et al. 1982; Westhoff und Graubner 2003). Diese Modelle und der hier skizzierte diagnostische Prozess stellen Versuche dar, das Vorgehen von Diagnostikerinnen und Diagnostikern in idealisierender Weise zu abstrahieren. Es wird nicht versucht, die reale Praxis mit all ihren Spezifika zu beschreiben (wie es deskriptive Ansätze tun),

sondern dargelegt, wie ein perfekter Ablauf theoretisch sein sollte (normative Ansätze). Die einzelnen Teilschritte des oben dargestellten diagnostischen Prozesses werden nachfolgend überblicksartig erläutert.

■ Schritt 1: Formulierung der globalen Fragestellung

In der Regel wird eine Klientin oder ein Klient mit einer Frage „aus ihrem oder seinem Leben“ an die Diagnostikerin oder den Diagnostiker herantreten. Im Sinne einer Auftragsklärung gilt es dann zunächst, diese Frage zu präzisieren, zu modifizieren oder ggf. auch abzulehnen.

Abzulehnende Fragestellungen

Es gibt einige Gründe dafür, eine Fragestellung abzulehnen (für weitere Ausführungen s. ▶ Abschn. 4.2):

- Der Diagnostikerin oder dem Diagnostiker fehlt die nötige Sachkunde; der Auftrag fällt nicht in ihren/seinen Kompetenzbereich. Eine Psychologin oder ein Psychologe in einer Personalabteilung soll vielleicht feststellen, ob Mitarbeitende psychisch krank sind. Da sie oder er nicht mit Diagnosesystemen wie dem ICD-11 oder DSM-5 vertraut ist, wird sie oder er auf klinisch erfahrene Kolleginnen und Kollegen verweisen.
- Der Auftrag ist mit dem eigenen Gewissen oder mit gesetzlichen Vorschriften nicht vereinbar. Beispielsweise könnte die Diagnostikerin oder der Diagnostiker um ein Gefälligkeitsgutachten gebeten werden.
- Die Diagnostikerin oder der Diagnostiker ist nicht neutral und kann den Auftrag daher vermutlich nicht hinreichend ergebnisoffen bearbeiten.
- Der Erkenntnisgewinn für die auftraggebende Person ist gemessen an der Belastung der Probandin bzw. des Probanden oder den anfallenden Kosten voraussichtlich gering. Die Diagnostikerin oder der Diagnostiker kann den Auftrag sofort ablehnen oder die auftraggebende Person darauf hinweisen, was in der Regel dazu führen wird, dass diese den Auftrag zurücknimmt.

In manchen Fällen ist es nötig, die Fragestellung zu präzisieren oder zu modifizieren. Beispielsweise kann der Umfang des Auftrags zunächst unklar sein. Soll etwa nur untersucht werden, ob die Probandin bzw. der Proband psychisch krank ist – die Frage wäre strenggenommen nur mit „Ja“ oder „Nein“ zu beantworten –, oder wünscht die Auftraggeberin bzw. der Auftraggeber auch eine Diagnose und/oder eine Einschätzung der Arbeitsfähigkeit? Es kommt auch vor, dass die Fragestellung zunächst so formuliert ist, dass sie nicht beantwortbar ist: Ob eine Straftäterin oder ein Straftäter nach ihrer bzw. seiner vorzeitigen Entlassung wieder rückfällig wird, kann niemand beantworten; es ist nur eine Wahrscheinlichkeitsaussage möglich. In solchen Fällen wird die Diagnostikerin bzw. der Diagnostiker mit der auftraggebenden Person Rücksprache halten und bei der präzisen Formulierung des Auftrags beraten.

Fragestellung bei Bedarf
modifizieren

■ Schritt 2: Differenzierung der globalen Fragestellung in dafür infrage kommende Teilfragen

Eine Fragestellung ist in der Regel so komplex, dass sie nicht direkt beantwortet werden kann. Die Diagnostikerin bzw. der Diagnostiker „zerlegt“ daher die globale Fragestellung in sog. „psychologische Fragen“, deren Beantwortung zur Lösung des in der Fragestellung formulierten Problems führt. Dazu knüpft sie bzw. er an den individuellen Fall an, nutzt aber auch allgemeingültige, wissenschaftliche und andere Erkenntnisse. Zum Anknüpfen an den individuellen Fall nutzt man am besten (Vor-)Informationen, beispielsweise das Protokoll eines Aufnahmegeräts, Gerichtsakten,

Hypothesengeleitetes Vorgehen

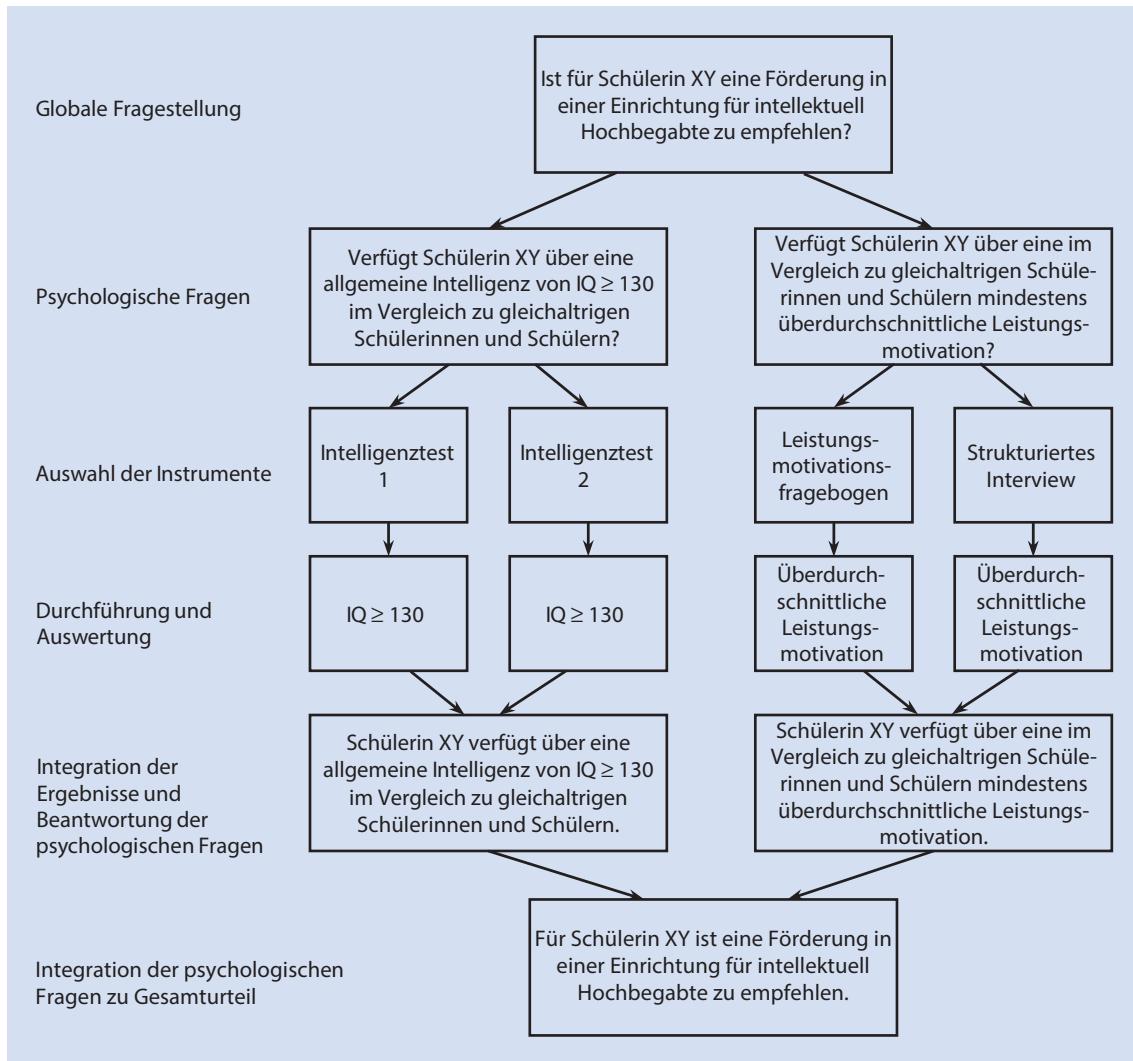


Abb. 1.3 Schematische und vereinfachte Darstellung des diagnostischen Prozesses

Hintergrundinformationen und wissenschaftliche Erkenntnisse beachten

Beschränkung auf wesentliche Faktoren

frühere Gutachten oder von der auftraggebenden Person übermittelte Hintergrundinformationen. Auch Informationen darüber, wie es zu der diagnostischen Untersuchung gekommen ist, also der Untersuchungsanlass, können nützlich sein. Insgesamt soll die Zerlegung der globalen Fragestellung in psychologische Fragen als hypothesenleitendes Vorgehen verstanden werden. Beispielhafte Formulierungen für psychologische Fragen sind in Abb. 1.3 aufgeführt.

Allerdings sind zur Formulierung von psychologischen Fragen auch Fachwissen und Berufserfahrung erforderlich. Beispielsweise liegen wissenschaftliche Erkenntnisse dazu vor, welche Faktoren zur Schulverweigerung („Schulschwänzen“) und zu einem Leistungsabbau in der Schule führen können oder welche Merkmale zur Vorhersage des Ausbildungserfolgs relevant sind. In anderen Fällen kann das Wissen über klinisch-psychologische Diagnosesysteme und deren Inhalte bei der Formulierung von psychologischen Fragen helfen. Im Bereich der Eignungsdiagnostik kann auf einschlägige Anforderungsanalysen (► Abschn. 6.1.2) zurückgegriffen werden.

Bei der Formulierung von konkreten psychologischen Fragen kommt es jedoch auch darauf an, nicht alle Faktoren zu berücksichtigen, die grundsätzlich relevant sein könnten. Vielmehr gilt es, die im vorliegenden Fall naheliegenden Faktoren zu erkennen (deshalb benötigt man Vorinformationen zum vorliegenden „Fall“). Die Maxime dabei ist: so viele wie nötig,

Einleitung

so wenige wie möglich. Die psychologischen Fragen müssen außerdem stets so gewählt werden, dass sie durch geeignete diagnostische Verfahren geklärt werden können. Falls nicht, würde man in Schritt 3 des diagnostischen Prozesses scheitern.

Im Zuge der Formulierung psychologischer Fragen ist auch zu klären, wie diese zu einem Gesamurteil hinsichtlich der globalen Fragestellung verrechnet werden (► Abschn. 5.1.3). Beispielsweise kann verlangt werden, dass alle psychologischen Fragestellungen als zutreffend bewertet werden müssen, sodass auch die globale Fragestellung bejaht wird. Oder es kann eine Mindestanzahl an positiv beurteilten psychologischen Fragen definiert werden, die nötig ist, um die globale Fragestellung positiv zu bewerten.

Festlegung, wie Informationen zum Gesamurteil verrechnet werden

■ Schritt 3: Auswahl der zur Beantwortung der Teilfragen bestmöglichen diagnostischen Instrumente

In diesem Schritt gilt es, die zur Beantwortung der Fragen am besten geeigneten diagnostischen Instrumente (Tests, Interviews, Verhaltensbeobachtungen) zu finden. Auswahlkriterien können einschlägige Validitätsbelege, aktuelle Normen, Zumutbarkeit, Zeitaufwand etc. sein (► Abschn. 2.6). Es sollten stets hohe Ansprüche an die Güte der verwendeten Instrumente gestellt werden, da sie die wesentliche Grundlage zur Beantwortung der psychologischen Fragen darstellen. Es empfiehlt sich, für jede psychologische Frage nach Möglichkeit mindestens 2 verschiedene Instrumente einzusetzen (multimethodales Vorgehen).

Auswahl der Verfahren anhand von Validitätsbelegen und anderen Kriterien

■ Schritt 4: Durchführung und Auswertung der diagnostischen Instrumente

Nun gilt es, die zur Beantwortung der Teilfragen ausgewählten Instrumente professionell anzuwenden und auszuwerten. Dazu ist es erforderlich, dass die Diagnostikerin bzw. der Diagnostiker die gängigen Praktiken zur Vorbereitung, Durchführung, Auswertung und Interpretation von diagnostischen Instrumenten beherrscht. Für psychologische Tests sind entsprechende Kompetenzen in Richtlinien der International Test Commission festgelegt (ITC 2013). Aber auch für diagnostische Interviews und Verhaltensbeobachtungen gibt es wissenschaftlich fundierte Empfehlungen zur Durchführung und Auswertung.

Professionelle Anwendung und Auswertung der Verfahren

■ Schritt 5: Integration der Ergebnisse zur Beantwortung der psychologischen Fragen und der globalen Fragestellung

Im Idealfall beantworten die in Schritt 4 anfallenden Ergebnisse die psychologischen Fragen. Dazu müssen die jeweils für die psychologischen Fragen relevanten Erkenntnisse zu einer einheitlichen Aussage zusammengeführt werden. Die Aussagen zu den psychologischen Fragen müssen wiederum in eine Antwort auf die globale Fragestellung integriert werden. Diesen Prozess nennt man diagnostische Urteilsbildung. Dabei können uneindeutige oder widersprüchliche Ergebnisse die Urteilsbildung erschweren oder sogar weitere Datenerhebung(en) notwendig machen. □ Abb. 1.3 visualisiert den diagnostischen Prozess.

Diagnostische Urteilsbildung

Es soll jedoch auch betont werden, dass der diagnostische Prozess zwar eine sinnvolle Abfolge diagnostischer Teilschritte, aber keineswegs eine Einbahnstraße darstellt. Eventuell muss man einen Schritt zurückgehen, um am Ende zum Ziel zu gelangen. Dies kann der Fall sein, wenn diagnostische Verfahren uneindeutige Ergebnisse liefern – hier könnten weitere Verfahren durchgeführt oder zusätzliche Informationsquellen einbezogen werden. Solche „Rückschleifen“ sind nicht als Ausdruck mangelnder diagnostischer Kompetenz zu werten, sondern stellen ein wesentliches Merkmal des diagnostischen Prozesses dar. Es soll hier auch darauf hingewiesen werden, dass spezifische Anwendungsfelder der Psychologischen Diagnostik spezifische Modifikationen des hier dargestellten diagnostischen Prozesses erfordern können.

Eventuell Nacherhebungen bei uneindeutigen Ergebnissen

Umsetzung hängt von der Art der diagnostischen Entscheidung ab

Randbedingungen diagnostischer Entscheidungen nach Cronbach und Gleser (1965)

Die konkrete Umsetzung des diagnostischen Prozesses hängt von der Art der diagnostischen Entscheidung ab. Eine umfangreichere Systematisierung diagnostischer Entscheidungen ist bereits 1965 von Cronbach und Gleser erstellt worden.

Arten diagnostischer Entscheidungen (nach Cronbach und Gleser 1965, S. 16):

1. Nutzen der Entscheidungen geht zugunsten	Institution	vs.	Individuum
2. Annahmequote	Festgelegt	vs.	Variabel
3. Testungen	Singulär	vs.	Multipel
4. Möglichkeit von Ablehnungen	Ja	vs.	Nein
5. Informationsdimensionen	Univariat	vs.	Multivariat
6. Entscheidungen	Terminal	vs.	Investigatorisch

Institutioneller vs. individueller Nutzen

Eine Entscheidung ist von institutioneller Art, wenn eine Organisation (z. B. ein Betrieb) nach einem standardisierten Vorgehen alle Personen in der gleichen Weise untersucht. So müssen etwa alle Personen ein und denselben Test bearbeiten oder an einem Vorstellungsgespräch teilnehmen. Gesucht wird eine Entscheidungsregel, die den Nutzen vieler (gleichartiger) Entscheidungen für die Institution maximiert. So hat der Betrieb ein Interesse daran, die bestgeeigneten Personen für eine Stelle zu finden. Ganz anders gelagert sind dagegen die Verhältnisse, wenn ein Individuum auf eine Diagnostikerin bzw. einen Diagnostiker oder eine Institution zugeht, um beispielsweise Rat bei der anstehenden Berufswahl einzuholen. Hierbei interessiert allein der individuelle Nutzen.

Annahmequoten

Festgelegte Annahmequoten liegen vor, wenn z. B. nur eine bestimmte Zahl von Therapie- oder Ausbildungsplätzen zur Verfügung steht. Ist die Zahl der Bewerberinnen und Bewerber größer als die der vorhandenen Plätze, erfolgt eine Auswahl. Hingegen ist bei nicht festgelegten oder variablen Annahmequoten bei jeder Entscheidung das Ergebnis offen. Beispielsweise erhalten (im Idealfall) alle Personen eine Therapie, die als therapiebedürftig beurteilt worden sind.

Singuläre vs. multiple Testung

Bei einstufigen Testungen erfolgt die diagnostische Entscheidung auf der Basis einer punktuell-einmaligen Diagnose. Bei mehrstufigen (sequenziellen) Testungen steht das Resultat erst nach einem gestuften Vorgehen in mehreren Schritten fest.

Möglichkeit von Ablehnungen

Sind Ablehnungen möglich, liegt die klassische Struktur von Selektionsparadigmen vor. Verbleiben hingegen alle Probandinnen bzw. Probanden im System und werden infolge der Diagnoseerstellung nur horizontal oder vertikal zu spezifischen Interventionen „verschoben“, spricht man von einer Platzierung; niemand wird von einer (positiven) Intervention ausgeschlossen.

Univariate vs. multivariate diagnostische Information

Die diagnostische Information kann sich auf eine Dimension beschränken (z. B. die Abiturnote), also univariat vorliegen, oder aus mehreren Dimensionen stammen und somit multivariat beschaffen sein (z. B. die Abiturnote und die allgemeine Intelligenz). Meist werden zur Erhöhung der Validität und damit auch der Entscheidungssicherheit mehrere Prädiktoren herangezogen, weil damit verschiedene Facetten des Kriteriums abgedeckt werden können.

Wird auf der Basis der diagnostischen Information eine Maßnahme eingeleitet, ohne dass eine weitere Psychologische Diagnostik vorgesehen ist und ohne dass die Entscheidung für oder gegen eine Maßnahme revidiert wird, handelt es sich um eine terminale Entscheidung. Eine investigatorische Entscheidung stellt den 1. Schritt in einem mehrstufigen Entscheidungsverfahren dar. Ihr folgt z. B. direkt oder nach einer Behandlung eine weitere diagnostische Untersuchung, die dann entweder zu einer weiteren investigatorischen oder zu einer terminalen Entscheidung führt.

Terminale vs. investigatorische Entscheidung

Es mag für viele Leserinnen und Leser naheliegend oder sogar selbstverständlich sein, dass Psychologische Diagnostik nach dem hier geschilderten Prozess erfolgt. Führt man sich jedoch gängige Beispiele aus der Praxis vor Augen, mögen die vorherigen Ausführungen weit weniger selbstverständlich erscheinen.

Praxis oft nicht idealtypisch

Beispiele, in denen der diagnostische Prozess nicht stringent vollzogen wird

- Alle Führungskräfte eines Unternehmens füllen einen umfangreichen Persönlichkeitstest aus; erst danach wird entschieden, was man mit den Ergebnissen macht.
- Kommissionen zur Auswahl von Professorinnen und Professoren nutzen eine Lehrprobe. Worauf dabei genau geachtet werden soll, ist jedem Kommissionsmitglied überlassen bzw. wird erst nach erfolgter Lehrprobe besprochen.
- Erst nach Vorliegen aller Informationen zu einer oder mehreren Personen wird über die Gewichtung und Verrechnung der Informationen zu einem Gesamturteil nachgedacht.

Die nachfolgenden Kapitel dieses Buchs befassen sich im Detail mit einzelnen Schritten des diagnostischen Prozesses. So wird in ► Kap. 2 besprochen, wie diagnostische Verfahren entwickelt und evaluiert werden sollten, um im Rahmen eines diagnostischen Prozesses brauchbare Informationen zu liefern. Es wird dort ebenfalls thematisiert, wie Ergebnisse diagnostischer Verfahren interpretiert werden können. ► Kap. 3 gibt einen Überblick über Leistungstests, Fragebögen, Interviewtechniken und andere Verfahren, mit deren Hilfe psychologische Fragen beantwortet werden können. In ► Kap. 4 wird gezeigt, was bei der Durchführung von diagnostischen Verfahren zu beachten ist und wie der gesamte diagnostische Prozess in einem Gutachten dokumentiert wird. ► Kap. 5 beschäftigt sich mit der Integration von diagnostischen Ergebnissen und zeigt Möglichkeiten der Planung und Evaluation des diagnostischen Prozesses auf. Die ► Kap. 6 bis 9 skizzieren Spezifika der Psychologischen Diagnostik in verschiedenen Anwendungsfeldern. Zuvor sollen jedoch in diesem Kapitel noch einige Meilensteine der Geschichte der Psychologischen Diagnostik sowie rechtliche und ethische Aspekte dargestellt werden.

Wilhelm Wundt: Begründer der Psychologie als wissenschaftliche Disziplin

Erste Tests vor 3000 Jahren in China

1.6 Meilensteine in der Geschichte der Psychologischen Diagnostik

Der Beginn der wissenschaftlichen Psychologie wird üblicherweise auf das Ende des 19. Jahrhunderts datiert. Ein Meilenstein war die Gründung des ersten Labors zur Erforschung psychologischer Phänomene im Jahre 1879 durch Wilhelm Wundt an der Universität Leipzig. Zuvor hatten sich Philosophen und Naturforscher bereits mit Fragen befasst, die im Prinzip auch heute noch die psychologische Forschung beschäftigen. Auch die Psychologische Diagnostik hat ihre Wurzeln in der frühen Vorgeschichte, und erste wissenschaftlich bedeutsame Ereignisse sind seit Ende des 19. Jahrhunderts zu vermelden.

Tests wurden bereits vor rund 3000 Jahren in China entwickelt und eingesetzt. Ein richtiges Prüfungssystem zur Auswahl von Beamten (Beamten gab es zu dieser Zeit nicht) wurde in der Sui-Dynastie (581–618 n. Chr.) entwickelt und in der Tang-Dynastie (618–907 n. Chr.) perfektioniert. Es bestand aus 3 Teilen: einer gewöhnlichen Prüfung, in der u. a. Kenntnisse über klassische Schriften erfasst wurden, einer Prüfung durch eine Kommission vor dem Kaiser sowie einer Prüfung kriegerischer Fertigkeiten (Bogenschießen, Reiten etc.). Die Kommission setzte u. a. Interviews ein, um die Fähigkeit zum Planen und Verwalten zu erfassen. Aus heutiger Perspektive würden wir von einem multimethodalen Ansatz sprechen, also dem Einsatz unterschiedlicher Methoden, um (hier) die Eignung einer Bewerberin oder eines Bewerbers festzustellen. Dieses Prüfungssystem bestand Hunderte von Jahren und beeinflusst noch heute die Praxis der Personalauswahl und von Prüfungen in China (Wang 1993). Es ist jedoch zu hoffen, dass die Prozedur in Umfang und Belastung substanzell reduziert wurde, denn für die historischen Auswahlprozeduren beschreiben Wainer et al. (2000, S. 2), dass „Kandidaten manchmal im Laufe der Tests verstorben seien“.

Für die heutige Diagnostik prägend sind Ereignisse und Entwicklungen, die in  Tab. 1.4 aufgelistet sind (für weitere Informationen s. Gregory 2004). Der Fokus liegt bei dieser Betrachtung auf psychologischen Tests.

Weiterführende Literatur

Zur Geschichte der Psychologischen Diagnostik (vorwiegend der Verwendung von Tests) kann das Buch von Gregory (2004) und dort besonders Kap. 1 empfohlen werden. Das von Lück und Miller (1993) herausgegebene Buch zur Geschichte der Psychologie enthält mehrere kurze und reich bebilderte Kapitel zur Entwicklung der Diagnostik in verschiedenen Anwendungsfeldern. Auch in dem von Lamberti (2006) herausgegebenen Buch finden sich bebilderte Abhandlungen zur historischen Entwicklung einzelner Anwendungsfelder der Diagnostik.

■ Tab. 1.4 Wichtige Ereignisse in der Geschichte der Psychologischen Diagnostik

Jahr	Ereignis	Kommentar
1884	Sir Francis Galton stellt auf der internationalen Gesundheitsausstellung in London ein psychometrisches Labor vor, das auch kognitive Tests umfasst	Dies ist vermutlich der erste systematische Versuch in der Neuzeit, interindividuelle Unterschiede geistiger Fähigkeiten zu messen. Bereits ein Jahr zuvor hatte Galton in einer Publikation dargelegt, dass sich Menschen in Bezug auf ihre kognitiven Fähigkeiten unterscheiden. Mit klug ausgedachten Tests zur Reaktionszeit, Tonhöhenwahrnehmung etc. versuchte er, biologische Grundlagen geistiger Fähigkeiten zu messen.
1891–1893	Erste Versionen der International Classification of Diseases (ICD) entstehen	Die Internationale Klassifikation psychischer Störungen (ICD) dient der Klassifikation körperlicher und psychischer Störungen. Sie liegt mittlerweile in der 11. Revision vor (ICD-11; WHO 2018). Die von 1891 bis 1893 entwickelten Listen systematisierten die damals bekannten Todesursachen. Sie wurden als Bertillon'sche Klassifikation der Todesursachen (BCCD) bzw. International List of Causes of Death (ILCD) bezeichnet (DIMDI 2019).
Beginn des 20. Jahrhunderts bis zum Zweiten Weltkrieg	Blütezeit der Psychotechnik	Zu Beginn des 20. Jahrhunderts bis zum Zweiten Weltkrieg blühte die Psychotechnik. Der deutschstämmige Psychologe Hugo Münsterberg (1863–1916), der später an der Harvard-Universität arbeitete, gilt als deren Begründer (van Drunen 1993). Münsterberg sah das Verhältnis von Psychologie zur diagnostischen Anwendung ähnlich wie das Verhältnis der Naturwissenschaften zur Technik. Diagnostiker hatten in diesem Denken eine ähnliche Funktion wie die Ingenieure im Bereich der Technik. Man entwickelte mit großem Einfallsreichtum technische Geräte. Die apparative Messung von Eigenschaften versprach eine hohe Genauigkeit. Psychotechnische Geräte kamen u. a. beim Militär, der Eisenbahn, der Straßenbahn, in der Sport- und in der Arbeitspsychologie zum Einsatz (Lück und Miller 1993). Dies illustrieren die Beispiele für Methoden der Psychotechnik (van Drunen 1993).
1901	Clark Wissler führt die erste systematische Validierungsstudie zu kognitiven Tests durch	Wissler führte mit über 300 Studierenden kognitive Tests der Art, wie sie Galton propagiert hatte, durch und setzte die Testleistungen mit Studiennoten in Bezug. Die Korrelationen waren so niedrig (die höchste betrug $r = .16$), dass der Versuch, mit solchen Tests geistige Fähigkeiten zu messen, als gescheitert galt.
1905	Alfred Binet und Theodore Simon veröffentlichen den ersten Intelligenztest	Der Test war völlig anders konzipiert als die Tests Galtons; er entsprach eher heutigen Intelligenztests. Entwickelt wurde er im Auftrag des französischen Unterrichtsministeriums mit dem Ziel, geistig zurückgebliebene Kinder zu entdecken, um sie angemessen zu beschulen. Der Test wurde bald in anderen Ländern adaptiert (1916 der Stanford-Binet-Test von Lewis M. Terman als amerikanische Version) und verbreitete sich schnell. Die Aufgaben dienten zudem als Vorbild für andere Tests. Noch heute ist ein Nachfolgetest in Gebrauch.
1912	William Stern schlägt den Begriff „Intelligenzquotient“ vor und gibt eine Formel dafür an	Bei den ersten Intelligenztests wurde lediglich das „Intelligenzalter“ bestimmt, das angibt, welchen Entwicklungsstand ein Kind erreicht hat. Stern schlug folgende Formel vor: Intelligenzquotient = $100 \times (\text{Intelligenzalter}/\text{Lebensalter})$. Ein Beispiel: Intelligenzalter = 6 Jahre, Lebensalter = 8 Jahre; IQ = $100 \times (6/8) = 75$. Heute wird der IQ über die Abweichung vom Populationsmittelwert bestimmt (vgl. ▶ Abschn. 2.6.4).
1917/18	Entwicklung und Einsatz des ersten Gruppen- tests (Army Alpha and Beta Examination)	1917 waren die USA in den Ersten Weltkrieg eingetreten. Die beiden Tests wurden von einer Arbeitsgruppe unter Leitung von Robert M. Yerkes entwickelt, um Rekruten zu untersuchen (es sollten geistig inkompente ausgesondert und bei den anderen die Platzierung optimiert werden). Der Alpha-Test bestand aus 8 Subtests (z. B. rechnerisches Denken, Synonyme/Antonyme, Analogien). Der Beta-Test bestand aus weitgehend sprachfreien Aufgaben für den Einsatz bei wenig sprachkompetenten Rekruten. Mitarbeitende, die an der Entwicklung der Army-Tests beteiligt waren, konstruierten später Intelligenztests für den Bildungsbereich oder die Wirtschaft. Die Army-Tests dienten auch vielen anderen Testautorinnen und -autoren als Vorbild.
1917/18	Entwicklung des ersten modernen Persönlichkeitstests (Personal Data Sheet)	Der harmlos als „Personal Data Sheet“ etikettierte Fragebogen diente ebenfalls zur Beurteilung von Rekruten, die von den USA in den Ersten Weltkrieg geschickt wurden. Er bestand aus 116 nach empirischen Kennwerten ausgewählten Fragen, die mit „Ja“ oder „Nein“ zu beantworten waren (Beispiel: „Gehen Ihnen Gedanken durch den Kopf, sodass Sie nicht schlafen können?“). Damit sollten neurotische Rekruten entdeckt werden, um sie dann gründlich psychiatrisch zu untersuchen. Der Fragebogen war Vorbild für andere Persönlichkeitsskalen.

(Fortsetzung)

■ Tab. 1.4 (Fortsetzung)

Jahr	Ereignis	Kommentar
1921	Der Rorschach-Test wird publiziert	Der Schweizer Psychiater Hermann Rorschach veröffentlichte den ersten projektiven Test, der später nach ihm benannt wurde. Jede der 10 Tafeln zeigt Gebilde, die aus schwarzen oder farbigen Tintenklecksen bestehen. Die Testperson soll angeben, was das sein könnte. Damit wurde ein völlig anderes Testkonzept verfolgt als mit den Persönlichkeitsfragebögen in den USA. Der Rorschach-Test (► Abschn. 3.5.1.1) wird noch heute eingesetzt, und es liegen Tausende von Publikationen dazu vor.
1939	Der erste Wechsler-Test erscheint	David Wechsler, ein Psychologe am Bellevue Hospital in New York, publizierte nach mehreren Jahren Vorbereitung die „Wechsler-Bellevue Intelligence Scales“. Er hatte nicht die Absicht, einen völlig neuen Test zu entwickeln. Die Items sind teilweise stark angelehnt an die Binet- und Army-Alpha- und Beta-Tests. Neu war Wechslers Formel zur Berechnung des Intelligenzquotienten, in der er den Testwert einer Person in Relation zum Mittelwert der Altersgruppe setzte. Der Test wurde 1955 zur bekannten Erwachsenenversion „Wechsler Adult Intelligence Scale“ (WAIS) weiterentwickelt. Für Kinder und schließlich auch für Vorschulkinder kamen ähnlich aufgebauten Tests auf den Markt. Die Tests wurden von Wechslers Nachfolgern kontinuierlich weiterentwickelt und in viele Sprachen übersetzt (für die Wechsler-Tests ► Abschn. 3.2.3.2).
1943	Das Minnesota Multiphasic Personality Inventory (MMPI) wird publiziert	Mit dem MMPI bringen der Psychologe Starke R. Hathaway und der Psychiater J. Charnley McKinley einen neuartigen Persönlichkeitsfragebogen auf den Markt. Wie beim Personal Data Sheet wurden die Items durch Vergleich von psychiatrischen und normalen Personen gewonnen. Das MMPI hat jedoch viele Skalen und – das war neu – Validitätsskalen, die verschiedene Formen der Verfälschung erfassen. Das Verfahren ist nach einer Überarbeitung (MMPI-2; ► Abschn. 3.3.3.1) heute noch weitverbreitet und wurde in Tausenden von Untersuchungen intensiv beforscht.
1952	Das Diagnostic and Statistical Manual of Mental Disorders (DSM) wird herausgegeben	Das Diagnostische und Statistische Manual Psychischer Störungen (DSM) der American Psychiatric Association (2018) dient der Klassifikation psychischer Störungen. Mittlerweile liegt es in der 5. Revision vor (DSM-5).
1961	Das Buch <i>Testaufbau und Testanalyse</i> erscheint	Nach dem Zweiten Weltkrieg war die Psychologische Diagnostik in der Bundesrepublik Deutschland durch die Entwicklung von Papier-und-Bleistift-Tests geprägt. Ein Pionier war Gustav A. Lienert (► Abb. 1.4). Sein Buch <i>Testaufbau und Testanalyse</i> erschien 1961 und erlebte mehrere Neuauflagen (Lienert und Raatz 1998). Generationen von Psychologiestudierenden lernten mit diesem Buch, wie man auf der Grundlage der Klassischen Testtheorie Tests konstruiert. Das Buch gehört immer noch zum Handwerkzeug von Testentwicklerinnen und -entwicklern.
1962	Erstes Computerauswertungsprogramm für einen Test verfügbar	Einen weiteren Meilenstein stellt der Einsatz des Computers im Rahmen der Psychologischen Diagnostik dar. Die ersten Anwendungen beschränkten sich zunächst auf die automatisierte Auswertung von papierhaft generierten Testdaten. Eines der ersten Auswerteprogramme wurde 1962 genutzt, um Daten des bereits erwähnten MMPI zu analysieren (Rome et al. 1962). Mit fortschreitender Technologie wurden Computer bald darauf in vielen Anwendungsfeldern der Psychologischen Diagnostik eingesetzt, sowohl zur Auswertung von Test als auch bald darauf zur Administration von Tests (Wainer et al. 2000). Bereits in den frühen 1980er-Jahren beschreiben Pawlik und Buse (1982) die Möglichkeiten, mit tragbaren Mikroprozessoren Verhalten im Alltag der Testpersonen aufzuzeichnen – eine Methode, die sich unter der Bezeichnung „ambulantes/ambulatorisches Assessment“ auch heute großer Beliebtheit in der psychologischen Forschung erfreut. In den 1980er-Jahren entstanden auch die ersten ernsthaft einsetzbaren computerisierten adaptiven Tests (Wainer et al. 2000). Hierbei ermittelt eine Software auf Basis vorangegangener Richtig- oder Falsch-Antworten einer Testperson in einem Leistungstest, welche Itemschwierigkeit am ehesten dem Leistungsvermögen der Testperson entspricht und administriert dieses Item.

(Fortsetzung)

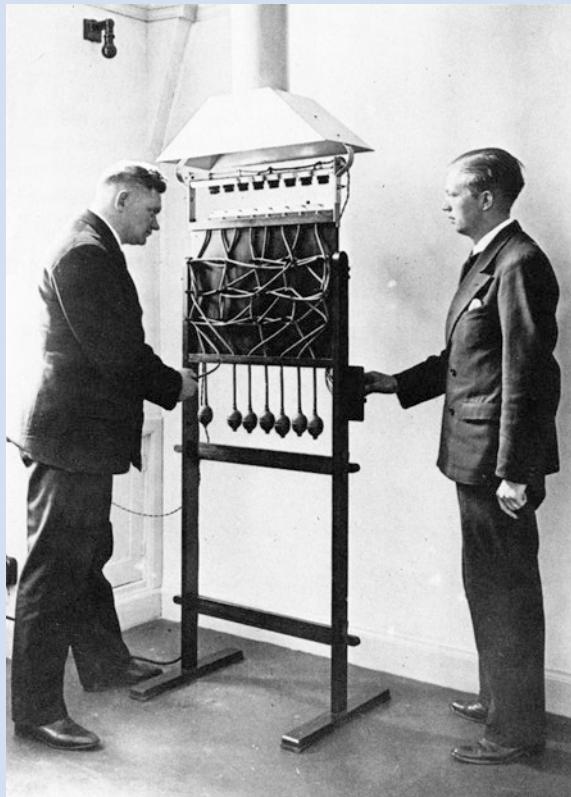
Tab. 1.4 (Fortsetzung)

Jahr	Ereignis	Kommentar
1976	Gründung der International Test Commission (ITC)	Die International Test Commission wurde 1976 während des Congress of the International Union of Psychological Sciences (IUPsyS) gegründet (Oakland et al. 2001). Ihr Ziel ist, auf eine angemessene und wirksame Testnutzung sowie eine wissenschaftlich fundierte Testentwicklung hinzuwirken.
2000	Erste Erhebung im Rahmen von PISA	Im Zuge einer fortschreitenden Globalisierung kamen international angelegte diagnostische Studien – vorwiegend im Bildungssektor – in Mode. Der bekannteste Vertreter sog. „Large-Scale-Assessments“ im Bildungsbereich ist das Programme for International Student Assessment, kurz: PISA (► Abschn. 7.4.2.1). Im Jahr 2000 fand die erste Erhebung statt, an der 180.000 Jugendliche aus 32 Ländern teilnahmen (Stanat et al. 2002). Seitdem werden im Abstand von 3 Jahren weltweit Leistungen in den Bereichen Mathematik, Lesen und Naturwissenschaften erhoben. Die Ergebnisse zwischen und innerhalb von teilnehmenden Ländern werden mit anderen Bildungsindikatoren in Bezug gesetzt, um daraus Maßnahmen zur Optimierung der Schulbildung abzuleiten.
2015	Psychologische Diagnostik auf Basis von Daten aus sozialen Medien	Eine weitere gesellschaftliche Veränderung beeinflusst maßgeblich die Psychologische Diagnostik: die massenhafte Nutzung sozialer Medien. Die daraus entstehenden Daten (z. B. „Likes“ auf Facebook) können zusammen mit anderen Daten (z. B. einem ausgefüllten Persönlichkeitsfragebogen) dazu genutzt werden, Algorithmen zu trainieren. Liegen genügend Daten zum Training der Algorithmen vor, gelingt meist eine erstaunliche gute Vorhersage persönlicher Merkmale (z. B. der Persönlichkeitseigenschaften, aber auch sexueller Präferenzen) auf Basis der in sozialen Medien hinterlassenen „Spuren“ (z. B. Kosinski et al. 2015).

Beispiele für Methoden der Psychotechnik

„Kerzentest“ zur Messung von überlegtem Handeln

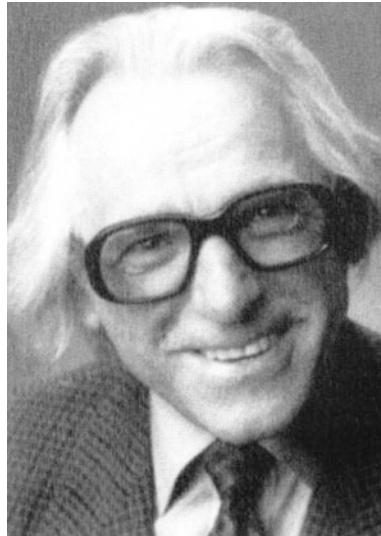
In Augenhöhe waren 8 elektrische Kerzen nebeneinander angeordnet. Von jeder Kerze führte ein Schlauch zu einem Gummiball in Handhöhe. Der Weg der Schläuche war verschlungen. Wenn der Versuchsleiter eine Kerze anschaltete, musste der Proband so schnell wie möglich den richtigen Gummiball drücken (was zum Erlöschen der Kerze führte).



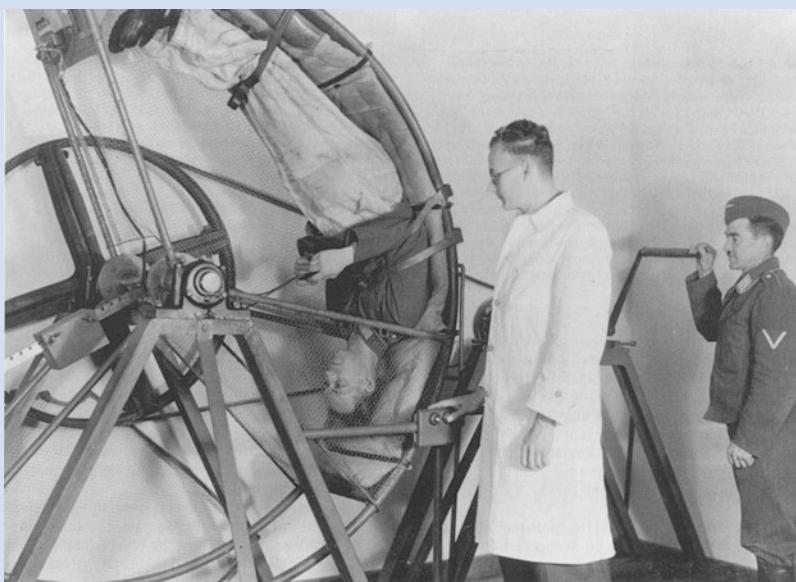
Kerzentest. (Aus van Drunen 1993, S. 256, © Quintessenz)

Untersuchung zur Stabilität des Nervensystems In Frankreich wurde im Ersten Weltkrieg ein „Kymograph“ zur Auswahl von Piloten eingesetzt; mittels einer rußgeschwärzten, sich drehenden Walze wurden Herzschlag, Atmung und Handdruck registriert. Der Test bestand darin, dass der Untersuchungsleiter unerwartet hinter der Testperson eine Pistole abfeuerte.

Untersuchung der geistigen Leistungsfähigkeit in einem Rhönrad In Deutschland führte die Wehrmacht im Zweiten Weltkrieg mit angehenden Piloten Tests durch. Die Probanden wurden in einem Rhönrad fixiert. Durch Drehung des Rads änderte sich ihre Körperlage. Dabei mussten sie verschiedene Aufgaben lösen, beispielsweise Rechenaufgaben.



■ Abb. 1.4 Prof. Dr. Dr. Gustav A. Lienert. (Aus Tent 2001, S. 242)



Rhönrad. (Aus Lück und Miller 1993, S. 281, © Quintessenz)

1.7 Gesetzliche Rahmenbedingungen und ethische Richtlinien

Wie die meisten Bereiche des öffentlichen und beruflichen Lebens unterliegt auch die Psychologische Diagnostik rechtlichen Bestimmungen. Diese rechtlichen Rahmenbedingungen folgen einer Systematik. Joussen (2004) spricht von einer „Normenpyramide“. Ein Recht, das in der Pyramide über einem anderen steht, hat immer Vorrang. Besteht ein Widerspruch zwischen 2 Hierarchieebenen, so ist immer das ranghöchste Gesetz entscheidend. Notwendigerweise ist ein ranghohes Gesetz allgemeiner formuliert als

„Normenpyramide“

Europäisches Recht steht an oberster Stelle

ein rangniedriges. So zeichnet sich das Grundgesetz durch allgemeine und abstrakte Formulierungen aus. Im Strafgesetzbuch oder im Betriebsverfassungsgesetz finden sich dagegen sehr konkrete Regelungen.

An oberster Stelle steht das Recht der Europäischen Gemeinschaft, gefolgt vom deutschen Grundgesetz. Eine Ebene tiefer stehen die einfachen Gesetze, die in Deutschland unter Beachtung des höheren Rechtes vom Parlament beschlossen werden. In diese Kategorie fallen etwa das Strafgesetzbuch oder das Bürgerliche Gesetzbuch. Rechtsverordnungen sind rangniedriger und werden von Ministerien und ihnen nachgeordneten Behörden erlassen. In der Pyramide ganz unten stehen weitere Rechtsnormen. Dazu gehören beispielsweise Satzungen von Organisationen und Richtlinien.

1.7.1 Menschenwürde und Privatsphäre

Im Recht der Europäischen Union ist Artikel 8 Absatz 1 der Konvention zum Schutz der Menschenrechte und Grundfreiheiten für die Psychologische Diagnostik relevant.

Europäische Menschenrechtskonvention

Artikel 8 Recht auf Achtung des Privat- und Familienlebens

(1) Jedermann hat Anspruch auf Achtung seines Privat- und Familienlebens, seiner Wohnung und seines Briefverkehrs.

Schutz von Privat- und Familienleben

In Artikel 8 Absatz 1 der Europäischen Menschenrechtskonvention wird explizit das Privat- und Familienleben geschützt. Mit bestimmten diagnostischen Verfahren kann man Informationen erlangen, die Betroffene nicht preisgeben möchten. Man denke an verdeckte Videoaufnahmen im Rahmen einer Verhaltensbeobachtung oder an projektive Verfahren. Auch die Weitergabe von persönlichen Informationen an andere kann problematisch sein (s. auch ► Abschn. 1.7.2).

Im Grundgesetz sind 2 Werte genannt, die für die Psychologische Diagnostik unmittelbar relevant sind: der Schutz der Menschenwürde und das Recht auf freie Entfaltung der Persönlichkeit.

Grundgesetz (GG) für die Bundesrepublik Deutschland

Artikel 1

(1) Die Würde des Menschen ist unantastbar. Sie zu achten und zu schützen ist Verpflichtung aller staatlichen Gewalt ...

Artikel 2

(1) Jeder hat das Recht auf die freie Entfaltung seiner Persönlichkeit, soweit er nicht die Rechte anderer verletzt und nicht gegen die verfassungsmäßige Ordnung oder das Sittengesetz verstößt.

(2) Jeder hat das Recht auf Leben und körperliche Unversehrtheit. Die Freiheit der Person ist unverletzlich. In diese Rechte darf nur auf Grund eines Gesetzes eingegriffen werden.

Neutrale Formulierungen wählen

Artikel 1 Absatz 1 gebietet etwa, in einem Gutachten keine herabsetzenden Formulierungen über die untersuchte Person zu machen. Der Sachverhalt kann stattdessen mit neutralen Formulierungen beschrieben werden. Also: „Die Intelligenz Herrn Müllers ist im Vergleich zu anderen Erwachsenen seines Alters weit unterdurchschnittlich ausgeprägt“ und nicht: „Herr Müll-

ler ist ein Idiot.“ Der Begriff „Idiot“ war übrigens früher ein anerkannter Fachbegriff, ist heute aber eine herabsetzende Bezeichnung. Aus Artikel 2 Absatz 1 wurde das Recht auf informationelle Selbstbestimmung hergeleitet. Psychologinnen und Psychologen wirken unter Umständen durch diagnostische Untersuchungen daran mit, dass die in Artikel 2 Absatz 2 garantierte Freiheit der Person eingeschränkt wird. Dies ist etwa der Fall, wenn über eine Sicherheitsverwahrung oder über eine Zwangseinweisung entschieden wird. Hier ist besonders darauf zu achten, dass dieser Eingriff aufgrund entsprechender Gesetze zulässig ist.

1.7.2 Geheimnisse, Schweigepflicht und Datenschutz

Im Strafgesetzbuch (StGB) sind die Paragrafen zum Geheimnisverrat für die Tätigkeit von Psychologinnen und Psychologen bedeutsam.

§ 203 StGB Verletzung von Privatgeheimnissen

- (1) Wer unbefugt ein fremdes Geheimnis, namentlich ein zum persönlichen Lebensbereich gehörendes Geheimnis oder ein Betriebs- oder Geschäftsgeheimnis, offenbart, das ihm als
[...]

(2) Berufspsychologen mit staatlich anerkannter wissenschaftlicher Abschlussprüfung

[weitere Berufsgruppen werden genannt]

anvertraut worden oder sonst bekanntgeworden ist, wird mit Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafe bestraft.

Die Verletzung von Privatgeheimnissen ist also alles andere als ein Bagatelldelikt. Im Extremfall kann man für ein Jahr ins Gefängnis geschickt werden, wenn man persönliche Informationen, die man etwa im Interview oder durch eine Testuntersuchung gewonnen hat, unbefugt weitergibt. Dabei ist unter einem Geheimnis „jede Information zu verstehen, die nur einer beschränkten Anzahl Personen bekannt ist und an deren Geheimhaltung der Betroffene ein persönliches oder wirtschaftliches Interesse hat“ (Joussen 2004, S. 87). Dies inkludiert in aller Regel die im Rahmen von diagnostischen Untersuchungen gewonnenen Erkenntnisse. Wenn durch deren Weitergabe ein Schaden entsteht, beispielsweise jemand deshalb seine Anstellung verliert, können zudem zivilrechtliche Forderungen folgen.

Verletzung von Privatgeheimnissen

Schweigepflicht

Bei der Einhaltung der Schweigepflicht sind einige Details zu beachten (ausführlich: Joussen 2004):

- Nicht geschützt sind Geheimnisse, die Berufspsychologinnen und -psychologen im privaten Bereich anvertraut werden. Die Schweigepflicht bezieht sich auf die Ausübung der Berufstätigkeit.
 - „Offenbaren“ bedeutet, dass eine Identifizierung der betroffenen Person möglich ist. Wer also Daten in anonymisierter Form weitergibt, offenbart kein Geheimnis.
 - Die Schweigepflicht gilt auch gegenüber Personen, die selbst der Schweigepflicht unterliegen (Kolleginnen bzw. Kollegen, Ärztinnen bzw. Ärzten etc.).
 - Zulässig ist die Weitergabe persönlicher Informationen, wenn Betroffene dem zustimmen. Es genügt die mündliche oder sogar die stillschweigende Einwilligung. Angesichts der Konsequenzen, die bei einer Klage drohen, kann eine schriftliche Erklärung sinnvoll sein.

Geheimnisse und Schweigepflicht

- Auch Kinder werden durch die Schweigepflicht geschützt. Da die Eltern auch ein Informationsrecht haben, sind im Einzelfall Schweigepflicht und Informationsrecht gegeneinander abzuwägen.
- Vor Gericht besteht in zivilrechtlichen Prozessen ein Zeugnisverweigerungsrecht. Berufspsychologinnen und -psychologen haben das Recht, Aussagen über ihnen anvertraute Geheimnisse zu verweigern. In Strafprozessen besteht dieses Schweigerecht nur für psychologische Psychotherapeutinnen bzw. -therapeuten und Kinder- und Jugendlichenpsychotherapeutinnen bzw. -therapeuten und bei ihnen auch nur für Informationen, die sie im Rahmen einer Untersuchung oder Heilbehandlung erfahren haben.

Datenschutzgrundverordnung

Die Datenschutzgrundverordnung (DSGVO) der Europäischen Union regelt die Verantwortlichkeiten und Pflichten bei der Sammlung und Speicherung personenbezogener Daten. Unter personenbezogenen Daten versteht diese Verordnung „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen“. Sofern Psychologische Diagnostik nicht zu Forschungszwecken durchgeführt wird, werden Daten meist so gespeichert, dass die betreffenden Personen identifizierbar sind. In diesen Fällen sind die Forderungen der DSGVO zu beachten. Diese sind nachfolgend auszugsweise aufgelistet und werden ausführlicher in ▶ Kap. 4.5.1 behandelt.

Einige Forderungen der DSGVO

Personenbezogenen Daten

- dürfen nur in einem Maß erhoben werden, wie es der Zweck der Erhebung erfordert.
- müssen vor unrechtmäßiger Verarbeitung oder Nutzung geschützt werden.
- dürfen nur mit Einwilligung der betreffenden Person erhoben und gespeichert werden, es sei denn andere wichtige Gründe erfordern die Datenerhebung und -speicherung (z. B. zum Schutz lebenswichtiger Interessen Dritter).
- dürfen nur nach ausführlicher Information der betreffenden Personen erhoben und gespeichert werden (Details regelt Artikel 13 der DSGVO).

Betroffene Personen haben ein

- Auskunftsrecht (z. B. über Verarbeitungszweck und Dauer der Speicherung),
- Recht auf Berichtigung,
- Recht auf Löschung,
- Recht auf Einschränkung der Verarbeitung (unter bestimmten Randbedingungen, s. Artikel 18 der DSGVO),
- Widerspruchsrecht.

Internationale Richtlinien für die Testanwendung der International Test Commission

Auch die „Internationalen Richtlinien für die Testanwendung“ der International Test Commission (ITC 2013) fordern die vertrauliche Behandlung von diagnostischen Erkenntnissen, genauer von Testergebnissen. Eine deutsche Fassung dieser Richtlinien entstand in Zusammenarbeit mit dem Berufsverband Deutscher Psychologinnen und Psychologen e. V. (BDP; ZPID 2001). Zur sicheren Verwahrung und vertraulichen Behandlung von Testmaterial heißt es dort:

Sichere Verwahrung und vertrauliche Behandlung von Testmaterial gemäß International Test Commission (ITC 2013; Auszüge)

Fachkompetente Testanwendende...

Gewährleisten die sichere Verwahrung von Testmaterial

- Stellen sicher, dass Testmaterial sicher verwahrt wird, und kontrollieren dessen Verfügbarkeit.

Gewährleisten die vertrauliche Behandlung von Testergebnissen

- Spezifizieren, wer Zugang zu Testergebnissen hat, und definieren Abstufungen dieser Datensicherung.
- Erläutern den Probandinnen und Probanden die Abstufungen der Datensicherung, bevor Tests vorgegeben werden.
- Gewähren nur berechtigten Personen Zugang zu Testergebnissen.
- Holen die entsprechenden Einverständniserklärungen ein, bevor sie Ergebnisse an andere weitergeben.
- Schützen Daten in Akten, so dass sie nur für befugte Personen zugänglich sind.
- Stellen klare Richtlinien auf, wie lange Testdaten in Akten aufbewahrt werden sollen.
- Entfernen Namen und andere identifizierende Daten aus Datensammlungen von Testergebnissen, die archiviert oder für Forschungszwecke, zur Normierung oder für andere statistische Zwecke verwendet werden.

1.7.3 Offenbarungspflicht

Unter bestimmten Bedingungen besteht eine Offenbarungspflicht. Wer von bestimmten Straftaten erfährt, die geplant sind oder gerade ausgeführt werden, kann mit bis zu 5 Jahren Haft bestraft werden, wenn er diese Kenntnisse nicht offenbart. Dieses Gesetz betrifft nicht nur Berufspsychologinnen und Berufspsychologen, sondern ist generell gültig.

Offenbarungspflicht

Offenbarungspflicht laut Strafgesetzbuch (StGB)

§ 138 StGB Nichtanzeige geplanter Straftaten

(1) Wer von dem Vorhaben oder der Ausführung

1. (*weggefalen*)
2. eines Hochverrats [...],
3. eines Landesverrats oder einer Gefährdung der äußeren Sicherheit [...],
4. einer Geld- oder Wertpapierfälschung [...] oder einer Fälschung von Zahlungskarten mit Garantiefunktion und Vordrucken für Euroschecks [...],
5. eines Mordes [...] oder Totschlags [...] oder eines Völkermordes [...] oder eines Verbrechens gegen die Menschlichkeit [...] oder eines Kriegsverbrechens [...],
6. einer Straftat gegen die persönliche Freiheit [...] soweit es sich um Verbrechen handelt [...],
7. eines Raubes oder einer räuberischen Erpressung [...] oder
8. einer gemeingefährlichen Straftat [...]

zu einer Zeit, zu der die Ausführung oder der Erfolg noch abgewendet werden kann, glaubhaft erfährt und es unterlässt, der Behörde oder dem Bedrohten rechtzeitig Anzeige zu machen, wird mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe bestraft.

(2) Ebenso wird bestraft, wer

1. von der Ausführung einer Straftat nach § 89a [Vorbereitung einer schweren staatsgefährdenden Gewalttat] oder
2. von dem Vorhaben oder der Ausführung einer Straftat nach § 129a [Bildung terroristischer Vereinigungen] [...]

zu einer Zeit, zu der die Ausführung noch abgewendet werden kann, glaubhaft erfährt und es unterlässt, der Behörde unverzüglich Anzeige zu erstatten. [...]

(3) Wer die Anzeige leichtfertig unterlässt, obwohl er von dem Vorhaben oder der Ausführung der rechtswidrigen Tat glaubhaft erfahren hat, wird mit Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafe bestraft.

Offenbarungspflicht, wenn Straftat noch abzuwenden ist

Entscheidend bei der Offenbarungspflicht ist, dass die Straftat, von der man erfährt, noch abgewendet werden kann. Berichtet beispielsweise eine Klientin oder ein Klient, dass sie bzw. er gerade jemanden im Affekt umgebracht hat, so besteht keine Offenbarungspflicht.

1.7.4 Rechtliche Regelungen für spezifische Anwendungsfelder der Psychologischen Diagnostik

Psychologische Diagnostik findet in ganz unterschiedlichen Kontexten statt und unterliegt damit den für diese Kontexte jeweils spezifischen rechtlichen Rahmenbedingungen.

■ Psychotherapie

So gilt für Psychologische Diagnostik im Kontext einer Psychotherapie das Gesetz über die Berufe des Psychologischen Psychotherapeuten und des Kinder- und Jugendlichenpsychotherapeuten (kurz: Psychotherapeutengesetz). Dieses Gesetz regelt auch (in § 1 (2)), dass eine „mittels wissenschaftlich anerkannter psychotherapeutischer Verfahren vorgenommene Tätigkeit zur Feststellung [...] von Störungen mit Krankheitswert“ (also Psychologische Diagnostik) eine Ausübung von Psychotherapie darstellt und damit Psychologischen Psychotherapeuten vorbehalten ist.

■ Eignungsdiagnostik und berufliche Leistungsbeurteilung

Im Kontext der Eignungsdiagnostik und der beruflichen Leistungsbeurteilung ist das Betriebsverfassungsgesetz (BetrVG) relevant. Darin werden diagnostische Verfahren und allgemeine Beurteilungsgrundsätze direkt angeprochen:

§ 94 BetrVG Personalfragebögen, Beurteilungsgrundsätze

- (1) Personalfragebögen bedürfen der Zustimmung des Betriebsrats. [...]
- (2) Absatz 1 gilt entsprechend [...] für die Aufstellung allgemeiner Beurteilungsgrundsätze.

Betriebsverfassungsgesetz

Personalfragebogen

Unter einem Personalfragebogen ist keineswegs ein Persönlichkeitstest zu verstehen, sondern Fragen zur Person, die etwa den Familienstand, das bisherige Einkommen, Krankheiten etc. betreffen. In der Rechtsprechung ist im Einzelnen geklärt, welche Fragen dabei überhaupt zulässig sind und dass Bewerberinnen und Bewerber unzulässige Fragen (auch solche im Einstellungsgespräch) nicht wahrheitsgemäß beantworten müssen. Unzulässig ist beispielsweise die Frage nach dem Vorliegen einer Schwangerschaft oder einer Gewerkschaftsmitgliedschaft. Persönlichkeitsfragebögen sind zulässige

Einleitung

Verfahren, soweit sie helfen, die Eignung von Bewerberinnen und Bewerbern festzustellen und diese nicht unangemessen ausforschen. Entscheidend ist, dass Arbeitgeber nicht erfahren, wie eine Bewerberin bzw. ein Bewerber die einzelnen Fragen beantwortet. Wenn eine Psychologin bzw. ein Psychologe die Auswertung vornimmt und nur das Ergebnis mitteilt, handelt es sich definitiv nicht um einen Personalfragebogen (von Hoyningen-Huene 1997).

Beurteilungsgrundsätze sind allgemeine Grundsätze, nach denen alle Bewerberinnen und Bewerber oder auch bereits eingestellte Personen in fachlicher oder persönlicher Hinsicht beurteilt werden. Unter Auswahlrichtlinien versteht man üblicherweise abstrakt formulierte Regeln. Darin wird festgelegt, welche Voraussetzungen bei Bewerberinnen und Bewerbern vorliegen müssen oder nicht vorliegen dürfen. Dabei kommen fachliche, persönliche und soziale Kriterien infrage.

Beurteilungsgrundsätze

§ 95 BetrVG Auswahlrichtlinien

- (1) Richtlinien über die personelle Auswahl bei Einstellungen, Versetzungen, Umgruppierungen und Kündigungen bedürfen der Zustimmung des Betriebsrats.
[...]
- (2) In Betrieben mit mehr als 500 Arbeitnehmern kann der Betriebsrat die Aufstellung von Richtlinien über die bei Maßnahmen des Absatzes 1 Satz 1 zu beachtenden fachlichen und persönlichen Voraussetzungen und sozialen Gesichtspunkte verlangen. [...]

■ Verkehrspsychologische Diagnostik

Im Bereich der verkehrspsychologischen Diagnostik finden Begutachtungen der Kraftfahreignung statt (► Abschn. 9.3.1). Hierzu sind die Begutachtungsleitlinien der Bundesanstalt für Straßenwesen relevant. Sie regeln beispielsweise die notwendige Qualifikation der Gutachterinnen und Gutachter sowie die begutachtungsrelevanten Fähigkeitsaspekte (z. B. Konzentrationsfähigkeit). Diese Leitlinien sind unter ► <https://www.bast.de/> einsehbar.

Begutachtungsleitlinien der Bundesanstalt für Straßenwesen

■ Internationale Regelungen

Neben den rechtlichen Rahmenbedingungen in den hier genannten und weiteren Anwendungsbereichen muss auch beachtet werden, dass in manchen Fällen Psychologische Diagnostik über Länder und damit über juristische Geltungsbereiche hinweg vorgenommen wird. So müssen internationale Konzerne bei der Gestaltung der Personalauswahl ggf. bedenken, dass diese in allen Ländern, für deren Niederlassungen die Personalauswahl stattfindet, rechtssicher erfolgt. Dies ist nicht trivial. So gelten für Personalauswahlverfahren in den USA deutlich strengere Vorgaben als in Deutschland.

Diagnostik über juristische Geltungsbereiche hinweg

1.7.5 Ethische Richtlinien

Ethische Richtlinien haben die Deutsche Gesellschaft für Psychologie e.V. (DGPs) und der BDP herausgegeben. Für den BDP dienen sie auch als Berufsordnung. Diese Richtlinien sollen verbindliche Regeln für das professionelle Verhalten von Psychologinnen und Psychologen vorgeben. Sie können unter ► <https://www.dgps.de/> eingesehen werden. Hier wird nur auf die Passagen Bezug genommen, die für die Diagnostik besonders relevant sind.

Ethische Richtlinien der deutschen Psychologenverbände

Unter der Überschrift „Umgang mit Daten“ finden sich mit Verweis auf § 203 des Strafgesetzbuchs (► Abschn. 1.7.2) Hinweise auf die Einhaltung der Schweigepflicht sowie zum Umgang mit Daten. Ein eigener Abschnitt befasst sich mit Gutachten und Untersuchungsberichten.

Bei der Erstellung von Gutachten und Untersuchungsberichten ist zu beachten:

- Sorgfaltspflicht: Sachliche und wissenschaftliche Fundiertheit sowie Sorgfalt und Gewissenhaftigkeit.
- Transparenz für Adressaten: Für die Adressatin bzw. den Adressaten sind das Gutachten oder der Bericht inhaltlich nachvollziehbar.
- Einsichtnahme gewähren: Einsichtnahme durch die Klientin bzw. den Klienten ermöglichen bzw. darauf hinwirken. Wenn keine Einsichtnahme möglich sein sollte, vorab darüber informieren.
- Keine Gefälligkeitsgutachten.
- Ebenso sind Gutachten nicht zulässig, die ohne eigene Mitwirkung zu stande gekommen sind.

Psychologinnen und Psychologen haben eine besondere Verantwortung gegenüber ihren Klientinnen und Klienten. Konkret werden in diesem Zusammenhang verlangt:

- Vertrauensverhältnis: Wenn das Vertrauensverhältnis gestört ist, können Psychologinnen und Psychologen einen Auftrag ablehnen oder beenden. Haben Klientinnen bzw. Klienten nicht selbst den Auftrag erteilt (beispielsweise bei forensischen Fragestellungen), besteht eine besondere Verpflichtung, im wohlverstandenen Interesse aller Beteiligten zu handeln.
- Aufklärung und Einwilligung: Klientinnen und Klienten über alle wesentlichen Maßnahmen unterrichten und Einwilligung dazu einholen.

Besondere Verantwortung gegenüber den eigenen Klienten

Ehrengericht

Bei Verstößen gegen die ethischen Richtlinien kann das Ehrengericht einer der beiden Berufsverbände eingeschaltet werden. Dieses kann im Extremfall den Ausschluss aus dem Berufsverband beschließen (die Mitgliedschaft im Berufsverband ist freiwillig, daher führt ein Ausschluss nicht zu einem Berufsverbot o. Ä.).

Die gemeinsamen Ethischen Richtlinien der DGPs und des BDP sind zum Teil an denen des großen amerikanischen Berufsverbandes American Psychological Association (APA); ► <https://www.apa.org/ethics/code/index>) angelehnt. Diese Richtlinien sind detaillierter und enthalten auch weitergehende Forderungen wie etwa die nach Beachtung der eigenen Kompetenzen beim Anbieten von Dienstleistungen, die Aufrechterhaltung und die Weiterentwicklung dieser Kompetenzen oder die Minimierung des Eindringens in die Privatsphäre.

Auch die bereits angesprochenen „Internationalen Richtlinien für die Testanwendung“ der International Test Commission können als ethische Richtlinien verstanden werden. Darin wird Testanwenderinnen und -anwendern die Verantwortung für eine ethisch korrekte Testanwendung zugesprochen.

Ethisch korrekte Testanwendung gemäß International Test Commission (ITC 2013; Auszüge)

Fachkompetente Testanwendende ...

- Stellen sicher, dass die mit ihnen oder für sie arbeitenden Personen die angemessenen professionellen und ethischen Standards einhalten.
- Beachten in der Kommunikation mit Testpersonen und anderen Beteiligten in gebührender Weise deren Empfindlichkeiten.
- Vermeiden Situationen, in denen sie selbst möglicherweise ein berechtigtes Interesse am Ergebnis des Tests haben oder zu haben scheinen oder in denen der Test die Beziehung zu ihrer Klientin bzw. ihrem Klienten schädigen könnte.
- Arbeiten auf der Grundlage und innerhalb der Grenzen wissenschaftlicher Prinzipien und empirischer Befunde.

Einleitung

- Wissen um die Grenzen ihrer eigenen Kompetenz und handeln innerhalb dieser.
- Stellen den am Testprozess Beteiligten klare und angemessene Informationen über die ethischen Prinzipien und rechtlichen Bestimmungen zur Verfügung, die psychologische Tests regeln.

Seit einiger Zeit nutzen manche Bereiche der psychologischen Diagnostik Daten, die durch sog. „neue Technologie“ erhoben werden. So können Bewegungsdaten, die eine App auf dem Smartphone von Studienteilnehmenden sammelt, Aufschluss über deren Alltagsaktivität geben. Die technologiebasierte Sammlung von Daten stellt jedoch, sofern nicht verantwortungsvoll vorgenommen, einen erheblichen Eingriff in die Privatsphäre von Personen dar. Man stelle sich dazu nur vor, Forschende hätten über entsprechende Apps jederzeit Zugriff auf den Standort einer Person. Um die Möglichkeiten, die das Sammeln von Daten über neue Technologien bietet, zu nutzen und gleichzeitig verantwortungsvoll mit der Privatsphäre von Personen umzugehen, hat der Rat für Sozial- und Wirtschaftsdaten entsprechende Empfehlungen formuliert (RatSDW 2020).

Empfehlungen des Rats für Sozial- und Wirtschaftsdaten zum angemessenen Umgang mit Daten aus neuen Informationstechnologien (RatSDW 2020; Auszüge)

- Daten sollten nur auf sicherem Wege gespeichert, verarbeitet und transferiert werden, z. B. unter Rückgriff auf Verschlüsselungstechniken.
- Von der Speicherung von Daten, welche die Reidentifizierung von Personen relativ leicht ermöglicht (z. B. Ortsinformationen) ist abzusehen, wenn sie nicht zur Beantwortung der konkreten Fragestellung benötigt werden.
- Daten (wie z. B. Ortsdaten, Audiodaten) können weniger präzise oder bereits weiterverarbeitet (z. B. durch Behavioral Signal Processing) abgespeichert und die Rohdaten noch auf dem Gerät gelöscht werden, um die Möglichkeit der Reidentifikation zu verringern.
- Daten mit identifizierbaren Informationen (z. B. volle Namen, Adressen, Kontonummern), die bspw. mittels Audio- oder Videoaufnahmen erhoben werden, sollten im Kodierprozess/ weiteren Verlauf gelöscht bzw. herausdientiert werden.
- Personen, die sich im regelmäßigen Kontakt mit den an einer Untersuchung teilnehmenden Personen befinden (z. B. Familie, befreundete Personen, Arbeitskolleginnen und -kollegen) und von denen Daten erhoben werden, sind über die Datenerhebung zu informieren und ihre Zustimmung ist einzuholen.
- Das Aufzeichnungsrisiko von Daten Dritter sollte sichtbar gemacht werden, um die Privatheitserwartung bei Drittpersonen zu minimieren.
- Für die Veröffentlichung von Daten, die eine Identifikation Dritter erlauben, ist es notwendig, deren Zustimmung einzuholen.
- Sofern möglich, sollten Untersuchungsteilnehmende private Daten (z. B. Audio- und Bildaufnahmen) löschen können, bevor die Forschungsdaten an die Forschenden weitervermittelt werden.
- Das Aufnahmegerät sollte die Möglichkeit einer proaktiven Zensur, z. B. in Form eines Privacy-Buttons bieten, um die Datenaufnahme zu unterbrechen.

Die vollständigen Empfehlungen sind unter ► <https://doi.org/10.17620/02671.47> abzurufen.

(Abdruck mit freundlicher Genehmigung des Rats für Sozial- und Wirtschaftsdaten)

Weiterführende Literatur und Internetressourcen

Zur weiteren Vertiefung in das Thema „Rechtsfragen psychologischer Diagnostik“ eignen sich besonders die Bücher von Joussen (2004) und Zier (2002).

Hilfreiche Internetressourcen:

- Zur Relevanz von Psychologischer Diagnostik in psychologischen Berufsbildern:
► <https://www.onetonline.org/>
- Leitlinien zur verkehrspychologischen Diagnostik: ► <https://www.bast.de/>
- Kompetenzen zur Testanwendung in den Richtlinien der International Test Commission: ► https://www.intestcom.org/files/guideline_test_use.pdf sowie ► https://www.psindex.de/pub/tests/ite_richtlinien.pdf (deutsche Fassung)
- Ethische Richtlinien der APA und der DGPs: ► <https://www.apa.org/ethics/code/index> und ► <https://www.dgps.de/>

1.8 Zusammenfassung

Psychologische Diagnostik befasst sich mit der Beantwortung von Fragestellungen, die sich auf die Beschreibung, Klassifikation, Erklärung oder Vorhersage menschlichen Verhaltens und Erlebens beziehen. Im Rahmen des diagnostischen Prozesses kommen Methoden zum Einsatz, die wissenschaftlichen Standards genügen. Als diagnostischer Prozess wird die Abfolge von Maßnahmen zur Gewinnung diagnostisch relevanter Informationen und deren Integration zur Beantwortung einer Fragestellung bezeichnet. Ganz allgemein gesagt bestehen die Ziele der Psychologischen Diagnostik im Beschreiben, Klassifizieren, Erklären und Vorhersagen.

Praktisch tätige Psychologinnen und Psychologen verwenden einen substantiellen Teil ihrer Arbeitszeit auf Tätigkeiten, die der Psychologischen Diagnostik zuzurechnen sind. Viele Fragestellungen aus Anwendungsfeldern der Psychologie (z. B. Klinische Psychologie, Gesundheitspsychologie) lassen sich ohne fundierte Psychologische Diagnostik nicht beantworten. Umgekehrt gilt: Die Psychologische Diagnostik muss sich Erkenntnissen aus diesen Anwendungsfeldern zunutze machen, um gute Methoden und Vorgehensweisen bereitzustellen. Zudem gilt es, die für die jeweiligen Anwendungsbereiche geltenden Gesetze und Vorschriften, sowie insgesamt Richtlinien des professionellen ethischen Handelns zu beachten.

?

Übungsfragen

- **Abschn. 1.1:**
 - Was sind zentrale Elemente der Definition von Psychologischer Diagnostik?
 - Wie ist Psychologische Diagnostik von Testen, medizinischer Diagnostik und Evaluation abgrenzen?
- **Abschn. 1.2:**
 - Warum bedarf es vor psychologischen Interventionen einer Psychologischen Diagnostik?
 - Nennen Sie typische diagnostische Fragestellungen in psychologischen Anwendungsfeldern!
- **Abschn. 1.3:**
 - Wie profitiert Psychologische Diagnostik von Grundlagendisziplinen der Psychologie? Wie profitieren Grundlagendisziplinen der Psychologie von Psychologischer Diagnostik?
- **Abschn. 1.4:**
 - Welche allgemeinen Ziele verfolgt die Psychologische Diagnostik?
 - Wovon gehen interaktionistische Ansätze zur Erklärung (und Prognose) von Verhalten aus?
- **Abschn. 1.5:**
 - Was versteht man unter einem diagnostischen Prozess?
 - Welche wesentlichen Teilschritte beinhaltet ein diagnostischer Prozess?

Einleitung

- Bei welchen Gründen sollte die Beantwortung einer Fragestellung abgelehnt werden?
- **Abschn. 1.6:**
 - Welche Relevanz haben soziale Medien für die Psychologische Diagnostik?
- **Abschn. 1.7:**
 - Welche im Grundgesetz verankerten Werte sind für die Psychologische Diagnostik unmittelbar relevant?
 - Wie ist die Schweigepflicht gesetzlich verankert und welche Details sind im Umgang damit zu beachten?
 - Was sind zentrale, für die Psychologische Diagnostik relevante Forderungen der Datenschutzgrundverordnung?
 - Was versteht man unter der Offenbarungspflicht?
 - Nennen Sie aus den Ethischen Richtlinien der Deutschen Gesellschaft für Psychologie e.V. (DGPs) einzelne Forderungen zur Erstellung von Gutachten!

Literatur

- American Psychiatric Association. (2018). *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5. Deutsche Ausgabe herausgegeben von Peter Falkai und Hans-Ulrich Wittchen* (2. Aufl.). Göttingen: Hogrefe.
- Bundesanstalt für Straßenwesen. (2019). Anzahl der medizinisch-psychologischen Untersuchungen (MPU) in Deutschland 2003 bis 2018. ► <https://de.statista.com/statistik/daten/studie/76153/umfrage/anzahl-der-mpu-idiotentest-seit-2003/>. Zugegriffen: 05. Juni 2020.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Oxford, England: University of Illinois Press.
- Deutsches Institut für Medizinische Dokumentation und Information (DIMDI). (2019). Von der ILCD zur ICD-10. ► <https://www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-who/historie/ilcd-bis-icd-10/index.html>. Zugegriffen: 18. März 2020.
- Dilling, H., Freyberger, H. J., & Cooper, J. E. (Hrsg.) (2010). *Taschenführer zur ICD-10-Klassifikation psychischer Störungen* (5. Aufl.). Bern: Huber.
- van Drunen, P. (1993). Von der Psychotechnik zur Psychodiagnostik. In H. E. Lück & R. Miller (Hrsg.), *Illustrierte Geschichte der Psychologie* (S. 254–256). München: Quintessenz.
- Eid, M., & Petermann, F. (2006). Aufgaben, Zielsetzungen und Strategien der Psychologischen Diagnostik. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 15–25). Göttingen: Hogrefe.
- Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin* 83, 956–974.
- Fissen, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik* (3. Aufl.). Göttingen: Hogrefe.
- Funder, D. C. (2016). Taking situations seriously: The situation construal model and the riverside situational Q-sort. *Current Directions in Psychological Science* 25, 203–208.
- Gesellschaft für Evaluation (2008). *Standards für Evaluation* (4. Aufl.). Mainz: Gesellschaft für Evaluation.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Boston: Pearson.
- von Hoyningen-Huene, G. (1997). *Der psychologische Test im Betrieb: Rechtsfragen für die Praxis*. Heidelberg: Sauer-Verlag.
- International Test Commission (ITC). (2013). ITC Guidelines on Test Use. 8th October, 2013, Version 1.2. Final Version. Document reference: ITC-G-TU-20131008. ► https://www.intest-com.org/files/guideline_test_use.pdf. Zugegriffen: 15. April 2020.
- Jäger, A. O., Breetz, E., Erler, R., & Habersang-Walther, R. (1982). *Mannheimer Schuleingangsdiagnosikum (MSD)*. Göttingen: Hogrefe.
- Joussen, J. (2004). *Berufs- und Arbeitsrecht für Diplom-Psychologen*. Göttingen: Hogrefe.
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 543–556.
- Lamberti, G. (2006). *Intelligenz auf dem Prüfstand: 100 Jahre Psychometrie*. Göttingen: Vandenhoeck & Ruprecht.
- Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID). (2001). Internationale Richtlinien für die Testanwendung, Version 2000. Deutsche Fassung. ► https://www.zpid.de/pub/tests/itec_richtlinien.pdf. Zugegriffen: 15. April 2020.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.
- Lück, H. E., & Miller, R. (1993). *Illustrierte Geschichte der Psychologie*. München: Quintessenz.

- Matarazzo, J. D. (1980). Behavioral health and behavioral medicine: Frontiers for a new health psychology. *American Psychologist* 35, 807–817.
- Meyer, C. S., Arx, P. H.-v., Lemola, S., & Grob, A. (2010). Correspondence between the general ability to discriminate sensory stimuli and general intelligence. *Journal of Individual Differences* 31, 46–56.
- Oakland, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International Test Commission: Its history, current status, and future directions. *International Journal of Testing* 1, 3–32.
- Occupational Information Network (O*NET). (2019). Summary Report for: 19-3031.02 – Clinical Psychologists. ► <https://www.onetonline.org/link/summary/19-3031.02>. Zugegriffen: 14. April 2020.
- Pawlak, K. (2006). *Handbuch Psychologie: Wissenschaft – Anwendung – Berufsfelder*. Berlin, Heidelberg: Springer.
- Pawlak, K., & Buse, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. *Zeitschrift für Differentielle und Diagnostische Psychologie* 3, 101–118.
- Rat für Sozial- und Wirtschaftsdaten (RatSWD). (2020). Datenerhebung mit neuer Informati-onstechnologie. Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz. RatSWD Output 6(6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). doi: ► <https://doi.org/10.17620/02671.47>. ► https://www.ratswd.de/dl/RatSWD_Output6.6_Datenerhebung-neueIT.pdf. Zugegriffen: 09. April 2020.
- Rauthmann, J. F., & Sherman, R. A. (2015). Measuring the Situational Eight DIAMONDS Char-acteristics of Situations. *European Journal of Psychological Assessment* 32, 155–164.
- Rauthmann, J. F. & Sherman, R. A. (2017). S8* – Situational Eight DIAMONDS – deutsche Fassung [Fragebogen]. In Leibniz-Zentrum für Psychologische Information und Dokumen-tation (ZPID) (Hrsg.), *Elektronisches Testarchiv (PSYNDEX Tests-Nr. 9007478)*. Trier: ZPID.
- Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology* 107, 677–718.
- Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology* 7, 331–363.
- Rome, H. P., Swenson, W. M., Mataya, P., McCarthy, C. E., Pearson, J. S., Keating, F. R., & Hat-haway, S. R. (1962). Symposium on automation techniques in personality assessment. Paper presented at the Proceedings of the Staff Meetings of the Mayo Clinic.
- Rost, D. H. (2009). *Intelligenz: Fakten und Mythen*. Weinheim, Basel: Beltz.
- Roth, M. & Herzberg, P. Y. (2008). Psychodiagnostik in der Praxis: State of the Art? *Klinische Diagnostik und Evaluation* 1, 5–18.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psycholo-gical Bulletin* 124, 262–274.
- Schuler, H., Hell, B., Trapmann, S., Schaar, H., & Boramir, I. (2007). Die Nutzung psycholo-gischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie* 6, 60–70.
- Stanat, P., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., & Weiß, M. (2002). PISA 2000: Die Studie im Überblick. Grundlagen, Methoden und Ergebnisse. ► https://www.mpib-berlin.mpg.de/Pisa/PISA_im_Ueberblick.pdf. Zugegriffen: 18. März 2020.
- Stemmler, G., Hagemann, D., Amelang, M., & Bartussek, (2010). *Differentielle Psychologie und Persönlichkeitsforschung* (7. Aufl.). Stuttgart: Kohlhammer.
- Tent, L. (2001). Nachruf auf Gustav A. Lienert. *Zeitschrift für Gerontologie und Geriatrie* 34, 242–244.
- Westhoff, K., & Graubner, J. (2003). Konstruktion eines komplexen Konzentrationstests. *Diagno-stica* 49, 110–119.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized ad-aptive testing: A primer*. New York: Routledge.
- Wang, Z.-M. (1993). Psychology in China: A review dedicated to Li Chen. *Annual Review of Psy-chology* 44, 87–116.
- World Health Organization (WHO). (2018). *International Statistical Classification of Diseases and Related Health Problems (ICD-11)*. Genf: World Health Organization.
- Zier, J. (2002). *Recht für Diplom-Psychologen: Eine Einführung*. Stuttgart: Kohlhammer.



Grundlagen diagnostischer Verfahren

Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang

Inhaltsverzeichnis

- 2.1 Allgemeines zu psychologischen Tests – 41**
 - 2.1.1 Was versteht man unter einem psychologischen Test? – 41
 - 2.1.2 Bandbreite psychologischer Tests – 43
 - 2.1.3 Rückschluss auf ein latentes Merkmal – 46
- 2.2 Die Klassische Testtheorie – 49**
 - 2.2.1 Zentrale Annahmen – 49
 - 2.2.2 Reliabilität von Messungen – 52
 - 2.2.3 Grenzen der Klassischen Testtheorie – 54
- 2.3 Item-Response-Theorien – 55**
 - 2.3.1 Item-Response-Theorien für dichotome Antwortformate – 57
 - 2.3.2 Item-Response-Theorien für ordinale Antwortformate – 76
 - 2.3.3 Item-Response-Theorien zur Klassifikation von Personen – 80
- 2.4 Konstruktionsprinzipien psychologischer Tests – 84**
 - 2.4.1 Ziel der Messung und Messgegenstand – 84
 - 2.4.2 Generieren von Testitems – 87
- 2.5 Grundzüge von Itemanalysen – 109**
 - 2.5.1 Itemschwierigkeit (nach der Klassischen Testtheorie) – 109
 - 2.5.2 Itemstreuung bzw. Itemvarianz – 112
 - 2.5.3 Trennschärfe – 113
 - 2.5.4 Itemladungen auf Faktoren – 118
 - 2.5.5 Itemvalidität – 126
 - 2.5.6 Itemanalysen nach Probabilistischen Testtheorien – 126
- 2.6 Testgütekriterien – 132**
 - 2.6.1 Objektivität – 133
 - 2.6.2 Reliabilität – 138
 - 2.6.3 Validität – 157

2.6.4 Nebengütekriterium: Normierung – 182

2.6.5 Weitere Nebengütekriterien – 190

2.7 Zusammenfassung – 198

Literatur – 201

Dieses Kapitel befasst sich mit grundlegenden Anforderungen an diagnostische Instrumente. Der Schwerpunkt liegt dabei auf psychologischen Tests. Diese stellen zwar nur eine von mehreren Möglichkeiten dar, diagnostisch relevante Informationen über Personen zu erheben. Im Vergleich zu anderen Möglichkeiten, beispielsweise Interviews oder Verhaltensbeobachtungen, sind die grundlegenden Anforderungen an psychologische Tests deutlich genauer spezifiziert und formalisiert. Diese Anforderungen können jedoch – mit wenigen Einschränkungen – auf diagnostische Interviews und Verhaltensbeobachtungen übertragen werden.

Anforderungen an psychologische Tests

2.1 Allgemeines zu psychologischen Tests

2.1.1 Was versteht man unter einem psychologischen Test?

Der Begriff „Test“ ist schon lange in unsere Alltagssprache und unser Alltagsleben eingedrungen. Bevor wir einen neuen Staubsauger, ein Auto, ein Fernsehgerät oder vielleicht auch nur ein Haarwaschmittel kaufen, suchen wir nach einem Testbericht über dieses Produkt. Einige Zeitschriften befassen sich mit Produkttests, d. h., sie prüfen Autos, Handys oder Versicherungen hinsichtlich diverser Aspekte, die für Verbraucherinnen und Verbraucher kaufentscheidend sein könnten. Banken werden einem „Stresstest“ unterzogen, um ihre Funktionsfähigkeit unter widrigen Randbedingungen abzuschätzen. In der Apotheke kann man Tests kaufen, die eine Schwangerschaft, hohe Blutzuckerwerte oder Eiweiß im Urin erkennen. Dies alles sind natürlich keine psychologischen Tests.

Der Begriff „Test“

Definition

Für einen **psychologischen Test** gilt:

- Es handelt sich um eine Messmethode, bei der Personen auf standardisierte Reizvorlagen (Aufgaben, Fragen etc.) reagieren.
- Reaktionen werden durch die spezifischen, im Test realisierten Bedingungen hervorgerufen.
- Die Reaktionen erlauben einen wissenschaftlich begründbaren Rückschluss auf die individuelle Ausprägung eines psychologischen Merkmals (oder auch mehrere Merkmale).
- Das Vorgehen ist standardisiert.
- Ziel ist eine quantitative (Ausprägung des Merkmals) und/oder eine qualitative Aussage (Vorhandensein oder Art des Merkmals) über das psychologische Merkmal.

(a bis c zitiert nach Heidenreich 1993, S. 389)

Wie man sieht, werden psychologische Tests durch diese Definitionselemente deutlich von Tests im allgemeinsprachlichen Sinne abgegrenzt (lediglich Definitionselement d wird ggf. von anderen Tests, beispielsweise Produkttests, ebenfalls erfüllt). Von offensichtlich frei konstruierten Selbsterkenntnis- und anderen „psychologischen“ Tests, die man gelegentlich in populären Zeitschriften antrifft, lässt Definitionsmerkmal c eine Unterscheidung zu.

Abgrenzung zur allgemeinsprachlichen Bedeutung

Fragebögen als psychologische Tests

Hingegen können Fragebögen nach dieser Definition auch als psychologische Tests verstanden werden. Die standardisierte Reizvorlage besteht in der zu beurteilenden Aussage, die eine Reaktion in Form einer Zustimmung oder Ablehnung (als die eigene Person beschreibend) hervorruft (s. u.). Die Reaktion erfolgt durch Ankreuzen des entsprechenden Kästchens.

Ein wissenschaftlich begründeter Rückschluss von den Reaktionen – in folgendem Fragebogen ist dies die Zahl der als zutreffend markierten Kästchen – auf Eigenschaften der Person, im Beispiel etwa die Bekümmерtheit einer Person, muss für die jeweiligen Fragebögen nachgewiesen werden (s. auch ▶ Abschn. 2.1.3 und 2.6.3). Eine standardisierte Durchführung ist möglich. In der Regel erfolgt, wie unter Definitionselement e beschrieben, eine quantitative oder qualitative Aussage über ein psychologisches Merkmal. Wenn wir von psychologischen Tests ohne weitere Eingrenzung sprechen, beziehen wir also stets auch Fragebögen ein.

Reiz 1	„Ich sorge mich oft um meine Zukunft“	
Reaktion 1	<input checked="" type="checkbox"/> trifft zu	<input type="checkbox"/> trifft nicht zu
Reiz 2	„Ich bin häufig voller Sorgen über das was kommt“	
Reaktion 2	<input checked="" type="checkbox"/> trifft zu	<input type="checkbox"/> trifft nicht zu
Reiz 3	„Wenn ich an meine Zukunft denke, habe ich große Ängste“	
Reaktion 3	<input type="checkbox"/> trifft zu	<input checked="" type="checkbox"/> trifft nicht zu

Reaktion auf Reize

Diese Reaktionen sollen durch die spezifischen, im Test realisierten Bedingungen hervorgerufen werden, und nicht oder nur in geringem Ausmaß durch andere Bedingungen (s. Definitionselement b), beispielsweise störende Umwelteinflüsse. In der Regel wird die Reaktion durch eine präzise Instruktion (z. B. „Kreuzen Sie an, ob die Aussage auf Sie zutrifft oder nicht“) eingegrenzt. Eine systematische Beobachtung von Alltagsverhalten oder die Beurteilung von Verhalten in Assessment-Center-Übungen ist demnach nicht als Test anzusehen.

Messgegenstand

Das psychologische Merkmal, dessen Ausprägung durch einen Test beschrieben werden soll, wird auch als Messgegenstand des Tests bezeichnet. Messgegenstand können Persönlichkeitsmerkmale (einschließlich Intelligenz, Interessen, Motivation etc.), aber auch emotionales Erleben (Emotionen, Gefühle etc.), Beziehungen zwischen Menschen (etwa die Qualität einer Paarbeziehung) oder situative Merkmale (z. B. belastende Faktoren am Arbeitsplatz) sein. Es muss begründbar sein, warum konkrete Reaktionen im Test einen sinnvollen Rückschluss auf das intendierte psychologische Merkmal erlauben (s. Definitionselement c).

Standardisiertes Vorgehen

Das standardisierte Vorgehen ist ein wesentliches Merkmal von psychologischen Tests (Definitionselement d). Die Bedingungen für die Durchführung müssen genau spezifiziert sein, ebenso die Auswertung und Interpretation der Antworten bzw. Ergebnisse.

Die Quantifizierung eines Merkmals bedeutet, dass die Ausprägung des intendierten Merkmals durch eine Zahl ausgedrückt wird. Dass die Ausprägung zwecks Interpretation auch in Kategorien wie „durchschnittlich“ oder „überdurchschnittlich“ übersetzt werden kann, schränkt die Forderung nach einer zahlenmäßigen Abbildung nicht ein. Manchmal ist es das Ziel, auf die Zugehörigkeit zu einer bestimmte Klasse oder Kategorie zu schließen – es wird also eine qualitative Aussage gemacht (s. Definitionselement e).

Ausprägung des Zielmerkmals als Zahl

2.1.2 Bandbreite psychologischer Tests

Alleine im deutschsprachigen Raum gibt es Hunderte von psychologischen Tests. So ergab eine Auszählung im April 2018 bei einem Verlag für deutschsprachige Tests über 900 Einträge unter der Rubrik „Tests“. Je nach Messanspruch, Zielgruppe und theoretischem Hintergrund kann sich die konkrete Ausgestaltung von Tests deutlich unterscheiden.

Große Bandbreite verfügbarer Tests

Unterschiedliche Testbeispiele

Aufgabe: Verbale Analogien erkennen und anwenden

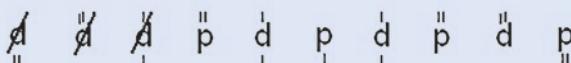
Wald: Bäume = Wiese: ?

- | | | | | |
|------------|---------|------------|----------|-----------|
| (a) Gräser | (b) Heu | (c) Futter | (d) Grün | (e) Weide |
|------------|---------|------------|----------|-----------|

Übungsaufgabe aus dem I-S-T 2000 R (Liepmann et al. 2007)

Aufgabe der Testpersonen ist es, die Beziehung zwischen dem 1. und 2. Wort zu erkennen und anzuwenden, indem das 3. um ein 4., aus 5 Alternativen auszuwählendes Wort ergänzt wird. Im Beispiel gilt: Ein Wald besteht aus Bäumen, eine Wiese aus Gräsern, daher ist (a) die richtige Lösung. Diese und ähnliche Aufgaben messen verbales Schlussfolgern, einen Aspekt der allgemeinen Intelligenz.

Aufgabe: Alle d's, die mit 2 Strichen versehen sind, durchstreichen



Test d2-R. (Aus Brickenkamp et al. 2010, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.test-zentrale.de).

Aufgabe der Testpersonen ist es, so schnell wie möglich, aber gleichzeitig möglichst fehlerfrei alle d's, die mit 2 Strichen versehen sind, durchzustreichen. Im Beispiel sind bereits 3 solcher d's markiert. Dieser Test beansprucht, die Konzentrationsfähigkeit zu messen.

Aufgabe: Tasten drücken und dadurch auf dem Monitor so schnell wie möglich einen Weg „entlanglaufen“

Anleitung...

Bei den folgenden Aufgaben müssen Sie einen vorgegebenen Weg möglichst rasch zurücklegen.
Oben sehen Sie einen Ausschnitt dieses Weges. Der Weg besteht aus 100 einzelnen Feldern, mit **roten** und **grünen** Pfeilen. Die Pfeile zeigen an, in welche Richtung der Weg führt und welche Taste daher gedrückt werden muss, um weiter zu kommen.

Drücken Sie nun auf die **grüne Taste!**

Weiter

(Auszug aus der Instruktion zum OLMT. Aus Schmidt-Atzert 2007, mit freundlicher Genehmigung von Schuhfried. Bildrechte: Schmidt-Atzert, L. (2007). *Objektiver Leistungsmotivationstest OLMT – Software und Manual* (2. Aufl., unter Mitarbeit von M. Sommer, M. Bühner & A. Jurecka).

Aufgabe in diesem Test ist es, durch wiederholtes und schnelles Drücken einer von 2 möglichen Tasten eine Strecke auf dem Monitor entlangzulaufen. Die Anzahl zurückgelegter Felder innerhalb des Zeitlimits gilt als Indikator für die Leistungsmotivation der Testpersonen.

Aufgabe: Ankreuzen, inwiefern eine Aussage die eigene Person beschreibt (typische Aufgabe in Persönlichkeitsfragebögen)

	1	2	3	4
Ich fühle mich wohl, wenn ich im Mittelpunkt stehe				x

Die Skala von 1 bis 4 könnte beispielsweise für „trifft gar nicht zu“, „trifft eher nicht zu“, „trifft eher zu“ und „trifft vollkommen zu“ stehen. Aufgabe von Testpersonen ist es, durch ein Kreuz anzugeben, in welchem Ausmaß eine Aussage auf sie zutrifft. Die hier genannte Aussage könnte indikativ für das Merkmal Extraversion sein. Von einer extravertierten Person würde man erwarten, dass sie bei dieser und bei ähnlichen Aussagen ihr Kreuz so wie im Beispiel gezeigt setzt.

Aufgabe: Ankreuzen, was man in der geschilderten Situation tun würde

An Ihrem Arbeitsplatz gibt es einige Orte, die Sie gefährlich finden. Sie denken, dass es zu Unfällen kommen könnte. Um die Gefahrenquellen dauerhaft zu beseitigen, müssten kostspielige Veränderungen vorgenommen werden. Leider ist Ihre Abteilung knapp bei Kasse. Wie würden Sie sich verhalten?

Trifft am wenigsten zu		Trifft am ehesten zu
<input type="checkbox"/>	Ich mache meinen Vorgesetzten auf die Gefahrenquellen aufmerksam. Er soll entscheiden, ob kostspielige Veränderungen notwendig sind	<input type="checkbox"/>
<input type="checkbox"/>	Ich kümmere mich darum, dass die Gefahrenquellen dauerhaft beseitigt werden, auch wenn dadurch Kosten für die Abteilung entstehen	<input type="checkbox"/>
<input type="checkbox"/>	Ich höre auf, mir darüber Gedanken zu machen. Mit etwas Vorsicht wird nichts passieren	<input type="checkbox"/>
<input type="checkbox"/>	Ich bringe Zettel mit Warnhinweisen (Vorsicht Stufe, etc.) an, um die anderen auf die Gefahr aufmerksam zu machen	<input type="checkbox"/>

(Aufgabe aus dem Situational-Judgment-Test von Bledow und Frese 2009, Abdruck mit freundlicher Genehmigung von John Wiley and Sons; deutsche Übersetzung aus Bledow und Frese 2005)

Aufgabe ist es hierbei, sich in die geschilderte Situation hineinzuversetzen und anzukreuzen, was man tun bzw. nicht tun würde. Dieser Test beansprucht die Messung von Proaktivität.

Weitere Testbeispiele finden sich in ▶ Abschn. 2.4.2.6 und in ▶ Kap. 3.

Um bei der Menge an verfügbaren Testverfahren den Überblick zu behalten, ist eine Systematik der Tests hilfreich. Das wichtigste Kriterium für eine Einteilung von Tests ist der **Messanspruch** (welches Merkmal soll erfasst werden?). Aber auch andere Unterteilungen sind geläufig.

Messanspruch als
Einteilungskriterium

Einteilungen der Testverfahren

- Nach Messanspruch (z. B. Persönlichkeitsfragebögen, Intelligenztests)
- In Leistungs- vs. Persönlichkeitstests bzw. Persönlichkeitsfragebögen (▶ Kap. 3)
- Nach psychologischer Disziplin (z. B. neuropsychologische Tests)
- Nach Zielgruppe (z. B. Tests für Kinder)
- Nach Administrationsform (z. B. Onlinetest, Paper-Pencil-Test)

Eindeutige Definition des Messanspruchs notwendig

Bisweilen ist die Forschungslage unbefriedigend

Anwendungsbereiche und Zielgruppen

Rückschluss auf latentes Merkmal

Alle wissenschaftlichen psychologischen Tests sollten das, was sie zu messen beanspruchen, klar benennen; dazu gehört auch eine Aussage darüber, wie sich das zu messende Merkmal in eine Theorie anderer, mehr oder weniger ähnlicher Merkmale einbetten lässt (Ziegler 2014). Also: Welche inhaltliche Nähe weist beispielsweise das Konstrukt Integrität zu den Dimensionen breiter Persönlichkeitsmodelle auf? Alleine durch das Label – in diesem Beispiel „Integrität“ – ist ein eindeutiges Verständnis des Messanspruchs nur schwer herzustellen.

Es sollte also der Messanspruch für jeden Test klar benannt sein. Es ist aber bei Weitem nicht so, dass es für jeden denkbaren Messanspruch einen Test gäbe. Neben mangelnder Nachfrage kann dafür auch eine unbefriedigende Forschungslage verantwortlich sein: Was man messen möchte, ist konzeptuell noch nicht hinreichend präzisiert worden, und oft mangelt es an empirischer Forschung, die ein theoretisches Modell stützt. Solche Bedenken werden manchmal beiseitegeschoben. Verschärft könnte man daher auch behaupten, dass es Tests gibt, die etwas messen (sollen), über das man kaum etwas weiß. Eine stark zugespitzte Bemerkung dazu lautet: „Sie wissen nicht, was es ist – aber messen können sie es.“ Dies kann ggf. dann akzeptiert werden, wenn trotz der Unklarheit des erfassten Merkmals zutreffende Prognosen für ein relevantes Kriterium möglich sind (z. B. beruflicher Erfolg).

Wichtige Anwendungsfelder, in denen häufig Tests eingesetzt werden, sind Berufseignungsdiagnostik, Klinische Psychologie, Neuropsychologie und Schul- und Erziehungsberatung. Für Anwenderinnen und Anwender stellt oft die Zielgruppe, für die ein Test aufgrund seiner Aufgaben und seiner Normen geeignet ist, ein wichtiges Auswahlkriterium dar. Es liegen Tests für Kinder, Jugendliche und Erwachsene vor, wobei oftmals der Altersbereich noch genauer festgelegt bzw. eingeschränkt ist. Anwenderinnen und Anwender haben manchmal eine Präferenz für Papier-und-Bleistift-Tests oder computergestützte Tests. Letztere haben den Vorteil, dass die Auswertung automatisch erfolgt. Sie setzen aber die Verfügbarkeit von Computerarbeitsplätzen und teilweise die Anschaffung von Basissoftware für ein Testsystem voraus. Aus pragmatischer Sicht stellt sich manchmal die Frage, ob ein Test im Einzelversuch durchgeführt werden muss oder ob auch Gruppenuntersuchungen möglich sind. Letzteres ist bei der Untersuchung vieler Probandinnen und Probanden äußerst ökonomisch. Weiterführende Informationen zu den unterschiedlichen Arten von Tests finden sich in ▶ Kap. 3.

2.1.3 Rückschluss auf ein latentes Merkmal

Neben inhaltlichen Theorien und Modellen, die den Messanspruch eines Tests spezifizieren, gilt es auch, eine weitere grundsätzliche Frage zu klären: Wie kann aus Reaktionen, die auf standardisierte Reizvorlagen (▶ Abschn. 2.1.1) erfolgen, auf eigentlich nicht sichtbare Merkmale von Personen – z. B. Intelligenz oder Verträglichkeit – geschlossen werden? Diese Frage adressiert ein Grundproblem der Psychologischen Diagnostik: Eine direkte Messung ist nicht möglich – es muss aufgrund von beobachtetem Verhalten, also den Reaktionen auf Testaufgaben, auf ein latentes Merkmal geschlossen werden. Anders als bei vielen anderen Messungen ist keine unmittelbare Beschreibung der relevanten Gegebenheit möglich.

Physikalische Größen

Messungen der Körpergröße oder des Körpergewichts sind vergleichsweise einfach. Es ist klar bestimmbar, was mit „Körpergröße“ und „Körpergewicht“ gemeint ist. Zudem handelt es sich um physikalische Größen, die sowohl für Menschen wie auch für Objekte (also z. B. ein Maßband, einen Zollstock oder ein Gewicht) gelten. Daher können Objekte genutzt werden, um Menschen zu beschreiben. Beispielsweise hat man noch bis 2019 das menschliche Gewicht mit einem Objekt, das als Urkilogramm bezeichnet wird, verglichen. Dies ist ein Zylinder mit einer Höhe von 39 mm und einem Durchmesser von 39 mm, der aus einer Legierung von 90 % Platin und 10 % Iridium besteht. Mittlerweile lässt sich das Kilogramm auf eine unveränderliche, universelle Naturkonstante, die sog. „Planck-Konstante“, zurückführen und damit von einem Objekt unabhängig bestimmen.

Für psychologische Tests braucht es Testtheorien, die grundlegende Annahmen darüber beinhalten, wie aufgrund von beobachtetem Verhalten in einem Test auf ein latentes Merkmal geschlossen werden kann.

Psychologische Tests brauchen Testtheorien

Definition

„Die **Testtheorie** als Teilgebiet der Psychometrie beschäftigt sich mit der Entwicklung und Formalisierung von psychometrischen Modellen für psychologische Tests und mit ihrer Nutzung für die Konstruktion und Evaluation psychologischer Tests“ (Eid und Schmidt 2014, S. 34).

Psychometrie bezeichnet ein Forschungsgebiet, das sich mit der Messung psychologischer Merkmale beschäftigt (vgl. Wirtz 2013).

Das Ziel – aus beobachtetem Verhalten, in Form von Reaktionen auf Testaufgaben, auf ein latentes Merkmal zu schließen (s. o.) – stellt eine grundsätzliche Anforderung an Testaufgaben: Sie sollten so gestaltet sein, dass Antworten von Personen zu einem möglichst hohen Anteil deren Ausprägung auf dem latenten Merkmal reflektieren (reflexive Messungen). □ Abb. 2.1 veranschaulicht diesen Gedanken. In dem darin dargestellten, idealen Fall beeinflusst nur ein latentes Merkmal die Ergebnisse der 6 Testaufgaben. Andere latente Merkmale oder situative Variablen (Müdigkeit der Testpersonen, Verständnisprobleme etc.) haben keinen nennenswerten Einfluss auf die Ergebnisse der 6 Aufgaben. In diesem Fall reflektiert das Testergebnis sehr gut die Ausprägung der Testpersonen in dem zu messenden latenten Merkmal.

Reflexive Messungen

Auch formative Messungen sollen hier nicht unerwähnt bleiben. Hier ist die Annahme konträr zu der reflexiven Messung: Nicht die zugrunde liegende latente Variable wirkt auf die Itemantworten, sondern die Itemantworten bilden die latente Variable. Dies könnte beispielsweise bei der Messung des wöchentlichen Bewegungspensums der Fall sein. Antworten auf Fragen nach der Laufleistung, der mit dem Rad zurückgelegten Kilometer, der gestiegenen Treppen usw. bilden gemeinsam das wöchentliche Bewegungspensum. Die einzelnen Bewegungsbereiche können dabei völlig unkorreliert sein – Personen können viel laufen, aber kaum Rad fahren (vgl. Hoyle 2012).

Formative Messungen

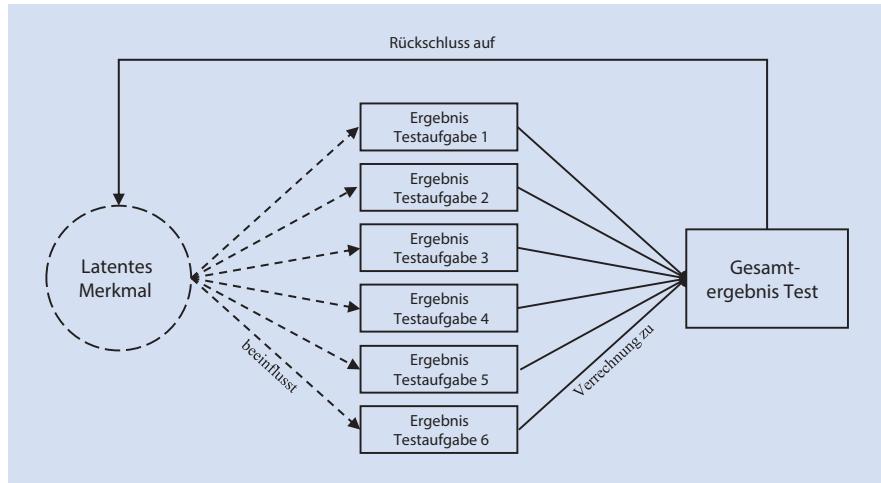


Abb. 2.1 Antworten auf Testaufgaben reflektieren das latente Merkmal

Rückschluss auf latente Merkmale nicht nur durch Testverfahren möglich

Es soll an dieser Stelle keineswegs der Eindruck vermittelt werden, dass nur durch psychologische Testverfahren Rückschlüsse auf latente Eigenschaften möglich sind. Eine Unterscheidung von Informationsquellen, auf deren Basis Aussagen über Personen möglich sind, wurde bereits 1946 von Cattell vorgenommen. Er unterschied L-Daten (life record data), d. h. Fremdbeurteilungen, Informationen aus der Biografie etc., Q-Daten (questionnaire data), d. h. Selbstberichte anhand von Fragebögen oder Interviews, und T-Daten (test data), d. h. objektive Daten, die anhand von (Leistungs-)Tests oder Verhaltensproben in standardisierten Situationen gewonnen wurden (vgl. Stemmler et al. 2016). Es gilt für alle Datenquellen, dass Rückschlüsse auf latente Eigenschaften fehlerbehaftet sind und damit nicht als deterministisch aufgefasst werden dürfen. Optimal ist daher eine Nutzung von Daten bzw. Informationen aus verschiedenen Quellen. Für jede genutzte Datenquelle gilt – ebenso wie für Test- und Fragebogendaten –, dass der gezogene Rückschluss auf latente Eigenschaften theoretisch fundiert und durch empirische Belege gestützt werden muss (► Abschn. 2.6.3).

Folgende Methoden sind in der Praxis zu beobachten, für die derzeit keine Evidenz vorliegt, dass sie haltbare Rückschlüsse auf Eigenschaften erlauben (als Referenz sind jeweils exemplarisch Studien oder Übersichtsarbeiten aufgeführt, die gegen die Nutzung der jeweiligen Methode sprechen):

- Grafologie, d. h., man schließt aufgrund der Handschrift einer Person auf ihre psychologischen Eigenschaften (vgl. Schmidt und Hunter 1998)
- Psycho-Physiognomik, d. h., man schließt aufgrund von körperlichen Merkmalen (z. B. Gesichtszügen) auf psychologische Eigenschaften (Kanning 2012)
- Horoskope, d. h., man schließt vom Geburtszeitpunkt in einem Jahr auf psychologische Eigenschaften (Hartmann et al. 2006)

2.2 Die Klassische Testtheorie

Nachfolgend stellen wir zunächst die Klassische Testtheorie und in ▶ Abschn. 2.3 die Item-Response-Theorien dar. Diese Unterteilung ist in der Praxis der Testentwicklung üblich. Es ist uns allerdings wichtig, zu betonen, dass es deutliche Gemeinsamkeiten zwischen den Testtheorien gibt und diese auch in einen gemeinsamen, übergeordneten Rahmen integriert werden können (Mellenbergh 1994). Für eine ausführliche Darstellung von Testtheorien in einem solchen integrativen Rahmen sei auf Eid und Schmidt (2014) verwiesen.

2.2.1 Zentrale Annahmen

Die sog. „Klassische Testtheorie“ (kurz: KTT) hat eine lange Tradition. Gullicksen (1950) hat frühere Forschungsarbeiten, darunter auch Arbeiten von Spearman aus den Jahren von 1904 bis 1913, zusammengefasst und aufgearbeitet. Eine mathematisch fundierte Fassung haben Lord und Novick (1968) vorgelegt. Dieses Buch gilt als Grundlage der Klassischen Testtheorie (vgl. Krauth 1996). Entsprechend ihrer langen Historie, aber auch wegen ihrer einfachen Anwendbarkeit basieren die meisten psychologischen Tests auf der Klassischen Testtheorie.

Viele psychologische Tests basieren auf der Klassischen Testtheorie

! Die zentrale Annahme der Klassischen Testtheorie ist, dass Messungen fehlerbehaftet sind. Sie nimmt an, dass eine einzelne Messung aufgrund von unsystematischen Einflussfaktoren ein höheres oder niedrigeres Ergebnis liefert als aufgrund der tatsächlichen Merkmalsausprägung zu erwarten wäre.

Zentrale Annahme der Klassischen Testtheorie

Diese zentrale Annahme lässt sich gut anhand von Beispielen aus dem Sport erläutern. Sportlerinnen und Sportler erzielen gelegentlich unerwartet gute oder schlechte Ergebnisse – eine Kugelstoßerin stößt weiter als üblich, ein Weitspringer springt weniger weit, als von ihm zu erwarten wäre, eine Sprinterin läuft schneller, als ihre übliche Zeit erwarten lässt.

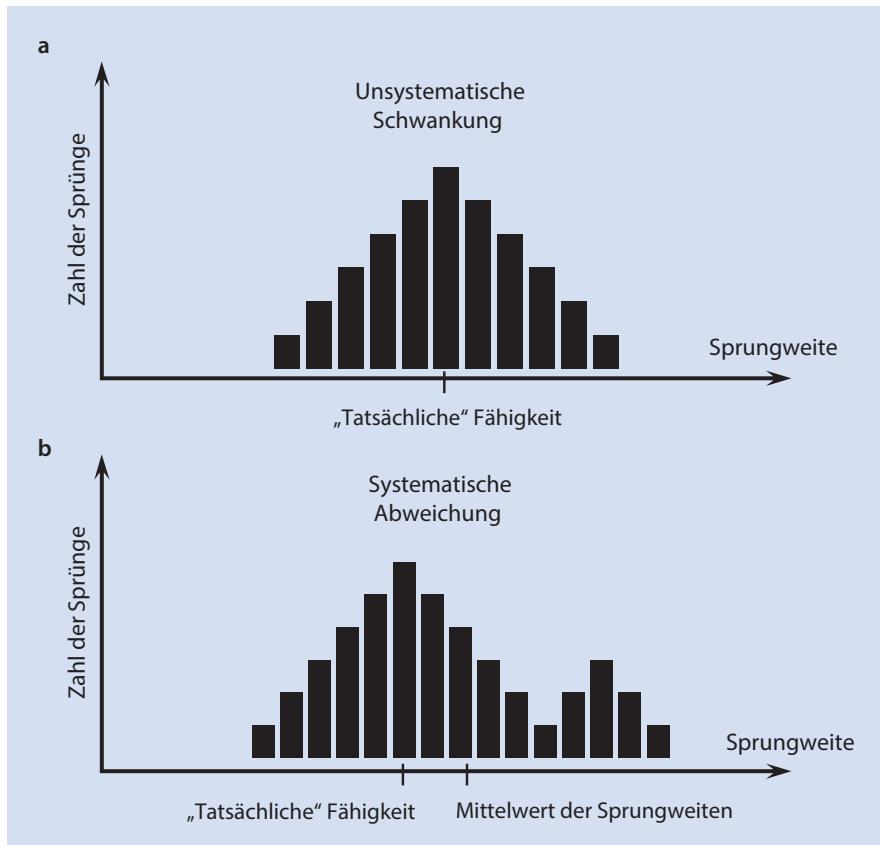
Messungen sind stets fehlerbehaftet – auch im Sport

Ebenso können Antworten auf Testaufgaben variieren und die eigentliche Merkmalsausprägung über- oder unterschätzt werden – ein hochintelligenter Mensch kann gelegentlich eine einfache Intelligenztestaufgabe falsch lösen, ein wenig gewissenhafter Mensch kann sich in einem Fragebogenitem als hoch gewissenhaft einstufen etc. Ein zentrales Rational der Klassischen Testtheorie ist daher, Messungen so oft wie möglich unter identischen Bedingungen zu wiederholen. Es wird angenommen, dass sich bei vielen Wiederholungen der gleichen (d. h. parallelen) Messung unsystematische Fehler „herausmitteln“ (s. „Äquivalenz von Messungen“, ▶ Abschn. 2.6.2.1): Während es bei einer Messung vielleicht zu einer leichten Abweichung vom wahren Wert nach unten kommt, ergibt sich bei einer anderen Messung eine Abweichung nach oben. Über unendlich viele Messungen hinweg gleichen sich die Messfehler aus, d. h., sie addieren sich zu 0.

Annahme: Messfehler gleichen sich über unendlich viele Messung aus

Fehler lassen sich nur „herausmitteln“, wenn sie tatsächlich unsystematisch sind, also einmal zufällig über, einmal zufällig unter der eigentlichen Merkmalsausprägung liegen, und stets gilt: Je weiter weg von der eigentlichen Merkmalsausprägung die „Ausreißer“ liegen, desto seltener kommen sie vor. Man erwartet also eine Verteilung wie in □ Abb. 2.2a dargestellt.

Nur unsystematische Messfehler gleichen sich über viele Messwiederholungen aus



■ Abb. 2.2 Mögliche Verteilung von Sprungweiten einer Person bei **a** unsystematischer Schwankung und **b** systematischer Abweichung

Systematische Fehler

Messungen können natürlich auch systematische Fehler enthalten. Im Sport könnten Athletinnen und Athleten einen Teil ihrer Leistungen unter Dopingeinfluss vollbracht haben. Bei psychologischen Testungen könnten sich Personen einen Teil der Testaufgaben und deren Lösungen vorab aus dem Internet heruntergeladen haben. Man würde dann eine Verteilung wie in ■ Abb. 2.2b erwarten – anhand des Mittelwertes wäre hier keine sinnvolle Aussage über die tatsächliche Leistungsfähigkeit der Personen möglich. Denn die systematischen Fehler würden sich nicht durch häufige Wiederholung der Messung „herausmitteln“.

Personen können natürlich nicht nur systematisch besser, sondern auch schlechter abschneiden – beispielsweise wenn sie durch Testbedingungen unsicher werden oder einen Teil der Aufgaben unter hohem Zeitdruck lösen müssen. Zentral ist an dieser Stelle: Die Klassische Testtheorie berücksichtigt zunächst einmal nur unsystematische Fehler.

Eine formalisierte Zusammenfassung der bisherigen Annahmen wird durch 2 Axiome der Klassischen Testtheorie vorgenommen (vgl. Bühner 2011, S. 44 ff.).

Zwei Axiome der Klassischen Testtheorie

Es gibt einen wahren Wert der Merkmalsausprägung einer Person v , definiert als der Erwartungswert unendlich häufiger Messungen unter identischen Bedingungen (Existenzaxiom).

$$\tau_v = E(X_v)$$

Der beobachtete Wert einer Person v in einem Testitem i setzt sich zusammen aus dem wahren Wert der Person v und einem zufälligen Messfehler (Verknüpfungsaxiom).

$$x_{vi} = \tau_{vi} + \varepsilon_{vi}$$

Diese Annahme gilt nicht nur für einzelne Testitems. Sie gilt auch für Testteile - etwa wenn man einen Test in 2 Hälften aufteilt und einen halben, verkürzten Test betrachtet. Und sie gilt auch für ganze Tests.

$$X_{vt} = \tau_v + \varepsilon_{vt}$$

τ_v = wahrer Wert einer Person v

τ_{vi} = wahrer Wert einer Person v in Item i

E = Erwartungswert

X_{vt} = beobachtete Werte einer Person v in Test t

x_{vi} = beobachtete Werte einer Person v in Item i

ε_{vt} = Messfehler für Person v in Test t

ε_{vi} = Messfehler für Person v in Item i

Die Klassische Testtheorie nimmt also an, dass der wahre Wert einer Person im Test unveränderlich ist. Das heißt, er ist bei jeder Durchführung des Tests gleich groß. Er könnte theoretisch ermittelt werden, indem man den Test extrem (genau genommen unendlich) oft durchführt und dabei sicherstellt, dass keine Erinnerungs- und Übungseffekte oder andere systematische Verzerrungen auftreten. Der Erwartungswert aller Messergebnisse (also aller beobachteten Werte) wäre dann der wahre Wert. Um einem möglichen Missverständnis vorzubeugen, soll hier klargestellt werden: Der wahre Wert bezieht sich in der Klassischen Testtheorie stets auf eine bestimmte Messung sowie die dabei angenommene Verteilung der Messwerte über viele Wiederholungen und nicht auf das zugrunde liegende Merkmal. Geht es beispielsweise um den wahren Wert einer Person in einem bestimmten Intelligenztest, ist dies nicht zu verstehen als die „wahre Intelligenz“ dieses Menschen.

Es folgt weiterhin aus der Annahme unsystematischer Messfehler, dass diese mit wahren Werten, beobachteten Werten oder Messfehlern aus anderen Messungen unkorreliert sind – sonst wären sie nicht unsystematisch. Diese Annahme gilt nicht nur auf Test-, sondern auch auf Itemebene. Übrigens: Dass die Messfehler unabhängig von den wahren Werten sind, bedeutet nichts anderes, als dass ein Test im unteren Bereich (niedrige Fähigkeit) ebenso genau misst wie im mittleren oder im oberen Bereich.

Wahrer Wert als Erwartungswert aller Messergebnisse

Messfehler sind mit wahren Werten, beobachteten Werten oder Messfehlern aus anderen Messungen unkorreliert

Unsystematischer Messfehler

Die Messfehler eines Tests/Items sind unabhängig von den

- wahren Werten von Personen in demselben Test/Item;
- wahren Werten von Personen in einem anderen Test/Item;
- beobachteten Werten von Personen in demselben Test/Item;
- beobachteten Werten von Personen in einem anderen Test/Item;
- Messfehlern eines anderen Tests/Items.

2.2.2 Reliabilität von Messungen

Aus den oben aufgeführten Axiomen lässt sich eine Aussage über die Reliabilität (Messgenauigkeit) einer Messung ableiten. Diese Formel stellt die wichtigste Ableitung aus den Annahmen der Klassischen Testtheorie dar:

Definition

Die **Reliabilität** einer Messung $r_{tt'}$ ist der Anteil der Varianz der wahren Werte τ an der Varianz der beobachteten Werte x .

$$r_{tt'} = \frac{s_\tau^2}{s_x^2}$$

s_τ^2 = Varianz der wahren Werte

s_x^2 = Varianz der beobachteten Werte

Reliabilität als Korrelation des Tests mit seiner Parallelversion

Ein Reliabilitätskoeffizient von beispielsweise .80 bedeutet demzufolge, dass die beobachtete Varianz der Testwerte zu 80 % auf Unterschiede zwischen den wahren Werten der Testpersonen zurückzuführen ist und zu 20 % auf Fehlervarianz beruht.

Wie lässt sich nun abschätzen, wie hoch der Anteil der Varianz der wahren Werte und wie hoch der Anteil der messfehlerbedingten Varianz ist? Man nutzt dazu die Korrelation des Tests „mit sich selbst“ – präziser müsste man sagen: die Korrelation des Tests mit einer (essenziell) parallelen Version (Äquivalenz von Messungen; ▶ Abschn. 2.6.2.1) des gleichen Tests (daher auch die Bezeichnung $r_{tt'}$ für die Reliabilität). Es gibt eine einfache Begründung dafür, warum diese Korrelation als Schätzung der Reliabilität herangezogen werden kann. Wahre Werte von mindestens essenziell parallelen Messungen (▶ Abschn. 2.6.2.1) sind höchstens um eine Konstante voneinander verschoben und perfekt korreliert. Würde also ein Test nur die Varianz wahrer Werte abbilden, also messfehlerfrei sein, wäre die Korrelation des Tests mit sich selbst bzw. mit einer essenziell parallelen Version des Tests $r_{tt'} = 1$. Wir haben ebenfalls bereits festgehalten, dass Messfehler unsystematisch sind, d. h., dass Messfehler eines Tests weder mit wahren Werten noch mit anderen Messfehlern (also auch nicht den bei Testwiederholung entstehenden) korrelieren. Somit mindert der Anteil der messfehlerbedingten Varianz an der Testvarianz unmittelbar dessen Korrelation mit sich selbst bzw. mit einer (essenziell) parallelen Version des gleichen Tests. Wenn – wie die Klassische Testtheorie annimmt – Testwerte nur aus wahren Wert und Messfehler bestehen, so reflektiert die Korrelation eines Tests mit einer essenziell parallelen Testversion zwingend den Anteil der Varianz der wahren Werte an der Gesamtvianz.

Herleitung der Reliabilität

Die Korrelation der beobachteten Werte eines Tests mit den beobachteten Werten einer essenziell parallelen Testversion (hier beschrieben als X_t und $X_{t'}$) ist gemäß der Formel für Korrelationen

$$r_{Xt,Xt'} = \frac{\text{cov}(X_t X_{t'})}{s_{Xt} \times s_{Xt'}}. \quad (2.1)$$

$r_{Xt,Xt'}$ = Korrelation beobachteter Werte eines Tests t mit denen der Parallelversion t'

$\text{cov}(X_t X_{t'})$ = Kovarianz beobachteter Werte eines Tests t mit denen der Parallelversion t'

s_{Xt} bzw. $s_{Xt'}$ = Standardabweichung der beobachteten Werte eines Tests t bzw. der Parallelversion t'

Da es sich bei t und t' um den gleichen Test handelt (präziser: um essenziell parallele Tests), sind die Standardabweichungen s_{Xt} und $s_{Xt'}$ gleich, sodass ▶ Gl. 2.1 wie folgt vereinfacht werden kann:

$$r_{Xt,Xt'} = \frac{\text{cov}(X_t X_{t'})}{s_x^2} \quad (2.2)$$

s_x^2 = Varianz der beobachteten Werte (in Test t und t' gleichermaßen)

Aufgrund der Annahmen der Klassischen Testtheorie, dass für jede Person v gilt

$$X_{vt} = \tau_v + \varepsilon_{vt}, \quad (2.3)$$

lässt sich die Kovarianz der beobachteten Werte aus den Tests t und t' (X_t und $X_{t'}$) gemäß der Regeln zur Kovarianz von Summen (hier: $\tau + \varepsilon$) beschreiben als

$$\text{cov}(X_t X_{t'}) = \text{cov}(T_t T_{t'}) + \text{cov}(T_t \varepsilon_{t'}) + \text{cov}(\varepsilon_t T_{t'}) + \text{cov}(\varepsilon_t \varepsilon_{t'}). \quad (2.4)$$

$\text{cov}(X_t X_{t'})$ = Kovarianz der beobachteten Werte aus t und t'

$\text{cov}(T_t T_{t'})$ = Kovarianz der wahren Werte aus t und t'

$\text{cov}(T_t \varepsilon_{t'})$ bzw. $\text{cov}(\varepsilon_t T_{t'})$ = Kovarianz der wahren Werte aus t und der Messfehler aus t' (bzw. der wahren Werte aus t' und der Messfehler aus t)

$\text{cov}(\varepsilon_t \varepsilon_{t'})$ = Kovarianz der Messfehler aus t und t'

Wie zuvor dargestellt, definiert die Klassische Testtheorie, dass

$$\text{cov}(T_t \varepsilon_{t'}) = \text{cov}(\varepsilon_t T_{t'}) = \text{cov}(\varepsilon_t \varepsilon_{t'}) = 0. \quad (2.5)$$

Somit kann die Kovarianz der beobachteten Werte X_t und $X_{t'}$ aus ▶ Gl. 2.4 einfacher beschrieben werden als

$$\text{cov}(X_t X_{t'}) = \text{cov}(T_t T_{t'}). \quad (2.6)$$

Setzt man ▶ Gl. 2.6 in ▶ Gl. 2.2 ein, erhält man

$$r_{Xt,Xt'} = \frac{\text{cov}(T_t T_{t'})}{s_x^2} \quad (2.7)$$

Für die Kovarianz der wahren Werte T_t und $T_{t'}$ bzw. deren Korrelation $r_{Tt,Tt'}$ gilt analog zu ▶ Gl. 2.1:

$$r_{Tt,Tt'} = \frac{\text{cov}(T_t T_{t'})}{s_{Tt} \times s_{Tt'}} \quad (2.8)$$

Dies lässt sich umstellen zu

$$\text{cov}(T_t T_{t'}) = r_{T_t T_{t'}} \times s_{T_t} \times s_{T_{t'}} \quad (2.9)$$

und kann wegen

$$r_{T_t T_{t'}} = 1 \quad \text{und} \quad s_{T_t} \times s_{T_{t'}} = s_T^2 \quad (2.10)$$

vereinfacht werden zu

$$\text{cov}(T_t T_{t'}) = s_T^2 \quad (2.11)$$

Setzt man ▶ Gl. 2.11 in ▶ Gl. 2.7 ein, so ist ersichtlich, dass die Korrelation eines Tests mit seiner essenziell parallelen Version den Anteil der Varianz der wahren Werte an der Gesamtvarianz – also die Reliabilität – reflektiert:

$$r_{X_t X_{t'}} = \frac{s_T^2}{s_X^2} \quad (2.12)$$

Dies wird vereinfacht geschrieben als

$$r_{tt'} = \frac{s_T^2}{s_X^2} \quad (2.13)$$

(Herleitung angelehnt an Moosbrugger 2012b).

Viele Testautorinnen und -autoren stützen sich auf die Klassische Testtheorie. Daher basieren viele der in Testmanualen berichteten Schritte der Testkonstruktion auf Annahmen der Klassischen Testtheorie. Diese Schritte werden in ▶ Abschn. 2.4 näher betrachtet. In ▶ Abschn. 2.6.2 gehen wir auf die Praxis der Reliabilitätsschätzung näher ein.

2.2.3 Grenzen der Klassischen Testtheorie

Wesentliche Kritikpunkte

Gegenüber der Klassischen Testtheorie sind verschiedene Einwände vorgebracht worden. Im Folgenden werden wesentliche Kritikpunkte genannt und kommentiert.

Messfehler verteilen sich nicht immer zufällig um den wahren Wert Ein immer wieder vorgebrachter Kritikpunkt besagt, dass sich Messfehler nicht unbedingt zufällig, d. h. so wie in □ Abb. 2.2a dargestellt, um den wahren Wert verteilen. Es existieren in der Tat Belege für die Existenz systematischer Fehler, etwa die systematische Verzerrung von Testwerten in Richtung eines sozial erwünschten Ergebnisses oder die Erhöhung von Testwerten durch Übungseffekte. Allerdings definiert die Klassische Testtheorie Messfehler als unsystematische Fehler (die also in jede Richtung wirken können und sich insgesamt aufheben). Sie bestreitet nicht, dass es daneben auch systematische Fehler gibt.

Die Parameter der Klassischen Testtheorie sind populations- und stichprobenabhängig Wie man in ▶ Abschn. 2.6.2.1 zur Reliabilitätsschätzung sehen wird, basieren Reliabilitätskoeffizienten (direkt oder indirekt) auf Korrelationen. Daher gilt: Je größer die Streuung des gemessenen Merkmals in der Untersuchungsstichprobe ist, desto höher fallen Reliabilitätsschätzungen aus. Das bedeutet auch, dass heterogenere Personenstichproben in Bezug auf das untersuchte Merkmal höhere Reliabilitätskoeffizienten hervorbringen. Damit sind

Reliabilitätsschätzungen – und weitere Parameter der Klassischen Testtheorie, die auf Korrelationen basieren – stichprobenabhängig. Diese Kritik stellt die Klassische Testtheorie nicht grundsätzlich infrage. Sie macht aber deutlich, dass die Kennwerte nicht ohne Weiteres auf andere Populationen übertragbar sowie von Stichproben auf eine Population generalisierbar sind. Kennwerte, die von Erhebungen mit Patientinnen und Patienten stammen, gelten nicht unbedingt für Gesunde. Ergebnisse, die an einer Stichprobe gewonnen wurden, dürfen nur dann als gültig für die Population angenommen werden, wenn es sich um eine repräsentative und zudem hinreichend große Stichprobe handelt. In der Praxis werden aber immer wieder Gelegenheitsstichproben zur Testentwicklung und zur Schätzung der Reliabilität (besonders der Re-test-Reliabilität) herangezogen.

Das Skalenniveau wird häufig missachtet In der Klassischen Testtheorie geht man davon aus, dass sich bei häufiger Wiederholung von Messungen, etwa durch mehrere parallele Items, Messfehler „herausmitteln“. Damit geht man auch davon aus, dass Ergebnisse von einzelnen Messungen aggregiert werden können – etwa zu einem Summen- oder Mittelwert. Dies setzt jedoch mindestens ein Intervallskalenniveau und die Prüfung, ob Summen- und Mittelwerte angemessene Berechnungen des Testwertes sind, voraus. Eine explizite Prüfung dessen bleibt jedoch in der Praxis häufig aus.

Ungeachtet dessen ist die Klassische Testtheorie weitverbreitet und wird bei der Testkonstruktion deutlich häufiger als andere Testtheorien genutzt. Vermutlich liegt dies auch daran, dass die Klassische Testtheorie eine sparsame, leicht verständliche Theorie ist, die mit wenigen Grundannahmen auskommt.

Trotz aller Kritik ist die Klassische Testtheorie sparsam, leicht verständlich und weitverbreitet

2.3 Item-Response-Theorien

Item-Response-Theorien (engl. item response theory, IRT) werden häufig auch als „Probabilistische Testtheorien“ bezeichnet. Diese Begriffe verraten bereits, worauf es bei diesen Theorien im Kern ankommt: Sie beschäftigen sich mit Antworten auf Items und beschreiben das auf ein oder mehrere Items bezogene Antwortverhalten von Personen (item response) als Wahrscheinlichkeitsfunktion, d. h. als probabilistische Funktion.

Antwortverhalten von Personen als Wahrscheinlichkeitsfunktion

Es stehen unterschiedliche Item-Response-Theorien zur Verfügung, die je nach Art des Tests und abhängig vom diagnostischen Ziel zur Anwendung kommen. Besteht ein Test beispielsweise aus einem Richtig-Falsch-Antwortformat, also einem dichotomen Antwortformat, können zur Quantifizierung von Personenmerkmalen dichotome Rasch-Modelle (► Abschn. 2.3.1.1) herangezogen werden. Ein Beispiel dafür ist ein Wissenstest, in dem jedes Item aus einer Frage und 4 Antwortalternativen besteht, wovon nur eine richtig und die übrigen 3 falsch sind.

Unterschiedliche Item-Response-Theorien vorhanden

Persönlichkeitsfragebögen verwenden meist Ratingskalen, z. B. wenn auf einer Skala von 0 („trifft gar nicht zu“) bis 5 („trifft vollkommen zu“) der Grad der Zustimmung zu einer Aussage angegeben werden soll. Hier kann das ordinale Rasch-Modell (► Abschn. 2.3.2) angewendet werden.

Ordinales Rasch-Modell

Nicht immer dienen Testungen dem Zweck der Quantifizierung von Merkmalen (z. B. „wie liberal ist eine Person?“), sondern der Klassifizierung von Personen (z. B. „welche politische Partei präferiert eine Person?“) – hierzu eignen sich Latent-Class-Analysen (► Abschn. 2.3.3). Darüber hinaus gibt es eine Vielzahl weiterer Item-Response-Theorien (Cohen et al. 2013). In diesem Abschnitt soll vor allem auf das dichotome Rasch-Modell ausführlich eingegangen werden. Das ordinale Rasch-Modell und Latent-Class-Modelle werden nur kurz skizziert. □ Tab. 2.1 gibt eine Übersicht über Testmodelle für verschiedene Antwortformate und diagnostische Ziele (angelehnt an Rost 2004).

Latent-Class-Modelle

Tab. 2.1 Übersicht über gängige probabilistische Testmodelle (angelehnt an Rost 2004)

Antwortformat	Beispielitem	Diagnostisches Ziel	Infrage kommende Testmodelle
Dichotom	<p>Welche Person gehört nicht zur Zero-Künstlergruppe?</p> <p><input type="checkbox"/> Günther Uecker <input type="checkbox"/> Neo Rauch <input type="checkbox"/> Heinz Mack <input type="checkbox"/> Otto Piene</p>	Quantifizierung von Merkmalen (z. B. Ausprägung des Allgemeinwissens)	Dichotome Rasch-Modelle: <ul style="list-style-type: none"> - Einparametrisches logistisches Modell - Zweiparametrisches logistisches Modell bzw. Birnbaum-Modell - Dreiparametrisches logistisches Modell
		Klassifikation von Personen (z. B. zur Identifikation von verschiedenen Wissenstypen)	Latente Klassenanalyse
		Quantifizierung und Klassifikation (z. B. zur Identifikation von verschiedenen Wissenstypen und der Ausprägung des Allgemeinwissens innerhalb jeden Typs)	Mixed-Rasch-Modelle für dichotome Daten
Ordinal	<p>Ich reagiere leicht ange- spannt.</p> <p><input type="checkbox"/> sehr zutreffend <input type="checkbox"/> eher zutreffend <input type="checkbox"/> weder noch <input type="checkbox"/> eher unzutreffend <input type="checkbox"/> unzutreffend</p>	Quantifizierung von Merkmalen (z. B. Ausprägung der emotionalen Stabilität)	Ordinalale Rasch- Modelle: <ul style="list-style-type: none"> - Ratingskalenmodell - Äquidistanzmodell - Dispersionsmodell
		Klassifikation von Personen (z. B. Identifikation von verschiedenen Belastungs-Typen)	Klassenanalyse ordinaler Daten
		Quantifizierung und Klassifikation (z. B. zur Identifikation von verschiedenen Belastungs-Typen und der Ausprägung der Belastung innerhalb jeden Typs)	Mixed-Rasch- Modelle für ordinale Daten
Nominal	<p>Einem Umweltverband für den Schutz bedrohter Arten Geld spenden</p> <p><input type="checkbox"/> Habe ich schon getan bzw. tue ich bereits.</p> <p><input type="checkbox"/> Kann ich mir gut vorstellen.</p> <p><input type="checkbox"/> Würde ich tun, wenn geeignete Bedingungen geschaffen würden.</p> <p><input type="checkbox"/> Ich halte das für ungeeignet um die Umwelt zu schützen.</p> <p>(aus Rost 2004, S. 184)</p>	Quantifizierung von Merkmalen (z. B. zum Ausmaß der Bereitschaft, sich für die Umwelt zu engagieren)	Mehrdimensionale Rasch-Modelle
		Klassifikation von Personen (z. B. zur Identifikation verschiedener Typen des umweltbewussten Handelns)	Klassenanalyse nominaler Daten

2.3.1 Item-Response-Theorien für dichotome Antwortformate

2.3.1.1 Herleitung der Grundannahmen und einparametrisches logistisches Modell

Um die Grundgedanken von Item-Response-Theorien für dichotome Antwortformate zu illustrieren, bieten sich Analogien aus dem Sport an. Das typische, dichotome Ergebnis in vielen Sportarten ist „gewinnen vs. verlieren“ – analog zu „richtiger vs. falscher Itemantwort“. In der Folge wird die Sportart Tennis (Damen) zur Verdeutlichung der Grundgedanken des dichotomen Rasch-Modells herangezogen.

Wir nehmen zunächst einmal an, dass mehrere Tennisspielerinnen (Spielerinnen A bis E) gegen dieselbe Gegnerin antreten (Tab. 2.2). Die Gegnerin ist eine gute Tennisspielerin (Weltranglistenplatz 50). Gegen sie tritt zunächst die wesentlich schwächere Spielerin E (Weltranglistenplatz 100) an. Der Ausgang des Tennismatches scheint klar: Mit großer Wahrscheinlichkeit wird Spielerin E verlieren; ein Sieg von Spielerin E wäre eine große Überraschung. Danach tritt die etwas stärkere Spielerin D (Weltranglistenplatz 80) an. Auch für Spielerin D ist eine Niederlage gegen die Gegnerin weiterhin wahrscheinlicher als ein Sieg. Der Ausgang dieses Tennismatches scheint jedoch im Vergleich zum vorherigen (Spielerin E gegen Gegnerin) nicht mehr ganz so klar zu sein. Als Nächstes tritt Spielerin C an; sie ist ebenso gut wie die Gegnerin (beide teilen sich den Weltranglistenplatz 50). In diesem Fall spricht man von einem sog. „50:50-Spiel“. Man meint damit: Die Wahrscheinlichkeit, dass Spielerin C gewinnt, ist ebenso groß wie die Wahrscheinlichkeit, dass Spielerin C verliert. Im Falle von dichotomen Ergebnissen sind beide Wahrscheinlichkeiten also .50. Im nächsten Match tritt die noch stärkere Spielerin B (Weltranglistenplatz 20) gegen die Gegnerin an. Für Spielerin B sollte nun die Wahrscheinlichkeit eines Siegs deutlich größer sein als die Wahrscheinlichkeit einer Niederlage. Schließlich folgt danach noch ein Match zwischen der Ersten der Weltrangliste (Spielerin A) und der Gegnerin (zur Erinnerung: die Gegnerin belegt Weltranglistenplatz 50). Nun ist die Siegeswahrscheinlichkeit von Spielerin A als extrem hoch zu bewerten.

Bei der Bewertung der Erfolgswahrscheinlichkeit der Spielerinnen A bis E wurden einige Annahmen gemacht, die intuitiv sinnvoll sind und auch für Item-Response-Theorien für dichotome Antwortformate gelten.

Gewinnen und verlieren im Sport
analog zu Item lösen vs. nicht lösen

Intuitiv sinnvolle Annahmen

1. Asymptotische Annäherung der Erfolgswahrscheinlichkeit an 0 und an 1 Wenn gleich die hier als „Gegnerin“ bezeichnete Spielerin grundsätzlich deutlich besser ist als Spielerin E (Weltranglistenplatz 50 vs. 100), so wird dennoch niemand so kühn sein, zu behaupten, dass Spielerin E *definitiv* verliert, also eine Erfolgswahrscheinlichkeit von 0 hat. Ein Sieg von Spielerin E ist zwar sehr unwahrscheinlich, aber keineswegs ausgeschlossen. Auch wenn noch schwächere Spielerinnen (z. B. eine Spielerin auf Weltranglistenplatz 250) gegen die Gegnerin antreten, bleibt eine – wenn auch sehr kleine – Wahrscheinlichkeit, dass diese gewinnen. Dies könnte z. B. dann geschehen, wenn die Gegnerin sich während des Matches verletzt oder einen „schwarzen Tag erwischt“, oder aber die schwächere Spielerin „über sich hinauswächst“. Diesen Überlegungen folgend sollte also die Erfolgswahrscheinlichkeit auch bei großer Unterlegenheit von Spielerinnen niemals 0 sein, da dies bedeuten würde, dass solche Spielerinnen definitiv verlieren müssten und damit erst gar nicht antreten bräuchten. Vielmehr sollte sich die Erfolgswahrscheinlichkeit asymptotisch der 0 annähern. Gleicher gilt für Spielerinnen, die stärker sind als die Gegnerin. Egal wie stark diese Spielerinnen sind, ihre Erfolgswahrscheinlichkeit wird niemals 1 sein (was hieße, dass diese Spielerinnen zwingend gewinnen müssen). Auch im Fall einer sehr großen Überlegenheit von Spielerinnen gegenüber der Gegnerin nähert sich daher die Erfolgswahrscheinlichkeit der 1 asymptotisch an.

Asymptotische Annäherung der
Erfolgswahrscheinlichkeit an 0 und 1

Tab. 2.2 Wie wird das Match gegen die Gegnerin ausgehen?

	Spielerin E (100)	Spielerin D (80)	Spielerin C (50)	Spielerin B (20)	Spielerin A (1)
Gegnerin „G“ (50)	Wahrscheinlichkeit eines Siegs von Spielerin E ist sehr gering	Wahrscheinlichkeit eines Siegs von Spielerin D ist eher gering	Wahrscheinlichkeit eines Siegs von Spielerin C ist genauso hoch wie Wahrscheinlichkeit einer Niederlage	Wahrscheinlichkeit eines Siegs von Spielerin B ist eher hoch	Wahrscheinlichkeit eines Siegs von Spielerin A ist sehr hoch

In Klammern sind die Weltranglistenplätze angegeben

Serena Williams verliert gegen Virginie Razzano

Bis zum 29.05.2012 hatte Serena Williams von 46 Erstrunden-Matches bei großen Tennisturnieren alle gewonnen. Als eine der besten Tennisspielerinnen ihrer Zeit verwundert dies auch nicht, da gute Spielerinnen bei Tennisturnieren meist gesetzt werden und eher schwächere Spielerinnen als Gegnerin in der 1. Runde zugelost bekommen. Dies war auch am 29.05.2012 so, als Serena Williams gegen die Weltranglisten 111., Virginie Razzano antrat. Nach 3 h und 3 min verwandelte Virginie Razzano schließlich, zur Überraschung aller, ihren 8. Matchball und gewann gegen Serena Williams (ESPN 2012). Deren Erstrundenstatistik verschlechterte sich somit auf 46 : 1 (Siege: Niederlagen) bzw. $p_{(x = \text{Sieg})} = .979$.



French Open 2012: Virginie Razzano besiegt Serena Williams. (© Dubreuil Corinne/Abaca / picture alliance)

Differenz der Fähigkeit von Spielerin und Gegnerin

2. Differenz der Fähigkeit von Spielerin und Gegnerin beeinflusst das Verhältnis von Erfolgs- zu Misserfolgswahrscheinlichkeit Für die intuitive Beurteilung, ob und in welchem Ausmaß die Erfolgswahrscheinlichkeit über der Misserfolgswahrscheinlichkeit liegt, ist es sinnvoll, die Fähigkeit einer Spielerin (als Schätzung der Fähigkeit lässt sich hier der Weltranglistenplatz verwenden) mit der Fähigkeit der Gegnerin zu vergleichen. Es gilt: Je deutlicher eine Spielerin der Gegnerin in der Fähigkeit überlegen ist, desto größer ist die Erfolgswahrscheinlichkeit (der Spielerin) im Vergleich zur Misserfolgswahrscheinlichkeit. Je deutlicher eine Spielerin der Gegnerin in der Fähigkeit unterlegen ist, desto geringer ist die Erfolgs- und desto größer ist die Misserfolgswahrscheinlichkeit (der Spielerin).

3. Bei gleicher Fähigkeit von Spielerin und Gegnerin sind Erfolgs- und Misserfolgswahrscheinlichkeit gleich groß Aus der Annahme, dass die Differenz der Fähigkeit von Spielerin und Gegnerin die Erfolgswahrscheinlichkeit beeinflusst, ergibt sich zwangsläufig die Frage: Wie verhalten sich Erfolgs- und Misserfolgswahrscheinlichkeit, wenn Spielerin und Gegnerin exakt gleich gut sind? Hier ist es sinnvoll, von einer gleich großen Erfolgs- und Misserfolgswahrscheinlichkeit auszugehen. Spielerin und Gegnerin würden in solchen Fällen ebenso wahrscheinlich gewinnen wie verlieren. Da es keine anderen Ereignisse als „gewinnen“ und „verlieren“ gibt, muss in solchen Fällen für beide Ereignisse (gewinnen bzw. verlieren) eine Wahrscheinlichkeit von .50 angenommen werden.

Aus diesen Annahmen lässt sich ein erstes Modell ableiten, das die Erfolgswahrscheinlichkeit verschiedener Spielerinnen in Abhängigkeit von ihren eigenen Fähigkeiten und der Fähigkeit einer Gegnerin beschreibt.

Die blaue Linie in Abb. 2.3 beschreibt den Verlauf der Wahrscheinlichkeit, gegen die als „Gegnerin“ bezeichnete Person zu gewinnen, in Abhängigkeit von der Fähigkeit der jeweiligen Spielerin. Alle 3 zuvor genannten Annahmen sind berücksichtigt. Man kann erkennen, dass sich die Erfolgswahrscheinlichkeiten im Bereich von ① asymptotisch der 0 und im Bereich von ② asymptotisch der 1 annähern. Im Bereich von ① beschreibt die Linie die Erfolgswahrscheinlichkeit für Spielerinnen mit vglw. geringer Fähigkeit (z. B. Spielerin E). Wie man am Verlauf der Linie sieht, ist deren Erfolgswahrscheinlichkeit gering. Im Bereich von ② beschreibt die Linie die Erfolgswahrscheinlichkeit von Spielerinnen mit vglw. hoher Fähigkeit (z. B. Spielerin A). Wie man am Verlauf der Linie sieht, ist deren Erfolgswahrscheinlichkeit hoch.

Die Annahme, dass die Differenz der Fähigkeit einer Spielerin und der Gegnerin die Erfolgswahrscheinlichkeit beeinflusst, ist durch den Verlauf der blauen Linie ebenfalls berücksichtigt. Je weiter die Fähigkeit der Spielerin unter der Fähigkeit der Gegnerin liegt, desto geringer wird deren Erfolgswahrscheinlichkeit. Je weiter die Fähigkeit der Spielerin über der Fähigkeit der Gegnerin liegt, desto größer wird ihre Erfolgswahrscheinlichkeit.

Bei gleicher Fähigkeit von Spielerin und Gegnerin sind die Erfolgs- und die Misserfolgswahrscheinlichkeit gleich groß

Verlauf der Erfolgswahrscheinlichkeit in Abhängigkeit von der Fähigkeit

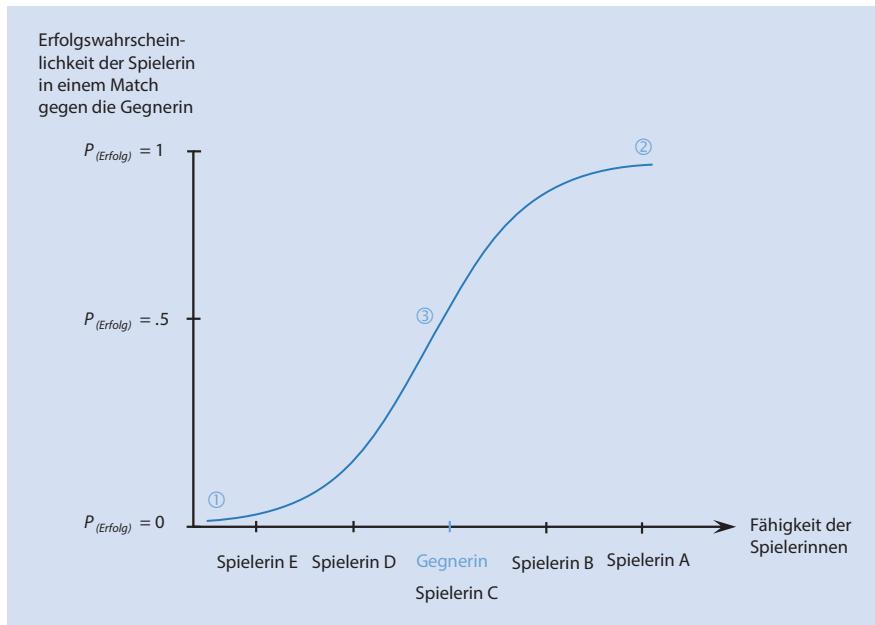


Abb. 2.3 Verlauf der Wahrscheinlichkeit, gegen die Gegnerin zu gewinnen, in Abhängigkeit von der Fähigkeit der Spielerinnen

Testitems als „Gegnerinnen“

Items und Personen können auf dem gleichen Kontinuum miteinander verglichen werden

Bei ③ verläuft die blaue Linie durch $p_{(\text{Erfolg})} = .50$, was der Annahme entspricht, dass für Spielerinnen, deren Fähigkeit exakt der Fähigkeit der Gegnerin entspricht (z. B. Spielerin C), die Erfolgs- und die Misserfolgswahrscheinlichkeit gleich groß sind.

Das Beispiel aus dem Bereich des Sports ist übertragbar auf Testitems, deren Antworten dichotom sind (z. B. richtige vs. falsche Lösung). Ein Testitem fungiert nun als „Gegnerin“ – es „verlangt“ eine gewisse Fähigkeit, um gelöst zu werden. Personen, die den Test bearbeiten und über eine deutlich höhere Fähigkeit als die von einem Item verlangte Fähigkeit verfügen, werden es mit größerer Wahrscheinlichkeit lösen als Personen, die nur eine leicht höhere als die verlangte Fähigkeit aufweisen. Personen, deren Fähigkeit exakt der vom Item verlangten Fähigkeit entspricht, haben eine genauso große Lösungs- wie Nichtlösungswahrscheinlichkeit (also von .50). Je weiter die Fähigkeit von Personen die vom Item verlangte Fähigkeit unterschreitet, desto weiter nähert sich die Lösungswahrscheinlichkeit der 0 an.

Man könnte also auch formulieren, dass Items und Personen gegeneinander antreten und sich miteinander messen. „Sich miteinander messen“ bedeutet auch, dass Items und Personen auf dem gleichen Kontinuum rangieren und miteinander verglichen werden. Daher ist in Abb. 2.4 auch das Item auf der gleichen Achse wie die Personen abgetragen. Obwohl nachfolgend zur Beschreibung von Items von „Itemschwierigkeit“ und zur Beschreibung von Personen von „Personenfähigkeit“ gesprochen wird, handelt es sich hierbei eigentlich um den gleichen Parameter. Daher spricht man auch vom einparametrischen logistischen Modell (kurz: 1PL-Modell, engl. one-parameter logistic model).

Die unter 2. und 3. zuvor formulierten Annahmen – dass die Differenz zwischen Personenfähigkeit und Itemschwierigkeit das Verhältnis von Erfolgs- zu Misserfolgswahrscheinlichkeit beeinflusst und dass bei gleicher Personenfähigkeit und Itemschwierigkeit die Erfolgs- und Misserfolgswahrscheinlichkeit gleich groß sind (s. o.) – lassen sich zunächst einmal zu einer einfachen Formel zusammenfassen. (Es soll vorab betont werden, dass diese Formel in der hier gewählten Form nicht zutrifft; sie dient lediglich der Herleitung der eigentlichen Formel des einparametrischen dichotomen Rasch-Modells.)

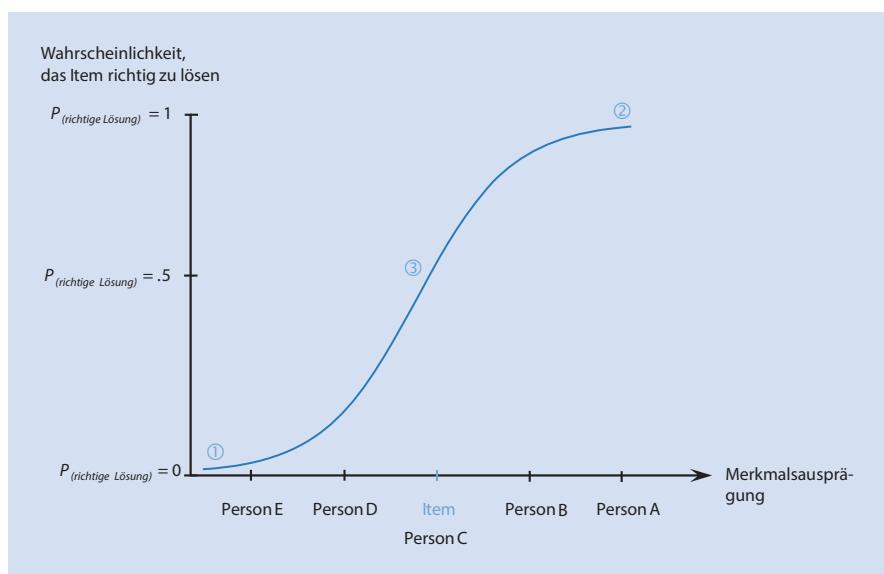


Abb. 2.4 Verlauf der Wahrscheinlichkeit, ein Item richtig zu lösen, in Abhängigkeit von der Merkmalsausprägung (Fähigkeit) der Personen

Differenz von Personenfähigkeit und Itemschwierigkeit als Funktion des Quotienten aus Erfolgs- und Misserfolgswahrscheinlichkeit

$$\xi_v - \sigma_i = \frac{p(X_{vi}=1)}{p(X_{vi}=0)}$$

ξ_v = Fähigkeit einer Spielerin v oder Merkmalsausprägung einer Person v

σ_i = Fähigkeit der Gegnerin i oder die von einem Item i verlangte Merkmalsausprägung (d. h. dessen Schwierigkeit)

$p(X_{vi}=1)$ = Erfolgswahrscheinlichkeit (in Match x für Spielerin v gegenüber Gegnerin i oder bei Itemantwort x von Person v in Item i)

$p(X_{vi}=0)$ = Misserfolgswahrscheinlichkeit

$$\frac{p(X_{vi}=1)}{p(X_{vi}=0)} = \text{Wettquotient}$$

Wettquotient = Verhältnis von Erfolgs- zu Misserfolgswahrscheinlichkeit

Die obige Formel beschreibt somit die Annahme, dass die Differenz der Fähigkeit von Spielerin und Gegnerin das Verhältnis von Erfolgs- zu Misserfolgswahrscheinlichkeit beeinflusst. Das Verhältnis von Erfolgs- zu Misserfolgswahrscheinlichkeit wird auch als Wettquotient bezeichnet (Rost 2004). Dieser Quotient wird bei Wetten gerne verwendet. Wenn die Chancen 3 : 1 stehen, dass Spielerin A gewinnt, meint dies, dass die Erfolgswahrscheinlichkeit von Spielerin A $3 \times$ größer ist als ihre Misserfolgswahrscheinlichkeit. Allerdings wurde in den Abb. 2.3 und 2.4 auf der Ordinate (y-Achse) nicht das Verhältnis zweier Wahrscheinlichkeiten bzw. der Wettquotient abgetragen, sondern nur eine Wahrscheinlichkeit: die Erfolgs- bzw. Lösungswahrscheinlichkeit. Daher ist die oben genannte Formel nicht unmittelbar zur Beschreibung der Kurven in Abb. 2.3 und 2.4 geeignet – sie muss zunächst nach $p(x_{vi}=1)$ umgestellt werden.

Umstellen des Wettquotienten

$$\xi_v - \sigma_i = \frac{p(X_{vi}=1)}{p(X_{vi}=0)}$$

$$(\xi_v - \sigma_i) \times p(X_{vi}=0) = p(X_{vi}=1)$$

$$(\xi_v - \sigma_i) \times (1 - p(X_{vi}=1)) = p(X_{vi}=1)$$

$$\xi_v - \sigma_i - (\xi_v - \sigma_i) \times p(X_{vi}=1) = p(X_{vi}=1)$$

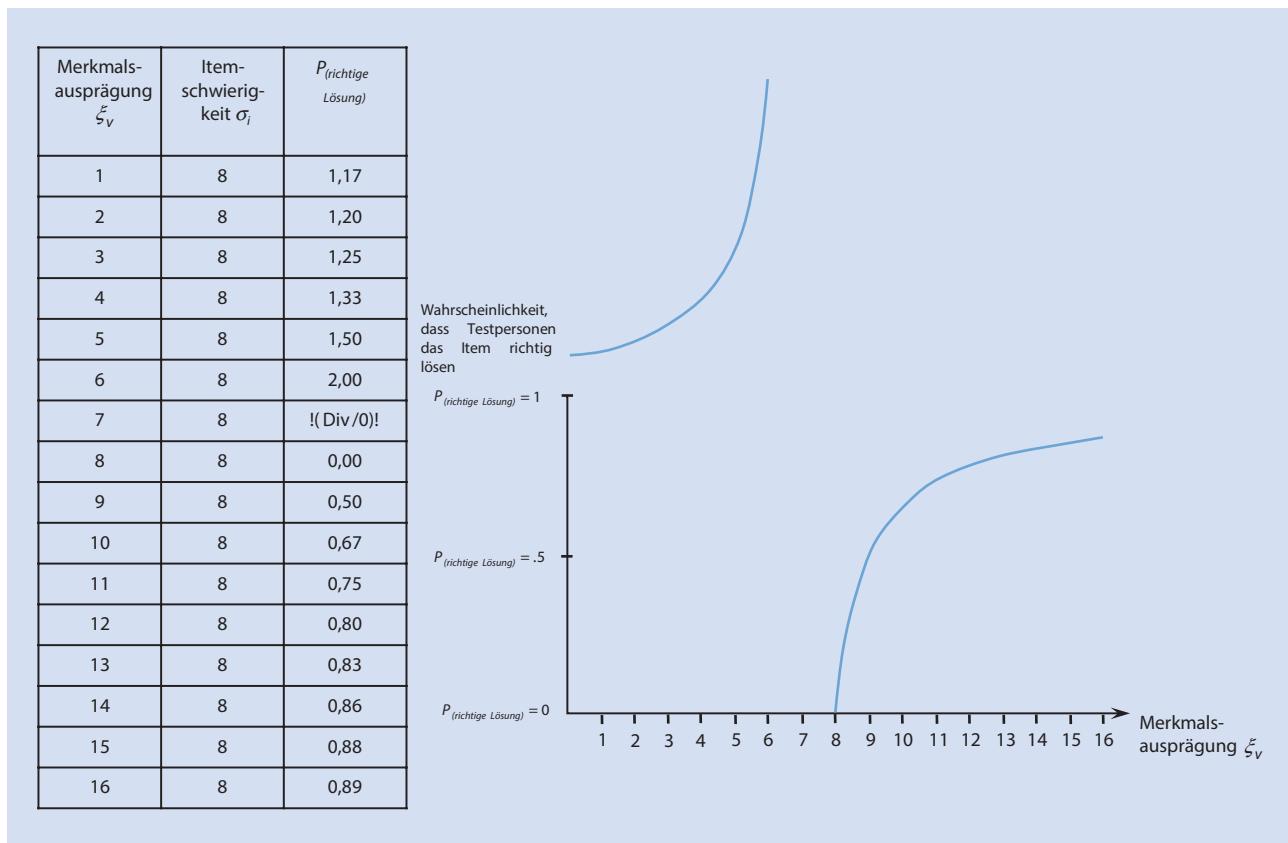
$$\xi_v - \sigma_i = p(X_{vi}=1) + (\xi_v - \sigma_i) \times p(X_{vi}=1)$$

$$\xi_v - \sigma_i = p(X_{vi}=1) \times (1 + (\xi_v - \sigma_i))$$

$$p(X_{vi}=1) = \frac{\xi_v - \sigma_i}{1 + (\xi_v - \sigma_i)}$$

(vgl. Rost 2004, S. 119)

Mit der so umgestellten Formel ist man schon einen Schritt näher an der adäquaten Beschreibung des Verlaufs der Kurven in Abb. 2.3 und 2.4, aber noch nicht am Ziel. Dies wird deutlich, wenn man ein Item mit einer Schwierigkeit σ_i von 8 sowie mehrere Fähigkeitsparameter ξ_v annimmt und sich den daraus resultierenden Kurvenverlauf ansieht (Abb. 2.5). Wie man sieht, ergibt diese Formel noch keinen Sinn. Dies wird an den Wahrscheinlichkeitswerten > 1 , einer nicht berechenbaren Wahrscheinlichkeit (Division durch 0) und einem Verlauf der Kurve, der von den bisher angenommenen Verläufen (Abb. 2.3 und 2.4) und der damit verbundenen Logik abweicht, deutlich.



■ Abb. 2.5 Unsinniger Verlauf der Wahrscheinlichkeit, ein Item richtig zu lösen, in Abhängigkeit von der Fähigkeit der Personen bei Annahme der vorherigen Formel (umgestellter Wettquotient)

Auch die Annahme, dass die Lösungswahrscheinlichkeit .50 ist, wenn Personenfähigkeit und Itemschwierigkeit gleich sind, ist durch diese Formel nicht berücksichtigt. Der Wettquotient kann also nicht angemessen die Differenz aus Personenfähigkeit und Itemschwierigkeit beschreiben; nicht zuletzt da er nur Werte von 0 bis $+\infty$ annehmen kann, die Differenz aus Personenfähigkeit und Itemschwierigkeit aber auch negativ sein kann.

Anders verhält es sich jedoch, wenn man statt des Wettquotienten den logarithmierten Wettquotienten verwendet (Rost 2004, S. 117).

Grundgleichung des einparametrischen dichotomen Rasch-Modells

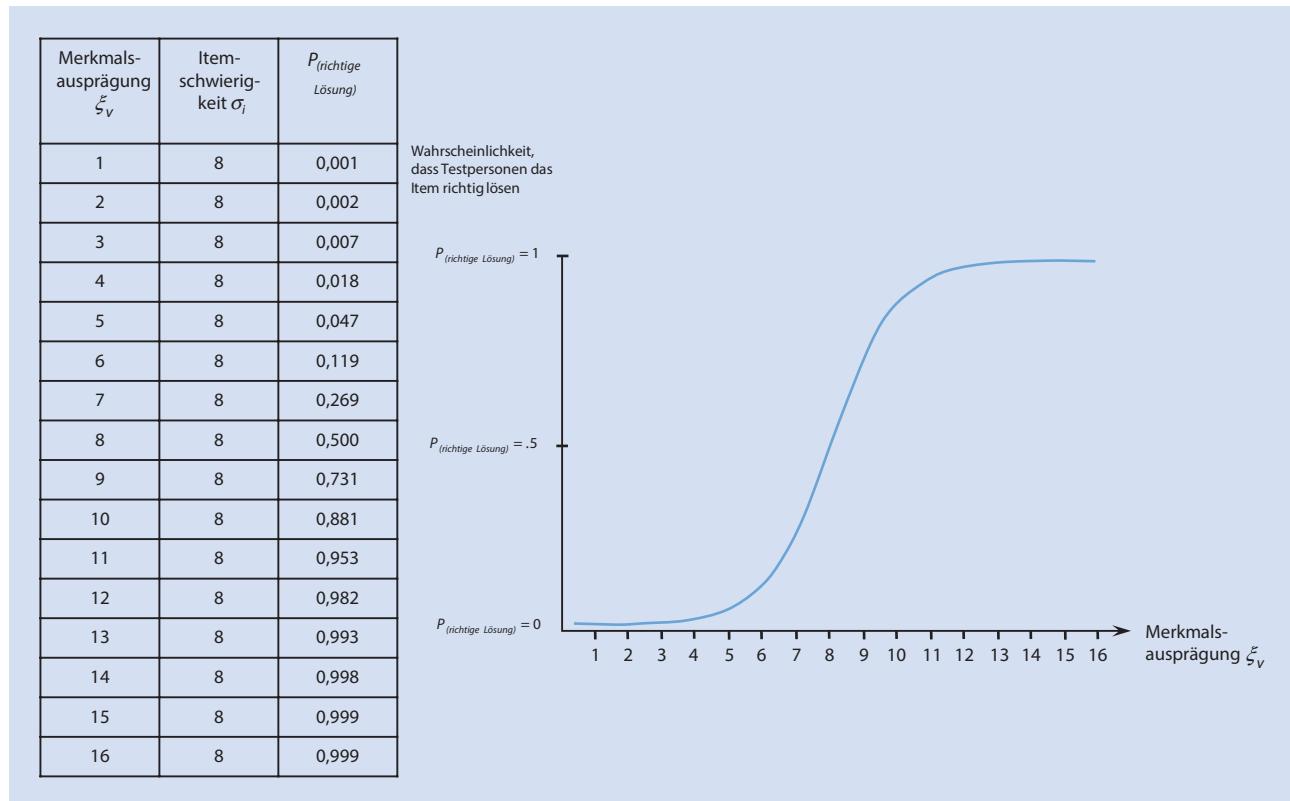
$$\ln\left(\frac{P(X_{vi}=1)}{P(X_{vi}=0)}\right)$$

Durch Umformung von

$$\xi_v - \sigma_i = \ln\left(\frac{P(X_{vi}=1)}{P(X_{vi}=0)}\right)$$

erhält man die Grundgleichung des einparametrischen dichotomen Rasch-Modells (Umformung analog zur vorherigen Umstellung des Wettquotienten):

$$P(X_{vi}=1) = \frac{e^{(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$



■ Abb. 2.6 Verlauf der Wahrscheinlichkeit, ein Item richtig zu lösen, in Abhängigkeit von der Merkmalsausprägung (Fähigkeit) der Personen

Die *Grundgleichung des einparametrischen dichotomen Rasch-Modells* beschreibt den in ■ Abb. 2.3 und 2.4 dargestellten Verlauf der Lösungswahrscheinlichkeit in Abhängigkeit von der Fähigkeit der Personen und der Schwierigkeit eines Items (vormals auch als „von Items verlangte Fähigkeit“ bezeichnet). Dies lässt sich an dem bereits in ■ Abb. 2.6 gewählten Zahlenbeispiel gut verdeutlichen. Setzt man beispielsweise in die Formel des dichotomen Rasch-Modells

$$P(X_{vi}=1) = \frac{e^{(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

für ξ_v und σ_i die gleichen Werte ein, so erhält man

$$P(X_{vi}=1) = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = .50,$$

was entsprechend der zuvor erfolgten Argumentation Sinn ergibt.

Die Grundgleichung des dichotomen Rasch-Modells lässt sich natürlich auch so umformen, dass sie die Misserfolgwahrscheinlichkeit, d. h. die Wahrscheinlichkeit, ein Item nicht zu lösen, beschreibt.

Grundgleichung des dichotomen Rasch-Modells

Misserfolgwahrscheinlichkeit

$$P(X_{vi}=0) = \frac{1}{1 + e^{(\xi_v - \sigma_i)}}$$

(Rost 2004, S. 119)

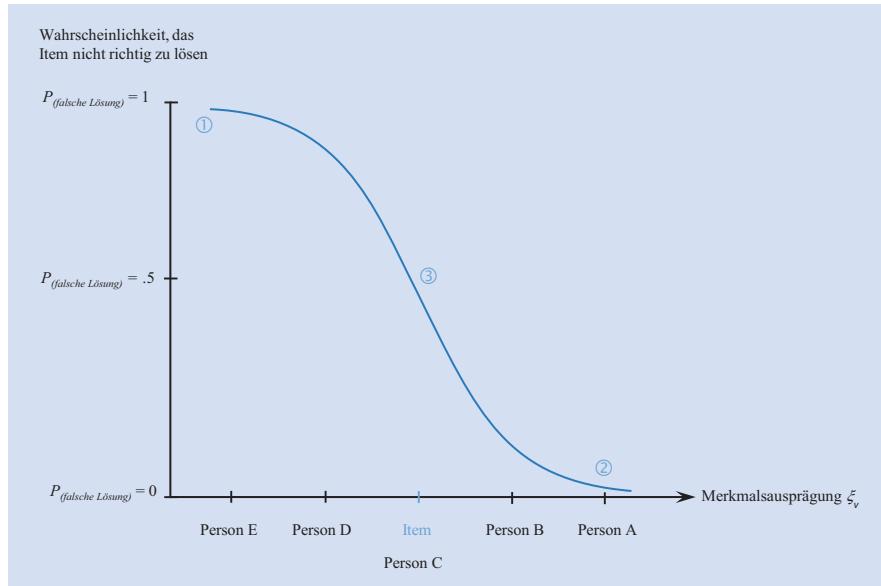


Abb. 2.7 Verlauf der Wahrscheinlichkeit, ein Item nicht richtig zu lösen, in Abhängigkeit von der Merkmalsausprägung (Fähigkeit) der Personen

Verlauf der Misserfolgswahrscheinlichkeit

Der Verlauf der Misserfolgswahrscheinlichkeit für ein fiktives Item i ist in **Abb. 2.7** dargestellt. Sinnvollerweise ist die Wahrscheinlichkeit, ein Item nicht zu lösen, umso höher, je weiter die Fähigkeit der Personen unter der vom Item verlangten Fähigkeit liegt. Folglich nähert sich die Kurve im Bereich von ① asymptotisch der 1. Die Wahrscheinlichkeit, ein Item nicht zu lösen, nimmt ab, je weiter die Fähigkeit der Testperson über der vom Item verlangten Fähigkeit liegt. Daher nähert sich im Bereich von ② die Kurve asymptotisch der 0 an. Bei ③ verläuft die blaue Linie durch $P_{(\text{falsche Lösung})} = .50$, was der Annahme entspricht, dass für Personen, deren Fähigkeit exakt der Schwierigkeit des Item entspricht, Lösungs- und Nichtlösungswahrscheinlichkeit gleich groß sind.

Allgemeinere Formulierung der Grundgleichung des dichotomen Rasch-Modells

Bislang haben wir 2 scheinbar unterschiedliche Formeln für die Lösungswahrscheinlichkeit, $p(x_{vi}=1)$, und für die Nichtlösungswahrscheinlichkeit, $p(x_{vi}=0)$, verwendet. Eine allgemeinere Formel, die sowohl Lösungs- als auch Nichtlösungswahrscheinlichkeiten abbildet, ist (Rost 2004, S. 119):

$$p(X_{vi}) = \frac{e^{x_{vi}(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

Durch Einsetzen von $X_{vi}=1$ (Item gelöst) oder $X_{vi}=0$ (Item nicht gelöst) entstehen die bisher verwendeten Formeln für die Lösungswahrscheinlichkeit

$$p(X_{vi}=1) = \frac{e^{(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

bzw. für die Nichtlösungswahrscheinlichkeit

$$p(X_{vi}=0) = \frac{1}{1 + e^{(\xi_v - \sigma_i)}}.$$

Die Fähigkeit, die ein Item verlangt, d. h. die Itemschwierigkeit, ist definiert als der Punkt auf dem Fähigkeitsspektrum, an dem Lösungs- und Nichtlösungswahrscheinlichkeit gleich groß, also jeweils .50 sind. Die Sinnhaftigkeit dieser Festlegung lässt sich an unserem Tennisbeispiel gut erläutern. Angenommen, der Weltranglistenplatz (d. h. die Fähigkeit) der als „Gegnerin“ bezeichneten Person wäre unbekannt, die Fähigkeit der Spielerinnen A bis E wäre jedoch bekannt. Würde man die Spielerinnen E und D immer wieder gegen die Gegnerin antreten lassen, so sollten beide Spielerinnen häufiger verlieren als gewinnen. Ihre Fähigkeit, d. h. ihr Weltranglistenplatz, wäre keine gute Schätzung der Fähigkeit der Gegnerin – die Gegnerin ist besser. Würde man die Spielerinnen B und A immer wieder gegen die Gegnerin antreten lassen, so sollten beide Spielerinnen häufiger gegen die Gegnerin gewinnen als verlieren. Ihre Fähigkeit, d. h. ihr Weltranglistenplatz, wäre auch keine gute Schätzung der Fähigkeit der Gegnerin – sie ist schlechter. Würde man Spielerin C mehrfach gegen die Gegnerin antreten lassen und feststellen, dass in 50 % der Matches Spielerin C und in 50 % der Matches die Gegnerin gewinnt, so würde man sinnvollerweise schließen, dass beide gleich gut sind. Somit entspräche die Fähigkeit der Gegnerin dem Punkt, an dem gleich fähige Spielerinnen eine Erfolgswahrscheinlichkeit von .50 haben (Punkt ③ in Abb. 2.3 und 2.4). Die gleiche Logik verwendet man für ein Testitem. Seine Schwierigkeit entspricht der Fähigkeit der Personen, die für dieses Item eine Lösungswahrscheinlichkeit von .50 aufweisen.

Itemschwierigkeit im dichotomen Rasch-Modell

Definition

Die **Itemschwierigkeit** ist definiert als der Punkt auf dem Merkmalskontinuum, an dem die Lösungswahrscheinlichkeit $p(X_{vi} = 1) = .50$ beträgt.

Was ist eine itemcharakteristische Kurve?

Die blaue Linie in Abb. 2.3 beschreibt die Erfolgswahrscheinlichkeit in Abhängigkeit von der Fähigkeit der Spielerinnen und der Fähigkeit einer (!) Gegnerin. Gleichermassen beschreibt die blaue Linie in Abb. 2.4 die Lösungswahrscheinlichkeit in Abhängigkeit von der Fähigkeit der Testpersonen und der Schwierigkeit eines (!) Items. Diese Linie ist also *charakteristisch für ein bestimmtes Item*. Dies wird auch durch die Bezeichnung der Linie als itemcharakteristische Kurve (engl. item characteristic curve, ICC) verdeutlicht. Für weitere Items – und ein Test besteht ja in den allermeisten Fällen aus mehr als einem Item – benötigt man also weitere itemcharakteristische Kurven.

Itemcharakteristische Kurven

In der Grundgleichung des einparametrischen dichotomen Rasch-Modells werden Items durch deren Schwierigkeit σ_i beschrieben. Um also mithilfe der selben Modellgleichung mehrere itemcharakteristische Kurven abzubilden, kann nur dieser Parameter verändert werden – die Fähigkeit der Testpersonen ändert sich ja nicht durch das Hinzufügen von Testitems. Eine Veränderung von σ_i in der Modellgleichung bewirkt, dass sich die itemcharakteristischen Kurven parallel zueinander nach rechts oder links auf dem Merkmalskontinuum verschieben. In Abb. 2.8 sind die itemcharakteristischen Kurven von 3 unterschiedlich schweren Items dargestellt. Da die Modellgleichung keinen Parameter für die Steigung der itemcharakteristischen Kurven vorsieht, sind diese alle gleich steil und damit parallel zueinander.

Parallele itemcharakteristische Kurven

Die Items 2 und 3 in Abb. 2.8 sind schwerer als Item 1, d. h., der Punkt, an dem Testpersonen mit einer bestimmten Fähigkeit eine Lösungswahrscheinlichkeit von .50 erreichen, hat sich weiter nach rechts verschoben (s. gepunktete und gestrichelte Linien).

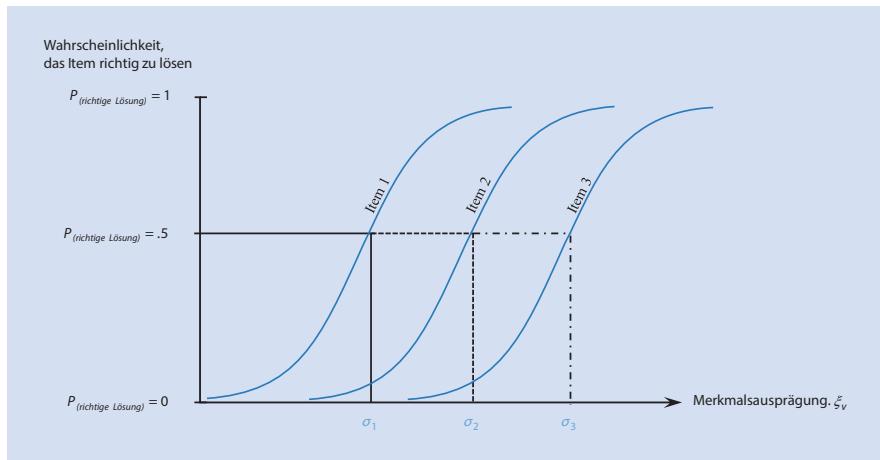


Abb. 2.8 Itemcharakteristische Kurven für 3 Items

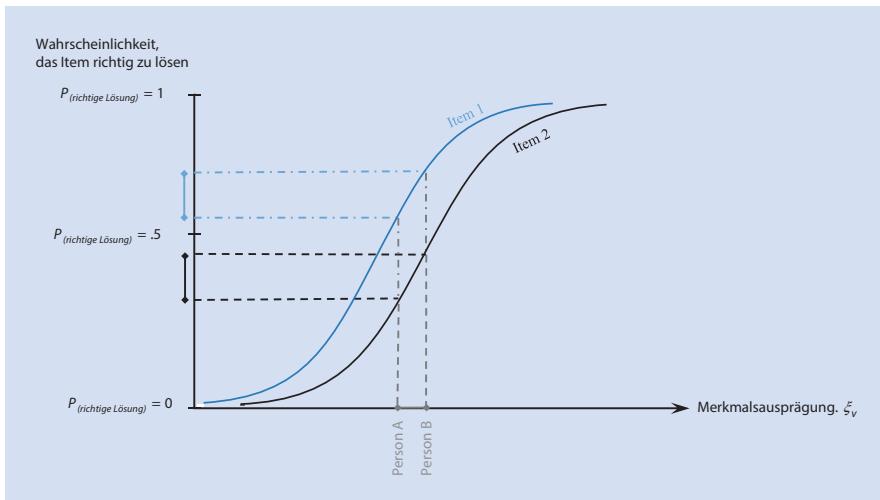
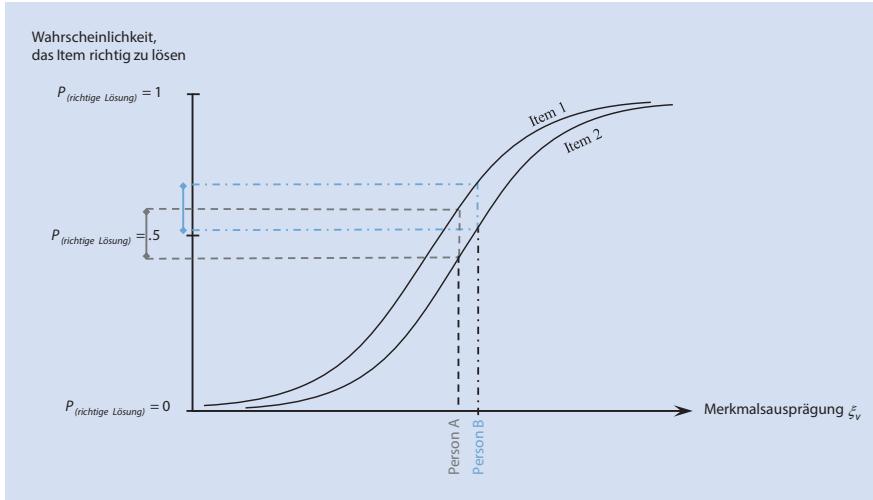


Abb. 2.9 Veranschaulichung der spezifischen Objektivität: Der Unterschied zwischen Person A und B auf dem Merkmalskontinuum (grau) resultiert in gleichen Abständen der Lösungswahrscheinlichkeiten (blau und schwarz) (Angelehnt an Bühner 2011, S. 505, © Pearson)

Parallel zueinander verlaufende, also gleich steile itemcharakteristische Kurven haben einen großen Vorteil: Unterschiede zwischen den Lösungswahrscheinlichkeiten von Personen verändern sich nicht in Abhängigkeit von dem gewählten Item.

Wir erinnern uns: Eine „fähigere“ Person sollte eine höhere Lösungswahrscheinlichkeit für ein Item aufweisen als eine weniger „fähige“ Person. Das führt dazu, dass der Abstand auf der Abszisse (x-Achse) auch zu einem Abstand auf der Ordinate (y-Achse) führt (s. gestrichelte blaue und schwarze Linien in Abb. 2.9). Allerdings bleibt der Abstand auf der Ordinate auch dann noch gleich groß, wenn man ihn anhand einer parallel verschobenen itemcharakteristischen Kurve abträgt. Das bedeutet, dass der Abstand der Lösungswahrscheinlichkeiten zwischen 2 Personen unabhängig von den

Spezifische Objektivität der Vergleiche



■ Abb. 2.10 Veranschaulichung der spezifischen Objektivität: Der Unterschied zwischen Item 1 und 2 ist unabhängig davon, ob man dafür Person A (grau) oder B (blau) heranzieht. (Angelehnt an Bühner 2011, S. 505, © Pearson)

verwendeten Items ist, sofern alle Items durch parallel zueinander verlaufenden itemcharakteristische Kurven beschrieben werden können. Oder anders gesagt: Zur Ermittlung des Fähigkeitsunterschieds zwischen Person A und B (■ Abb. 2.9) ist es egal, ob Item 1 oder Item 2 herangezogen wird (s. gleich lange blaue und schwarze Balken neben der Ordinate). Diese Eigenschaft des dichotomen Rasch-Modells nennt man *spezifische Objektivität der Vergleiche*.

Spezifische Objektivität der Vergleiche meint auch, dass Vergleiche von Items nicht davon abhängen, welche Personen man dazu heranzieht. Auch dies lässt sich am besten grafisch veranschaulichen. In ■ Abb. 2.10 sieht man, dass die Abstände der Lösungswahrscheinlichkeiten für beide Items gleich ausfallen, unabhängig ob man dafür Person A oder B betrachtet.

Eine einfache Möglichkeit, zu prüfen, ob für einen vorliegenden Test tatsächlich von spezifischer Objektivität der Vergleiche ausgegangen werden kann, bietet der grafische Modelltest. Bei diesem teilt man die Stichprobe in 2 Substichproben (z. B. ältere und jüngere Personen), berechnet die Itemschwierigkeiten getrennt für beide Stichproben und inspiziert dann anhand eines Streudiagramms, ob die getrennt geschätzten Schwierigkeitsparameter der Items konvergieren. Dies sieht man im Streudiagramm daran, dass sich die Schwierigkeiten größtenteils entlang einer Winkelhalbierenden anordnen (■ Abb. 2.11).

Den grafischen Modelltest könnte man auch so durchführen, dass entlang der Winkelhalbierenden Personen statt Items abgetragen werden. Hierbei teilt man die Items in 2 Teilmengen und schätzt die Personenparameter in diesen Teilmengen getrennt. So ließe sich prüfen, ob die spezifische Objektivität der Vergleiche auch für Teilmengen gilt (sog. „Itemhomogenität“).

Vergleiche von Items hängen nicht von den gewählten Personen ab

Grafischer Modelltest

Itemhomogenität

Wesentliche Annahmen des dichotomen Rasch-Modells

- Eindimensionalität
- Lokale stochastische Unabhängigkeit

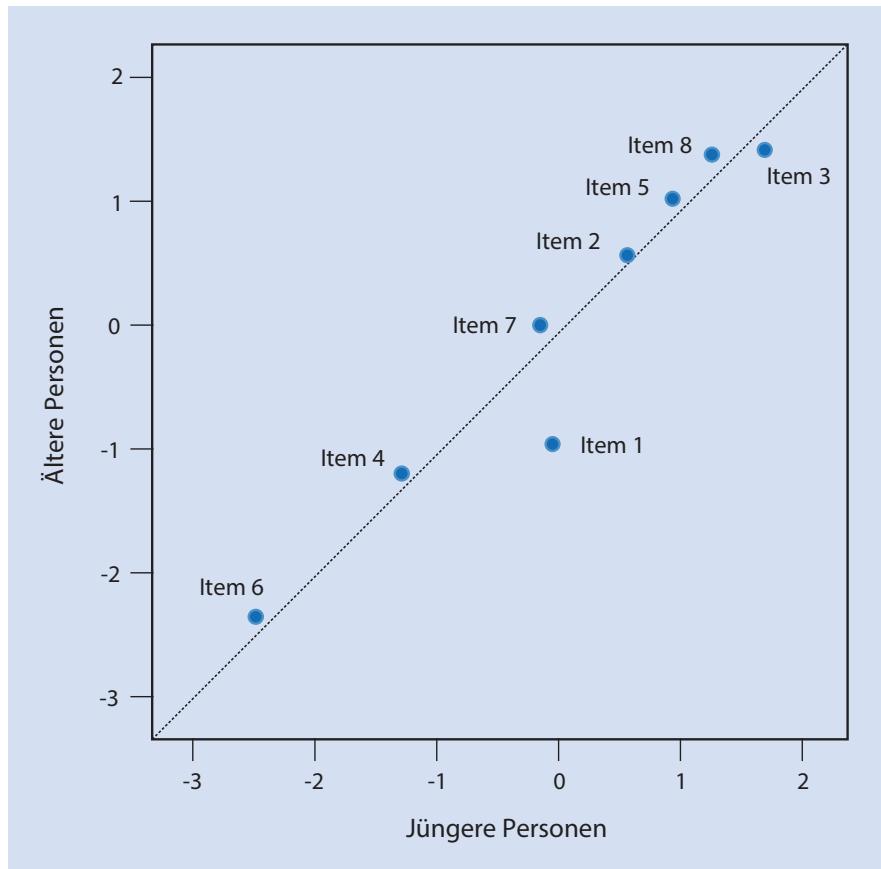


Abb. 2.11 Ergebnis eines grafischen Modelltests. (vgl. Eid und Schmidt 2014)

Annahme: Eindimensionalität

Um das Antwortverhalten von Personen bei der Beantwortung von Items auf nur einer Achse (d. h. nur anhand eines einzigen Kontinuums) beschreiben zu können, ist es zwingend notwendig, dass das Antwortverhalten auch nur aufgrund einer (latenten) Fähigkeit oder Eigenschaft zustande kommt. Diese Annahme des Rasch-Modells bezeichnet man als *Eindimensionalität* oder *Rasch-Homogenität* (vgl. Eid und Schmidt 2014). Sollte eine Spielerin aus unserem Tennisbeispiel nicht nur aufgrund ihrer Spielstärke (so könnte man die hinter der Abszisse in Abb. 2.3 liegende, latente Fähigkeit nennen) gegen die Gegnerin gewinnen oder verlieren, sondern noch weitere Eigenschaften eine Rolle spielen, wäre ein einziges Kontinuum nicht ausreichend, um die Erfolgswahrscheinlichkeiten der Spielerinnen abzutragen. Dies wäre beispielsweise der Fall, wenn die Matches auf unterschiedlichen Tennisplatzbelägen (Hartplatz, Sand, Rasen etc.) ausgetragen würden und die Spielerinnen für diese Beläge unterschiedlich spezialisiert sind. Nehmen wir zur Verdeutlichung nur einmal an, Spielerin D (Weltranglistenplatz 80, Tab. 2.2) sei eine ausgesprochene Rasenplatzspezialistin, die im obigen Beispiel als „Gegnerin“ bezeichnete Person jedoch nicht. Die gesamte Logik des obigen Beispiels geriete ins Wanken, wenn das nächste Match zwischen Spielerin D und der Gegnerin auf einem Rasenplatz stattfände. Entgegen der obigen Annahmen würde man nun eher darauf wetten, dass Spielerin D gegen die Gegnerin gewinnt. Dabei hat man jedoch nicht nur eine einzige Eigenschaft (die allgemeine Spielstärke, ausgedrückt in Weltranglistenplätzen) berücksichtigt, sondern zusätzlich noch eine weitere Eigenschaft (die Präferenz für Tennisplatzbeläge) hinzugezogen. Es lässt sich also festhalten, dass die bisherigen Annahmen nur dann haltbar sind, wenn „gewinnen oder verlieren“ bzw.

„richtige oder falsche“ Itemantworten nur auf eine einzige Eigenschaft der Personen zurückzuführen sind.

Eine weitere Annahme des dichotomen Rasch-Modells ist die der *lokalen stochastischen Unabhängigkeit*. Grundsätzlich sind 2 Ereignisse dann stochastisch unabhängig, wenn die Wahrscheinlichkeit eines Ereignisses nicht durch ein anderes Ereignis beeinflusst wird. Dies lässt sich ebenfalls anhand des Tennisbeispiels verdeutlichen: Zwei Ereignisse könnten hier darin bestehen, dass alle Spielerinnen A bis E gegen die Gegnerin und danach gegen eine weitere Person (nennen wir sie „Gegnerin 2“) antreten. Bei stochastischer Unabhängigkeit hieße dies, dass die Siegchancen der Tennisspielerinnen A bis E in ihren Matches gegen Gegnerin 2 unbeeinflusst davon sein sollten, wie deren vorherige Matches gegen die ursprüngliche Gegnerin ausgefallen sind. Eine offensichtliche Verletzung der stochastischen Unabhängigkeit bestünde dann, wenn man bei einem Sieg gegen die Gegnerin eine längere Erholungspause und bei einer Niederlage eine kürzere Erholungspause bis zum Match gegen Gegnerin 2 erhielte. Dann würde der Ausgang des vorherigen Matches das Ergebnis des nachfolgenden Matches beeinflussen.

Aber auch ohne solch offensichtliche Verletzungen ist die Annahme der stochastischen Unabhängigkeit keineswegs trivial. Nehmen wir an, dass Spielerin A und Spielerin E zuerst gegen die Gegnerin antreten. Spielerin A gewinnt, Spielerin E verliert gegen die Gegnerin. Wenn man nun wetten müsste, welche der beiden Spielerinnen (A oder E) in ihrem nächsten Match gegen Gegnerin 2 gewinnt, so würde man (ohne weitere Informationen zu haben) sicherlich eher auf Spielerin A wetten. Somit hat man – zu Recht – das Ergebnis des vorherigen Matches berücksichtigt, um auf das Ergebnis des kommenden Matches zu wetten. Bei stochastischer Unabhängigkeit sollte das vorherige Match jedoch irrelevant für das Ergebnis des kommenden Matches sein. Natürlich ist es dies nicht: Spielerin A sollte zuvor gegen die Gegnerin aufgrund ihrer hohen Spielstärke gewonnen haben, diese kann sie natürlich auch gegen Gegnerin 2 einsetzen. Spielerin E sollte zuvor gegen die Gegnerin aufgrund ihrer geringen Spielstärke verloren haben. Diese wird ihr nun wahrscheinlich auch gegen Gegnerin 2 zum Nachteil gereichen. Somit entsteht eine stochastische Abhängigkeit der Ereignisse durch die Eigenschaft bzw. Fähigkeit, um die es bei diesem Spiel geht. Im Falle des Lösens von Testitems entsteht diese Abhängigkeit durch die Fähigkeit, die mit den Testitems gemessen wird. *Lokale* stochastische Unabhängigkeit besagt, dass Ereignisse dann stochastisch unabhängig sein sollten, wenn die interessierende Fähigkeit konstant gehalten wird (Rost 2004). Nehmen wir dazu an, dass Spielerin A und eine exakt gleich gute Spielerin A' gegen die Gegnerin antreten. Während Spielerin A wie erwartet gewinnt, verliert Spielerin A' – jedoch nicht aufgrund mangelnder Spielstärke, sondern „weil sie zufällig einen schwarzen Tag erwischt hat“. Auf wen sollte man nun in den danach folgenden Matches gegen Gegnerin 2 wetten, auf Spielerin A oder Spielerin A'? Wenn Spielerin A' sich von der vorherigen Niederlage erholt hat, so sollten ihre Chance genauso hoch sein wie die der gleich guten Spielerin A. Im Falle von Testitems besagt lokale stochastische Unabhängigkeit also: Ist die zu messende Eigenschaft bzw. Fähigkeit konstant, so sollte die Lösungswahrscheinlichkeit eines Items unabhängig von dem Ergebnis bei einem anderen Item sein.

Annahme: lokale stochastische Unabhängigkeit

Stochastische Unabhängigkeit bei konstanter Fähigkeit

- ! Bei der initialen Verwendung eines Tests sind die Itemschwierigkeiten unbekannt und müssen geschätzt werden. Die Personenparameter sind ebenfalls zunächst unbekannt – Sinn des Testens ist es ja gerade, diese zu ermitteln.

Bislang haben wir zur Veranschaulichung der zugrunde liegenden Logik des dichotomen Rasch-Modells so getan, als wäre die Schwierigkeit der Items und/oder die Fähigkeit von Personen bekannt. (Im Falle der verwendeten

Personeneigenschaft und Itemschwierigkeit gleichzeitig ermitteln

Tennisanalogie haben wir so getan, als würden wir die Spielstärke der Spielerinnen und der Gegnerin kennen). In Wahrheit geht es jedoch gerade darum, durch einen Test oder Fragebogen eine bislang unbekannte Eigenschaft möglichst gut einzuschätzen. Daher haben wir die wesentliche Frage bislang ausgespart: Wie kann man nun mithilfe der Grundgleichung des einparametrischen dichotomen Rasch-Modells die latente Eigenschaft von Personen möglichst gut ermitteln? (Bei Leistungstests wird die interessierende latente Eigenschaft häufig als Personenfähigkeit bezeichnet.) Diese Frage wäre einfach zu beantworten, wenn wir die Schwierigkeit der im Test enthaltenen Items kennen würden. Der bereits beschriebenen Logik entsprechend würden wir einer Person die Fähigkeit zuschreiben, die der Schwierigkeit der Items entspricht, die diese Person in 50 % der Fälle löst und in 50 % der Fälle nicht löst. Leider ist die Schwierigkeit der Items aber ebenfalls unbekannt. (Zudem wäre es selbst bei bekannter Itemschwierigkeit nicht praktikabel, das gleiche Item wiederholt zu applizieren und zu prüfen, in wie viel Prozent der Fälle es von einer Person gelöst würde.) Es wird deutlich, dass eine zentrale Herausforderung darin besteht, die Ausprägung der latenten Personeneigenschaft und der Itemschwierigkeit *gleichzeitig* zu ermitteln. Dazu stehen uns nur die Grundgleichung des dichotomen Rasch-Modells und die tatsächlichen Antworten der getesteten Personen auf die verwendeten Items zur Verfügung (s. hierzu z. B. □ Tab. 2.3).

Likelihood-Funktion

Zur Schätzung der Personeneigenschaft und der Itemschwierigkeit wird eine *Likelihood-Funktion* herangezogen (Rost 2004). Diese hilft uns dabei, Personen- und Itemparameter so zu schätzen, dass unter Annahme der Geltung des dichotomen Rasch-Modells das vorliegende Datenmuster (wie etwa das Beispiel für 4 Personen und 4 Items in □ Tab. 2.3) möglichst wahrscheinlich ist.

Definition

Unter **Likelihood** versteht man die Wahrscheinlichkeit der vorliegenden Daten unter Annahme der Geltung des zugrunde liegenden Modells (vgl. Rost 2004, S. 112).

Konkret heißt das: Wie wahrscheinlich ist das vorliegende Datenmuster, wenn wir zu dessen Berechnung die Formel des dichotomen Rasch-Modells

$$P(X_{vi}) = \frac{e^{x_{vi}(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

verwenden?

Zur Berechnung der Likelihood setzen wir die Grundgleichung des einparametrischen dichotomen Rasch-Modells zunächst in jede Zelle der □ Tab. 2.3 ein, wobei wir für gelöste Items (in diesen Zellen steht eine 1) $x_{vi}=1$ einsetzen

□ Tab. 2.3 Beispiel für 4 Itemantworten von 4 Personen

	Item 1	Item 2	Item 3	Item 4	Zeilensumme
Person 1	1	1	0	1	3
Person 2	1	1	1	0	3
Person 3	1	0	0	0	1
Person 4	0	1	0	0	1
Spaltensumme	3	3	1	1	

1 = Item gelöst, 0 = Item nicht gelöst

Tab. 2.4 Beispiel für 4 Itemantworten von 4 Personen mit eingefügter Grundgleichung des einparametrischen dichotomen Rasch-Modells. (Angelehnt an Moosbrugger 2012a)

	Item 1	Item 2	Item 3	Item 4	Zeilensumme
Person 1	$\frac{e^{(\xi_1-\sigma_1)}}{1+e^{(\xi_1-\sigma_1)}}$	$\frac{e^{(\xi_1-\sigma_2)}}{1+e^{(\xi_1-\sigma_2)}}$	$\frac{1}{1+e^{(\xi_1-\sigma_3)}}$	$\frac{e^{(\xi_1-\sigma_4)}}{1+e^{(\xi_1-\sigma_4)}}$	3
Person 2	$\frac{e^{(\xi_2-\sigma_1)}}{1+e^{(\xi_2-\sigma_1)}}$	$\frac{e^{(\xi_2-\sigma_2)}}{1+e^{(\xi_2-\sigma_2)}}$	$\frac{e^{(\xi_2-\sigma_3)}}{1+e^{(\xi_2-\sigma_3)}}$	$\frac{1}{1+e^{(\xi_2-\sigma_4)}}$	3
Person 3	$\frac{e^{(\xi_3-\sigma_1)}}{1+e^{(\xi_3-\sigma_1)}}$	$\frac{1}{1+e^{(\xi_3-\sigma_2)}}$	$\frac{1}{1+e^{(\xi_3-\sigma_3)}}$	$\frac{1}{1+e^{(\xi_3-\sigma_4)}}$	1
Person 4	$\frac{1}{1+e^{(\xi_4-\sigma_1)}}$	$\frac{e^{(\xi_4-\sigma_2)}}{1+e^{(\xi_4-\sigma_2)}}$	$\frac{1}{1+e^{(\xi_4-\sigma_3)}}$	$\frac{1}{1+e^{(\xi_4-\sigma_4)}}$	1
Spaltensumme	3	3	1	1	

Statt der bisher gewohnten, allgemeinen Bezeichnung von Personeneigenschaft und Itemschwierigkeit für eine beliebige Person v und ein beliebiges Item i sind hier spezifische Parameter für die Personen 1 bis 4 und Items 1 bis 4 eingefügt

und für nicht gelöste Items (in diesen Zellen steht eine 0) $x_{vi}=0$ in die Formel einsetzen (Tab. 2.4).

Die Likelihood lässt sich durch Multiplizieren der in allen Zellen der Matrix enthaltenen Wahrscheinlichkeiten ermitteln.

Berechnung der Likelihood

$$L = \prod_{v=1}^N \prod_{i=1}^k p(x_{vi}) = \prod_{v=1}^N \prod_{i=1}^k \frac{e^{x_{vi}(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

Angewendet auf das obige Beispiel der Personen v 1 bis 4 und der Items i 1 bis 4 bedeutet dies:

$$L = \frac{e^{(\xi_1-\sigma_1)}}{1+e^{(\xi_2-\sigma_1)}} \times \frac{e^{(\xi_1-\sigma_2)}}{1+e^{(\xi_1-\sigma_2)}} \times \frac{1}{1+e^{(\xi_1-\sigma_3)}} \times \frac{e^{(\xi_1-\sigma_4)}}{1+e^{(\xi_1-\sigma_4)}} \times \frac{e^{(\xi_2-\sigma_1)}}{1+e^{(\xi_2-\sigma_1)}} \times \frac{e^{(\xi_2-\sigma_2)}}{1+e^{(\xi_2-\sigma_2)}} \\ \times \frac{e^{(\xi_2-\sigma_3)}}{1+e^{(\xi_2-\sigma_3)}} \times \frac{1}{1+e^{(\xi_2-\sigma_4)}} \times \frac{e^{(\xi_3-\sigma_1)}}{1+e^{(\xi_3-\sigma_1)}} \times \frac{1}{1+e^{(\xi_3-\sigma_2)}} \times \frac{1}{1+e^{(\xi_3-\sigma_3)}} \\ \times \frac{1}{1+e^{(\xi_3-\sigma_4)}} \times \frac{1}{1+e^{(\xi_4-\sigma_1)}} \times \frac{e^{(\xi_4-\sigma_2)}}{1+e^{(\xi_4-\sigma_2)}} \times \frac{1}{1+e^{(\xi_4-\sigma_3)}} \times \frac{1}{1+e^{(\xi_4-\sigma_4)}}$$

Es fällt auf, dass der Nenner aller hierbei verwendeten Brüche nicht von den tatsächlichen Itemantworten (richtig oder falsch) beeinflusst ist. Man könnte ihn auch ohne die Itemantworten zu kennen, in diese Funktion einsetzen.

Wenn wir für den Moment einmal nur den Zähler beachten, dann steht dort:

$$e^{(\xi_1-\sigma_1)} \times e^{(\xi_1-\sigma_2)} \times 1 \times e^{(\xi_1-\sigma_4)} \times e^{(\xi_2-\sigma_1)} \times e^{(\xi_2-\sigma_2)} \times e^{(\xi_2-\sigma_3)} \\ \times 1 \times e^{(\xi_3-\sigma_1)} \times 1 \times 1 \times 1 \times e^{(\xi_4-\sigma_2)} \times 1 \times 1$$

oder kürzer:

$$e^{(\xi_1-\sigma_1)} \times e^{(\xi_1-\sigma_2)} \times e^{(\xi_1-\sigma_4)} \times e^{(\xi_2-\sigma_1)} \times e^{(\xi_2-\sigma_2)} \times e^{(\xi_2-\sigma_3)} \\ \times e^{(\xi_3-\sigma_1)} \times e^{(\xi_4-\sigma_2)}$$

wegen $e^x \times e^y = e^{x+y}$ lässt sich dies auch so formulieren:

$$e^{(\xi_1-\sigma_1) + (\xi_1-\sigma_2) + (\xi_1-\sigma_4) + (\xi_2-\sigma_1) + (\xi_2-\sigma_2) + (\xi_2-\sigma_3) + (\xi_3-\sigma_1) + (\xi_4-\sigma_2)} \\ = e^{(3 \times \xi_1) + (3 \times \xi_2) + (1 \times \xi_3) + (1 \times \xi_4) - (3 \times \sigma_1) - (3 \times \sigma_2) - (1 \times \sigma_3) - (1 \times \sigma_4)}$$

Es wird deutlich, dass der Personenparameter jeder Person so oft in die Berechnung der Likelihood eingeht, wie die Person Items gelöst hat. Zum Beispiel hat die 1. Person 3 Items gelöst, ihr Personenparameter geht $3 \times$ in die

Summenscore als erschöpfende Statistik

Berechnung der Likelihood ein. Ebenso geht jeder Itemparameter so oft in die Berechnung der Likelihood ein (wenn man den Nenner einmal außen vor lässt), wie das Item insgesamt gelöst wurde. Zum Beispiel wurde das 3. Item $1 \times$ gelöst, der Schwierigkeitsparameter dieses Items geht daher $1 \times$ in die Berechnung der Likelihood ein. Das bedeutet, dass es nicht mehr wichtig ist, welches Item eine Person gelöst hat. Die Randsumme, d. h. der *Summenscore*, ist ausreichend. Da auf Basis der Likelihood-Funktion die Personen- und Itemparameter geschätzt werden, ist auch hierfür nur die Randsumme relevant. Damit ist der Summenscore eine *erschöpfende Statistik*; es kommt nicht mehr darauf an, welche Items eine Person gelöst hat, sondern nur wie viele (vgl. Eid und Schmidt 2014). In unserem Beispiel ist es daher irrelevant, dass Person 1 andere Items löst als Person 2. Beide haben den gleichen Summenscore und damit auch den gleichen Personenparameter.

Noch lässt sich jedoch die Likelihood für unser Beispiel nicht berechnen: Dazu werden konkrete Werte für die Personen- und Itemparameter benötigt. Diese müssen geschätzt werden. Wie bereits erwähnt, ist der leitende Gedanke bei der Schätzung der Personen- und Itemparameter, diese so zu schätzen, dass die Likelihood-Funktion *maximiert* wird. Die dazu verfügbaren *Maximum-Likelihood-Schätzmethoden* lassen sich danach unterscheiden, ob

- a) zunächst nur die Itemparameter (Conditional-Maximum-Likelihood-Methode oder Marginal-Maximum-Likelihood-Methode) und danach erst die Personenparameter (Unconditional-Maximum-Likelihood-Methode) geschätzt werden oder
- b) Personen- und Itemparameter gleichzeitig geschätzt werden (Unconditional- oder Joint-Maximum-Likelihood-Methode).

Weiterführende Literatur

Diese und weitere Schätzverfahren sind ausführlich bei Bühner (2021), Eid und Schmidt (2014), Moosbrugger und Kelava (2020) und bei Rost (2004) beschrieben.

2.3.1.2 Erweiterung des einparametrischen logistischen Modells um Diskriminations- und Rateparameter

Die Variante des dichotomen Rasch-Modells, die wir bisher vorgestellt haben, ist ein sehr sparsame Modellvariante. Sie nimmt zwar 2 Parameter zur Spezifikation von Personenfähigkeit und Itemschwierigkeit an, streng genommen handelt es sich hierbei jedoch um den gleichen Parameter – Personen und Items werden auf dem gleichen Kontinuum abgetragen. Es wird daher auch als einparametrisches logistisches Modell (1PL-Modell) bezeichnet.

Es wurde bereits dargestellt, dass aus der Annahme nur eines Parameters auch folgt, dass alle itemcharakteristischen Kurven parallel verlaufen. Inhaltlich besagt dies, dass alle Items gleich gut zwischen Personen unterschiedlicher Fähigkeit diskriminieren – ein Anstieg in der Fähigkeit führt bei allen Items zum gleichen Anstieg der Lösungswahrscheinlichkeit. Dies mag jedoch eine etwas restriktive Annahme sein. Möglicherweise diskriminieren manche Items ja besser zwischen Fähigkeitsunterschieden der Personen als andere. Es ist durchaus denkbar, dass bei manchen Items ein kleiner Unterschied in der Fähigkeit von Personen zu einem deutlichen Unterschied in den Lösungswahrscheinlichkeiten führt. Es mag jedoch auch Items geben, bei denen ein deutlicher Unterschied in der Fähigkeit von Personen nur in einem kleinen Unterschied der Lösungswahrscheinlichkeiten resultiert. Die damit einhergehenden Verläufe der itemcharakteristischen Kurven sind in Abb. 2.12 exemplarisch dargestellt.

Maximum-Likelihood-Schätzmethoden

Sparsames Modell, da Itemschwierigkeit und Personenfähigkeit gleicher Parameter

Restriktive Annahme: gleiche Steigung der itemcharakteristischen Kurven

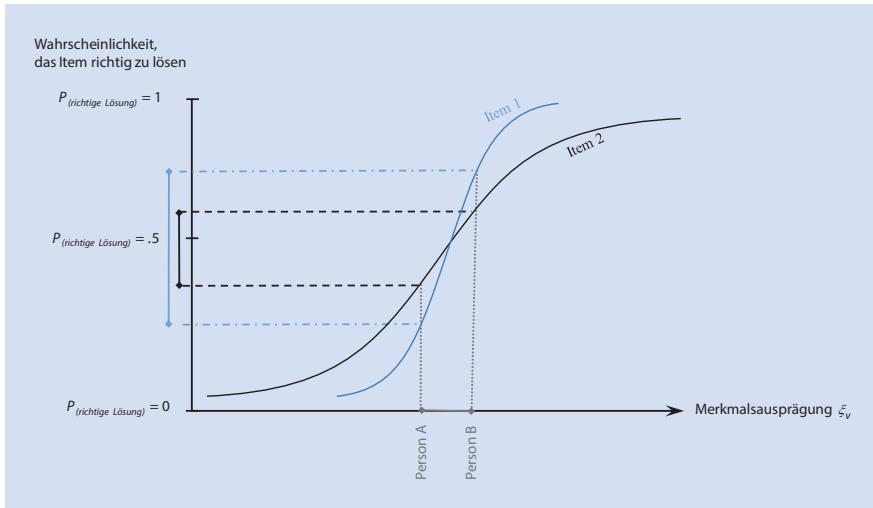


Abb. 2.12 Itemcharakteristische Kurven für Items mit unterschiedlichem Diskriminationsparameter. (Angelehnt an Bühner 2011, S. 504, © Pearson)

Wenn man zur Verdeutlichung wieder das Tennisbeispiel heranziehen will, so könnte man sich vorstellen, dass die blaue itemcharakteristische Kurve eine Gegnerin („Gegnerin 1“) beschreibt, die stets ihre normale Spielweise abruft. Die schwarze Linie könnte hingegen eine Gegnerin („Gegnerin 2“) beschreiben, die sich besonders stark anstrengt, je stärker die gegen sie antretenden Spielerinnen sind (Gegnerin ist hier analog zu Items zu verstehen, nicht zu Personen; vgl. ► Abschn. 2.3.1.1). Wenn stärkere Spielerinnen gegen die Gegnerinnen 1 und 2 antreten, ist es daher nachvollziehbar, dass die Wahrscheinlichkeit, gegen die Gegnerin 2 zu gewinnen, etwas geringer ist, als gegen Gegnerin 1 zu gewinnen. Es könnte weiterhin so sein, dass Gegnerin 2 sich nicht besonders anstrengt, wenn sie gegen schwächere Spielerinnen antritt. Schwächere Spielerinnen haben nun eine höhere Chance, gegen Gegnerin 2 als gegen Gegnerin 1 zu gewinnen.

Analogie: unterschiedlich motivierte Tennisspielerinnen

Wie man in Abb. 2.12 sieht, verlaufen itemcharakteristische Kurven von den Items, die besser zwischen kleinen Fähigkeitsunterschieden diskriminieren (Item 1), steiler. Wenn Items mit unterschiedlicher Steigung in ein und demselben Test enthalten sind, so muss unweigerlich davon ausgegangen werden, dass ein Parameter zur Beschreibung der itemcharakteristischen Kurven nicht mehr ausreicht. Es ist ein weiterer Parameter nötig, der die Steigung der itemcharakteristischen Kurven beschreibt. Diesen Parameter nennt man **Itemdiskriminationsparameter**.

Itemdiskriminationsparameter

Grundgleichung des dichotomen Rasch-Modells mit Diskriminationsparameter (Rost 2004, S. 133)

$$p_{(X_{vi})} = \frac{e^{x_{vi} \times \beta_i \times (\xi_v - \sigma_i)}}{1 + e^{\beta_i \times (\xi_v - \sigma_i)}}$$

β_i = Diskriminationsparameter des Items i

Dieses Modell wird auch als „zweiparametrisches logistisches Modell“ (kurz: 2PL-Modell) bezeichnet, da es einen Diskriminationsparameter als 2. Parameter vorsieht. Alternativ wird dieses Modell auch als „Birnbaum-Modell“ bezeichnet (da es 1968 erstmals von Allan Birnbaum diskutiert wurde; Rost 2004).

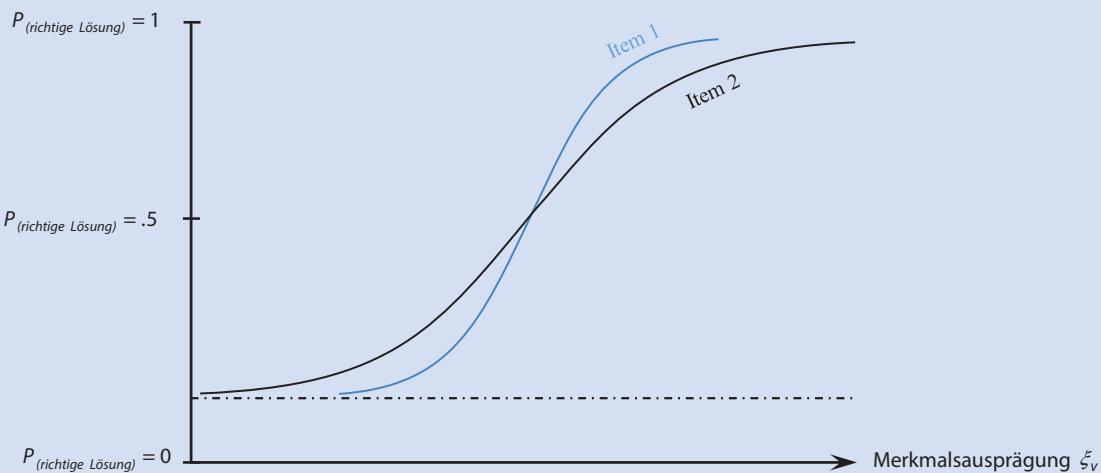
Manche Eigenschaften des 1PL-Modells gelten nicht mehr

Berücksichtigung der Ratewahrscheinlichkeit

In □ Abb. 2.12 sieht man auch sehr gut, dass eine Eigenschaft des einparametrischen logistischen Modells nun nicht mehr gilt: die spezifische Objektivität der Vergleiche. Der Unterschied zwischen den Personen A und B in der gemessenen Eigenschaft führt zu anderen Abständen auf der Ordinate (y-Achse), je nachdem welches Item man dazu heranzieht. Ebenso gilt nicht mehr, dass der Summenwert eine erschöpfende Statistik für die Itemantworten ist – es ist nun zu berücksichtigen, welche Items gelöst wurden.

Eine weitere Annahme der bisher vorgestellten dichotomen Rasch-Modelle mag für manche Testverfahren ein wenig zu kurz greifen. Bisher sind wir davon ausgegangen, dass sich die Lösungswahrscheinlichkeiten asymptotisch an 0 und 1 annähern. Dies erscheint dann sinnvoll, wenn eine richtige Lösung der Items nicht durch *Raten* zustande kommen kann (z. B. bei einem offenen Antwortformat) oder ausgeschlossen werden kann, dass Testpersonen von der Möglichkeit zu raten Gebrauch machen. Häufig muss jedoch bei einem Multiple-Choice-Antwortformat (► Abschn. 2.4.2.6) davon ausgegangen werden, dass Testpersonen raten, falls sie die richtige Antwort nicht wissen oder wenn zum Ende des Tests die Bearbeitungszeit verriegt. In diesen Fällen sollte sich die Lösungswahrscheinlichkeit nicht asymptotisch an 0 annähern, sondern sich an eine durch Raten bedingte höhere Wahrscheinlichkeit annähern (□ Abb. 2.13).

Wahrscheinlichkeit,
das Item richtig zu lösen



□ Abb. 2.13 Beispielhafte itemcharakteristische Kurven eines dreiparametrischen logistischen Modells. Die blaue und schwarze Linie stellen itemcharakteristische Kurven für 2 Items mit gleicher Ratewahrscheinlichkeit (gestrichelte horizontale Linie) dar

Dabei kann das Ausmaß, in dem Raten auftritt, über die Items hinweg variieren. Möglicherweise kommt bei manchen Items Raten nicht vor, und Personen, die die Antwort nicht kennen, kreuzen gar keine Alternative an. Die Ratewahrscheinlichkeit ist also keineswegs stets $1/4$ wenn 4 Antwortalternativen zur Verfügung stehen. Vielmehr verfügt jedes Item über einen eigenen Rateparameter.

Grundgleichung des dichotomen Rasch-Modells mit Diskriminations- und Rateparameter (Rost 2004, S. 135)

$$P(x_{vi}) = \gamma_i + (1 - \gamma_i) \frac{e^{x_{vi} \times \beta_i \times (\xi_v - \sigma_i)}}{1 + e^{\beta_i \times (\xi_v - \sigma_i)}}$$

γ_i = Rateparameter des Items i

Dieses Modell wird auch als dreiparametrisches logistisches Modell (kurz: 3PL-Modell) bezeichnet, da es einen Rateparameter als 3. Parameter vorsieht.

2.3.1.3 Vorteile und Nutzen dichotomer Rasch-Modelle

Möglicherweise fragt man sich, wozu diese, vielleicht kompliziert anmutenden Formeln und Modellannahmen nützlich sind. Daher soll an dieser Stelle auf einige Möglichkeiten der Nutzung dichotomer Rasch-Modelle eingegangen werden.

Dichotome Rasch-Modelle machen explizite Annahmen über das Zustandekommen von Itemantworten in Abhängigkeit von der zu messenden Eigenschaft von Personen und der Schwierigkeit der Testitems. Es stehen verschiedene Modelltests zur Verfügung, mit denen geprüft werden kann, ob diese Annahmen für einen vorliegenden Test haltbar sind. Bestätigt ein Modelltest die Geltung des einparametrischen dichotomen Rasch-Modells, so kann davon ausgegangen werden, dass spezifische Objektivität der Vergleiche, Eindimensionalität und lokale stochastische Unabhängigkeit für dieses Testverfahren gegeben sind sowie dass der Summenwert eine erschöpfende Statistik für Itemantworten darstellt.

Es ist zudem möglich, unterschiedlich restriktive Modelle (1PL-Modell vs. 2PL-Modell vs. 3PL-Modell) gegeneinander zu testen. Dadurch erhält man einen tieferen Einblick in Prozesse, die bei der Beantwortung eines Tests relevant sind. Zum Beispiel zeigt eine Überlegenheit eines Modells, das einen Rateparameter vorsieht, dass Raten für einen vorliegenden Test tatsächlich eine Rolle spielt.

Weiterhin können mit dem grafischen Modelltest und anderen, hier nicht weiter erläuterten Tests „auffällige“ Items und Personen identifiziert werden. So ist es möglich, z. B. im Rahmen einer Testrevision solche Items zu identifizieren und zu eliminieren, für die die Annahmen des dichotomen Rasch-Modells nicht gelten. Ebenso können Personen mit auffälligem Antwortmuster, das nicht zu den Annahmen des dichotomen Rasch-Modells passt, identifiziert und deren Testergebnisse mit Vorsicht interpretiert werden.

Aufgrund der Beschreibung von Personen und Items auf dem gleichen Kontinuum ist es sehr gut möglich, die Passung von Items zu den getesteten Personen zu prüfen. Sollten die meisten Personen eine deutlich höhere oder niedrigere Fähigkeit als die von den verwendeten Items geforderte Fähigkeit aufweisen, wäre dies keine gute Passung zwischen Personen und Items.

Die Beschreibung von Personen und Items auf einem Kontinuum kann zudem sehr gut im Rahmen des adaptiven Testens genutzt werden. Beim

Nutzen dichotomer Rasch-Modelle

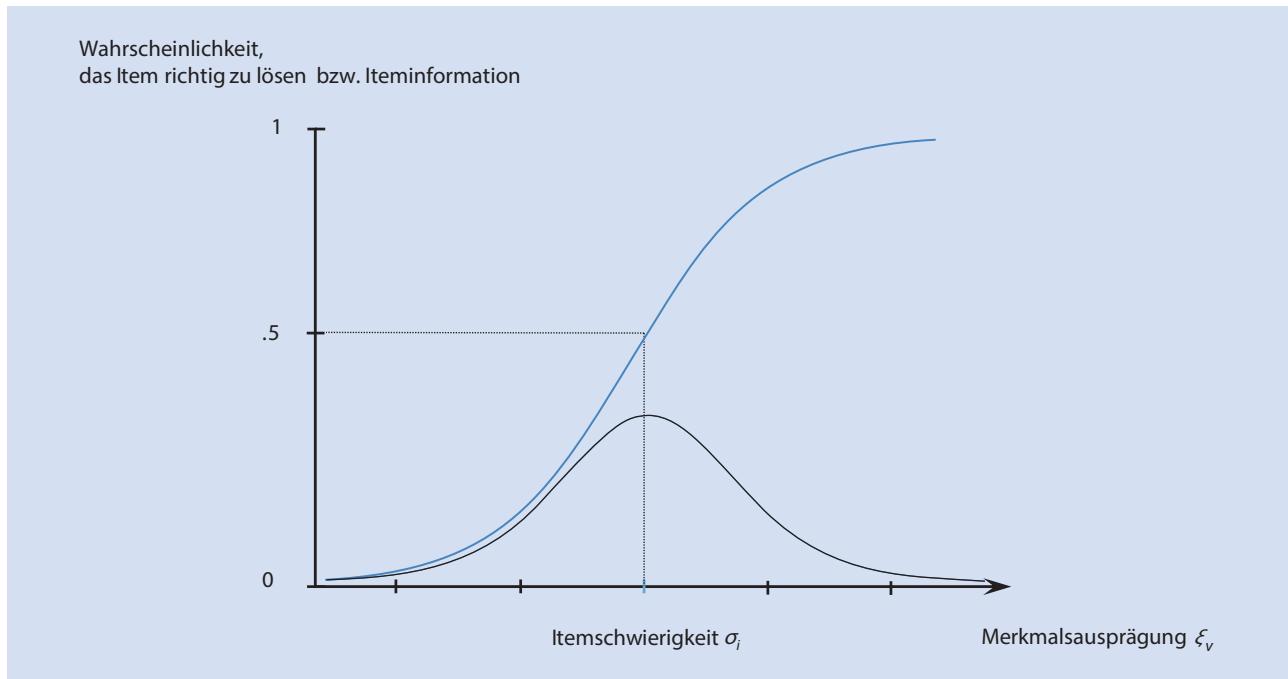
Annahmen über das Zustandekommen von Itemantworten

Modelle gegeneinander testen

Identifikation von auffälligen Items oder Personen

Passung der Items zu Personen

Iteminformationsfunktion



■ Abb. 2.14 Iteminformationsfunktion. (Angelehnt an Kelava und Moosbrugger 2020b, S. 395)

adaptiven Testen erhalten Testteilnehmerinnen und Testteilnehmer keineswegs alle die gleichen Testitems, wie es bei traditioneller, d. h. nicht adaptiver Testvorgabe üblich ist. Vielmehr existiert ein großer Pool an Items, deren Schwierigkeit bekannt ist (da sie beispielsweise im Zuge der Testentwicklung ermittelt wurde). Während der Testbearbeitung wird kontinuierlich, d. h. nach jeder Itemantwort die Fähigkeit einer Person neu geschätzt. Aus dem Pool an bestehenden Items werden dann diejenigen ausgewählt und appliziert, deren Schwierigkeiten möglichst nahe an der aktuell geschätzten Fähigkeit der Person liegen. Diese Items bieten die meisten Informationen, da die itemcharakteristischen Kurven hier die größte Steigung haben und somit kleine Unterschiede auf dem Merkmalskontinuum zu größeren Unterschieden der Lösungswahrscheinlichkeit führen. Die Iteminformationsfunktion (■ Abb. 2.14) beschreibt die Bereiche auf dem Merkmalskontinuum, für die ein bestimmtes Item besonders „informativ“ ist.

2.3.2 Item-Response-Theorien für ordinale Antwortformate

■ Grundannahmen

Darstellung der
Lösungswahrscheinlichkeit bei
dichotomen Items ausreichend

Addition beider Kurven zu 1

Zur Beschreibung dichotomer Antwortformate sind wir bislang mit einer itemcharakteristischen Kurve zur Beschreibung eines Items ausgekommen. Dies liegt daran, dass es nur 2 mögliche Ergebnisse gibt: Das Item wird gelöst oder das Item wird nicht gelöst. Es reicht dann, nur die itemcharakteristische Kurve für das Ereignis „Item wird gelöst“ darzustellen. Die Wahrscheinlichkeit für das Ereignis „Item wird nicht gelöst“ ergibt sich dann aus 1 minus der Wahrscheinlichkeit, das Item zu lösen. In ■ Abb. 2.15 sind dennoch beide Wahrscheinlichkeitsverläufe für ein dichotomes Item eingezeichnet.

Wie man sieht, ist links des Schnittpunkts beider Kurven die Wahrscheinlichkeit höher, das Item nicht richtig zu beantworten. Rechts des Schnittpunkts ist die Wahrscheinlichkeit höher, das Item richtig zu beantworten. Für jeden Punkt auf der Abszisse (x-Achse) addieren sich beide Kurven zu 1, da die Wahrscheinlichkeit, das Item entweder richtig oder falsch zu lösen, 1 sein muss.

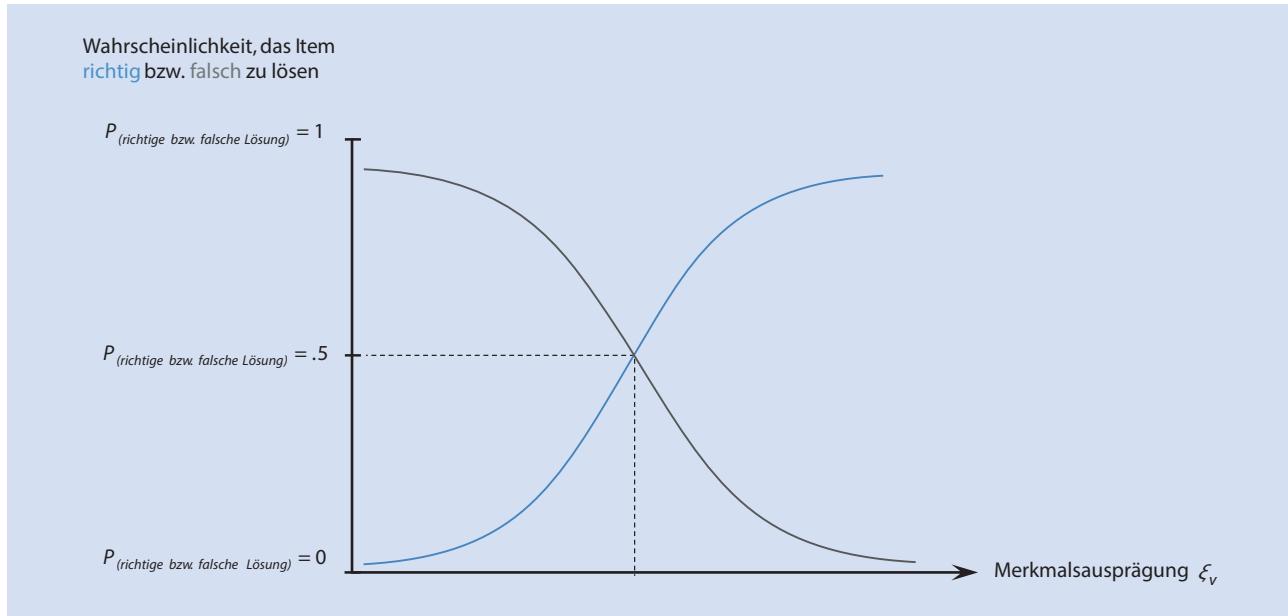


Abb. 2.15 Lösungs- und Nichtlösungswahrscheinlichkeit eines dichotomen Items. (In Anlehnung an Eid und Schmidt 2014, S. 146, mit freundlicher Genehmigung des Hogrefe Verlages)

Nehmen wir nun an, ein Item bestünde nicht aus einem dichotomen Antwortformat, sondern aus einer 3-stufigen Beurteilungsskala. Beispielsweise könnten Testpersonen in einem Test zur Leistungsmotivation aufgefordert sein, Aussagen wie „Ich gebe stets mein Bestes“ oder „Es ist mir wichtig, gute Leistungen zu bringen“ mit 0=„Ablehnung“, 1=„weder Ablehnung noch Zustimmung“ oder 2 „Zustimmung“ zu beantworten. Für solche ordinalen Antwortformate genügt es nun nicht mehr, das Antwortverhalten mit einer Kurve pro Item zu beschreiben; es sind vielmehr 3 Kurven nötig.

Die 3 Kurven in Abb. 2.16 beschreiben nun die Wahrscheinlichkeiten, eine der 3 Antworten der 3-stufigen Beurteilungsskala zu wählen – in Abhängigkeit von der Ausprägung des zu messenden Merkmals. Diese 3 Wahrscheinlichkeiten sind mit p_0 , p_1 und p_2 bezeichnet. Beispielsweise beschreibt p_0 die Wahrscheinlichkeit, die unterste Antwortkategorie (0=Ablehnung) zu wählen, abhängig von der latenter Merkmalsausprägung. Personen, deren latente Eigenschaft unter der 1. Schwelle liegt, haben die höchste Wahrscheinlichkeit, die 0. Antwortkategorie zu wählen. Je weiter die latente Eigenschaft einer Person unter Schwelle 1 liegt, desto deutlicher übersteigt die Wahrscheinlichkeit für die 0. Antwortkategorie die Wahrscheinlichkeiten der beiden anderen Antwortkategorien. Im Gegensatz dazu wird die Wahrscheinlichkeit für Kategorie 1 (p_1) größer, je größer die latente Eigenschaft ist – aber nur bis zu dem in der Grafik dargestellten Maximum. An der Schwelle 1 schneiden sich die Wahrscheinlichkeitskurven der Kategorien 0 und 1. Ab hier übersteigt die Wahrscheinlichkeit von Kategorie 1 die von Kategorie 0. Es ist inhaltlich sinnvoll, dass Personen mit höherer Ausprägung der Leistungsmotivation (s. obiges Beispiel) irgendwann eher auf das Item mit „weder Ablehnung noch Zustimmung“ antworten als mit „Ablehnung“. Wie bereits erwähnt steigt die Wahrscheinlichkeit, Kategorie 1 zu wählen, jedoch nicht kontinuierlich an. Übersteigt die Leistungsmotivation auch Schwelle 2, so ist nun die Wahrscheinlichkeit, Kategorie 2 zu wählen, am größten.

Ordinalen Antwortformate erfordern mehr Kurven je Item

Wahrscheinlichkeiten einer Antwort in Abhängigkeit von der Merkmalsausprägung

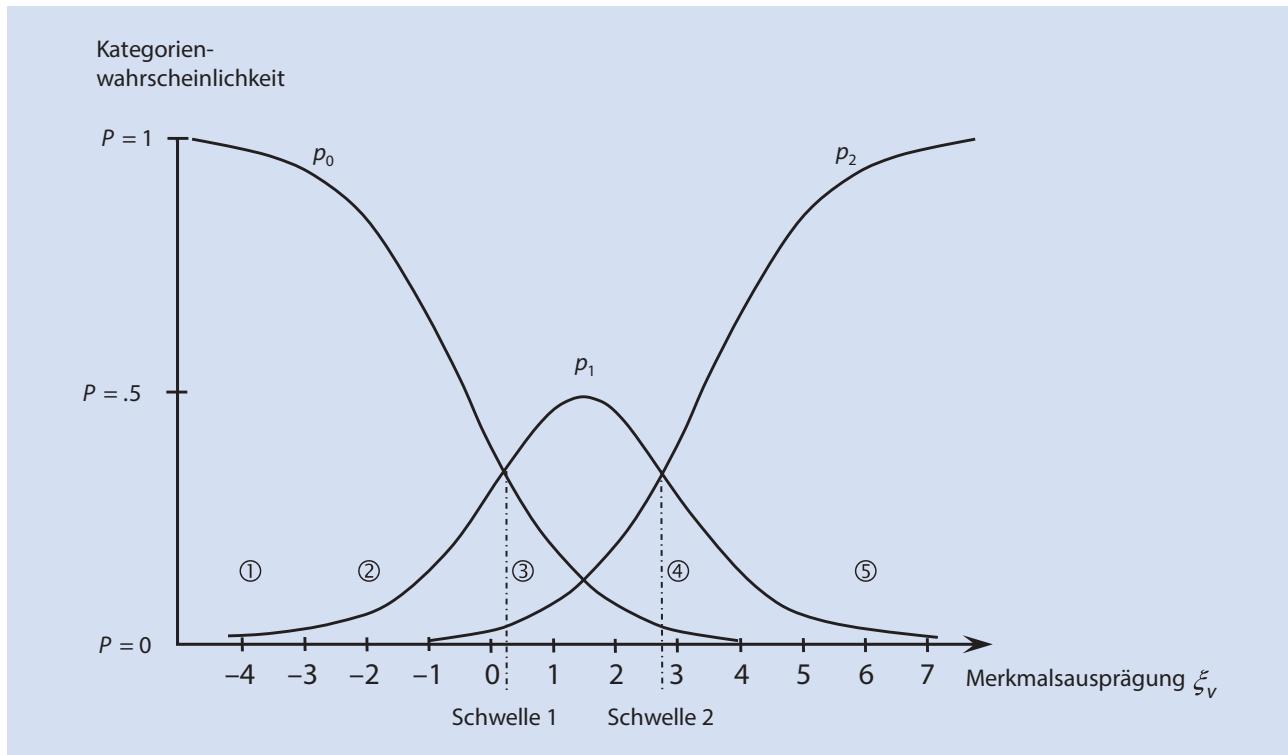


Abb. 2.16 Antwortwahrscheinlichkeiten bei einem Item mit 3 Antwortkategorien. Die Wahrscheinlichkeit für Antwortkategorie 0 = p_0 , für Antwortkategorie 1 = p_1 und für Antwortkategorie 2 = p_2 . (Angelehnt an Rost 2004, S. 144, © Hogrefe)

Betrachten wir nochmals die Aussagen, die sich anhand dieser Kurven über die Antwortwahrscheinlichkeiten verschiedener Personen machen lassen.

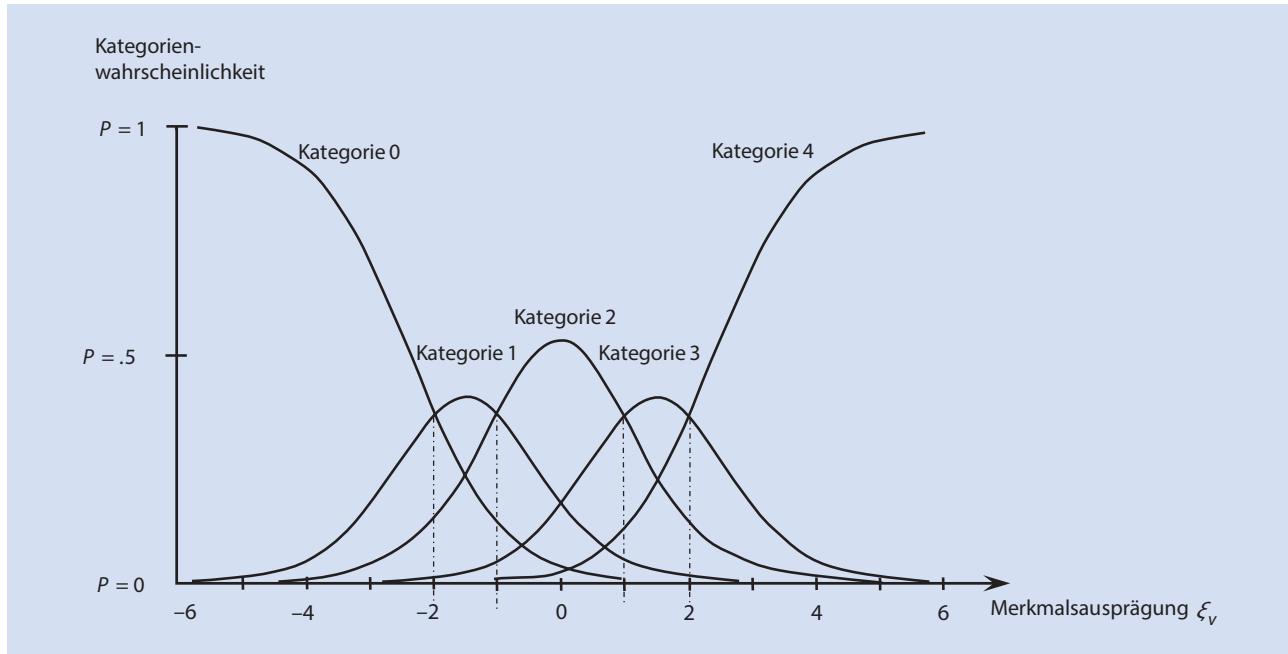
Wann ist welche Antwortkategorie am wahrscheinlichsten?

1. Bei extrem niedriger Merkmalsausprägung (① in Abb. 2.16) ist es am wahrscheinlichsten, dass Personen Kategorie 0 wählen. Diese Wahrscheinlichkeit nähert sich mit immer niedriger werdender Merkmalsausprägung asymptotisch der 1 an.
2. Bei etwas höheren Merkmalsausprägungen (②) steigt die Wahrscheinlichkeit, dass Personen Kategorie 1 wählen. Dennoch ist nach wie vor die Wahrscheinlichkeit, Kategorie 0 zu wählen, am höchsten.
3. Liegt die Merkmalsausprägung über dem Schnittpunkt der ersten beiden Kurven (③), so übersteigt nun die Wahrscheinlichkeit für Kategorie 1 die von Kategorie 0. Dieser Schnittpunkt wird auch als Schwelle bezeichnet.
4. Liegt die Merkmalsausprägung über dem Schnittpunkt von Kurve 1 und 2 (④), übersteigt die Wahrscheinlichkeit für Kategorie 2 die für Kategorie 1.
5. Bei extrem hoher Merkmalsausprägung (⑤) nähert sich die Wahrscheinlichkeit für Kategorie 2 asymptotisch der 1 an.

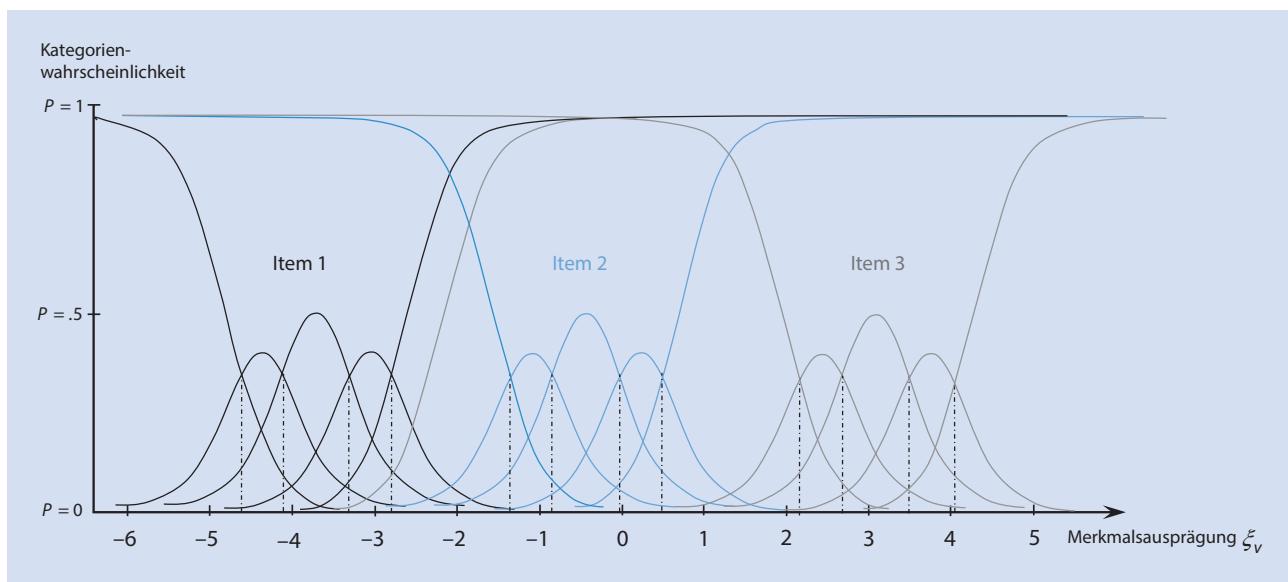
Abstände der Schwellen

Die Abstände der Schwellen innerhalb eines Items beschreiben die Breite des Merkmalsbereichs, für den letztlich bei der Beantwortung des Items wahrscheinlich die gleiche Antwort gewählt wird. Abb. 2.17 zeigt unterschiedliche Schwellenabstände innerhalb eines Items mit 5 Antwortkategorien (von 0 bis 4). Wie man sieht, ist der Merkmalsbereich, in dem die Antwortkategorie 1 am wahrscheinlichsten ist, deutlich kleiner als der Merkmalsbereich, in dem die Antwortkategorie 2 am wahrscheinlichsten ist.

Anders als bei den itemcharakteristischen Kurven, die wir im Rahmen des dichotomen Rasch-Modells kennengelernt haben (► Abschn. 2.3.1), werden ordinale Items nicht durch eine, sondern durch so viele Kurven, wie es



■ Abb. 2.17 Item mit 5 unterschiedlich breiten Antwortkategorien. (Angelehnt an Rost 2004, S. 221, © Hogrefe)



■ Abb. 2.18 Darstellung der Kategorienwahrscheinlichkeiten, Schwellen und Schwierigkeiten von 3 Items mit jeweils 5 Antwortkategorien

Antwortkategorien gibt, beschrieben. Zur Erinnerung: In ■ Abb. 2.16 und in ■ Abb. 2.17 wird nur ein einziges Item beschrieben.

Mehrere Items eines Tests können sich sowohl in Bezug auf ihre Schwellenabstände als auch auf die Position der Schwellen auf dem latenten Kontinuum unterscheiden. ■ Abb. 2.18 zeigt 3 Items unterschiedlicher Schwierigkeit. Die Schwellenabstände sind innerhalb der Items unterschiedlich, aber über die Items hinweg gleich. Das heißt, bei allen Items ist der Abstand von Schwelle 1 zu Schwelle 2 gleich groß, bei allen Items ist der Abstand von Schwelle 2 zu Schwelle 3 gleich groß usw. In diesem Fall spricht man von einem sog. „Ratingskalenmodell“.

Ratingskalenmodell

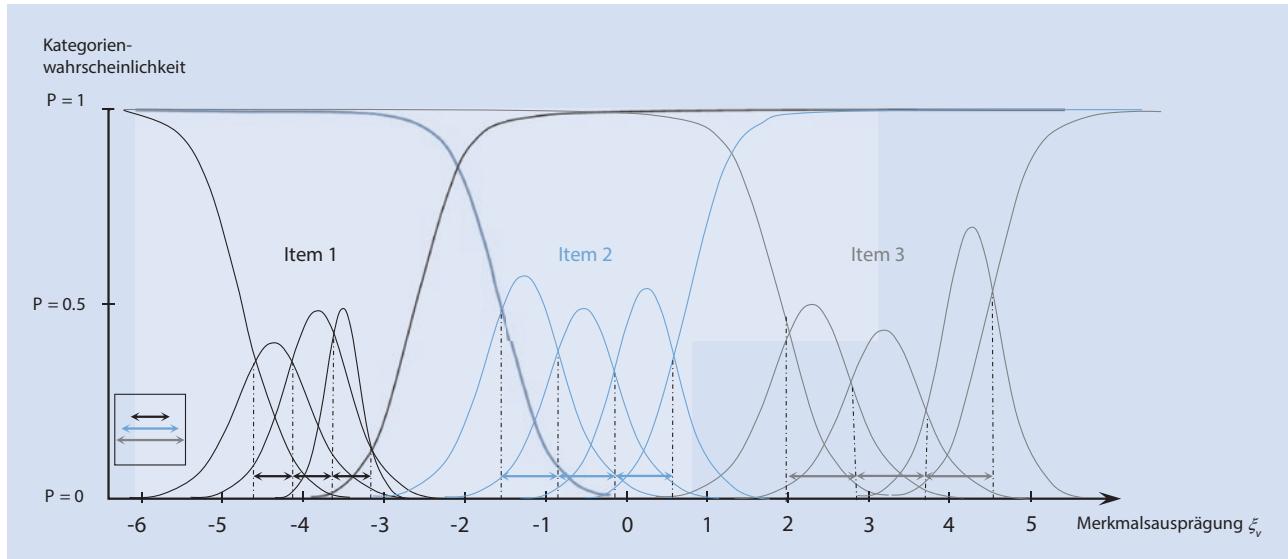


Abb. 2.19 Schematische Darstellung eines Äquidistanzmodells

Äquidistanzmodell

Sind die Schwellenabstände innerhalb jedes einzelnen Items gleich, nicht aber zwischen den Items eines Tests, spricht man von einem Äquidistanzmodell. Abb. 2.19 stellt dies schematisch dar.

Weiterführende Literatur

Für die Ableitung der Modellgleichung des ordinalen Rasch-Modells und die Schätzung der Parameter wird auf Rost (2004) oder Eid und Schmidt (2014) verwiesen.

2.3.3 Item-Response-Theorien zur Klassifikation von Personen

Klassifikation von Personen statt Quantifizierung der Merkmalsausprägungen

In manchen Fällen besteht das Ziel einer Testung nicht darin, die Merkmalsausprägung von Personen zu quantifizieren, sondern darin, Personen verschiedenen Klassen zuzuordnen. Als Beispiel soll hier der Fragebogen zum Arbeitsbezogenen Verhaltens- und Erlebensmuster (AVEM; Schaarschmidt und Fischer 2008) herangezogen werden. Er erfasst 11 Dimensionen des beruflichen belastungsrelevanten Verhaltens und Erlebens (z. B. beruflicher Ehrgeiz, innere Ruhe/Ausgeglichenheit, erlebte soziale Unterstützung). Der AVEM nimmt an, dass Personen sich so systematisch auf diesen 11 Dimensionen unterscheiden, dass sie 4 Klassen bzw. Typen zugeordnet werden können. Diese Typen sind: Typ A – Selbstüberforderung, Typ B – Erschöpfung und Resignation, Typ G – Gesundheit und Typ S – Schonung. Jeder Typ weist ein für ihn charakteristisches Muster an Ausprägungen der 11 Dimensionen auf. In Abb. 2.20 sind typische Ausprägungen des Typs A auf einigen der 11 Dimensionen dargestellt.

Randbedingungen der Klassifikation von Personen

Die Annahme, dass sich Personen aufgrund ihrer Ausprägungen auf kontinuierlichen Dimensionen (im Falle des AVEM auf 11 Dimensionen) in Kategorien einteilen lassen, ist nur unter bestimmten Randbedingungen sinnvoll:

1. Erstens muss es tatsächlich eine Häufung der für die Kategorien typischen Antwortmuster geben, davon abweichende Antwortmuster sollten eher selten vorkommen. Mit anderen Worten: Die Kategorien sollten hinreichend viel Varianz in den Antwortmustern erklären.
2. Daraus ergibt sich, dass Personen aufgrund ihrer Antworten möglichst klar einer Kategorie zugeordnet werden können. Personen, die keiner Kategorie zugeordnet werden können, sollten selten sein.



Abb. 2.20 Auszug aus dem Profilbogen zum AVEM des Typs A – Selbstüberforderung. (Aus Schaaerschmidt und Fischer 2008, © Pearson)

3. Die Kategorien sollten distinkt genug sein, sodass kaum Personen zu 2 oder mehr Kategorien zugeordnet werden.
4. Die Zuordnung von Personen zu Kategorien sollte über die Zeit stabil sein. Wenn zum Testzeitpunkt 1 viele Personen der Kategorie A und diese zu Testzeitpunkt 2 der Kategorie B zugeordnet werden (ohne dass sich an den zugrunde liegenden Merkmalen etwas geändert hätte), spricht dies gegen die Sinnhaftigkeit der Kategorienbildung.

Diese Fragen sind jedoch nicht einfach zu beantworten. Stellen wir uns die 11 Dimensionen des AVEM vor, für die uns Ausprägungen von 100 Personen vorlägen. Wie stellen wir fest, ob eine Bildung von Klassen sinnvoll ist? Und wenn ja, wie viele Klassen sollten es sein?

Hier kommt die Latent-Class-Analyse (engl. latent class analysis, LCA) ins Spiel. Damit können Gruppen identifiziert werden, die sich bezüglich der Antworten auf die applizierten Items unterscheiden. So könnte eine Gruppe identifiziert werden, die alle Items mit Leistungsbezug und alle Items zum Belastungserleben bejaht (Typ A im AVEM). Eine andere Gruppe könnte alle Items mit Leistungsbezug verneinen, aber dennoch ein hohes Belastungserleben berichten (Typ B im AVEM).

Latent-Class-Analyse

Wendet man Latent-Class-Modelle auf dichotome Daten an (also richtig gelöste vs. falsch gelöste Items), so ist die zentrale Annahme, dass sich zwischen den Gruppen die Lösungswahrscheinlichkeiten der Items unterscheiden. Innerhalb einer Gruppe erhält jedes Item eine Lösungswahrscheinlichkeit, die für alle Personen der Gruppe gleich ist. Werden also Personen einer Gruppe zugeordnet, die wir beispielhaft als „Auswendiglernende“ bezeichnen, so sollen diese Personen in den Items eines Tests (z. B. einer Klausur) gleiche Lösungswahrscheinlichkeiten haben. Sinngemäß würden in diesem Beispiel die Lösungswahrscheinlichkeiten für Items, die man durch Auswendiglernen gut lösen kann, hoch ausfallen (in Tab. 2.5 wird angenommen,

Gleiche Lösungswahrscheinlichkeit aller Items innerhalb einer Gruppe

Tab. 2.5 Hypothetisches Ergebnis einer Latent-Class-Analyse

Gruppe	Prozentuale Verteilung der Testpersonen auf Gruppen	p_{1G}	p_{2G}	p_{3G}	p_{4G}	p_{5G}
1 („Auswendiglernende“)	64	.82	.76	.28	.19	.31
2 („Verstehende“)	36	.54	.46	.90	.75	.71

p_{1G} = Wahrscheinlichkeit aller Personen einer bestimmten Gruppe, Item 1 zu lösen.

Gruppen überlappen nicht

Visualisierung der Lösungswahrscheinlichkeit pro Gruppe

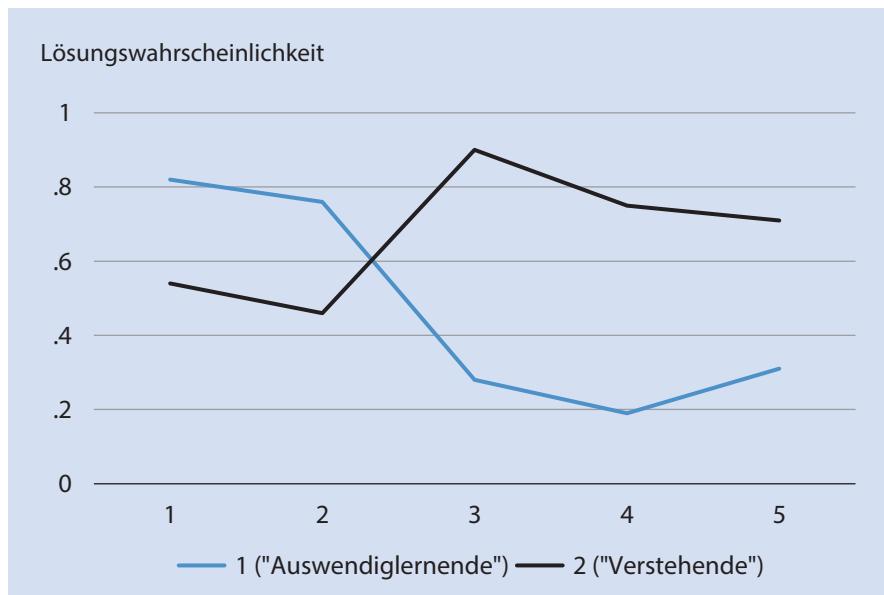
Anzahl der Gruppen a priori zu definieren

dass dies die Items 1 und 2 sind) und für Items, die man nicht gut durch Auswendiglernen lösen kann, niedrig (Items 3, 4 und 5). Die andere in □ Tab. 2.5 aufgenommene Gruppe könnte man als beispielsweise als „Verstehende“ bezeichnen, die Items 3, 4 und 5 wären von dieser Personengruppe mit höherer Wahrscheinlichkeit zu lösen.

Eine weitere wichtige Annahme von Latent-Class-Modellen wird ebenfalls in □ Tab. 2.5 deutlich. Die prozentuale Klassengröße addiert sich zu 100. Das heißt, die gebildeten Gruppen überlappen nicht – jede Person wird auf Basis ihres Antwortmusters genau einer Gruppe zugeordnet, und zwar der, für die das Antwortmuster am wahrscheinlichsten ist (Rost 2004).

Natürlich sieht man den so identifizierten Gruppen nicht direkt an, wodurch sie sich unterscheiden und wie sie zu bezeichnen sind – es sind latente Gruppen bzw. Klassen. Dass es sich im Beispiel aus □ Tab. 2.5 um „Auswendiglernende“ und „Verstehende“ handeln könnte, muss aufgrund von Itemmerkmalen und den jeweiligen, gruppenspezifischen Lösungswahrscheinlichkeiten geschlussfolgert werden. So könnten die Items 1 und 2 reine Wissensabfragen sein und die Items 3 bis 5 den Transfer des erlernten Wissens verlangen. Eine grafische Verarbeitung der Lösungswahrscheinlichkeiten aus □ Tab. 2.5 und damit eine typische Visualisierung im Rahmen von Latent-Class-Modellen findet sich in □ Abb. 2.21. Man beachte, dass statt einer Linie pro Item (itemcharakteristische Kurve im Rahmen der dichotomen Rasch-Modelle, ▶ Abschn. 2.3.1.1) oder Linien pro Antwortkategorie eines Items (Kategorienwahrscheinlichkeiten im Rahmen ordinaler Testmodelle, ▶ Abschn. 2.3.2) nunmehr eine Linie pro Gruppe dargestellt wird. Sie beschreibt den Verlauf der Lösungswahrscheinlichkeiten pro Gruppe über alle Items hinweg (Rost 2004).

Da die Gruppen latent sind, ist vorab natürlich auch nicht bekannt, wie viele Gruppen sich sinnvoll unterscheiden lassen. Möglicherweise müsste man neben den „Auswendiglernenden“ und „Verstehenden“ noch die „Nichtlernden“ als separate Gruppe aufführen? Im Rahmen von Latent-Class-Modellen muss die Zahl der Gruppen zunächst vorgegeben werden. Anhand von Modell-Passungs-Indizes kann geprüft werden, ob sich die Modelle durch Hinzunahme einer weiteren Gruppe substanzial verbessern.



□ Abb. 2.21 Beispielhaftes Ergebnis einer Latent-Class-Analyse (siehe Text für weitere Erläuterungen)

Weiterführende Literatur

Weitere, gut verständliche Ausführungen zu Latent-Class-Modellen sowie insgesamt zu probabilistischen Testmodellen finden sich bei Rost (2004) sowie in Moosbrugger und Kelava (2020). Eine ebenso gute Darstellung der Probabilistische Testtheorien ist bei Bühner (2021) sowie bei Eid und Schmidt (2014) zu finden.

Wie wichtig Methoden sind, um verlässlich latente Klassen zu identifizieren, soll hier nochmals verdeutlicht werden. Denn das Bestreben danach, Menschen hinsichtlich ihrer Persönlichkeit in Typen einzuteilen, scheint immer noch aktuell zu sein. Dies zeigt sich einerseits an der Vielzahl von Testverfahren, die eine Zuordnung der eigenen Person zu einem von mehreren Persönlichkeitstypen versprechen. Andererseits wird auch deutlich, dass wissenschaftliche Publikationen, die Persönlichkeitstypen identifizieren, meist große Beachtung in der Öffentlichkeit finden.

Eines der typologischen Verfahren, der Golden Profiler of Personality (GPOP; Golden et al. 2004), der eine Einteilung in 16 unterschiedliche Persönlichkeitstypen verspricht, wurde 2009 im Auftrag des Diagnostik- und Testkuratoriums bewertet (vgl. auch ► Abschn. 2.6). Das Ergebnis dieser Bewertung fällt wie folgt aus: „Zusammengefasst handelt es sich beim GPOP um ein in theoretischer, empirischer und pragmatischer Sicht unzureichendes Verfahren, das nicht den TBS-TK-Ansprüchen [das Rezensionssystem des Diagnostik- und Testkuratoriums; Anm. d. Autoren] an psychometrische Testverfahren genügt“ (Höft und Muck 2009, S. 323).

Eine Veröffentlichung, die große Beachtung erfahren hat, erschien 2018 in der renommierten Zeitschrift *Nature Human Behavior* (Gerlach et al. 2018). Unter dem Titel „A robust data-driven approach identifies four personality types across four large data sets“ publizierten die Autorin und die Autoren Ergebnisse auf Basis von Analysen, die insgesamt mehr als 1,5 Mio. Teilnehmende einschlossen. Sie identifizierten 4 Typen, die sie als „durchschnittlich“, „selbst-zentriert“, „reserviert“ und „Rollenmodell“ bezeichneten. In einer Reanalyse der Daten, die als Kommentar ebenfalls in *Nature Human Behavior* veröffentlicht wurde, legten Freudenstein et al. (2019) jedoch dar, dass lediglich 42 % der untersuchten Stichprobe einem der 4 Typen zugeordnet werden konnte und dass die durchschnittliche Zuordnungswahrscheinlichkeit der Personen zu einem Typ nur bei etwa 50 % lag. Sie betitelten ihre Replik mit „Four personality types may be neither robust nor exhaustive“.

Insgesamt sollen diese Ausführungen dafür sensibilisieren, dass zur Identifikation von Typen verlässliche diagnostische und statistische Methoden herangezogen werden sollten. Auch darf die Begeisterung für vermeintlich einleuchtende Einteilungen in Persönlichkeitstypen nicht über methodische Mängel hinwegtäuschen.

Fazit Wie man sieht, sind die Erläuterungen zu Probabilistischen Testtheorien deutlich umfangreicher als die zur Klassischen Testtheorie. Das liegt daran, dass Probabilistische Testtheorien sehr viel elaboriertere Annahmen darüber machen, wie Testantworten und -ergebnisse in Abhängigkeit von der infrage stehenden Merkmalsausprägung zustandekommen. Hingegen begnügt sich die Klassische Testtheorie mit wenigen Grundannahmen (Axiomen) und daraus folgenden Ableitungen. Dies scheint jedoch gleichzeitig ihr „Erfolgsprinzip“ zu sein: Sie ist immer noch die weitaus häufiger genutzte Basis für die Testkonstruktion. Aus diesem Grund werden in ► Abschn. 2.5 und 2.6.2.3 die auf der Klassischen Testtheorie beruhenden Analysen (zur Konstruktion und Evaluation von Tests) ausführlicher besprochen als analoge, auf Probabilistischen Testtheorien fußende Analysen.

2.4 Konstruktionsprinzipien psychologischer Tests

2

Grundlegende Fragen für Testentwicklung klären

Items formulieren, sodass sie das Merkmal in möglichst reiner Form erfassen

Mehrfache Itemrevision als Teil des Testentwicklungsprozesses

Testziel festlegen

Bevor ein Test entsteht, sind grundlegende Fragen zu klären. Zuerst stellt sich die Frage: Welches Merkmal soll gemessen werden (= Messgegenstand) und wie genau ist es definiert? Wichtig ist außerdem: Für wen soll der Test geeignet sein, wer soll ihn später bearbeiten? Für welchen Verwendungszweck soll der Test entwickelt werden?

Die konkreten Itemformulierungen werden maßgeblich von den Antworten auf diese Fragen abhängen. Insbesondere muss das bereits in Abb. 2.1 skizzierte Rational bedacht werden: Die Antworten der anvisierten Stichprobe auf ein bestimmtes Item soll in möglichst „reiner“ Form das intendierte Merkmal reflektieren. Es gilt also vor allem, Formulierungen und Aufgaben zu wählen, die dies leisten.

Da dies für viele psychologische Merkmale alles andere als leicht ist – man denke nur an Formulierungen und Aufgaben, die möglichst ausschließlich die Leistungsmotivation, Intelligenz oder Aggressivität der Testpersonen reflektieren – müssen initial formulierte Items empirisch überprüft und meist mehrfach revidiert werden. Der Prozess der Testentwicklung inklusive Item- bzw. Testrevision verläuft idealerweise – in Anlehnung an Eid und Schmidt (2014) – wie in Abb. 2.22 dargestellt. Dieser Prozess wird in den nachfolgenden Abschnitten näher erläutert.

2.4.1 Ziel der Messung und Messgegenstand

Vor der Definition des Messgegenstands steht die Festlegung eines Ziels. Dient das zu entwickelnde Instrument der Messung eines psychologischen Merkmals, wie etwa der Gewissenhaftigkeit, der subjektiven Beanspruchung oder der Lebenszufriedenheit? Oder geht es Testentwicklerinnen und -entwicklern darum, mithilfe des Tests zu prüfen, ob definierte Kriterien erfüllt sind, sodass Personen einer bestimmten Kategorie zugeordnet werden können? Solche Kriterien können vorgegebene Leistungsziele sein – beispielsweise „mindestens die

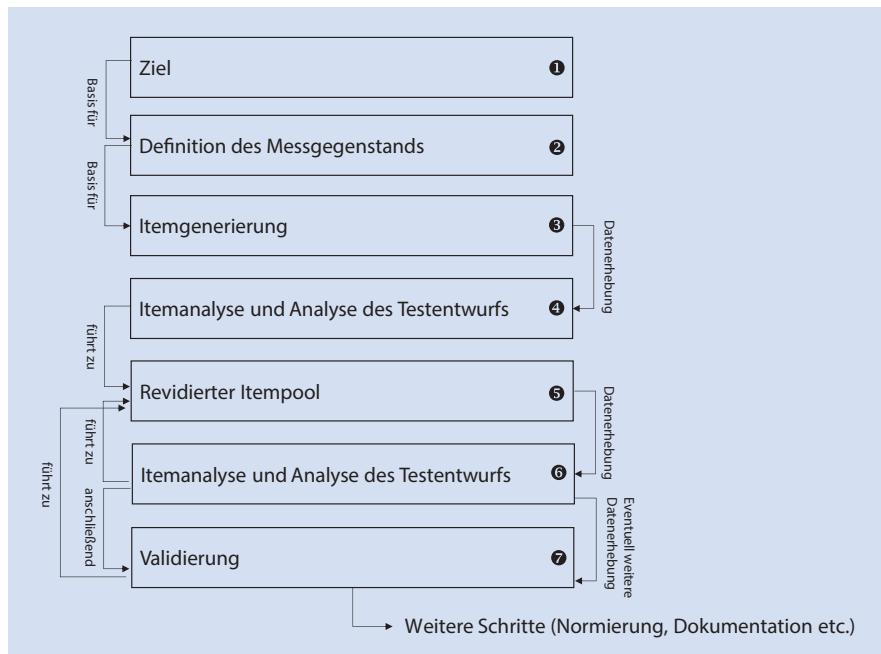


Abb. 2.22 Wesentliche Schritte der Testkonstruktion

Hälften aller curricular vorgesehenen Mathematik-Wissensbereiche richtig beantworten können“. Die entsprechenden Kategorien wären in diesem Fall „bestanden“ vs. „nicht bestanden“. Oder sollen Leistungen bzw. Gegebenheiten vorhergesagt werden, die außerhalb von Testsituationen entstehen und Teil der sozialen Realität sind – beispielsweise das erfolgreiche Absolvieren einer Ausbildung? Das Ziel eines Tests würde dann darin bestehen, möglichst gut zwischen erfolgreichen und nicht erfolgreichen Auszubildenden zu unterscheiden.

Ziele von Tests

1. Messung eines psychologischen Merkmals
2. Prüfung von definierten Kriterien und Zuordnung von Personen zu Kategorien
3. Vorhersage von Gegebenheiten außerhalb der Testsituation

Je nach Zielfestlegung lässt sich der Messgegenstand unterschiedlich definieren. Dient der Test zur Messung eines psychologischen Merkmals, so ist dieses präzise zu definieren. Testautorinnen und -autoren müssen genau benennen, was sie unter „Leistungsmotivation“, „Intelligenz“ oder „Aggressivität“ verstehen. Zu diesen und anderen Konstrukten finden sich in der Fachliteratur viele Definitionen. Die Aufgabe besteht darin, Definitionen oder Beschreibungen des interessierenden Merkmals zu sichten, eine geeignete Definition zu übernehmen oder eine eigene Arbeitsdefinition aufzustellen. Die auf einer klaren Definition des Merkmals bzw. einer Theorie aufbauende Itemkonstruktion wird als *deduktive Methode* beschrieben (► Abschn. 2.4.2.1). Aber auch eine Sichtung und statistische Analyse der Ähnlichkeit bereits verfügbarer Items (zu dem zu messenden Merkmal) kommt als Methode infrage, also eine *induktive Methode* der Itemgenerierung (► Abschn. 2.4.2.2).

Merkmal definieren

Empfehlungen des Diagnostik- und Testkuratoriums

Das Diagnostik- und Testkuratorium empfiehlt die Prüfung, ob im Manual folgende Fragen hinreichend beantwortet werden (Diagnostik- und Testkuratorium 2018b, S. 113, © Hogrefe):

- Schließt der Test an eine bestehende Theorie an oder entwickeln die Testautor(innen) eine eigene Theorie?
- Wird diese Theorie ausreichend beschrieben? Wird das Konstrukt hinlänglich beschrieben?
- Wird deutlich, was und was nicht zu dem zu messenden Bereich gerechnet wird?
- Wird beschrieben, was die Unterschiede und Gemeinsamkeiten gegenüber Tests mit überlappendem Geltungsanspruch sind?
- Wird angegeben, was auf theoretischer Ebene/auf der Ebene des Aufgabenmaterials der Mehrwert des neuen Instruments über bestehende Instrumente hinaus ist?
- Wird deutlich, ob ein beliebiges Item zum Test gehören könnte oder nicht?
- Werden das oder die zu messende(n) Konstrukt(e) auf solche Weise analysiert, sodass deutlich wird, welche Aspekte innerhalb des Konstrukt(e)s oder der Konstrukte unterschieden werden können?

Besteht das Ziel in der Festlegung, ob eine Person zu einer bestimmten Kategorie gehört, so muss die Zugehörigkeit zur Kategorie zunächst klar definiert werden. Kriterien für die Zugehörigkeit können z. B. auf Basis von Curricula vorgenommen werden, wenn es sich um einen Schulleistungstest handelt und die Zuordnung „bestanden“ vs. „nicht bestanden“ sein soll. Die PISA-Testungen (Programme for International Student Assessment) definieren beispielsweise für die Bereiche Lesen, Mathematik und Naturwissenschaften

Kriteriumsorientierte
Testkonstruktion

jeweils 7 Kompetenzstufen. Je nach Abschneiden im Test wird man einer dieser Kompetenzstufen zugeordnet. Insgesamt spricht man – wenn klare Kriterien zur Kategorienzugehörigkeit verwendet werden, um Items zu konstruieren – von der *kriteriumsorientierten Methode* der Testkonstruktion (► Abschn. 2.4.2.3). Die Definition des Messgegenstands besteht hierbei darin, die relevanten Kriterien zu identifizieren.

PISA: Lesekompetenzstufe I

PISA definiert die niedrigste Lesekompetenzstufe (für die Erhebungen zwischen 2006 und 2009) wie folgt:

- » „Jugendliche auf dieser Stufe können in einem kurzen, syntaktisch einfachen Text aus einem gewohnten Kontext, dessen Form vertraut ist (z. B. in einer einfachen Liste oder Erzählung), eine einzige, explizit ausgedrückte Information lokalisieren, die leicht sichtbar ist. Der Text enthält in der Regel Hilfestellungen für den Leser wie Wiederholungen, Bilder oder bekannte Symbole. Es gibt kaum konkurrierende Informationen. Bei anderen Aufgaben müssen einfache Zusammenhänge zwischen benachbarten Informationsteilen hergestellt werden“ (Naumann et al. 2010, S. 28). Wir besprechen PISA näher in ► Abschn. 7.4.2.

An den PISA-Studien teilnehmende Länder. Stand: 2015 (► <https://www.oecd.org/pisa/aboutpisa/>). OECD-Länder sind grau, weitere teilnehmende Länder blau gekennzeichnet (© OECD).



Externe Testkonstruktion

Bei dem gerade geschilderten Ziel entsteht die Realität durch das Testergebnis. Das heißt, man gehört zu der Kategorie „bestanden“, weil man im Test mehr als 50 % der Aufgaben richtig gelöst hat. Eigentlich bedeutet das, dass man anhand des Testergebnisses Kategorien bildet. Manchmal sind die Kategorien (oder auch Leistungen bzw. Gegebenheiten) bereits in der Realität vorhanden und sollen durch einen Test möglichst gut prognostiziert werden. In der Regel kann der Messgegenstand dann nicht auf Basis formaler Kriterien definiert werden. Vielmehr muss er durch theoretische und/oder empirische Ableitung der zur Prognose der Realität relevanten Bedingungen entwickelt werden. Im Bereich der Eignungsdiagnostik wird dieses Vorgehen als Anforderungsanalyse bezeichnet (► Abschn. 6.1.2). Erfolgt die Itemkonstruktion und -auswahl mit dem Ziel, ein durch die Realität vorgegebenen Messgegenstand abzubilden, so spricht man von einer *externalen Testkonstruktionsmethode* (► Abschn. 2.4.2.4).

Eine der ersten Testungen zu Unterscheidung existierender Gruppen

Wainer et al. (2000) nennen eine Bibelstelle (Richter 12:4–6), die Hinweise auf eine sehr frühe Anwendung eines (wohlgemerkt nicht psychologischen) Tests zur Unterscheidung existierender Gruppen liefert.

» „Und Jeftah sammelte alle Männer von Gilead und kämpfte gegen Ephraim. Und die Männer von Gilead schlugen Ephraim – denn diese hatten gesagt: Ihr seid Flüchtlinge aus Ephraim; denn Gilead liegt mittler zwischen Ephraim und Manasse –; und Gilead besetzte die Furten des Jordans vor Ephraim. Wenn nun einer von den Flüchtlingen Ephraims sprach: Lass mich hinübergehen!, so sprachen die Männer von Gilead zu ihm: Bist du ein Ephraimit? Wenn er dann antwortete: Nein!, ließen sie ihn sprechen: Schibboleth. Sprach er aber: Sibboleth, weil er's nicht richtig aussprechen konnte, dann ergriffen sie ihn und erschlugen ihn an den Furten des Jordans, sodass zu der Zeit von Ephraim fielen zweihundvierzigtausend.“ (Richter 12:4–6, LU, rev. 2017).

Nach Wainer et al. (2000) beschreibt diese Bibelstelle die Verfolgung der Ephraimiten durch die Gileaditen. Um nicht entdeckt zu werden, mischten sich Ephraimiten unter Gileaditen. Zur Unterscheidung beider Gruppen wurde ein einfacher Test verwendet: Alle musste ein bestimmtes Wort (shibboleth) laut aussprechen. Da Ephraimiten und Gileaditen den Anfangslaut dieses Wortes unterschiedlich aussprachen, konnten sie unterschieden werden. Bei falscher Aussprache wurde man als Ephraimit klassifiziert und getötet (vgl. Wainer et al. 2000, S. 2). Unter dem Begriff „shibboleth“ versteht man heute im Englischen Worte, anhand derer man Gruppen vermeintlich unterscheiden kann.

2.4.2 Generieren von Testitems

Abhängig vom Ziel der Messung können unterschiedliche Methoden der Itemgenerierung bzw. -auswahl zum Einsatz kommen (Schritt 3 in Abb. 2.22), die nachfolgend beschrieben werden.

2.4.2.1 Deduktive Methode

Geht es um die Messung eines Merkmals, stellt die deduktive Methode (häufig auch als rationale Methode bezeichnet) für viele Testentwicklerinnen und -entwickler die ideale Lösung dar. Man beruft sich auf eine Theorie, die eine gute Beschreibung des Merkmals liefert, und formuliert Items entsprechend der theoretischen Vorgaben. In der Tat gibt es für viele Merkmalsbereiche – für deren Messung es einen praktischen und/oder wissenschaftlichen Nutzen gibt, es also Sinn ergibt, einen Test zu entwickeln – elaborierte Theorien. Dies gilt für breite und spezifische Fähigkeitskonstrukte – also sowohl für allgemeine Intelligenz als auch für spezifische kognitive Fähigkeiten wie Aufmerksamkeit oder Kreativität (z. B. McGrew 2005). Dies gilt ebenso für breite und enge Persönlichkeitsmerkmale. Die am weitesten verbreitete Theorie breiter Persönlichkeitsmerkmale ist das Fünf-Faktoren-Modell der Persönlichkeit (McCrae und Costa 1987). Es unterscheidet und definiert die Merkmale Extraversion, Emotionalität, Verträglichkeit, Gewissenhaftigkeit und Offenheit für Erfahrungen und bildet somit eine erste Basis für die Itemkonstruktion. Aber auch engere Persönlichkeitsmerkmale werden intensiv beforscht, wodurch die Theoriebildung ebenfalls recht weit fortgeschritten ist.

Deduktive Methode: Theorie bildet Basis für Items

Jedes Item beinhaltet Fehleranteile

Möchte man beispielsweise einen Narzissmusfragebogen entwickeln, so könnte man auf ein elaboriertes Modell von Back et al. (2013) zurückgreifen. Anhand dieses Modells würde man Narzissmus als mehrdimensionales Konstrukt verstehen, bestehend aus den Subdimensionen „Bewunderung“ und „Rivalität“. Für beide Dimensionen definieren die Autorin und die Autoren jeweils affektiv-motivationale, kognitive und behaviorale Komponenten (► Tab. 2.6), anhand derer sich Items entwickeln lassen.

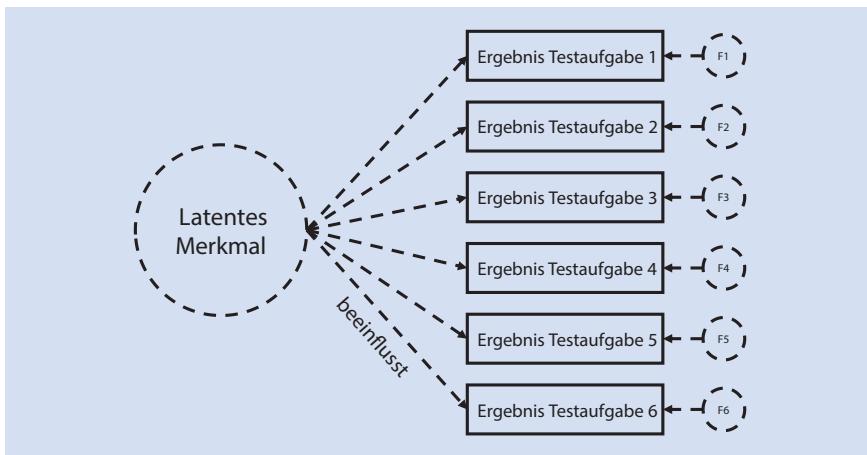
Natürlich gilt: Jede der beispielhaft genannten Aussagen kann durchaus auch indikativ für andere Aspekte sein. Beispielsweise kann der Wunsch, dass Rivalinnen bzw. Rivalen scheitern, schlicht existenziell notwendig sein (z. B. für Vertriebsmitarbeiterinnen und -mitarbeiter) und damit nicht Ausdruck des persönlichen Narzissmus. Entsprechend den Annahmen der Klassischen Testtheorie reflektiert jedes Item also stets – und in möglichst hohem Maße – das intendierte Merkmal, aber auch unsystematische Fehleranteile. ► Abb. 2.24 stellt dies grafisch dar. Solange Fehleranteile unsystematisch sind, sollten sie sich über viele Aussagen hinweg „herausmitteln“. Nichtsdestotrotz bedarf es empirischer Analysen, um sicher sein zu können, dass mit den entwickelten Items das jeweils interessierende (und nicht ein völlig anderes) Konstrukt gemessen wird (z. B. anhand von Faktorenanalysen, ► Abschn. 2.6.3.3). Würde das, was wir zuvor für Vertriebsmitarbeiterinnen und -mitarbeiter gesagt haben, über alle Items eines Fragebogens und für viele Befragte gelten, so wäre der Aspekt „sich so zu verhalten, wie es existenziell notwendig ist“ ein systematischer Teil der Messung und kein unsystematischer Fehleranteil mehr.

► Tab. 2.6 Dimensionen und Komponenten des Narzissmus nach Back et al. (2013, mit freundlicher Genehmigung der American Psychological Association)

	Komponente		
	Affektiv-motivational	Kognitiv	Behavioral
Bewunderung			
Beschreibung	Streben nach Einzigartigkeit	Überlegenheitsfantasien	Charmantes Verhalten
Itembeispiele	„Es gibt mir Kraft eine besondere Person zu sein.“	„Eines Tages werde ich berühmt sein.“	„Durch meine ausgefallenen Beiträge schaffe ich es, im Zentrum der Aufmerksamkeit zu stehen.“
Rivalität			
Beschreibung	Streben nach Überlegenheit	Abwertung anderer	Aggressivität
Itembeispiele	„Ich möchte, dass meine Rivalen scheitern.“	„Die meisten Menschen sind irgendwie Verlierer.“	„Es macht mich ärgerlich, wenn andere mir die Show stehlen.“
Die Itembeispiele sind dem Narcissistic Admiration and Rivalry Questionnaire (NARQ; Back et al. 2013) entnommen			



■ Abb. 2.23 Moderne Darstellung eines Narzissten. (© tinx/stock.adobe.com)



■ Abb. 2.24 Ein latentes Merkmal beeinflusst das Testverhalten; Fehlereinflüsse sollten möglichst gering sein. F = Fehlervarianz bzw. Residualvarianz

Fazit Die Itemgenerierung nach der deduktiven Methode erfolgt auf Basis theoretischer Überlegungen. Ein Fokus der deduktiven Methode liegt also auf dem 2. Schritt in ■ Abb. 2.22, der Definition des Messgegenstands. Das bedeutet nicht, dass alle folgenden Schritte nicht ebenfalls durchlaufen werden.

2.4.2.2 Induktive Methode

Bei der induktiven Methode der Testentwicklung stützen sich die Personen, die einen Test konstruieren, nicht primär auf eine bestimmte Theorie.

Induktive Methode: bestehende, korrelierende Items verwenden

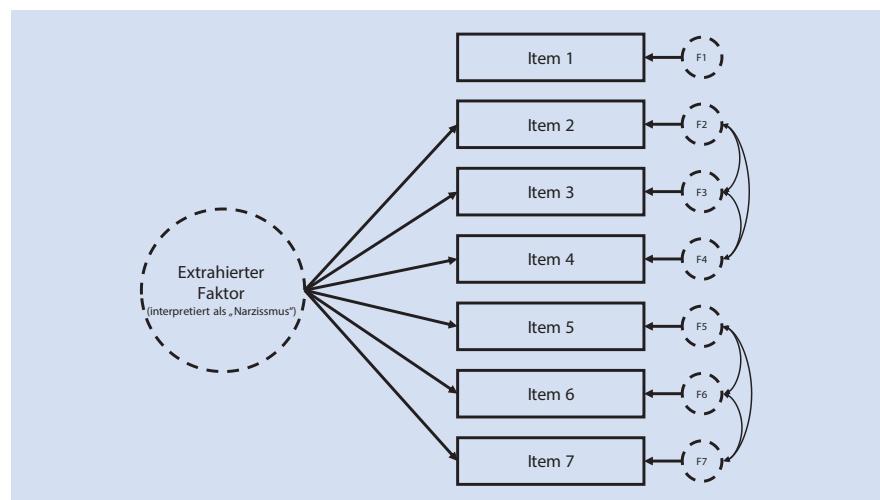
Auf einen gemeinsamen Faktor ladende Items auswählen

Die Strategie besteht vielmehr darin, diejenigen Items zu einem Testentwurf zusammenzufassen, die hoch miteinander korrelieren und damit (sehr wahrscheinlich) gemeinsam ein latentes Merkmal abbilden. Analog zu dem in □ Abb. 2.24 dargestellten Prinzip geht es also darum, aus vielen Items diejenigen herauszufiltern, die so hoch miteinander korrelieren, dass sie wahrscheinlich das gleiche latente Merkmal reflektieren und ansonsten nur noch durch unsystematische Fehler beeinflusst werden.

Die Itemgenerierung nach der induktiven Methode erfolgt durch Sichtung bereits bestehender Messinstrumente mit gleichem oder ähnlichem Messanspruch und der Zusammenstellung von Items aufgrund ihrer Ähnlichkeit. So könnte die induktive Methode bei der Entwicklung eines neuen Narzissmusfragebogens angewendet werden, indem zuerst einmal alle bereits verfügbaren Narzissmusfragebögen bzw. -items gesichtet werden. Dabei könnten redundante Items bereits aussortiert werden; alle anderen Items könnten Teil des initialen Itempools sein und einer ersten Stichprobe zur Bearbeitung vorgelegt werden. Vielleicht wird es, wie in □ Abb. 2.24 dargestellt, 6 Items geben, die gemeinsam auf einem Faktor (oben als latentes Merkmal bezeichnet) laden, während diverse andere Items dies nicht tun. Letztere würden dann aussortiert und nur die 6 auf dem Faktor ladenden Items einer weiteren Stichprobe zur erneuten Prüfung vorgelegt werden.

Möglicherweise bilden die Items des initialen Itempools nicht nur einen Faktor ab. Dann ist zu prüfen, ob das zu messende Merkmal mehrdimensional ist, also aus mehreren Teilespekten besteht. Im Fall des Narzissmus könnte dies bedeuten, dass zwar alle Items Narzissmus reflektieren, aber einige Items eher den Bewunderungsaspekt und andere eher den Rivalitätsaspekt des Narzissmus. Mithilfe der Methode der Faktorenanalyse (► Abschn. 2.5.4 und 2.6.3.3) könnte die induktive Methode bei der Entwicklung eines Narzissmusfragebogens also das Bild in □ Abb. 2.25 ergeben.

In diesem Fall würde Item 1 nicht weiter berücksichtigt, da es nicht auf dem extrahierten, gemeinsamen Faktor lädt. Durch die Korrelation der Fehler (F_2 , F_3 und F_4 sowie F_5 , F_6 und F_7) wird deutlich, dass die Items 2 bis



□ Abb. 2.25 Initiale Struktur eines fiktiven Narzissmusfragebogens mit 7 Items. F = Fehlervarianz bzw. Residualvarianz

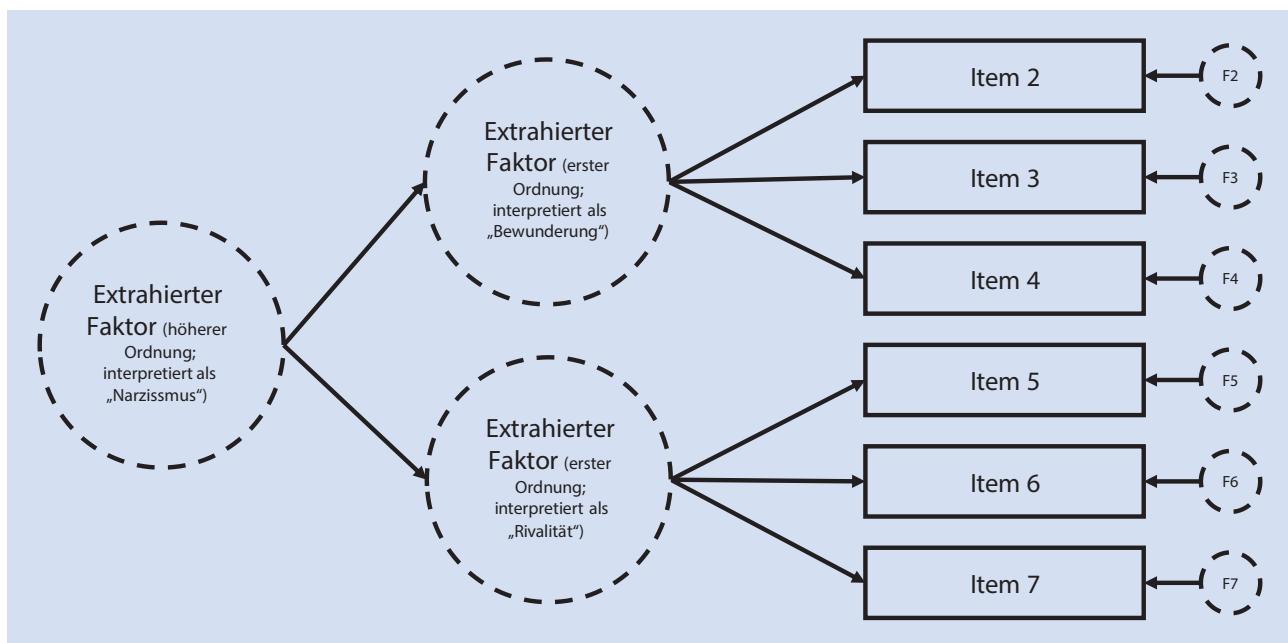
7 nicht nur den als Narzissmus interpretierten Faktor und unsystematische Fehlereinflüsse reflektieren – vielmehr haben die Fehlereinflüsse eine gewisse Systematik. Gegebenenfalls wird durch inhaltliche Betrachtung der Items klar, dass die Items 2 bis 4 eher den Bewunderungsaspekt des Narzissmus reflektieren und die Items 5 bis 7 eher den Rivalitätsaspekt. Eine erneute Analyse – ohne Item 1 – könnte dann das Bild in □ Abb. 2.26 ergeben.

Eine weitere Möglichkeit, Items nach ihrer Ähnlichkeit zueinander zu sortieren, bietet der Prototypenansatz. Hierbei erfolgt keine statistische Analyse der Dimensionalität. Vielmehr beurteilen Menschen Verhaltensweisen oder Adjektive hinsichtlich ihrer Prototypikalität für ein zu messendes Merkmal oder sie generieren zu einem Merkmal prototypische Verhaltensweisen bzw. Adjektive.

Broughton (1984) hat College-Studierenden in Wohnheimen die Eigenschaftswörterliste von Gough und Heilbrun (1980) mit der Instruktion vorgelegt, zu jedem Attribut anzugeben, wie prototypisch dieses für die Dimensionen „achievement“, „dominance“, „nurturance“, „affiliation“, „exhibition“, „autonomy“, „aggression“ und „deference“ ist. Mithilfe eines solchen Vorgehens wird die ursprünglich von den Testautorinnen und -autoren vorgenommene Zuordnung der Einzelitems auf die betreffenden Skalen überprüft (vgl. dazu auch den unter ▶ Abschn. 2.6.3.1 genannten Ansatz der quantitativen Inhaltsanalyse); außerdem bietet sich dadurch die Möglichkeit, ggf. kürzere Skalen zu formieren, die sich nur aus hochprototypischen Items zusammensetzen.

Prototypenansatz = Auswahl der Items nach Ähnlichkeit zueinander

Fazit Die Itemgenerierung nach der induktiven Methode erfolgt durch Sichtung existierender Items und ihrer Zusammenstellung nach Ähnlichkeit. Ein wesentlicher Fokus der induktiven Methode liegt auf dem 4. Schritt in □ Abb. 2.22 und dort vor allem auf der Analyse der Dimensionalität oder auch allgemein der Ähnlichkeit der Testitems (s. Prototypenansatz). Im 3. Schritt begnügt man sich in der Regel mit der Auswahl bestehender Items.



□ Abb. 2.26 Struktur eines fiktiven revidierten Narzissmusfragebogens mit 6 Items, vereinfachte schematische Darstellung. F = Fehlervarianz bzw. Residualvarianz

Die Definition des Messgegenstands (2. Schritt) steht nicht im Vordergrund. Natürlich können dennoch bei der initialen Itemgenerierung zusätzliche theoretische Überlegungen einfließen und somit die deduktive mit der induktiven Methode kombiniert werden.

2.4.2.3 Kriteriumsorientierte Methode

Definition

Kriteriumsorientierte Tests sind inhaltsvalide Testverfahren, mit denen nicht die Position einer Person in Relation zu einer Vergleichsnorm, sondern das Erreichen oder Verfehlen eines konkreten Kriteriums ermittelt werden soll.

Definition, welche Leistungen zu erbringen sind

Ist beispielsweise das Kriterium „bestanden“ vs. „nicht bestanden“, so ist zunächst klar zu definieren, welche Leistungen zum Bestehen ausreichen und welche nicht. Im Falle eines Schulleistungstests bedeutet dies, zu definieren, welche Wissensbereiche relevant sind und in welchem Umfang diese bekannt sein sollen. So sieht der Lehrplan der Jahrgangsstufe 7 für Gymnasien in Bayern Folgendes vor:

Lehrplan der Jahrgangsstufe 7 für Gymnasien in Bayern

- Sicher rechnen mit rationalen Zahlen und beherrschen der Grundlagen der Prozentrechnung.
- Terme aufstellen und analysieren können sowie elementare Term-Umformungen ausführen.
- Lineare Gleichungen auch im Anwendungszusammenhang aufstellen und lösen.
- Daten rechnerisch und grafisch auswerten.
- Mit grundlegenden Begriffen (u. a. Kongruenz) Zusammenhänge an geometrischen Figuren beschreiben und geometrische Sätze (u. a. Satz von Thales) bei Konstruktionen und Begründungen anwenden.
- Im algebraischen bzw. geometrischen Kontext argumentieren.

(Nach ISB 2004)

Kriteriumsorientierte Methode

Inhaltsvalide Items ermöglichen korrekte Schlüsse

Möchte man mit einem Test prüfen, ob diese Leistungsbereiche beherrscht werden, so sollten die inkludierten Items diese Inhalte repräsentativ abbilden. Das heißt, es sollte zu allen Bereichen Items formuliert werden und es sollten gleich viele sowie gleich schwere Items zu allen Bereichen generiert werden.

Warum sind inhaltsvalide Items bei der kriteriumsbezogenen Testkonstruktion besonders wichtig? Dies lässt sich an einem einfachen Gedankenexperiment verdeutlichen. Stellen wir uns dazu vor, es gäbe 2 Schülerinnen, die jeweils 5 der 6 obigen Kompetenzbereiche perfekt beherrschen. Schülerin A lernt „auf Lücke“ und lässt Bereich „Im algebraischen bzw. geometrischen Kontext argumentieren“ komplett aus. Schülerin B hingegen hat keine Ahnung vom Bereich „Daten rechnerisch und grafisch auswerten“. Da beide jeweils 5 der 6 Kompetenzbereiche perfekt beherrschen, ist ihre Mathematikkompetenz exakt gleich. Würden Lehrerinnen und Lehrer den Bereich „Daten rechnerisch und grafisch auswerten“ nicht prüfen, hätte Schülerin B 5 der 5 abgefragten Kompetenzbereiche gelöst, Schülerin A jedoch nur 4 der 5 (da „Im algebraischen bzw. geometrischen Kontext argumentieren“ im Test enthalten war). Somit unterscheiden sich das Testergebnis, obwohl faktisch beide Schülerinnen die gleiche Mathematikkompetenz aufweisen.

Für das bereits verwendete Narzissmusbeispiel würde sich eine kriteriumsorientierte Methode der Testentwicklung anbieten, wenn nicht die Ausprägung von Personen auf einer Narzissmusskala interessiert, sondern eine Einteilung in „narzisstische Persönlichkeitsstörung liegt vor“ vs. „liegt nicht vor“ gemäß den Kriterien des Diagnostischen und statistischen Manuals für mentale Störungen (DSM-5; American Psychiatric Association 2015) vorgenommen werden soll.

Kriterien der narzisstischen Persönlichkeitsstörung nach DSM-5

Mindestens 5 der folgenden Kriterien müssen erfüllt sein:

- Grandioses Gefühl der eigenen Wichtigkeit
- Eingenommenheit von Fantasien grenzenlosen Erfolgs, Macht, Glanz, Schönheit oder idealer Liebe
- Glaube, „besonders“ und einzigartig zu sein und nur von anderen besonderen oder angesehenen Personen verstanden zu werden oder nur mit diesen verkehren zu können
- Verlangen nach übermäßiger Bewunderung
- Anspruchsdenken
- Ausbeutung von zwischenmenschlichen Beziehungen
- Mangel an Empathie
- Neid auf andere oder Glaube, dass andere neidisch auf ihn/sie sind
- Arrogante, überhebliche Verhaltensweisen oder Haltungen

Im Rahmen eines kriteriumsorientierten Vorgehens wird das Vorliegen aller Symptomebereiche abgefragt. Für die Diagnostik von psychischen Störungen und Persönlichkeitsstörungen haben sich strukturierte Interviews als Standard etabliert. Im Strukturierte Klinische Interview für das DSM-IV (Achse II; Wittchen et al. 1997) finden sich beispielsweise die Fragen: „Denken Sie oft an die Macht, den Ruhm oder die Anerkennung, die Sie eines Tages haben werden?“ sowie „Denken Sie oft an die ideale Liebe, die Sie eines Tages finden werden?“ (s. Kriterium: Eingenommenheit von Fantasien grenzenlosen Erfolgs, Macht, Glanz, Schönheit oder idealer Liebe). Ein strukturiertes Interview eignet sich insbesondere deswegen, da auf Nachfrage („Erzählen Sie mehr davon?“) eine genauere Einschätzung durch klinisch geschulte Personen vorgenommen werden kann.

Kriteriumsorientiertes Vorgehen auch jenseits von Leistungstests möglich

- ! Bei der kriteriumsorientierten Methode werden Items so formuliert und ausgewählt, dass sie das definierte Kriterium repräsentativ abbilden.

2.4.2.4 Externe Methode

Ansatzpunkt der externalen Methode der Testentwicklung ist das Vorliegen verschiedener Gegebenheiten in der Realität. Dies können Gruppen von Personen als Teil der sozialen Realität sein. Dazu mögen etwa Gymnasial-, Gesamt- und Förderschüler/-innen, die Angehörigen verschiedener Berufe wie Architekten/Architektinnen, Kaufleute, Friseure/Friseurinnen, Maschinisten/Maschinistinnen und Verkäufer/-innen zählen, ebenso wie erfolgreiche Absolventen und Absolventinnen von Berufsausbildungen oder Studiengängen, aber auch bereits erfolgte – und nicht etwa durch den entwickelten Test erst vorgenommene – psychiatrische Klassifikationen wie eine Schizophrenie, eine manisch-depressive Störung oder eine narzisstische Persönlichkeitsstörung. An der diagnostischen Erfassung dieser Gruppen besteht ein berechtigtes Interesse, damit durch optimale Auswahl und ggf. Behandlung sowohl der

Externe Methode = Items sollen vorliegende Gegebenheiten der Realität vorhersagen

Ziel: Items diskriminieren zwischen Personengruppen

Theoretische Überlegungen fließen mit ein

Manchmal ist es schwer nachzu vollziehen, warum external konstruierte Items „funktionieren“

individuelle als auch der gesellschaftliche Nutzen nach Möglichkeit gefördert werden können. Allerdings bezieht sich die externe Testkonstruktion nicht nur auf die Identifikation von Gruppen; auch andere Gegebenheiten der Realität können Ziel der Vorhersage durch den Test sein. Dies könnte beispielsweise die Vorhersage der Schul-, Studiums- oder Ausbildungsnote sein.

Möchte man Instrumente zur Identifikation von Menschen entwickeln, die existierenden Gruppen angehören, so sieht eine puristische Anwendung der externalen Testentwicklungsstrategie wie folgt aus: Man identifiziert Personen, für die bekannt ist, ob sie der relevanten Gruppe angehören – für unser Narzissmusbeispiel könnten das Patientinnen und Patienten in stationärer klinischer Behandlung sein. Dann legt man sowohl Mitgliedern als auch Nichtmitgliedern dieser Gruppe eine möglichst große und inhaltlich breit gefächerte Zahl von Items vor. Es wird dann anhand der Antworten von Mitgliedern und Nichtmitgliedern geprüft, welche Items von beiden Personengruppen unterschiedlich beantwortet werden, also zwischen den Personengruppen diskriminieren. Diese Items werden für die vorläufige Testversion ausgewählt und einer neuen Stichprobe von Mitgliedern und Nichtmitgliedern der fraglichen Gruppe zur Kreuzvalidierung vorgelegt. Gelingt die Kreuzvalidierung, d. h., diskriminieren die zuvor ausgewählten Items auch in der neuen Stichprobe zwischen Mitgliedern und Nichtmitgliedern, ist die Testentwicklung nach der externalen Methode weitgehend abgeschlossen. Man kann dann den so entwickelten Test verwenden, um die wahrscheinliche Gruppenzugehörigkeit für Personen zu prognostizieren, bei denen diese (noch) unbekannt ist.

Ähnlich sieht das Vorgehen aus, wenn der Test zur Vorhersage von kontinuierlichen Gegebenheiten der Realität (wie etwa Schulnoten) entwickelt wird. Man prüft dann die initiale Itemauswahl darauf, ob durch sie eine gute Vorhersage (im statistischen Sinne) gelingt bzw. welche Items und Testteile zur Vorhersage beitragen und welche nicht. Der so revidierte Testentwurf wird dann einer Kreuzvalidierung unterzogen.

Idealerweise fließen bei der initialen Auswahl einer möglichst großen und inhaltlich breit gefächerten Zahl von Items theoretische Überlegungen mit ein. Das heißt, die voraussichtliche Differenzierungsfähigkeit von Items wird aufgrund theoretischer Überlegungen eingeschätzt. Für die Differenzierung zwischen Personen mit bereits diagnostizierter narzisstischer Persönlichkeitsstörung einerseits und diesbezüglich unauffälligen Personen andererseits wird man kaum nach politischen Präferenzen oder nach dem Ernährungsverhalten fragen. Es spielen also auch deduktive Gesichtspunkte bei der Anwendung der externalen Methode eine Rolle.

! Die nach der externalen Strategie entwickelten Instrumente können im Regelfall nur für eine Differenzierung im Sinne der vorab untersuchten Gruppen herangezogen werden. Für einzelne Personen können nur Aussagen dahingehend gemacht werden, wie wahrscheinlich sie der einen oder anderen Gruppe angehören.

Als Ergebnis einer externalen Testkonstruktion kann ein sehr heterogener Itempool entstehen. Wenngleich nach erfolgreicher Prüfung in der Testentwicklung feststehen mag, dass die Items „funktionieren“, weil sie zwischen den intendierten Gruppen differenzieren, so bleibt eventuell dennoch unklar, warum dies so ist. Wenn theoretische Überlegungen nur nachrangig in die initiale Itemauswahl eingeflossen sind, kann es Testleiterinnen und Testleitern schwerfallen, Testpersonen zu erläutern, warum die Zugehörigkeit zu einer Gruppe wahrscheinlich oder unwahrscheinlich ist. Tab. 2.7 zeigt Items, die Teil des Minnesota Multiphasic Personality Inventory-2 (MMPI-2;

Tab. 2.7 Items des MMPI-2 (Hathaway et al. 2000, © Hogrefe)

Skala	Inhaltlich kaum nachvollziehbare Items	Inhaltlich gut nachvollziehbare Items
Depression	Ich habe nie Blut erbrochen oder gehustet	Manchmal komme ich mir wirklich nutzlos vor
Schizophrenie	Ich war nie in jemanden verliebt	Ich höre seltsame Dinge, wenn ich alleine bin
Soziale Introversion	Es fällt mir schwer, meine Gedanken bei einer Aufgabe oder Arbeit zu behalten	Ich wünschte, ich wäre nicht so schüchtern

Hathaway et al. 2000) sind – einem Test, der nach der externalen Strategie entwickelt wurde (► Abschn. 3.3.3.1). Die Autoren hatten zunächst eine Liste von 1000 Items angelegt, die sich auf psychopathologische Symptome bezogen. Gruppen von klinisch auffälligen Personen, die von Psychiatern als Schizophrene, Hysteriker, Hypochondrer usw. diagnostiziert worden waren, bearbeiteten die Items ebenso wie „unauffällige Normale“. Jene 550 Fragen wurden schließlich zu Skalen vereinigt, die die Patientinnen und Patienten von den Kontrollpersonen am besten differenzierten. Wie man in □ Tab. 2.7 sieht, ist schwer zu erklären, warum manche dieser Items zwischen einer bestimmten Gruppe klinisch auffälliger Personen und Gesunden differenzieren.

! Bei der externalen Methode werden Items generiert oder ausgewählt, die

- von Mitgliedern der fraglichen, existierenden Gruppen wahrscheinlich unterschiedlich beantwortet werden oder
- mit kontinuierlich ausgeprägten externen Gegebenheiten korrelieren.

Wahl der Testkonstruktionsstrategie

Bei der Darstellung der 4 Testkonstruktionsstrategien klang mehrfach an, dass sie sich gegenseitig ergänzen können. Zwar steht ein theoriegeleitetes Vorgehen (s. deduktive Methode, ► Abschn. 2.4.2.1) nicht immer im Vordergrund, sollte aber dennoch bei allen Methoden einfließen. Würde eine Testautorin oder ein Testautor im Manual unter der Überschrift „Theoretische Grundlagen“ wichtige Theorien und Erkenntnisse über den Messgegenstand ignorieren und sich ausschließlich für ein induktives Vorgehen bei der Itemkonstruktion entscheiden, wäre dies deutlich zu kritisieren. Letztlich bedeutet der Verzicht auf ein explizit deduktives Vorgehen, dass man den Stand der Wissenschaft ignoriert und versucht, das Rad neu zu erfinden. Ein Verzicht auf das deduktive Vorgehen bedarf zumindest einer guten Begründung.

2.4.2.5 Zu beachtende Randbedingungen

Schon bei der Itemformulierung und initialen Itemselektion sind – losgelöst von der Konstruktionsstrategie – weitere Aspekte zu beachten, die in der Folge thematisiert werden.

Zielgruppe Wer soll den Test später bearbeiten? Ist der Test für Kinder, Jugendliche oder ältere Erwachsene vorgesehen? Soll er speziell für Patientinnen und Patienten entwickelt werden oder für die „Normalbevölkerung“? Richtet er sich an Personen mit einem niedrigen Bildungsniveau, an solche mit

Iteminhalt an Zielgruppe anpassen

Sprachproblemen oder vielleicht an spezielle Berufsgruppen? Aus der Festlegung auf bestimmte Zielgruppen ergeben sich Konsequenzen für die Formulierung der Items. Für Personen mit niedrigem Bildungsniveau müssen die Items in einer einfachen Sprache (kurze Sätze, geläufige Begriffe etc.) abgefasst werden. Bei Kindern ist zudem auf eine kindgerechte Sprache zu achten.

Die Verständlichkeit, aber auch die thematische Einbindung von Items wirken sich auf die Akzeptanz des Tests aus. Bei Jugendlichen, die sich um einen Ausbildungsplatz bewerben, kann in einer Aufgabe zum rechnerischen Denken die Berechnung von Bleistiftpreisen verlangt werden („1 Bleistift kostet 60 Cent, 10 Bleistifte 5 €. Wie viel Prozent ist ein Bleistift billiger, wenn man ihn im Zehnerpack kauft?“). Zur Auswahl von Führungskräften wäre dieses Item ungeeignet. Man könnte jedoch eine Textaufgabe entwerfen, die dem gleichen Denkschema folgt – nur eben mit anderen Gegebenheiten (z. B. Produktionsmengen statt Bleistiften) und ggf. anderen Zahlen (um die Schwierigkeit zu erhöhen). In Persönlichkeitsfragebögen können Fragen, die Patientinnen und Patienten völlig normal vorkommen, bei psychisch gesunden Personen Verwunderung und Ablehnung hervorrufen.

Regeln zur Itemformulierung für Persönlichkeitsfragebögen

Gute Übersichten dazu, was beim Formulieren von Items zu beachten ist, finden sich in verschiedenen Lehrbüchern. Eine besonders ausführliche Darstellung bieten Thielsch et al. (2012) an. Viele der dort und an anderen Stellen genannten Regeln lassen sich auf die beiden folgenden „Metaregeln“ herunterbrechen:

- Items leicht verständlich formulieren: Missverständnisse sollten ebenso vermieden werden wie ein generelles Unverständnis bei Testpersonen. Daher sollten Aussagen in Persönlichkeitsfragebögen nicht mehrdeutig sein sowie klare und kurze Aussagen enthalten. Ebenso sind doppelte Verneinungen zu vermeiden. Es sollten auch nicht mehrere Aussagen in einem Item enthalten sein („Ich gehe gerne auf Parties und trinke öfter mal Alkohol“).
- Fragen so formulieren, dass sich Unterschiede zwischen Personen zeigen: Wenn alle Testpersonen die gleiche Antwort auf ein Item geben, ist es für den Fragebogen wertlos. Formulierungen, die dies nahelegen, sollten vermieden werden. Auch Suggestivfragen sind nicht zu empfehlen.

Interessanterweise zeigt jedoch eine Studie von Pargent et al. (2019), in der die Autorin und die Autoren absichtlich viele Regeln der guten Itemkonstruktion verletzt haben, dass die daraus resultierende Testversion psychometrische Kennwerte aufwies, die sich nicht von der Originalversion oder einer optimierten Version (in der Regeln der guten Itemkonstruktion umgesetzt wurden) unterschieden. Möglicherweise lesen viele Testpersonen die Items nur flüchtig und oberflächlich und reagieren dann nur auf die Kernaussage. Dieses Phänomen bedarf der weiteren Klärung.

Anwendungsbereich spezifizieren

Anwendungsbereich Tests werden für bestimmte Verwendungszwecke entwickelt. Damit wird auch der Geltungsbereich festgelegt, also die Fragestellungen, für deren Beantwortung der Test einen Beitrag leisten soll. Beispiele für Anwendungsbereiche sind Berufsberatung, Personalauswahl, Personalentwicklung, Entdeckung von psychischen Störungen, Feststellung der Schwere einer psychischen Störung oder Erfassung von Verhaltensauffälligkeiten in der Schule. Ein Item zur Erfassung von Wettbewerbsorientierung, einem Teilaspekt der Leistungsmotivation, könnte lauten: „Ich liebe Computerspiele, in denen man gegen andere kämpfen muss.“ Im Rahmen der Berufsberatung

von Schulabgängerinnen und Schulabgängern passt das Item thematisch. Da keine Berufserfahrung vorhanden ist, kann man den Freizeitbereich ansprechen. In der Personalauswahl könnten bei einem solchen Item Akzeptanzprobleme auftreten („Wollen die herauskriegen, ob ich während der Arbeit am Computer spiele?“).

Einsatzbedingungen Wichtige Fragen zu den Einsatzbedingungen sind: Wer wird den Test vorgeben, wer wird ihn auswerten und wie wird der Test durchgeführt? Hinter der Frage nach dem „Wer“ steht die Frage, welche Expertise für die Testdarbietung erforderlich ist. Wird eine (eventuell sogar eigens dafür) geschulte Person den Test darbieten, oder muss der Test so beschaffen sein, dass auch ein Laie oder eine Hilfskraft die Darbietung übernehmen kann? Sieht man eine Interaktion zwischen Testleiterin bzw. Testleiter und Probandin bzw. Proband vor, etwa Nachfragen bei freien Antworten, sind entweder eine Schulung oder zumindest sehr klare Instruktionen nötig. Auch wenn die Antworten sofort zu bewerten sind und abhängig vom Ergebnis unterschiedlich fortgefahren wird, spricht dies für die Beschränkung auf geschulte Personen. Diese Bedingung kennen wir von strukturierten klinischen Interviews, aber auch von bestimmten Intelligenztests wie dem Wechsler-Intelligenztest für Kinder (WISC-V, ▶ Abschn. 3.2.3.2). Einfache Fragebögen können nach einer Einweisung auch von nicht psychologisch ausgebildeten Personen dargeboten werden. Allerdings muss sichergestellt sein, dass die Fragen verständlich sind, damit es nicht zu Nachfragen kommt (z. B. „Was ist bei Frage 15 mit ... gemeint“?), die einen Laien leicht überfordern könnten.

Einsatzbedingungen beachten

Mit dem „Wie“ der Durchführung ist gemeint, ob der Test als Papier- und Bleistift-Test oder computergestützt dargeboten wird. Eventuell sind bestimmte apparative Vorrichtungen, etwa zur Erfassung der Reaktionszeit oder der Feinmotorik, wünschenswert. Ist die Möglichkeit einer Gruppenuntersuchung vorzusehen, die eine sehr ökonomische Datenerhebung erlaubt? Schließlich ist die Durchführungszeit als ein wichtiger Aspekt der Testdurchführung zu bedenken. Unter Umständen besteht in dem Bereich, für den man den Test entwickeln will, ein großer Bedarf an ökonomischen, schnell durchzuführenden Verfahren. Ein solcher Test muss zwangsläufig aus relativ wenigen Items bestehen. Damit diese Items den ganzen Merkmalsbereich abdecken, sollten sie thematisch nicht zu eng gefasst sein (also zum Thema Ausgehen nicht „Ich gehe gerne mit Freunden in eine Kneipe“, sondern „Ich gehe gerne aus“).

Einsatzkontext gibt wichtige Randbedingungen vor

2.4.2.6 Antwortformate

In ▶ Abschn. 2.1.1 haben wir für Tests Folgendes definiert: Es handelt sich um eine Messmethode, bei der Personen auf standardisierte Reizvorlagen (Aufgaben, Fragen etc.) reagieren. Die Art und Weise, wie Personen reagieren können, wird durch das Antwortformat eines Tests vorgegeben.

Freie und gebundene Antwortformate

Man kann zunächst grundsätzlich zwischen freien und gebundenen Antwortformaten unterscheiden. Freie Formate sind dadurch gekennzeichnet, dass sie keine inhaltlichen Vorgaben machen. Ein weithin bekannter Test, der freie Antworten vorsieht, ist der Rorschach-Test (▶ Abschn. 3.5.1.1). Der Probandin bzw. dem Probanden wird ein Tintenklecks mit der Frage vorgelegt „Was könnte das sein?“. Es sind alle denkbaren Aussagen erlaubt, von z. B. „Das könnte eine Fledermaus sein“ bis „Dort oben, das sieht aus wie das Gesicht einer Hexe“. Freie Antworten müssen entweder von Probandinnen und Probanden selbst aufgeschrieben oder, wie beim Rorschach-Test, von der Testleiterin bzw. dem Testleiter protokolliert werden. Damit verbunden ist gleichzeitig ein wesentlicher Nachteil: Die Auswertung ist aufwendig. Im Gegensatz zu gebundenen Formaten ist ein reines Zählen von (richtigen)

Antworten, angekreuzten Kästchen oder eine computerisierte Auswertung kaum möglich. Ein weiterer Nachteil besteht darin, dass man nicht absolut sicher sein kann, dass die gegebenen Antworten den intendierten Messgegenstand reflektieren. Der wesentliche Vorteil freier Antworten liegt darin, dass das Antwortspektrum nicht eingegrenzt wird und die Antwort nicht durch die vorgesehenen Antwortalternativen gebahnt wird.

Vor- und Nachteile des offenen Antwortformats

Vor- und Nachteile

Die Vor- und Nachteile eines offenen Antwortformats lassen sich gut anhand des bereits gezeigten Items eines Situational-Judgment-Tests (► Abschn. 2.1.2) erläutern.

An Ihrem Arbeitsplatz gibt es einige Orte, die Sie gefährlich finden. Sie denken, dass es zu Unfällen kommen könnte. Um die Gefahrenquellen dauerhaft zu beseitigen, müssten kostspielige Veränderungen vorgenommen werden. Leider ist Ihre Abteilung knapp bei Kasse. Wie würden Sie sich verhalten?

Trifft am wenigsten zu		Trifft am ehesten zu
<input type="checkbox"/>	Ich mache meinen Vorgesetzten auf die Gefahrenquellen aufmerksam. Er soll entscheiden, ob kostspielige Veränderungen notwendig sind	<input type="checkbox"/>
<input type="checkbox"/>	Ich kümmere mich darum, dass die Gefahrenquellen dauerhaft beseitigt werden, auch wenn dadurch Kosten für die Abteilung entstehen	<input type="checkbox"/>
<input type="checkbox"/>	Ich höre auf, mir darüber Gedanken zu machen. Mit etwas Vorsicht wird nichts passieren	<input type="checkbox"/>
<input type="checkbox"/>	Ich bringe Zettel mit Warnhinweisen (Vorsicht Stufe etc.) an, um die anderen auf die Gefahr aufmerksam zu machen	<input type="checkbox"/>

(Aufgabe aus dem Situational-Judgment-Test von Bledow und Frese 2009, Abdruck mit freundlicher Genehmigung von John Wiley and Sons; deutsche Übersetzung aus Bledow und Frese 2005)

Würde man Personen auf die Situationsschilderung und die anschließende Frage „Wie würden Sie sich verhalten?“ frei antworten lassen, würden manche Personen vielleicht andere Antworten geben als die 4 hier genannten, z. B. „Ich recherchiere erst einmal die rechtlichen Randbedingungen und finanziellen Konsequenzen, die Unfälle am Arbeitsplatz nach sich ziehen können“. Solche Antworten entsprechen ggf. eher der Realität der befragten Personen. Zudem ist es bei einem freien Format nicht möglich, dass sich Personen die 4 Alternativen durchlesen und überlegen, was wohl am besten ist. Andererseits stellt sich die Frage, wie die freien Antworten auszuwerten sind und ob damit Erkenntnisse hinsichtlich des interessierenden Merkmals – in diesem Fall Proaktivität – gewonnen werden können.

Um Missverständnissen vorzubeugen soll hier auch betont werden, dass freie Antwortformate keine völlig freien Antworten erlauben. Testpersonen müssen ihre Antworten natürlich instruktionsgemäß entweder aufschreiben oder mündlich nennen – sie können nicht statt des einen (z. B. aufschreiben) das andere tun (z. B. mündlich nennen) oder gar ganz anders reagieren (die Antwort pantomimisch darstellen).

Freie Antworten nur bedingt frei

Gebundene Antwortformate geben feste Antwortmöglichkeiten vor. Im „Extremfall“ kann auf Aussagen wie „Ich bin ein ehrgeiziger Mensch“ lediglich durch Ankreuzen von „Ja“ oder „Nein“ geantwortet werden. Es gibt jedoch auch weniger extreme Varianten gebundener Antwortformate. □ Tab. 2.8 nennt gebräuchliche Antwortformate und deren Vor- und Nachteile. Gelegentlich finden auch weitere Antwortformate Verwendung, so etwa das Nachzeichnen von geometrischen Figuren (in Tests zur visuellen Merkfähigkeit), das Verbinden von Zahlen zur Messung der Informationsverarbeitungsgeschwindigkeit (Zahlen-Verbindungs-Test von Oswald und Roth 1997) oder das freie Zeichnen (z. B. eines Menschen; dies kommt in mehreren Einschulungstests vor).

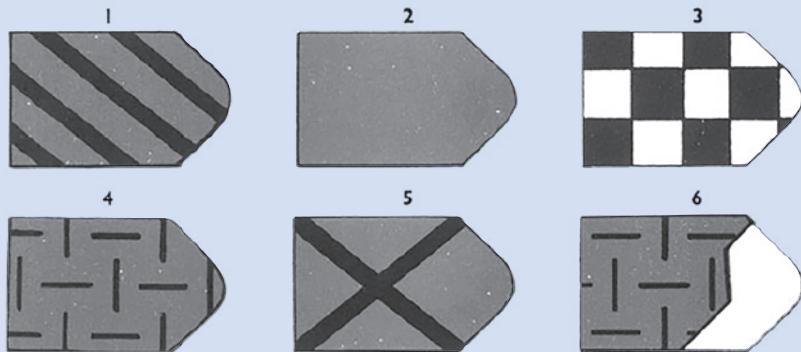
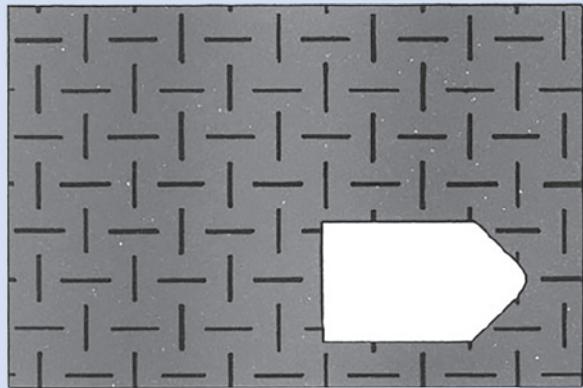
Beispiele für verschiedene Aufgabentypen

Zuordnungsaufgabe

Person		Entdeckung/Erfindung	
1)	Robert Koch	a)	Australien
2)	James Watt	b)	Penicillin
3)	Konrad Zuse	c)	Glühbirne
4)	Alfred Nobel	d)	Fernrohr
5)	Thomas Alva Edison	e)	Dampfmaschine
6)	Johannes Gutenberg	f)	Amerika
7)	Alexander Fleming	g)	Buchdruck mit beweglichen Lettern
8)	James Cook	h)	Dynamit
		i)	Radio
		j)	Tuberkuloseerreger
		k)	Computer

Kreuzen Sie jeweils **eine** Antwort an!

1)	a	b	c	d	e	f	g	h	i	j	k
2)	a	b	c	d	e	f	g	h	i	j	k
3)	a	b	c	d	e	f	g	h	i	j	k
4)	a	b	c	d	e	f	g	h	i	j	k
5)	a	b	c	d	e	f	g	h	i	j	k
6)	a	b	c	d	e	f	g	h	i	j	k
7)	a	b	c	d	e	f	g	h	i	j	k
8)	a	b	c	d	e	f	g	h	i	j	k

Multiple-Choice-Aufgabe**2**

Beispiel einer Multiple-Choice-Aufgabe aus dem Raven Progressiven Matrizen test zur Erfassung der Allgemeinen Intelligenz. (Raven 1965).

Forced-Choice-Aufgabe

Beispiel für eine Forced-Choice-Aufgabe (Teilnehmende bewerten nicht alle Antwortalternativen eines Fragebogens, sondern müssen eine auswählen):

Ich bin ein Mensch, der

- a) spielt, um zu gewinnen.
- b) andere in Entscheidungen einbezieht.
- c) das Verhalten anderer Leute analysiert.
- d) sich bei Fehlern schuldig fühlt.

(Aus Martin et al. 2002, S. 251, Übersetzung der Autoren)

■ Tab. 2.8 Itemformate und deren Vor- und Nachteile

Itemformat	Erläuterungen und Beispiel(e)	Vor- und Nachteile
Völlig freie Antworten (Erzählungen, Berichte)	Testleiterin bzw. Testleiter formuliert eine Frage oder gibt eine Aufgabenstellung vor; Probandin bzw. Proband antwortet schriftlich oder mündlich; im Persönlichkeitbereich gebräuchlich bei strukturierten klinischen Interviews sowie bei einigen projektiven Tests (► Abschn. 3.5); im Leistungsbereich bei Kreativitätstests und Problemlöseaufgaben	(+) Geeignet, wenn komplexes Denken, originelle Lösungen oder Praxistransfer erfasst werden sollen (-) Auswertung meist aufwendig (-) Auswertungsobjektivität meist eingeschränkt (-) Antwort ggf. abhängig von mündlicher bzw. schriftlicher Ausdrucksfähigkeit
Eingeschränkte freie Antworten	Auf eine Frage wie „Welche Länder grenzen an Deutschland?“ wird eine kurze Antwort verlangt; manchmal soll in einem Lückentext oder in einer Sprechblase eine Ergänzung vorgenommen werden	(+) Geeignet, wenn verfügbares Wissen und nicht bloßes Wiedererkennen erfasst werden soll, ebenso für originelle Lösungen (-) Auswertung eher aufwendig (-) Auswertungsobjektivität eventuell eingeschränkt
Zuordnungsaufgaben (und Sortieraufgaben)	Die Aufgaben bestehen aus 2 Spalten; jedes Element der einen Spalte muss einem Element der anderen Spalte zugeordnet werden; bei einer Sortieraufgabe müssen Items in die richtige Reihenfolge gebracht werden (quasi eine Zuordnung von Elementen zu Positionen).	(+) Zur Erfassung von Wissen und Kenntnissen geeignet (+) Objektiv und ökonomisch (mit Schablone, Auswertungsprogramm) auszuwerten (-) Erfasst nur Wiedererkennen und nicht freien Abruf von Gedächtnisinhalten
Multiple-Choice-Aufgaben (und Forced-Choice-Aufgaben)	Für eine Frage stehen mehrere Antwortalternativen zur Verfügung, von denen eine oder mehrere ausgewählt werden sollen; dieser Aufgabentyp findet bei Leistungstests sehr oft Verwendung. Bei Persönlichkeitfragebögen wird manchmal ein Forced-Choice-Format gewählt; von mehreren Antworten, die alle sozial ähnlich erwünscht sind, indiziert eine das Merkmal.	(+) Objektiv und ökonomisch (mit Schablone, Auswertungsprogramm) auszuwerten (-) Gute Distraktoren oft schwer zu finden (-) Erfasst bei Leistungstests nur Wiedererkennen und nicht freien Abruf von Gedächtnisinhalten (-) In Leistungstests: Ratewahrscheinlichkeit hoch (kann reduziert werden, wenn mehrere Antworten richtig sein können)
Beurteilungsaufgaben	Bei Persönlichkeits- und Interessentests soll die Testperson z. B. einstufen, wie gut die Aussage auf sie zutrifft oder wie häufig das Verhalten bei ihr vorkommt; es kommen unterschiedliche Antwortskalen zum Einsatz	(+) Objektiv und ökonomisch (mit Schablone, Auswertungsprogramm) auszuwerten (+) Liefert differenziertere Informationen als dichotome Antwortskala
Aufgaben mit dichotomen Antworten	Für eine Frage stehen nur 2 Antwortmöglichkeiten zur Auswahl: „Ja“ oder „Nein“ bzw. „Stimmt“ oder „Stimmt nicht“	(+) Objektiv und ökonomisch (mit Schablone, Auswertungsprogramm) auszuwerten (-) Entscheidung oft erzwungen (bei nicht klarer Antwort ist trotzdem eine der beiden Alternativen anzukreuzen)

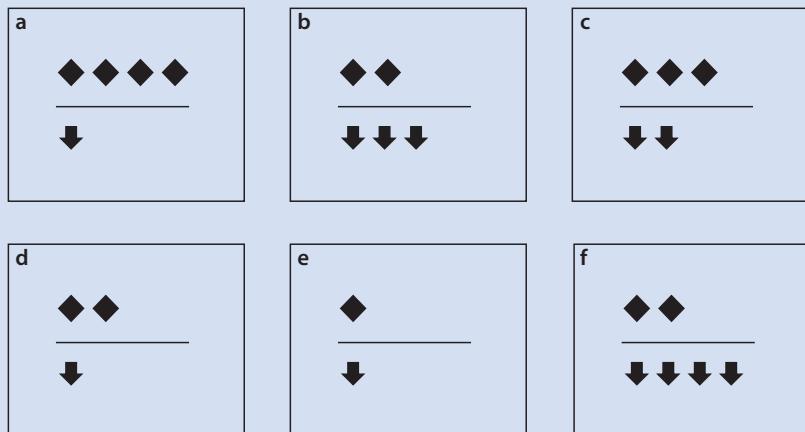
Zu dem bei Leistungstests bevorzugt verwendeten Multiple-Choice-Antwortformat sind ein paar zusätzliche Anmerkungen angebracht. Aus Sicht der Testperson ist die Aufgabe leicht zu verstehen („Ich soll ankreuzen, welche Antwort richtig ist“ oder „... welche Antwort am besten passt“). Kluge Probandinnen und Probanden, die die richtige Antwort nicht kennen, haben trotzdem gute Chancen, die richtige Antwort zu finden. Sie klammern beispielsweise zunächst offensichtlich unplausible Lösungen aus. Unter den verbleibenden Antworten treffen sie entweder eine Zufallswahl oder sie lassen sich von ihrer Intuition leiten. Bei der Testkonstruktion kann man dem nur mit einer sehr sorgfältigen Auswahl von Distraktoren entgegenwirken. Mittrig und Rost (2008) fanden heraus, dass bei 2 gängigen Tests zur Intelligenzmessung mit einer geschickten Strategie oft die Lösung zu finden ist, indem man die Antwortmöglichkeiten genau miteinander vergleicht. Mit anderen Worten: Die Distraktoren verraten manchmal, was die richtige Antwort

Vor- und Nachteile von Multiple-Choice-Antwortformaten

sein muss (s. u.). Ein anderes Problem besteht darin, dass Distraktoren eventuell nicht so falsch sind, wie man meint. Bei einem der meistgebrauchten Intelligenztests stellte sich nach vielen Jahren heraus, dass bei einigen Items ein Distraktor die bessere Antwort darstellte und nicht die vom Autor als „richtig“ angesehene Lösung (Schmidt-Atzert et al. 1995).

„Die verflixten Distraktoren“

Mittring und Rost (2008) fanden heraus, dass in 2 gängigen Intelligenztests viele Aufgaben alleine durch Inspektion der Antwortalternativen gelöst werden können. Die eigentliche Intelligenztestaufgabe muss dafür nicht beachtet werden. Um diese Strategie zu erläutern, nehmen wir an, dass bei einer Aufgabe die Antwortalternativen wie folgt aussehen:



Wie man sieht, bestehen die Antwortalternativen aus einer Kombination von 2 verschiedenen Symbolen (Raute und Pfeil), die in unterschiedlicher Anzahl vorhanden sind. Um die von Mittring und Rost (2008) entdeckte Strategie anzuwenden, zählt man zunächst einfach die gleichen Symbole, also: 4 Rauten kommen im Antwortset 1× vor, 3 Rauten ebenfalls 1×, 2 Rauten kommen 3× vor, 1 Raute wieder nur 1×. Das heißt, am häufigsten kommen 2 Rauten vor. Die häufigste Menge an Pfeilen ist 1 (sie kommen 3× vor). Kombiniert man nun beide Ergebnisse, so ist wahrscheinlich die Antwortalternative (d) richtig, da dort 2 Rauten mit 1 Pfeil kombiniert werden. Mittring und Rost (2008) konnten zeigen, dass diese Strategie bei 2 gängigen Intelligenztests sehr häufig zur richtigen Lösung führt (in einem der beiden Tests bei 75 % der Aufgaben!).

Wie kommt es dazu, dass diese Strategie „funktioniert“? Testautorinnen und -autoren versuchen natürlich, Distraktoren so zu konstruieren, dass sie nicht offensichtlich falsch sind. Vielmehr sollen sie prinzipiell als Lösung infrage kommen. Daher ergibt es für Aufgaben wie die im Beispiel gezeigte Sinn, eine teilweise richtige Lösung (also z. B. 2 Rauten) möglichst häufig in den Distraktoren unterzubringen.

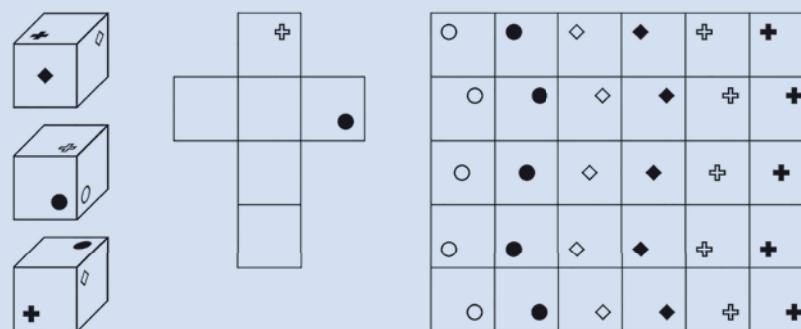
Etwas allgemeiner gefasst, würde man die Nutzung dieser Strategie als „Testschläue“ bezeichnen. Im Englischen ist der Begriff „test wiseness“ geläufig. Laut Millman et al. (1965, S. 707) bezeichnet dies die „Fähigkeit von Testpersonen, Gegebenheiten eines Tests oder der Testsituation so zu nutzen, dass ein gutes Testergebnis entsteht“. Zu diesen Strategien gehören u. a.

- Effektive Nutzung der Testzeit
- Raten
- Nutzung von Hinweisen (s. das obige Beispiel von Mittring und Rost 2008)
- Schlussfolgern (z. B. Eliminieren von eindeutig falschen Antworten)

Hausknecht et al. (2007) zeigten im Rahmen einer Metaanalyse, dass ein Coaching (inklusive der Vermittlung von Testbearbeitungsstrategien) zwischen 1. und 2. Durchführung des gleichen Tests zu einer deutlichen Verbesserung der Testleistung führt (für Stichproben- und Messfehler korrigierte Effektstärke von $d=0,70$).

Um den Einfluss der Strategie „Raten“ abzumildern, sollte in der Instruktion eines Tests unbedingt eine klare Aussage gemacht werden, wie man sich verhalten soll, wenn man die Lösung nicht findet oder sich nicht sicher ist. Die Aussage kann etwa lauten: „Wenn Sie die richtige Antwort nicht gefunden haben, kreuzen Sie diejenige an, die aus Ihrer Sicht am ehesten zu passen scheint.“ Ohne eine solche Anweisung besteht die Gefahr, dass einige Testpersonen durch Raten (oder eine Ausschlussstrategie mit Raten) Punkte erzielen, während andere auf diese Chance verzichten. So könnten zaghafte Menschen viele Aufgaben unbeantwortet lassen, während risikofreudige mit einer intelligenten Ratestrategie zusätzliche Punkte sammeln. In dem bereits vorgestellten probabilistischen 3PL-Modell (► Abschn. 2.3.1.2) kann das Ausmaß des Ratens pro Item berücksichtigt werden und daraufhin ein sinnvoller Umgang mit dem Phänomen des Ratens gewählt werden.

Raten bzw. geschickter Umgang mit Multiple-Choice-Aufgaben kann auch durch innovative Antwortformate verhindert werden. So entwickelten Thissen et al. (2016) eine Würfelabwicklungsaufgabe (s. nachfolgende Abbildung), bei der Testpersonen die richtige Lösung selbst konstruieren müssen. Aus einer Menge an Symbolen können diejenigen zusammengestellt werden, die in Kombination die richtige Lösung ergeben.



(Used with permission from *European Journal of Psychological Assessment* (2019), e-pub ahead of print © 2019 Hogrefe Publishing, ► www.hogrefe.com, ► <https://doi.org/10.1027/1015-5759/a000534>)

Die aus Thissen et al. (2019, S. 5) entnommene Abbildung zeigt ein Item, bei dem Testpersonen den aufgefalteten Würfel in der Mitte so mit Symbolen aus dem rechten Symbolsatz versehen sollen, dass (beim Zusammenfalten) einer der linken Würfel entsteht. Das Prinzip dieses Antwortformats besteht also darin, die Antwort durch eine beliebig kombinierbare Menge an einzelnen Elementen selbst zu erstellen.

Beurteilungsskalen für Persönlichkeitsfragebögen üblich

Für Persönlichkeitsfragebögen sind gebundene Antwortformate in Form von Beurteilungsskalen (häufig auch als Ratingskalen bezeichnet) gebräuchlich. Dabei beurteilen Testpersonen das Zutreffen von Aussagen (z. B. „Es bereitet mir Freude, Lehrbücher zu lesen“) auf einer mehrfach abgestuften oder kontinuierlichen Beurteilungsskala.

Wahl der Beurteilungsskala

Wetzel und Greiff (2018) fassen wichtige Fragen, die sich bei der Wahl der Ratingskala stellen, zusammen. Denn die Frage „Es bereitet mir Freude, Lehrbücher zu lesen“ lässt sich auf vielfältige Weise anhand von Ratingskalen beantworten:

unipolar	oder	bipolar
① ② ③ ④ nie selten häufig immer		-2 -1 +1 +2 Ab- lehnung Zu- stimmung

Unipolare Skalen beginnen mit einem Nullpunkt („nie“) und nehmen dann in der Ausprägung in eine Richtung zu (im Beispiel bis „immer“). Bipolare Skalen bezeichnen in ihren Enden den negativen und positiven Pol eines Kontinuums (hier von „Ablehnung“ bis „Zustimmung“). Die Darbietung einer Nummerierung, die zu den – in diesem Fall bipolaren – verbalen Bezeichnungen passt, wirkt sich positiv auf die Reliabilität der Messungen aus (Rammstedt und Krebs 2007).

anhand weniger Kategorien	oder	anhand vieler Kategorien
① ② nein ja		① ② ③ ④ ⑤ ⑥ ⑦ Ablehnung Zustimmung

Die optimale Zahl der Abstufungen einer Beurteilungsskala ist kaum pauschal zu benennen. Da das Ziel der meisten Fragebögen ist, Unterschiede zwischen Menschen abzubilden, sollten genügend Antwortoptionen vorgesehen werden, um zwischen unterschiedlichen Haltungen bezüglich der relevanten Aussage eines Fragebogens zu differenzieren. Gleichzeitig sollten nicht zu viele Abstufungen gewählt werden, da die Bedeutung der einzelnen Abstufung ggf. für Testpersonen nicht mehr erkennbar wird. Für eine allgemeine Empfehlung können zwischen 5 und 7 Kategorien als brauchbar erachtet werden (Scherpenzeel und Saris 1997). Es ist zudem ratsam – weil der Reliabilität der Messung zuträglich – alle Abstufungen einer Skala zu betiteln (z. B. „trifft teilweise zu“; vgl. DeCastellarnau 2017).

mit mittlerer Kategorie	oder	ohne mittlere Kategorie
① ② ③ ④ ⑤		① ② ③ ④
trifft nicht zu	weder/noch	trifft zu

Scherpenzeel und Saris (1997) konnten in einer Metaanalyse aufzeigen, dass sich das Vorhandensein einer mittleren Kategorie positiv auf die Validität der Messungen auswirkt.

ohne oder mit Nummerierung	oder	mit Symbolen
○ ○ ○ ① ② ③		🙁 😐 😊 trifft nicht zu trifft zu
trifft nicht zu	trifft zu	

DeCastellarnau (2017) fasst in ihrer Übersicht zusammen, dass die Mehrzahl der Studien für die Wahl nichtverbaler Bezeichnung von Antwortoptionen (also „Zahlen“ vs. „keine Zahlen“ vs. „Symbole“) keinen Einfluss auf das Antwortverhalten findet.

Darüber hinaus empfiehlt es sich, folgende Regel zu beachten: Die verbale Beschriftung sollte eine Gleichabständigkeit der Kategorien implizieren (Krosnick und Fabrigar 1997). Eine 4-stufige Skala sollte also eher nicht mit „trifft gar nicht zu“, „trifft zu“, „trifft sehr zu“ und „trifft absolut zu“ bezeichnet werden – zwischen „trifft gar nicht zu“ und „trifft zu“ scheint ein größerer Abstand zu liegen als zwischen „trifft zu“ und „trifft sehr zu“. Gängige Bezeichnungen der Skalenabstufungen finden sich bei Rohrmann (1978; zitiert nach Bühner 2021).

	○-----○-----○-----○-----○				
Häufigkeitsskala	nie	selten	gelegentlich	oft	immer
Intensitätsskala	gar nicht	wenig	mittelmäßig	überwiegend	völlig
Wahrscheinlichkeitsskala	keinesfalls	wahrscheinlich nicht	vielleicht	ziemlich wahrscheinlich	ganz sicher
Zustimmungsskala	trifft gar nicht zu	trifft wenig zu	trifft teils teils zu	trifft ziemlich zu	trifft völlig zu

Es muss an dieser Stelle auch auf die Frage der Itempolung eingegangen werden. Denn obwohl in Empfehlungen zur Itemformulierung von doppelten Verneinungen und (allgemeiner) von schwer verständlichen Formulierungen abgeraten wird (vgl. ▶ Abschn. 2.4.2.5), so raten Forscherinnen und Forscher andererseits durchaus dazu, Items auch negativ im Sinne des Konstrukt zu formulieren. Das heißt, Items zur Depressivität sollten nicht nur lauten: „Ich bin oft gedrückter Stimmung“, sondern auch: „Ich bin meist fröhlich“ oder „Ich bin selten gedrückter Stimmung“. In den letzten beiden Items würde

Positiv und negative gepolte Items

Vor- und Nachteile unterschiedlich gepolter Items

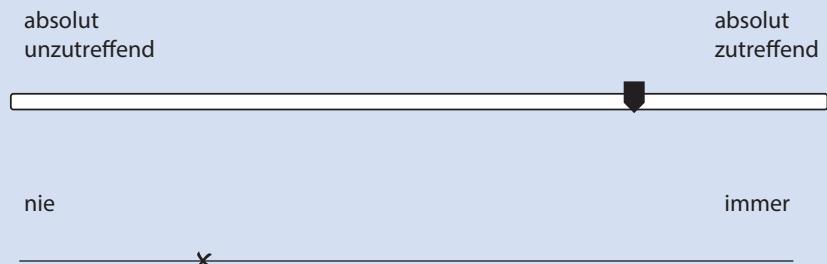
eine hohe Zustimmung für eine geringe Ausprägung des Merkmals (hier Depressivität) sprechen – deswegen spricht man von negativ gepolten Items: Der negative Pol der Skala spricht für eine hohe Merkmalsausprägung und umgekehrt. Forscherinnen und Forscher raten zu einer Mischung von positiv und negativ gepolten Items, da so verzerrte Ergebnisse aufgrund von Ja-Sage-Tendenzen (Akquieszenz) vermieden werden (z. B. Hinkin 1998).

Eine negative Itempolung muss natürlich bei der Auswertung von Antworten berücksichtigt werden. Es liegt nahe, die der Antwortskala zugeordneten numerischen Werte einfach umzudrehen. Wenn Reaktionen auf die Aussage „Ich bin meist fröhlich“ auf einer Skala von 1 („trifft gar nicht zu“) bis 5 („trifft vollkommen zu“) erfolgen, so macht man aus der 5 bei der Auswertung des Fragebogens eine 1, da „trifft vollkommen zu“ als Reaktion auf diese Aussage für eine geringe Depressivität spricht.

Andererseits mag man sich fragen, ob eine Mischung positiv und negativ gepolter Items innerhalb eines Fragebogens wirklich sinnvoll ist. Für eine klare Aussage dazu ist die derzeitige Forschungslage zu uneinheitlich. Es sollte jedoch bedacht werden, dass Items höher miteinander korrelieren, wenn sie in gleicher Weise gepolt sind. Dadurch können neben dem intendierten Merkmal weitere Einflüsse systematisch auf die Messung einwirken (vgl. Di Stefano und Motl 2006).

Visuelle Analogskalen

Es besteht auch die Möglichkeit, vollständig auf Kategorien zu verzichten und Testpersonen auf einem Kontinuum angeben zu lassen, wie zutreffend eine Aussage ist oder wie häufig sie eine bestimmte Verhaltensweise zeigen. Dazu werden visuelle Analogskalen verwendet. In der Regel werden dabei die Enden einer Linie oder eines Balkens benannt. Testpersonen können dann ein Kreuz an die Stelle des Kontinuums setzen, die am ehesten ihrer Antwort entspricht. Bei computergestützten Verfahren kann dies durch das Verschieben eines Schiebereglers geschehen. Es ist allerdings fraglich, ob Testpersonen tatsächlich mit der Feinheit bzw. Genauigkeit urteilen, wenn sie Fragebogenitems beantworten, die durch visuelle Analogskalen impliziert und ermöglicht wird (für eine Übersicht s. Wetzel und Greiff 2018).



Visuelle Analogskalen (im Falle papierhaft dargebotener visueller Analogskalen muss die Position der Kreuze mit einem Lineal ausgemessen werden)

Für die zuvor dargestellten Antwortformate existieren typische Methoden des Scorings der Itemantworten. Mit Scoring ist die Bewertung einer Antwort in Form einer Zahl gemeint. Wir besprechen nachfolgend Methoden des Scorings zu den Antwortformaten, die in ▶ Tab. 2.8 aufgeführt sind.

Als ein Nachteil von völlig freien Antwortformaten wurde bereits deren aufwendige Auswertung benannt. Der Aufwand besteht in der Regel darin, die gesamte freie Antwort zu sichten und deren Inhalt nach bestimmten Kriterien zu bewerten. Im einfachsten Fall kann eine Bewertung lediglich als „richtig“ oder „falsch“ erfolgen. Wenn etwa in einem Intelligenztest für Kinder gefragt wird „Was ist ein Meteorit?“ und Kinder darauf frei antworten sollen, kann das Scoring mit „1“ für richtige Antworten („Ein Gesteinsbrocken, der aus dem Weltall auf die Erde fällt“) und „0“ für falsche bzw. keine Antworten erfolgen. Allerdings muss das Testmanual auch in diesem vermeintlich einfachen Fall genau festlegen, welche Antworten als richtig gelten und welche nicht. In unserem Beispiel ließe sich etwa diskutieren, ob „Ein Stein, der vom Himmel fällt“ schon als richtig einzustufen ist. Denkbar ist auch ein Scoring in Form von 0, 1 und 2, wobei 1 Punkt für teilweise richtige Antworten vergeben wird. Ein Beispiel für noch aufwendigere Scorings freier Antworten ist die Einschätzung von Motivausprägungen (etwa des Leistungsmotivs) anhand frei erzählter Geschichten (siehe ▶ Abschn. 3.5.1.2). Dies gelingt nur mithilfe ausführlicher Kodiermanuale und aufwendiger Schulungen der Kodiererinnen und Kodierer (und soll an dieser Stelle nicht vertieft werden).

Eingeschränkt freie Antworten bestehen in der Nennung weniger Begriffe als Antwort auf eine Frage. Daher genügt es für das Scoring, zu bewerten, ob ein oder mehrere richtige Begriffe genannt wurden. Als Antwort auf eine Frage kann daher eine „1“ oder „0“ kodiert werden – für „richtig“ bzw. „nicht richtig gelöst“ – oder es wird die Zahl der richtig genannten Begriffe pro Aufgabe vermerkt. Gleiches gilt für Zuordnungsaufgaben: Infrage kommen entweder nur „richtig“ vs. „nicht richtig“ oder die Zahl der pro Aufgabe richtig vorgenommenen Zuordnungen.

Multiple-Choice-Aufgaben in Leistungstests werden ebenfalls in den allermeisten Fällen mit „1“ und „0“ kodiert. Dabei wird 1 Punkt vergeben, wenn die richtige Antwortalternative ausgewählt wurden und 0 Punkte für die Wahl jeder anderen, d. h. falschen Antwortalternative. Sind mehrere Antwortoptionen richtig, können Teilpunkte vergeben werden. Dichotome Antworten werden ebenfalls mit „1“ und „0“ kodiert.

Methoden des Scorings

Empirische Optionsgewichtung (empirical option weighting)

Es mag Lesenden zurecht etwas undifferenziert vorkommen, wenn alle falschen Antworten einer Multiple-Choice-Aufgabe mit „0“ bewertet werden. Möglicherweise verfügen Testpersonen, die eine falsche, aber prinzipiell plausible Antwort gewählt haben, über eine etwas höhere Ausprägung des zu messenden Merkmals als Testpersonen, die eine völlig unplausible Lösung gewählt haben. Dieser Gedanke kann an einer einfachen Beispielaufgabe illustriert werden.

„Wie viele Planeten gehören zu unserem Sonnensystem?“

- a) 7
- b) 8
- c) 9
- d) 10

Es erscheint unfair, wenn alle falschen Antworten (a, c und d) gleichermaßen mit „0“ bewertet werden. Denn bis vor einiger Zeit war 9 noch die richtige Lösung. Testpersonen, die c gewählt haben, verfügen möglicherweise nur über ein etwas veraltetes Wissen – sie haben nicht mitbekommen, dass Pluto nicht mehr zu den Planeten zählt. Die Antworten a und d sind jedoch schon immer falsch gewesen.

Das Beispiel zeigt, dass auch die Wahl von falschen Antwortoptionen in einem gewissen Zusammenhang mit dem zu messenden Merkmal stehen kann. Um dies im Scoring zu berücksichtigen, kann eine empirische Optionsgewichtung vorgenommen werden. Dabei gewichtet man jede Antwortoption mit ihrer Korrelation zum Gesamttestwert (s. auch Trennschärfe, ▶ Abschn. 2.5.3). So könnten die obigen Antwortoptionen wie folgt mit dem Gesamttestwert „Wissen Astronomie“ korrelieren:

- a) -.25
- b) .38
- c) .15
- d) -.30

Somit erhält eine Person, die die klar falschen Optionen a oder d wählt Minuspunkte, während man leichte Pluspunkte (.15) für c und deutlichere (.38) für die richtige Antwort b erhält. Mehrere Studien zeigen, dass ein solches Scoring die psychometrische Qualität von Multiple-Choice-Tests steigert (z. B. Diederhofen und Musch 2017).

Beurteilungsaufgaben werden entsprechend der Rangfolge der Beurteilungsabstufungen numerisch aufsteigend kodiert. Einige der zuvor aufgeführten Skalenbeispiele waren bereits mit Zahlen versehen und illustrieren mögliche Formen des Scorings – also beispielsweise in Form von Zahlen von 1 bis 5 bei einer 5-stufigen Beurteilungsaufgabe. Dabei ist es zunächst unerheblich, ob die Zahlenfolge bei 1 oder 0 beginnt. Die bereits erwähnten invers gepolten Items werden dabei mit Zahlen in der umgekehrten Abfolge versehen – also beispielsweise von 5 bis 1. Visuellen Analogskalen können zahlenmäßige Abstufungen in gleichförmigen Abständen hinterlegt werden, etwa von 0 bis 100. Alternativ kann auch der Abstand des Kreuzes bis zum Beginn der Skala gemessen werden und als Zentimeterwert in das Scoring eingehen.

Das ebenfalls bereits eingeführte Forced-Choice-Antwortformat hat den Vorteil, dass sich Testpersonen nicht beliebig positiv hinsichtlich eines Merkmals darstellen können (► Abschn. 2.6.5.4). Vielmehr müssen sich Testpersonen entscheiden, welche der (meist 3 oder 4) zur Verfügung stehenden eigenschaftsbezogenen Aussagen sie eher beschreibt und welche weniger. Das bedeutet aber auch, dass das Scoring schwieriger ist. Denn anders als bei Beurteilungsskalen liegt nicht zu jeder Aussage eine Bewertung vor. Bei Forced-Choice-Antwortformaten liegt lediglich eine Information darüber vor, welche Aussage gegenüber anderen Aussagen präferiert wurde (und/oder besonders abgelehnt wurde; sofern auch die am wenigsten zutreffende Aussage angekreuzt werden soll). Dementsprechend können keine Absolutwerte pro Eigenschaft ermittelt werden, sondern es liegt zunächst nur die relative Präferenz einer Eigenschaft der Person gegenüber anderen Eigenschaften der gleichen Person vor. Anders ausgedrückt: Ergebnisse können nicht über Personen hinweg verglichen werden. Glücklicherweise liegen seit einiger Zeit auch Auswertungsmodelle vor, die es – mit einem Rechneraufwand – möglich machen, mit Forced-Choice-Antwortformaten gemessene Eigenschaften über Personen hinweg zu vergleichen (Brown und Maydeu-Olivares 2013).

Weiterführende Literatur und Internetressourcen

Eine sehr gute Übersicht zur Gestaltung von Antwortskalen liegt von Menold und Bogner (2015) vor. Sie ist unter ► https://www.gesis.org/fileadmin/upload/SDMwiki/Archiv/Ratingskalen_MenoldBogner_012015_1.0.pdf (Stand: April 2020) abrufbar.

2.5 Grundzüge von Itemanalysen

In diesem Abschnitt werden gängige Methoden der empirischen Itemanalyse vorgestellt. Diese kommen in den Schritten 4 und 6 des in □ Abb. 2.22 dargestellten Ablaufs der Testentwicklung zum Einsatz.

2.5.1 Itemschwierigkeit (nach der Klassischen Testtheorie)

Definition

Die **Itemschwierigkeit** nach der Klassischen Testtheorie gibt an, wie groß der Anteil der Personen ist, die das Item im Sinne des Merkmals beantwortet haben. Je höher der Anteil der Personen ist, die ein Item im Sinne des Merkmals beantworten, desto leichter ist das Item.

Beachte: Im Rahmen einer Probabilistischen Testtheorie, dem einparametrischen dichotomen Rasch-Modell, ist die Itemschwierigkeit anders definiert – nämlich als der Punkt auf dem Fähigkeitsspektrum, an dem die Lösungswahrscheinlichkeit $p(x_{vi}=1)=.50$ ist (► Abschn. 2.3.1.1).

Unterschiedliche Definition von Itemschwierigkeit in Klassischer und Probabilistischer Testtheorie

„Im Sinne des Merkmals beantwortet“ ist bei Leistungstests die richtige Antwort und bei Fragebögen die Antwort, die eine maximale Merkmalsausprägung indiziert (also z. B. eine 5 auf einer Antwortskala von 1 bis 5 eines positiv gepolten Items zur Depressivität). In □ Abb. 2.27 ist (von oben nach unten) ein Item mit (im allgemeinsprachlichen Sinne) mittlerer Itemschwierigkeit, eines mit hoher Schwierigkeit und eines mit geringer Schwierigkeit dargestellt.

Die Item Leichtigkeit kann anhand folgender Formeln berechnet werden:

Berechnung der Itemleichtigkeit

Berechnung der Itemleichtigkeit

Für Leistungstestitems mit 0 (falsche Antworten) und 1 (richtige Antworten) sowie für Fragebogenitems, bei denen die Antwortskala bei 0 beginnt und der höchste Antwortskalenpunkt für die höchste Merkmalsausprägung steht, gilt:

$$P_i = \frac{\sum_{v=1}^n x_{vi}}{n \times \max(x_i)} \times 100$$

P_i = Itemleichtigkeitsindex für das Item i

x_{vi} = Itemantwort einer Person v in Item i

n = Anzahl aller Personen, die Item i beantwortet haben

$\max(x_i)$ = numerisch größtmögliche Antwort in Item i (dies wäre eine „1“ bei Richtig-Falsch-Antworten)

Dabei gilt: je größer P_i , desto leichter das Item (im allgemeinsprachlichen Sinne). Daher bevorzugen wir für P_i die Bezeichnung Itemleichtigkeitsindex, und verwenden hier bewusst nicht die Bezeichnung als Itemschwierigkeitsindex.

Für Fragebogenitems, bei denen die Antwortskala nicht bei 0 beginnt (s. auch das Beispiel in Abb. 2.27), gilt:

$$P_i = \frac{\sum_{v=1}^n (x_{vi} - \min(x_i))}{n \times (\max(x_i) - \min(x_i))} \times 100$$

$\min(x_i)$ = numerisch kleinstmögliche Antwort in Item i (dies wäre eine 1 bei Antworten auf einer Skala von 1 bis 5)

(Formeln aus Kelava und Moosbrugger 2020a, S. 146–147).

Unabhängig von der Art des Items (Leistungs- vs. Persönlichkeitstestitem) und der Nummerierung der Skala bewegt sich der Itemleichtigkeitsindex zwischen Werten von 0 bis 100.

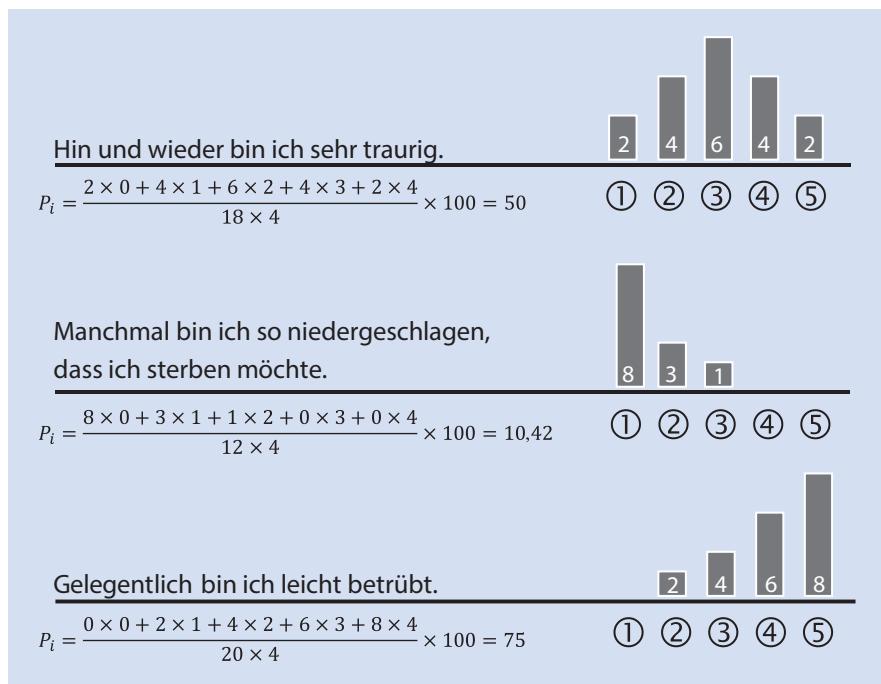


Abb. 2.27 Darstellung von 3 Items unterschiedlicher Leichtigkeit. Antworten von 18 bzw. 20 Personen auf einer Antwortskala von 1 bis 5 (Erläuterungen zur Berechnung s. Text)

Ein Itemleichtigkeitsindex von 0 oder 100 bedeutet, dass alle Personen das Item gleich beantwortet haben. Die Information eines solchen Items lautet: „Alle Personen sind gleich.“ Das Item trägt also nicht dazu bei, Unterschiede zwischen Personen aufzuzeigen. Solche Items sollten keinen Eingang in den initialen Testentwurf finden. Es ist günstig, bei der Testkonstruktion mit den ausgewählten Items ein möglichst breites Schwierigkeitsspektrum abzubilden.

Extreme Itemleichtigkeit bzw.
-schwierigkeit nicht wünschenswert

- ! Es gilt zu beachten, dass Itemschwierigkeiten bzw. -leichtigkeiten – falls sie nicht gemäß probabilistischer Testmodelle berechnet wurden (vgl. ▶ Abschn. 2.3) – ein Item relativ zu der *für die Analysen verwendeten Stichprobe* beschreiben. Daher ist es bereits bei der Testkonstruktion und für Analysen des 1. Testentwurfs wichtig, möglichst Stichproben zu rekrutieren, die repräsentativ für die intendierte Zielgruppe des Tests sind. Manchmal ist die intendierte Zielgruppe nur aufwendig zu rekrutieren (z. B. klinische Populationen oder Führungskräfte).

Man könnte stattdessen versuchen, erste Analysen des Testentwurfs anhand von Daten einer Studierendenstichprobe vorzunehmen, und sich die eigentliche Zielgruppe zur Prüfung des finalen Tests „aufheben“. Allerdings ist zu beachten, dass stichprobenabhängige Kennwerte je nach Stichprobe deutlich anders ausfallen und zu anderen (ggf. falschen) Entscheidungen bei der Itemselektion führen können. Beispielsweise können sich Items in einer Stichprobe aus Schülerinnen und Schülern als zu schwer herausstellen, wären jedoch für Studierende hinsichtlich ihrer Schwierigkeit absolut angemessen.

Itemleichtigkeitsindex von der Stichprobe abhängig

Es ist weiterhin zu beachten, dass einige Leistungstests mit einer Zeitbegrenzung versehen sind. Man spricht bei moderat zeitbegrenzten Tests auch von Tests mit einer Speedkomponente. Dahingegen spricht man von „reinen“ Speedtests, wenn diese aus sehr leichten Aufgaben und einer extrem anspruchsvollen Zeitbegrenzung bestehen. Als reine Powertests werden solche Leistungstests bezeichnet, die keine Zeitbegrenzung haben und stattdessen aus immer schwierigeren Aufgaben bestehen.

Speed- vs. Powertests

Bei der Berechnung der Itemschwierigkeit von Tests mit einer Speedkomponente ist zu beachten, dass Items am Ende eines Tests von einigen Personen aufgrund der Zeitbegrenzung erst gar nicht in Angriff genommen werden. Sie gelten als nicht gelöst. Sie werden daher häufig als (im allgemeinsprachlichen Sinne) schwierige Items erachtet. Dies ist insofern kritisch, da die Nichtinangriffnahme eines Items keine Eigenschaft dieses Items ist, sondern eine Eigenschaft der vorherigen Items: Dadurch dass die vorherigen Items zu schwierig waren, konnten viele Personen das spätere Item nicht in Angriff nehmen. Bei strikter Verwendung der zuvor erwähnten Formel zur Berechnung der Itemschwierigkeit löst sich dieses Problem jedoch. Dort gehen im Nenner nur die Personen ein, die das Item tatsächlich beantwortet haben. Dieses Vorgehen hat aber den Nachteil, dass die Itemkennwerte nur für eine stark selegierte Personenstichprobe (besonders leistungsfähige Personen) ermittelt werden. Um dies zu umgehen, kommen 2 weitere Ansätze infrage: Erstens kann man den Test unter Speedbedingung („so schnell wie möglich arbeiten“ und Zeitmessung) dennoch vollständig bearbeiten lassen. Zweitens ist es möglich, 2 Itemabfolgen zu verwenden, beispielsweise so dass alle Items auch einmal in der Mitte des Tests platziert sind (z. B. Item 1, 2, 3, 4, 5, 6 – Item 4, 5, 6, 1, 2, 3).

Itemschwierigkeit von selten in Angriff genommenen Items

2.5.2 Itemstreuung bzw. Itemvarianz

2

Definition

Die synonymen Begriffe **Itemstreuung** und **Itemvarianz** beschreiben das Ausmaß, in dem bei einem Item die Antworten zwischen den Testpersonen variieren.

Abb. 2.28 zeigt 3 Items mit unterschiedlicher Varianz. Während das 1. Item sehr leicht ist und daher kaum Varianz aufweist, streuen die Antworten bei dem 2. und 3. Item deutlich – und in unterschiedlicher Weise – über die Antwortskala. Das 2. ist ein eher schwieriges Item, das dennoch zwischen Personen differenziert. Dem 3. Item würde man eine mittlere Schwierigkeit attestieren – auch wenn die meisten Antworten „extrem“ sind, da besonders häufig „1“ oder „5“ als Antwort gewählt wurde.

Die Itemstreuung kann anhand folgender Formeln berechnet werden:

Berechnung der Itemstreuung

$$s_{xi}^2 = \frac{\sum_{v=1}^n (x_{vi} - \bar{x}_i)^2}{n}$$

s_{xi}^2 = Varianz der beobachteten Werte in Item i

x_{vi} = Itemantwort einer Person v in Item i

\bar{x}_i = Mittelwert der Antworten in Item i

n = Anzahl aller Personen, die Item i beantwortet haben

(Formel aus Kelava und Moosbrugger 2020a, S. 152)

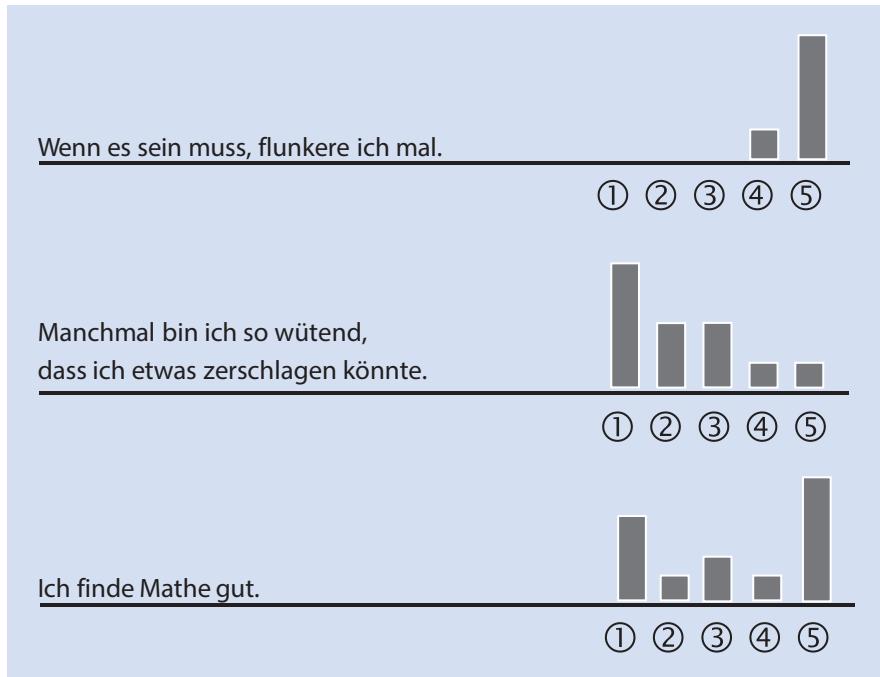


Abb. 2.28 Items mit unterschiedlicher Streuung bzw. Varianz

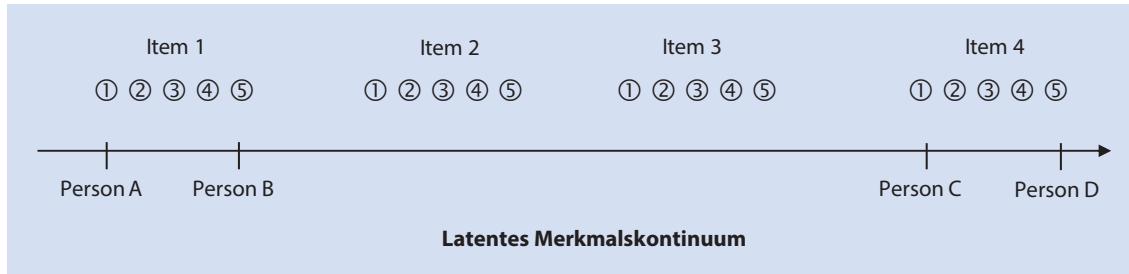


Abb. 2.29 Schematische Anordnung von Items unterschiedlicher Schwierigkeit auf einem Merkmalskontinuum

Wie bereits erwähnt, sind Itemstreuung und Itemschwierigkeit nicht unabhängig voneinander. Extreme Itemschwierigkeiten führen zu geringen Itemstreuungen. Mittlere Itemschwierigkeiten bieten die beste Voraussetzung für hohe Itemstreuungen. Allerdings darf daraus nicht geschlossen werden, dass ein Test nur Items mittlerer Schwierigkeit enthalten sollte. Denn dies würde bedeuten, dass der Test für Personen mit einer mittleren Merkmalsausprägung gut differenzieren würde; er würde aber schlecht differenzieren für Menschen mit hoher oder niedriger Merkmalsausprägung. Abb. 2.29 veranschaulicht dies, indem Personen und Items auf einem gedachten Kontinuum der Merkmalsausprägung bzw. der Schwierigkeit angeordnet sind. Die Items 2 und 3 haben eine mittlere Schwierigkeit, Item 4 ist ein sehr schwieriges, Item 1 ein sehr leichtes Item (im allgemeinsprachlichen Sinne).

In dieser hypothetischen Anordnung von Items und Personen auf einem Kontinuum würde Person B bei Item 1 eine „5“ (z. B. „stimme voll und ganz zu“) ankreuzen. Person B würde jedoch bei den Items 2, 3 und 4 jeweils eine „1“ ankreuzen (z. B. „stimme gar nicht zu“), da die Items zu extrem formuliert sind, als dass Person B mit ihrer geringen Merkmalsausprägung diesen zustimmen könnte. Würde hier beispielsweise Gewissenhaftigkeit gemessen, so könnte Person B der Aussage von Item 1 (z. B. „Ich erledige meine Aufgaben sorgfältig“) zustimmen, aber die Aussagen der Items 2 bis 4 (z. B. „Ich bringe alle meine Aufgaben zu Ende“, „Ich bin stets pünktlich“, „In allem, was ich tue, kann man sich 100 %ig auf mich verlassen“) ablehnen.

Bestünde der Test nur aus Items im mittleren Schwierigkeitsbereich (Items 2 und 3), so könnten weder Personen mit niedriger (A und B) noch solche mit hoher (C und D) Merkmalsausprägung gut unterschieden werden. Die Personen A und B würden bei den Items 2 und 3 jeweils die „1“ ankreuzen und sich damit in ihrem Antwortverhalten nicht unterscheiden. Personen C und D würden bei den Items 2 und 3 jeweils die „5“ ankreuzen und sich somit ebenfalls nicht unterscheiden. Erst die Hinzunahme der Items 1 und 4 ermöglicht eine Unterscheidung dieser Personen. Person A kreuzt bei Item 1 eine „1“ an, während Person B die „5“ wählt. Person C kreuzt bei Item 4 die „1“ an, Person D die „5“.

Abhängigkeit von Itemschwierigkeit und -streuung

Beispielhaftes Antwortverhalten auf Items unterschiedlicher Leichtigkeit

2.5.3 Trennschärfe

Definition

Die **Trennschärfe** eines Items ist definiert als die Korrelation des Items i mit dem Test oder Testteil t , zu dem dieses Item gehört (daher auch häufig als r_{it} bezeichnet).

Trennschärfe=Maß der Übereinstimmung zwischen Item und Skala

Trennschärfe sollte $> .30$ sein

Testinhalt bei der Bewertung der Trennschärfe beachten

Bei Leistungstestitems mit Richtig-Falsch-Antworten ist sie ein Kennwert dafür, in welchem Ausmaß die durch das Item erfolgte Differenzierung der Probandinnen und Probanden in Lösende und Nichtlösende mit derjenigen durch die Skala als Ganzes übereinstimmt. In Persönlichkeitsfragebogen gibt sie an, inwiefern ein Item das Ergebnis des Testteils, zu dem das Item gehört, reflektiert. Die Trennschärfe zeigt also an, in welchem Ausmaß das Item das Gleiche misst wie der Test bzw. bei mehrdimensionalen Tests die Subskala des Tests. Man könnte auch sagen: Die Trennschärfe zeigt an, wie gut das Item zwischen Personen mit hohem vs. niedrigem Testwert „trennt“.

Wie hoch sollten Trennschärfen sein? Als Korrelationskoeffizient kann die Trennschärfe zwischen -1 und +1 liegen. Negative Werte können tatsächlich vorkommen, sind aber eher selten. Ein negativer Wert würde bedeuten, dass Personen, die ein Item lösen, im Gesamttest eher schlecht abschneiden und umgekehrt. Ein solches Item sollte modifiziert oder ganz entfernt werden. Gängige Konventionen empfehlen, möglichst nur Items mit einer Trennschärfe $> .30$ in einem Test zu belassen. Eine Trennschärfe, die nur geringfügig $> .00$ ist, besagt in der Tat, dass das Item nicht gut zwischen Personen mit hohen vs. niedrigen Werten im Test trennt.

Bei der Beurteilung der Trennschärfen sollte jedoch auch die Homogenität des zu messenden Merkmals beachtet werden. Heterogene, vielschichtige Merkmale bestehen zumeist aus Items mit moderaten Trennschärfen. Bei homogenen Merkmalen sollten die Trennschärfen dagegen hoch sein. An einem einfachen Beispiel lässt sich die Logik dieser Argumentation nachvollziehen. Ein Schulleistungstest soll nur einen speziellen Aspekt der Rechenfertigkeit messen, nämlich das Addieren von einstelligen Zahlen ($4 + 5 = ?$, $8 + 4 = ?$ etc.). Alle Items erfassen etwas sehr Ähnliches – sie verlangen alle Additionen. Deshalb erwarten wir hohe Trennschärfen. Ein anderer Schulleistungstest soll die Beherrschung der Grundrechenarten prüfen. Er enthält daher Items zu Addition, Subtraktion, Multiplikation und Division. Diese Fertigkeit ist deutlich heterogener. Schülerinnen und Schüler, die gut addieren können, tun sich vielleicht bei der Subtraktion schwer. Daher erwarten wir niedrigere Trennschärfen als beim 1. Test. Für nach der externalen Testkonstruktionsstrategie entwickelte Tests, deren Items so ausgewählt wurden, dass sie eine real existierende Gruppenzugehörigkeit (z. B. „in Behandlung“ vs. „nicht in Behandlung“) vorhersagen, können Trennschärfen aufgrund der Heterogenität der Items gänzlich ungeeignet sein. Insgesamt ist zu beachten: Eine Optimierung eines Tests mit einem heterogenen Messanspruch anhand von Trennschärfen kann zu einer Verringerung der Validität führen. Wie hoch die Trennschärfen in verbreiteten Tests sind, kann Tab. 2.9 entnommen werden.

Notwendigkeit der Part-whole-Korrektur

Bei der Berechnung des Testwertes als Summe aller Antworten bleibt das jeweilige Item, für das die Trennschärfe bestimmt werden soll, unberücksichtigt. Für diese sog. „Part-whole-Korrektur“ gibt es einen einfachen Grund. Würde das jeweilige Item bei der Berechnung des Testwertes inkludiert werden, käme es bei der Berechnung der Korrelation wegen der bestehenden algebraischen Abhängigkeit zu einer Korrelation des Items mit sich selbst. Die Korrelation des Items mit dem Test wird dadurch künstlich erhöht. Dieser Effekt ist umso stärker, je weniger Items der Test hat. Bei 5 Items wird der Gesamtwert bereits zu 1/5 durch das zu analysierende Item definiert. Es sollte daher stets eine Part-whole-Korrektur erfolgen. Die einschlägigen Statistikprogramme nehmen diese Korrektur bei der Berechnung der Trennschärfe automatisch vor.

Tab. 2.9 Trennschärfen in ausgewählten Tests

Test, Skala	Anzahl Items	Trennschärfen		Itemschwierigkeit	
		Spanne	Mittelwert	Spanne	Mittelwert
FPI-R ^a , Soziale Orientierung	12	.28–.44	.34	0,31–0,75	0,54
FPI-R, Beanspruchung	12	.37–.66	.50	0,26–0,57	0,47
NEO-PI-R ^b , Facette „Offenheit für Werte“	8	.12–.30	.25	2,06–3,04	2,64
NEO-PI-R, Facette „Depression“	8	.46–.67	.58	1,16–2,51	1,75
I-S-T 2000 R ^c , Satzergänzung	20	.13–.39	.26	0,17–0,92	0,64
I-S-T 2000 R, Rechenzeichen	20	.15–.60	.43	0,14–0,95	0,60

^aFreiburger Persönlichkeitsinventar – revidierte Fassung (FPI-R), Fahrenberg et al. (2010, S. 35): Skalen mit den durchschnittlich niedrigsten und höchsten Trennschärfen ausgewählt

^bNEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R), Ostendorf und Angleitner (2004, S. 93 f.): Subskala mit dem niedrigsten und höchsten Cronbachs Alpha für die Gesamtgruppe (S. 105); bei Schwierigkeit Skala (hier nicht als Leichtigkeitsindex P , sondern als mittlere Itemantwort angegeben) von 0 bis 4

^cIntelligenz-Struktur-Test 2000 – Revision (I-S-T 2000 R), Liepmann et al. (2007, S. 24 ff.): aus Grundmodul Form A/B (ohne Merkfähigkeit) Skala mit dem niedrigsten und höchsten Cronbachs Alpha für die Gesamtgruppe

Von welchen Faktoren hängt die Trennschärfe ab? Die Höhe der Trennschärfe hängt von der inhaltlichen Passung des Items, der Verteilungsform von Itemantworten und Testwerten sowie von der Streuung des Items und der Testwerte ab. Unter „inhaltlicher Passung“ verstehen wir, dass das Item gut geeignet ist, das Merkmal zu messen, das der Test erfasst. Für mangelnde Passung gibt es die folgenden beiden Gründe. Der naheliegende ist, dass das Item schlecht ausgewählt oder schlecht formuliert wurde. Wenn ein Fragebogen Aggressivität erfasst, dann ist ein Item wie „Ich widerspreche manchmal meinen Gesprächspartnerinnen und -partnern“ vermutlich unpassend, da es eher etwas über Dominanz oder Selbstvertrauen aussagt als über Aggressivität. Eine Schädigung oder Schädigungsabsicht ist nicht zu erkennen. In seltenen Fällen ist ein Item eigentlich passend, die übrigen Items verfehlten aber die Messintention. Beispielsweise könnte ein Depressionsfragebogen viele Items zu körperlichen Beschwerden enthalten. Obwohl bestimmte körperliche Beschwerden wie Appetit- oder Gewichtsverlust zum Bild einer Depression gehören, müssen eine Reihe anderer Symptome vorhanden sein. Mit dem Schwerpunkt auf körperliche Beschwerden ist der Fragebogen vielleicht eher zu einem Instrument zur Erfassung von somatoformen Störungen geworden. Ein eigentlich zur Depression passendes Item zum Thema Schuldgefühle (z. B. „Ich habe oft Schuldgefühle“) korreliert eventuell mit dem Gesamtwert so niedrig, dass man das Item als ungeeignet ansehen wird. An diesem fiktiven Beispiel wird noch einmal deutlich, dass die Trennschärfe nur die Übereinstimmung mit dem (part-whole-korrigierten) Testwert prüft und nicht, wie gut das Merkmal gemessen wird.

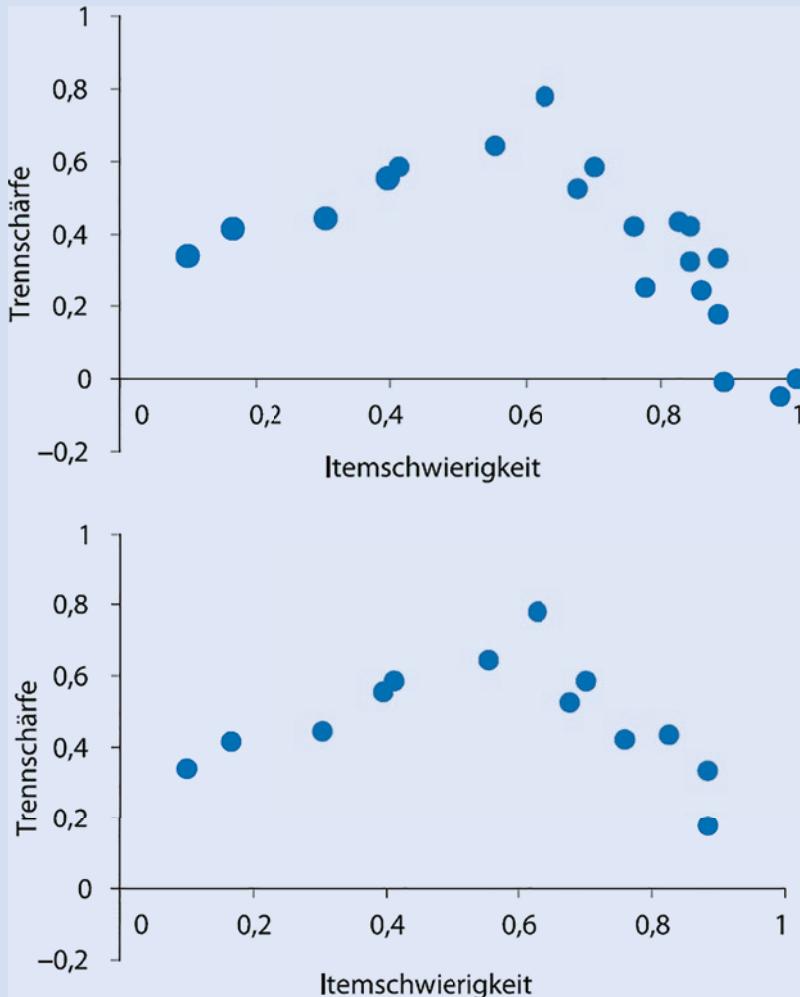
Die Verteilungsform der (part-whole-korrigierten) Testwerte und Itemantworten kann die Korrelation zwischen Item und Test mindern, und zwar immer dann, wenn Item- und Testwerte unterschiedliche Verteilungsformen aufweisen. Sind beide in gleicher Weise schief verteilt, wird die Korrelation dadurch nicht eingeschränkt. Empirisch stellt sich meist eine umgekehrte u-förmige Beziehung in dem Sinne dar, dass mit sehr niedrigen und sehr hohen Schwierigkeiten eher mäßige, mit mittleren Schwierigkeiten eher hohe Trennschärfen einhergehen. Dies liegt an den mit extremer werdenden Itemschwierigkeiten zunehmend geringeren Streuungen. Die Trennschärfe hängt als Korrelationskoeffizient von der Streuung der Items (und der Testwerte) ab. Ist sie klein, mindert das die Korrelation und damit die Trennschärfe.

Faktoren, die die Trennschärfe beeinflussen

Einfluss der Schiefe der Verteilung und der Streuung

Trennschärfe und Itemschwierigkeit

Mithilfe einer Grafik, in der die Items nach ihrer Schwierigkeit und ihrer Trennschärfe angeordnet sind, wird deutlich, welche Items gemessen an ihrer Schwierigkeit eine zu niedrige Trennschärfe haben. Folgende Grafik zeigt ein Beispiel:



Diese Abbildungen zeigen die Beziehung der Trennschärfe zur Itemschwierigkeit. Die obere Abbildung zeigt alle Items, in der unteren wurden Items eliminiert. Die Itemschwierigkeit wird hier nicht als Leichtigkeitsindex P_p sondern als Mittelwert der Itemantworten angegeben.

Es handelt sich um die 20 Items zum rechnerischen Denken aus dem Intelligenztest I-S-T 2000 R (Liepmann et al. 2007). Der Test wurde von 120 Studierenden bearbeitet. Diese Gruppe ist für diese Itemanalyse nicht adäquat, weil sie zu leistungsstark ist. Wir unterstellen zu Demonstrationszwecken, dass die Kennwerte von einer adäquaten Stichprobe stammen. Zunächst erkennt man gut die umgekehrt u-förmige Beziehung, wie sie typisch für die Beziehung

zwischen Trennschärfe und Itemschwierigkeit ist. Die höchsten Trennschärfen finden wir erwartungsgemäß im mittleren Bereich der Itemschwierigkeit (hier etwas nach rechts verschoben). Drei Items in der oberen Abbildung haben Trennschärfen von 0 oder sogar negative Trennschärfen und würden daher eliminiert. Weiterhin finden wir relativ viele leichte Items (Itemschwierigkeiten um .8 und höher), von denen man 4 eliminieren könnte – vorzugsweise die mit den niedrigeren Trennschärfen. Übrig blieben die in der unteren Abbildung gezeigten Items.

Nach Elimination der Items ist es angemessen, erneut die Trennschärfen zu berechnen und zu bewerten. Anders als die Itemschwierigkeiten werden sich die Trennschärfen verändern. Der Gesamttestwert wird nämlich nur noch über die 13 ausgewählten Items berechnet.

Alternativ ist ein Selektionskennwert berechenbar, der die Trennschärfen für die Itemstreuung (die sich aus der Itemschwierigkeit herleitet) korrigiert (s. Büchner 2021, S. 222). Je kleiner die Itemstreuung (beispielsweise aufgrund extremer Itemschwierigkeit) ist, desto stärker wird die Trennschärfe aufgewertet.

$$SK = \frac{r_{it}}{2 \times s_i}$$

SK = Selektionskennwert

r_{it} = Trennschärfe des Items i aus Test t

s_i = Standardabweichung des Items i

Selektionskennwert verrechnet
Trennschärfe und Itemstreuung

Die Trennschärfe hängt folglich auch von der Streuung der Testwerte ab. Wurde der Test einer sehr homogenen Stichprobe von Personen vorgegeben, bei der das Merkmal nicht stark variiert, müssen die Trennschärfen zwangsläufig niedriger ausfallen als bei einer heterogenen Stichprobe.

Einflussfaktoren auf die Trennschärfe

- Inhaltliche Passung von Item und Test zueinander
- Verteilungsform von Itemantworten und Testwerten
- Streuung der Item- und Testwerte

Trennschärfen von Distraktoren Bei Items im Multiple-Choice-Format ist es möglich, auch die Distraktoren (Falsch-Antworten) einer Itemanalyse zu unterziehen. Dazu werden anstelle der Richtig-Antworten die Antworten auf einen Distraktor analysiert (also Distraktor angekreuzt vs. nicht angekreuzt). Wenn ein Item 4 Distraktoren hat, sind 4 Analysen erforderlich. Auf diese Weise erfährt man zunächst einmal zunächst einmal, wie häufig ein Distraktor gewählt wurde. Zu selten gewählte Distraktoren sind zu leicht als Falsch-Antwort zu erkennen und können durch Umformulierung eventuell schwerer gemacht werden. Die Trennschärfe eines Distraktors sollte negativ sein, da gelten sollte: Personen, die den Distraktor gewählt haben, also das Item falsch beantwortet haben, sollten im Gesamttest schlechter abschneiden als Personen, die den Distraktor nicht gewählt haben. Unter Umständen

Distraktortrennschärfe

entdeckt man, dass eine scheinbar falsche Antwort positiv mit dem Testwert korreliert (positive Trennschärfe). Entweder handelt es sich um einen Zufallsbefund (der nicht replizierbar ist) oder diese Antwort ist doch richtig bzw. bei einer anderen Interpretation des Wortlauts richtig. Ein solcher Distraktor sollte ausgetauscht oder das Item umformuliert werden.

2.5.4 Itemladungen auf Faktoren

Faktorenanalyse

Die Faktorenanalyse stellt eine wichtige Methode im Rahmen der Testkonstruktion dar. Insbesondere bei der induktiven Strategie der Testkonstruktion „verlässt“ man sich zumeist auf Ergebnisse von Faktorenanalysen. Es würde jedoch den Rahmen dieses Lehrbuchs sprengen, die Methode der Faktorenanalyse im Detail vorzustellen. Die nachfolgenden Erläuterungen sollen dennoch ein Grundverständnis der (explorativen) Faktorenanalyse ermöglichen. Für weiterführende Erläuterungen und Details bzw. Varianten der Methode wird auf die am Ende dieses Abschnitts genannte Literatur verwiesen.

Definition

Die Faktorenanalyse stellt eine Methode dar, mithilfe derer korrelative Zusammenhänge zwischen beobachtbaren Variablen (d. h. Items) durch möglichst wenige zugrunde liegende Dimensionen (sog. „Faktoren“) beschrieben werden können.

Ziel: Beschreibung des Datenmusters durch wenige Faktoren

Das grundlegende Rational der Faktorenanalyse wurde bereits in □ Abb. 2.24 realisiert: Items eines Tests sollten nach Möglichkeit ausschließlich das zugrunde liegende Merkmal reflektieren. Beeinflusst ein Merkmal tatsächlich in hohem Maße die Antworten aller Testitems, so sollten diese hoch miteinander korrelieren. Mithilfe der Faktorenanalyse versucht man, dieses Rational nachzuvollziehen. Allerdings geht man quasi umgekehrt vor: Es liegen Itemantworten und Korrelationen der Items untereinander vor – man prüft nun, wie dieses Datenmuster durch einen oder mehrere Faktoren (dann interpretiert als zugrunde liegende Merkmale) sparsam beschrieben werden kann.

Konkret gelingt dies, indem zunächst alle beobachteten Variablen, also die Testitems, als Linearkombination der neu gebildeten Faktoren verstanden werden. Geht man, wie in dem in □ Abb. 2.24 dargestellten Fall, von nur einem Faktor aus, lauten die Linearkombinationen der Items:

$$Y_i = \alpha_i + \lambda_{i1} \times \eta_1 + \varepsilon_i$$

Sollen die Itemantworten durch 2 Faktoren erklärt werden, kommen folgende Linearkombinationen zur Anwendung:

$$Y_i = \alpha_i + \lambda_{i1} \times \eta_1 + \lambda_{i2} \times \eta_2 + \varepsilon_i$$

Y_i = Item i

α_i = Konstante für Item i

$\lambda_{i1}/\lambda_{i2}$ = Ladung des Items i auf Faktor 1 bzw. Faktor 2

η_1/η_2 = Faktor 1 bzw. Faktor 2

ε_i = Fehlervariable

(vgl. Eid et al. 2017, S. 880).

Man versucht also – analog zum Vorgehen bei einer multiplen Regression – Itemwerte durch eine Kombination von Faktoren möglichst gut zu erklären. Die Fehlervariable beschreibt das Ausmaß, um das die Itemwerte durch die jeweils berücksichtigten Faktoren *nicht* erklärt werden können. Der Anteil der Itemvarianz, der durch alle berücksichtigten Faktoren erklärt werden kann, wird als **Kommunalität** bezeichnet; sie errechnet sich aus der Summe der quadrierten Ladungen eines Items auf allen (unkorrelierten) Faktoren.

Definition

Die **Kommunalität** h_i^2 bezeichnet den Anteil der Varianz eines Items i , der durch alle berücksichtigten (unkorrelierten) Faktoren $j=1$ bis k erklärt wird:

$$h_i^2 = \sum_{j=1}^k \lambda_{ij}^2$$

Nun wäre es ziemlich unsinnig, ebenso viele Faktoren zu berücksichtigen wie Items in einem Test enthalten sind. Im Falle der in □ Abb. 2.24 dargestellten 6 Items könnten diese natürlich durch 6 Faktoren perfekt beschrieben werden. Allerdings erfolgt dann durch die Faktoren keine Abstraktion der Daten und das Ziel, Itemwerte durch wenige zugrunde liegende Faktoren zu beschreiben, würde verfehlt.

Doch wie viele Faktoren sind nötig, um die tatsächlich beobachteten Daten hinreichend gut zu beschreiben? Im Falle von 6 Items könnten vielleicht 5 Faktoren herangezogen werden? Oder würden auch 4 ausreichen? Oder wäre gar – im Sinne eines eindimensionalen Modells – ein einziger Faktor ausreichend? Um diese Fragen im Rahmen der explorativen Faktorenanalyse zu beantworten, geht man schrittweise vor (hätte man theoretisch begründbare Annahmen über die Faktorenzahl, wären konfirmatorische Faktorenanalysen angezeigt). Das heißt, man prüft, wie gut die vorliegenden Daten durch zunächst 1 Faktor, dann durch 2, durch 3 Faktoren usw. beschrieben werden. Wir veranschaulichen dies anhand des nachfolgenden Beispiels.

Nehmen wir vereinfachend an, dass uns für 5 Items Daten von 100 Personen vorliegen (□ Tab. 2.10; für Empfehlungen zu Stichprobenumfängen s. Fabrigar et al. 1999). Nehmen wir weiterhin an, die Korrelationsmatrix dieser Items entspricht der in □ Tab. 2.11.

Man sieht anhand der Korrelationen bereits, dass vor allem die Items 3 bis 5 viel „Gemeinsames“ abbilden. Vermutlich werden sie durch einen einzigen Faktor ausreichend zu beschreiben sein. Aber auch die Items 1 und 2 korrelieren substanzell miteinander. Sie korrelieren jedoch nicht bzw. kaum mit den Items 3, 4 und 5. Vermutlich wird also ein weiterer Faktor nötig sein, um die Items 1 und 2 zu beschreiben.

Nun kreieren wir also unsere 1. neue Variable (Faktor 1). Diese Variable spiegelt keine tatsächlich gemessenen Werte von Personen wider, sie ist auch nicht als Summenscore über (in unserem Beispiel) 5 Testitems zu verstehen. Es ist eine fiktive Variable, die so gebildet wird, dass sie – als Teil der zuvor genannten Linearkombination – die Daten der 5 Items möglichst gut reproduziert.

In unserem Beispiel ergibt sich – bei Verwendung einschlägiger Statistiksoftware – ein Faktor (F1), mit dem die 5 Items korrelieren (□ Tab. 2.12).

Definition

Faktorladung: Die Ladung eines Items auf einem Faktor (in der zuvor aufgeführten Linearkombination mit λ_i notiert) kennzeichnet den Zusammenhang des Items mit dem Faktor.

Die neue Variable F1 und die Ladungen der Items mit dieser Variablen können genutzt werden, um die Werte aller 5 Items (zuvor Y_i) und damit auch deren Interkorrelationen zu reproduzieren. In unserem Beispiel sieht die so reproduzierte Korrelationsmatrix wie in □ Tab. 2.13 gezeigt aus.

Nun wird die reproduzierte mit der ursprünglichen Korrelationsmatrix abgeglichen. Man sieht, dass die Interkorrelationen der Items 3 bis 5 recht gut reproduziert werden. Dies deckt sich mit den sehr hohen Ladungen dieser Items auf dem Faktor. Die Korrelation zwischen den Items 1 und 2 wird jedoch nicht gut reproduziert. Möglicherweise lässt sich dies durch die Berücksichtigung eines 2. Faktors (F2) optimieren?

Nun wiederholen wir das skizzierte Vorgehen für unser Beispiel, nehmen aber 2 (unkorrelierte) Faktoren an. Es resultieren die Ladungen in □ Tab. 2.14. Entsprechend unserer Erwartung weisen die Items 1 und 2 hohe Ladungen auf Faktor 2 auf.

□ Tab. 2.10 Auszug aus beispielhaften Daten von 100 Personen und 5 Testitems

	1	2	3	4	5
1	2	4	4	5	3
2	4	2	4	2	2
3	3	3	4	5	3
...
100	5	5	2	4	2

□ Tab. 2.11 Interkorrelationen der 5 Testitems

	1	2	3	4	5
1					
2	.638				
3	.074	.221			
4	.018	.462	.875		
5	-.050	.349	.955	.861	

□ Tab. 2.12 Ladungen der 5 Testitems auf Faktor 1 (Ergebnis der Hauptachsenanalyse mit dem Weighted-Least-Square-Schätzer)

	F1
1	.071
2	.424
3	.858
4	.914
5	.886

Tab. 2.13 Auf Basis der 1-Faktor-Lösung reproduzierte Interkorrelationen der 5 Test-items

	1	2	3	4	5
1					
2	.030				
3	.061	.364			
4	.065	.388	.784		
5	.063	.376	.760	.810	

Tab. 2.14 Ladungen der 5 Testitems auf den Faktoren 1 und 2 (Ergebnis der Hauptachsenanalyse mit dem Weighted-Least-Square-Schätzer, unrotierte Lösung)

	F1	F2
1	.110	.862
2	.503	.716
3	.900	-.251
4	.954	-.066
5	.929	-.179

Tab. 2.15 Auf Basis der 2-Faktoren-Lösung reproduzierte Interkorrelationen der 5 Testitems

	1	2	3	4	5
1					
2	.672				
3	-.117	.273			
4	.048	.433	.875		
5	-.053	.339	.881	.898	

Die anhand dieser 2 Faktoren reproduzierte Korrelationsmatrix spiegelt das tatsächliche Korrelationsmuster der 5 Items sehr gut wider (**Tab. 2.15**). Eine weitere Verbesserung der Passung zwischen empirischen und durch das Zwei-Faktoren-Modell vorhergesagten Werten scheint kaum mehr möglich.

Wie bereits erwähnt, indizieren die hohen Ladungen der Items 3 bis 5 auf Faktor 1, dass Faktor 1 eher deren gemeinsame Inhalte abbildet. Wären diese Items, die sich mit Angst und Sorge befassen, so würde man dem Faktor 1 diese inhaltliche Bedeutung zuordnen. Faktor 2 wird durch die Items 1 und 2 gebildet. Wenn sich diese Items eher mit körperlichen Belastungssymptomen befassen, so würde man Faktor 2 entsprechend inhaltlich interpretieren.

Definition

Der **Eigenwert** eines Faktors beschreibt den Anteil der Varianz aller in die Analyse einbezogener Items, der durch diesen Faktor erklärt wird.

■ Tab. 2.16 Ladungen von 8 Narzissmusitems auf den Faktoren 1 und 2

Itemnr	Theoretisch angenommene Dimension (auf der das Item laden sollte)	F1	F2
1	Bewunderung	.70	.09
2	Bewunderung	.68	.06
3	Bewunderung	.85	.10
4	Bewunderung	.79	.15
5	Rivalität	.02	.69
6	Rivalität	.02	.60
7	Rivalität	.18	.66
8	Rivalität	.20	.55
	Eigenwert	2,37	1,62

Der Eigenwert des 1. Faktors wurde wie folgt berechnet: $.70^2 + .68^2 + .85^2 + .79^2 + .02^2 + .02^2 + .18^2 + .20^2 = 2,37$

Eigenwerte als Basis zur Identifikation der Anzahl an Faktoren

Eigenwerte liefern wertvolle Hinweise für eine angemessene Zahl der zu extrahierenden Faktoren. Man erhält im Rahmen der Faktorenanalyse einen Eigenwert pro Faktor – sie werden (bei orthogonalen, d. h. unkorrelierten Faktoren) durch Addition der quadrierten Ladungen aller Items auf einem Faktor gebildet.

Nehmen wir an, dass für die Items eines Narzissmusfragebogens die Ladungen in ■ Tab. 2.16 auf 2 orthogonalen Faktoren vorliegen.

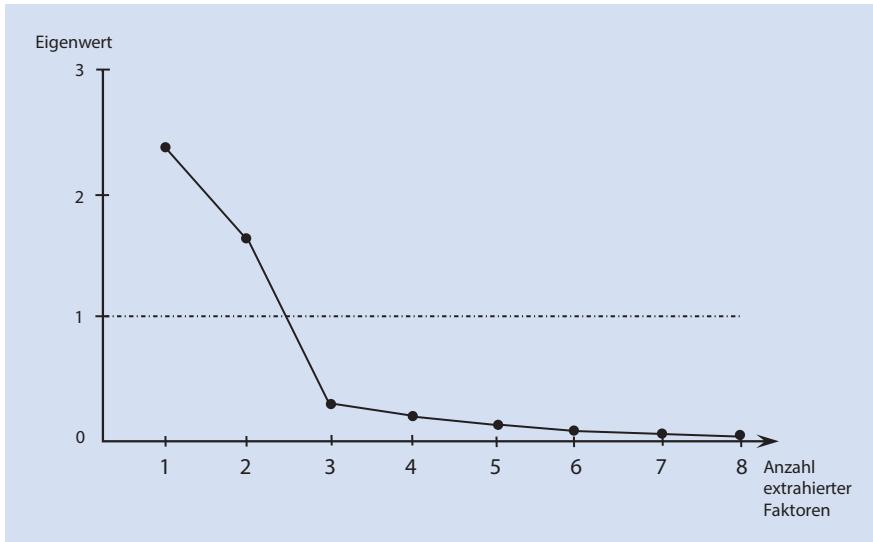
Wir vergegenwärtigen uns nochmals, dass die (standardisierte) Ladung eines Items auf einem Faktor maximal 1 sein kann. Daraus folgt, dass der Eigenwert eines Faktors maximal der Anzahl der Zahl der in die Faktorenanalyse eingehenden Items entsprechen kann. Im hier verwendeten Beispiel liegt der maximale Eigenwert als bei 8. Ein solcher Eigenwert würde bedeuten, dass der Faktor alle 8 Items perfekt beschreibt (die durch den Faktor erklärte Varianz wäre 100 %).

Ein gängiges Extraktionskriterium – das sog. „Kaiser-Kriterium“ – verlangt, dass Eigenwerte >1 sein sollen, da Faktoren ansonsten insgesamt nicht mehr Varianz der Items beschreiben als dies durch ein einzelnes Item gelänge. Ginge es um eine bloße Reduktion der Information, könnte man besser einzelne Items statt eines Faktors mit einem Eigenwert <1 interpretieren.

Ein weiteres Extraktionskriterium – das Scree-Kriterium – basiert auf einer Inspektion des Verlaufs der Eigenwerte über alle denkbaren Faktoren (von nur einem Faktor bis zu so vielen Faktoren, wie Items vorhanden sind) (■ Abb. 2.30). Weist der Verlauf der Eigenwertlinie einen Abfall von einem

Kaiser-Kriterium

Scree-Kriterium



■ Abb. 2.30 Eigenwerteverlauf. Auf der Ordinate sind Eigenwerte (von 0 bis 3, maximal könnten sie im vorliegenden Beispiel bis 8 reichen) dargestellt. Auf der Abszisse ist die Zahl der extrahierten Faktoren abgetragen. Bei Faktor 3 ist ein deutlicher Knick zu sehen, also ein Abfall der Eigenwerte vom 2. zum 3. Faktor. Es werden daher nur 2 Faktoren extrahiert

zum nächsten Faktor auf, bleibt man besser bei der Zahl der Faktoren, die vor dem Abfall der Eigenwerte gegeben war. Optisch stellt sich ein solcher Abfall als Knick im Eigenwerteverlauf dar. Allerdings liefert der Scree-Test nicht immer ein eindeutiges Ergebnis. Manchmal sieht man beispielsweise nicht nur einen, sondern 2 Knicke im Eigenwertverlauf, was den Interpretationsraum zusätzlich erhöht.

Scree-Plot als Darstellung eines Berges

Cattell (1966; zitiert nach Hoyle und Duval 2004) schlug vor, den Scree-Plot als Darstellung eines Berges zu betrachten, an dessen Fuß sich Geröll sammelt. Bei der Auswahl der zu extrahierenden Faktor ist zu prüfen, ob deren Eigenwerte noch Teil des Bergs – dann sind sie zu extrahieren – oder schon Teil des Gerölls und damit nicht zu extrahieren sind.

Schließlich kann der empirische Eigenwerteverlauf auch mit einem Eigenwerteverlauf verglichen werden (■ Abb. 2.31), der auf Basis von zufällig generierten (aus unkorrelierten Variablen bestehenden) Daten simuliert wurde. Es werden nur Faktoren extrahiert, deren Eigenwerte noch über dem „zufälliger Faktoren“ liegen. Diese Bestimmung der optimalen Zahl der Faktoren nennt man „Parallelanalyse“ (Horn 1965).

Parallelanalyse

Vorsicht

In der Praxis werden Faktorenanalysen und Hauptkomponentenanalysen häufig gleichgesetzt. Unter Hauptkomponentenanalyse (engl. principal component analysis, PCA) versteht man eine Methode der Datenreduktion. Anders als die Faktorenanalyse berücksichtigt sie jedoch zur Beschreibung der beobachteten Variablen, in unserem Fall also der Testitems, keine Fehlervariable (zuvor mit ε_i bezeichnet). Damit sind die resultierenden Komponenten keine Faktoren bzw. keine latenten Variablen, sondern lediglich lineare Komposita der beobachteten Variablen. Sie dürfen nicht als gemeinsame Faktoren der beobachteten Variablen interpretiert werden (Fabrigar et al. 1999).

Hauptkomponentenanalyse ist keine Faktorenanalyse

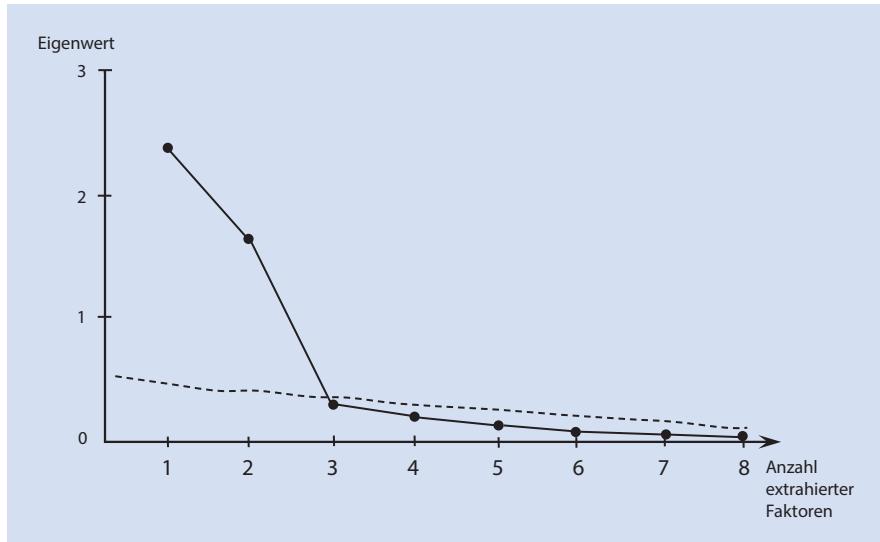


Abb. 2.31 Tatsächlicher und simulierter Eigenwerteverlauf. Auf der Ordinate sind Eigenwerte (von 0 bis 3, maximal könnten sie im vorliegenden Beispiel bis 8 reichen) dargestellt. Auf der Abszisse ist die Zahl der extrahierten Faktoren abgetragen. Der simulierte Eigenwerteverlauf ist als gestrichelte Linie dargestellt. Bei Faktor 3 fällt der empirische unter den simulierten Eigenwert

Inspizieren und Eliminieren von Items

Nach der Elimination: erneute Faktorenanalyse

Theorie- vs. datengetriebene Faktorenanalyse

Itemanalysen werden im Kontext von Faktorenanalysen dadurch vorgenommen, dass die Ladungen der Items auf den Faktoren inspiziert werden. Dabei muss zunächst sichergestellt werden, dass die intendierte Struktur eines Tests durch die Faktorenanalyse gut wiedergegeben wird. Also: Wenn nur ein Merkmal gemessen werden soll, so sollte die Faktorenanalyse auch nur einen Faktor als ausreichend identifizieren. Ist dies nicht der Fall, kann geprüft werden, ob Items, die auf anderen als dem 1. Faktor laden, eliminiert werden können. Dabei sind neben den Ergebnissen der Faktorenanalyse inhaltliche Überlegungen (z. B. zur Prototypikalität eines Items hinsichtlich des zu messenden Konstrukt) anzustellen. Zudem mag es Items geben, die keine substantiellen Ladungen (eine häufig verwendete Heuristik sieht Ladungen $> .30$ als substantiell an) aufweisen oder die nicht eindeutig zuzuordnen sind, da sie gleich hoch auf verschiedenen Faktoren laden. In diesen Fällen ist ebenfalls – nach inhaltlicher Prüfung der Iteminhalte und deren Relevanz – über einen Ausschluss der Items nachzudenken.

Nach Elimination von Items mit unpassenden Faktorenladungen wird man die Faktorenanalyse erneut durchführen und prüfen, ob die Daten für die Extraktion von nur einem Faktor sprechen und ob nun alle Items hinreichend hoch auf diesem einen Faktor laden.

Auf Basis eines initialen Itempools erfolgte explorative Faktorenanalysen sollten nach Revision des Testentwurfs an einer neuen Stichprobe repliziert werden. Idealerweise wird spätestens bei der Revision des Testentwurfs eine konfirmatorische, d. h. theorieprüfende Form der Faktorenanalyse angewendet. Dabei wird *a priori* spezifiziert, wie viele Faktoren zu identifizieren sind und welche Items auf diesen Faktoren laden. Es wird dann vorrangig geprüft, ob die vorliegenden Daten zu der *a priori* spezifizierten Faktorenstruktur „passen“ (für weitere Ausführungen s. Moosbrugger und Schermelleh-Engel 2012).

Beispielhafte Itemselektion nach Faktorenanalyse

Wir greifen das bereits in ► Abschn. 2.4.2.1 genannte Beispiel der Narzissmusdiagnostik auf. Die entwickelten Items sollten idealerweise die beiden relevanten Dimensionen „Bewunderung“ und „Rivalität“ abbilden. Nehmen wir an, folgendes Ergebnis der Faktorenanalyse läge vor (wir gehen hier nicht auf das Skalenniveau der Daten und die damit verbundene Wahl der entsprechenden Faktorenanalyse ein):

Itemnr	Theoretisch angenommene Dimension (auf der das Item laden sollte)	F1	F2	F3
1	Bewunderung (a-m)	.65		
2	Bewunderung (a-m)	.56		
3	Bewunderung (k)	.62		
4	Bewunderung (k)	.49	.32	
5	Bewunderung (b)			.62
6	Bewunderung (b)			.55
7	Rivalität (a-m)		.72	
8	Rivalität (a-m)		.77	
9	Rivalität (k)		.65	
10	Rivalität (k)		.69	
11	Rivalität (b)			
12	Rivalität (b)			.40

Ladungen <.30 sind in diesem Beispiel nicht dargestellt; im Sinne transparenterer Ergebnisdarstellungen sollten bei realen Analysen jedoch stets alle Ladungen angegeben werden. a-m = affektiv-motivationale Komponente; k = kognitive Komponente; b = behaviorale Komponente.

Im Falle des hier dargestellten, fiktiven Ergebnisses einer Faktorenanalyse (bei der 3 orthogonale Faktoren identifiziert wurden; wir gehen hier vereinfachend auf die Frage der Faktorenrotation nicht näher ein) wäre man mit den Items 1 bis 3 und 7 bis 10 zufrieden, da sie hoch auf den intendierten Faktoren und nur gering (<.30) auf anderen Faktoren laden. Item 4 zeigt eine substantielle Nebenladung, d. h. eine Ladung auf einem weiteren als dem intendierten Faktor. Es reflektiert also teilweise auch „Rivalität“. Würde man Item 4 bei der Bildung eines Gesamtwertes für „Bewunderung“ inkludieren, nähme man in Kauf, dass Testwerte für „Bewunderung“ durch „Rivalität“ verunreinigt werden. Item 11 fällt dadurch auf, dass es auf keinem der Faktoren lädt. Die Items 5, 6 und 12 bilden zusammen einen weiteren Faktor. Diesen könnte man als behavioralen Faktor bezeichnen, da die auf ihm ladenden Items allesamt die behaviorale Komponente des Narzissmus abbilden. Es wird deutlich, dass es mit einer schlichten Aussonderung dieser Items – da sie nicht auf den intendierten Rivalitäts- bzw. Bewunderungsfaktoren laden – nicht getan ist. Die verbleibenden Items würden nur unzureichend die behaviorale Komponente abbilden. Man hätte dann den Inhalt der Messung verändert: Narzissmus wäre plötzlich ein weitgehend affektiv-motivationales und kognitives Konstrukt.

Wie bereits erwähnt, lässt sich die theoretisch angenommene Struktur eines Tests auch explizit prüfen. Dazu werden konfirmatorische Faktorenanalysen herangezogen. Hierbei werden die angenommenen Faktoren, die zu diesen Faktoren gehörigen Items sowie die Interkorrelation der Faktoren vorab

Konfirmatorische Faktorenanalyse

spezifiziert. Es kann dann geprüft werden, ob die tatsächlichen Korrelationen der Testitems zu den durch die implizierte Struktur erwarteten Korrelationen passen.

Weiterführende Literatur

Diese vereinfachte Darstellung der Faktorenanalyse soll nur deren Grundgedanken illustrieren. Weitere Ausführungen finden sich bei Eid et al. (2017) und bei Moosbrugger und Schermel-leh-Engel (2012).

2.5.5 Itemvalidität

Itemvalidität = Zusammenhang eines Items mit einem Außenkriterium

Itemvalidität über alle Items hinweg vergleichen

Unter Itemvalidität versteht man den Zusammenhang eines Items mit einem Außenkriterium. Dieser Zusammenhang ist bei der externalen Testkonstruktion besonders wichtig. Geht es um die Identifikation von real existierenden Kategorien, so sollten sich Itemantworten je nach Kategorienzugehörigkeit unterscheiden. Geht es beispielsweise um die Unterscheidung von Personen mit oder ohne narzisstische Persönlichkeitsstörung (► Abschn. 2.4.2.4), so sollten sich die Itemantworten zwischen diesen Gruppen substanzial unterscheiden. Die Narzissmusgruppe wird den Items stärker zustimmen als die Vergleichsgruppe und nicht nur höhere Gesamttestwerte, sondern auch bei allen Items höhere Itemausprägungen aufweisen. Soll ein Test zur Vorhersage einer kontinuierlich ausgeprägten Gegebenheit genutzt werden, so sollten die Items mit diesem Kriterium korrelieren.

Bei der Analyse der Itemvalidität ist zu beachten, dass hierbei in der Regel wichtige Symmetrieverbedingungen (für weitere Erläuterungen s. ► Abschn. 2.6.3.4) verletzt sind. Man bedenke, dass *ein* Item mit einem *breiten* Kriterium aus der Realität korreliert wird. Es sind daher keine hohen Korrelationen zu erwarten. Wichtiger als der Blick auf die absolute Höhe der Korrelationen ist vielmehr die vergleichende Betrachtung der Itemvalidität über alle Items eines Testentwurfs hinweg.

Itemanalyse

Analysen, die Erkenntnisse auf Itemebene liefern, betreffen die

- Itemschwierigkeit bzw. -leichtigkeit,
- Itemstreuung,
- Trennschärfe,
- Ladung auf intendierten Faktoren,
- Itemvalidität.

Gute Items zeichnen sich dadurch aus, dass sie

- eine hinreichende Streuung bei der intendierten Zielgruppe „produzieren“;
- ein der Zielgruppe angemessenes Schwierigkeitsspektrum abdecken;
- eine der Homogenität bzw. Heterogenität des zu messenden Konstrukt an gemessene Trennschärfe aufweisen;
- in intendierter Weise auf Faktoren laden;
- reale Gegebenheiten abbilden, sofern dies der Messanspruch des Tests ist.

2.5.6 Itemanalysen nach Probabilistischen Testtheorien

Probabilistische Itemanalyse

Die Grundannahmen Probabilistischer Testtheorien wurden in ► Abschn. 2.3 dargestellt. Itemanalysen, die darauf aufbauen, prüfen in aller Regel, welche Items zu den in Probabilistischen Testtheorien inhärenten Annahmen passen

und welche nicht. Beispielsweise kann im Rahmen des einparametrischen dichotomen Rasch-Modells zwar eine globale Modellgeltung festgestellt werden, aber dennoch können einzelne Items Antwortmuster aufweisen, die bei Modellgeltung unwahrscheinlich sind. Solche Items lassen sich mithilfe von *Itemfitmaßen*, z. B. dem Q-Index (vgl. Rost 1999), identifizieren und können dann ggf. ausgesondert werden.

Itemfitmaß „Q-Index“

Für Itemanalysen nach Probabilistischen Testtheorien stehen verschiedene Itemfitmaße zur Verfügung. Eines dieser Fitmaße, der Q-Index, folgt der Logik, dass Items dann gut zum angenommenen Modell passen, wenn sie von den „richtigen“ Personen gelöst wurden (Rost 2004). Das heißt, schwierige Items sollten vornehmlich von Personen mit hoher Fähigkeitsausprägung gelöst worden sein. Items mittlerer Schwierigkeit sollten vornehmlich von Personen mit mittlerer und hoher Fähigkeitsausprägung gelöst worden sein, seltener von Personen mit niedriger Fähigkeitsausprägung.

Der Q-Index prüft daher, wie wahrscheinlich das vorliegende Muster der Antworten (aller Personen bei diesem Item) ist, gegeben die Randsumme des Items. Die Randsumme des Items entspricht der Zahl der Personen, die dieses Item gelöst haben. Das heißt, es wird geprüft, ob ein vorliegendes Antwortmuster bei einer gegebenen Schwierigkeit des Items wahrscheinlich ist.

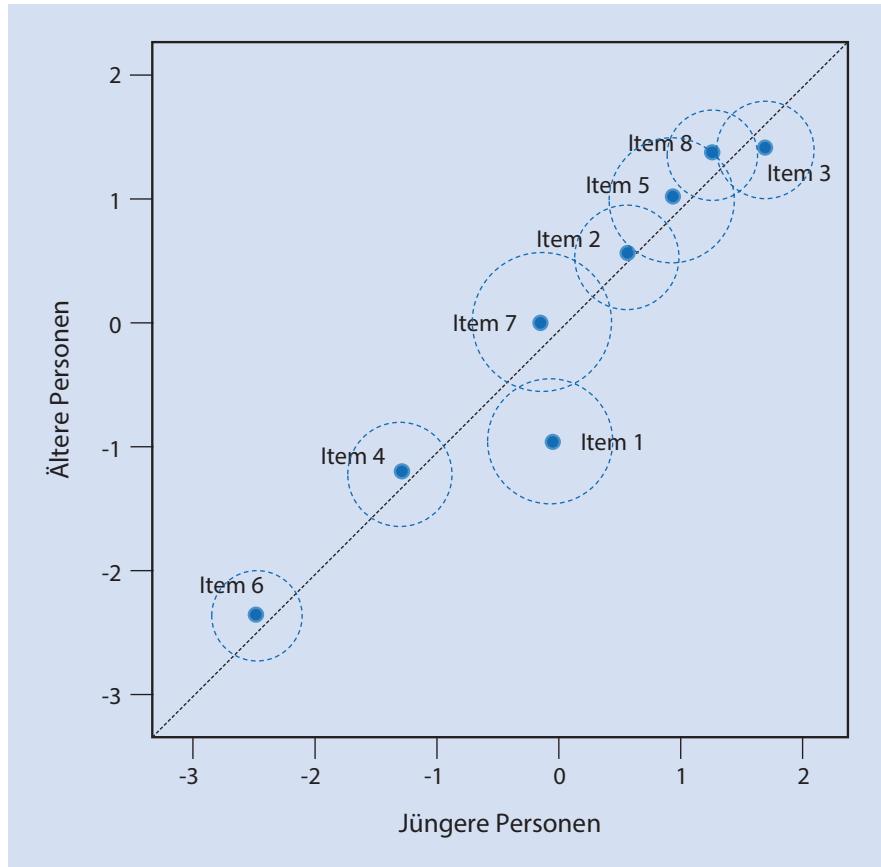
Eine z -Standardisierung dieser bedingten Wahrscheinlichkeiten ermöglicht eine einfache Interpretation des Q-Indexes. Bei z -Werten größer als 1,96 ist von einem sog. „Underfit“ des Items auszugehen, d. h., dass viele Personen das Item lösen, obwohl ihre Fähigkeit dies eigentlich nicht erwarten ließe, und/oder dass viele Personen das Item nicht lösen, obwohl dies aufgrund ihrer Fähigkeit zu erwarten wäre. Für das zuvor gewählte Beispiel der Tennisspielerinnen hieße das: Die Gegnerin verliert unerwartet oft gegen schwächere Spielerinnen und gewinnt unerwartet oft gegen bessere Spielerinnen (► Abschn. 2.3.1.1).

In ► Abschn. 2.3 haben wir ebenfalls den *grafischen Modelltest* kennengelernt. Mit diesem lässt eine Eigenschaft des dichotomen Rasch-Modells prüfen: die der spezifischen Objektivität der Vergleiche. Das bedeutet, dass Itemparameter auch in Teilstichproben gleich ausfallen sollten. Somit kann der grafische Modelltest ebenfalls genutzt werden, um Items zu identifizieren, für die diese Annahme nicht haltbar ist. Solche Items fallen dadurch auf, dass ihre Itemparameter deutlich von der Diagonalen abweichen, also in 2 Substichproben ungleich sind. Ab wann man von einer „deutlichen“ Abweichung spricht oder nicht, lässt sich anhand von Konfidenzintervallen um die geschätzten Itemparameter objektivieren (Kreise in □ Abb. 2.32). Berührt das Konfidenzintervall die Diagonale nicht mehr, sollte über einen Ausschluss des betreffenden Items nachgedacht werden. In □ Abb. 2.32 beträfe dies Item 1. Neben der optischen Inspektion kann der Wald-Test (s. Eid et al. 2017) zur Itemselektion genutzt werden, mit dem ebenfalls die Annahme gleicher Itemparameter in 2 Substichproben geprüft wird.

Eine weitere Möglichkeit der Identifikation von Items, für die Annahmen des gewählten probabilistischen Modells nicht haltbar sind, besteht darin, die Modellannahmen zu lockern. So könnte man statt eines 1PL-Modells ein 2PL-Modell (► Abschn. 2.3.1.2) testen und die entsprechenden Itemparameter schätzen. Dadurch lässt sich prüfen, welche Items sich in dem hinzugekommenen Itemparameter (genauer dem Diskriminationsparameter; dieser kommt im 2PL-Modell hinzu) von den restlichen Items unterscheiden und kann diese – falls theoretisch sinnvoll – eliminieren. Nach Elimination dieser Items kann geprüft werden, ob nun das Modell mit den strengerem

Grafischer Modelltest

Erweiterung um weitere Parameter



■ Abb. 2.32 Grafischer Modelltest mit Konfidenzintervallen um Itemparameter. (Vgl. Eid und Schmidt 2014)

Annahmen passt. Dies kann durch Berücksichtigung des Rateparameters (s. 3PL-Modell, ▶ Abschn. 2.3.1.2) realisiert werden: Items, deren Rateparameter eine kritische Höhe übersteigt oder sich deutlich von den übrigen Items unterscheidet, können ggf. aussortiert werden.

Wie das dichotome macht auch das ordinale Rasch-Modell (▶ Abschn. 2.3.2) prüfbare Annahmen über das Zustandekommen von Itemantworten. Darüber hinaus erlaubt es, die Ordnung und Breite der Kategorien zu überprüfen. Bezuglich der Ordnung der Kategorien wird angenommen, dass mit zunehmender Merkmalsausprägung die jeweils nächsthöhere Antwortkategorie wahrscheinlicher wird. Stellen wir uns zur Veranschaulichung vor, ein Item habe 4 Antwortmöglichkeiten: „starke Ablehnung“, „Ablehnung“, „Zustimmung“ und „starke Zustimmung“. Nehmen wir zudem an, die Kategorieneinwahrscheinlichkeiten verliefen so, wie in ■ Abb. 2.33 dargestellt.

Mit zunehmender Merkmalsausprägung wird nach der untersten Antwortkategorie (z. B. „starke Ablehnung“) direkt die übernächste Antwortkategorie (z. B. „Zustimmung“) am wahrscheinlichsten. Es gibt keinen Punkt auf dem Merkmalskontinuum, an dem die 2. Antwortkategorie (z. B. „Ablehnung“) am wahrscheinlichsten ist. Ein solches Item sollte ausgeschlossen werden.

Schwellen müssen sowohl innerhalb als auch zwischen Items nicht notwendigerweise gleiche Abstände haben. Testautorinnen und -autoren können dies jedoch mithilfe des ordinalen Rasch-Modells überprüfen und, falls gewünscht, Items so selektieren, dass die Antwortkategorien über die Items hinweg jeweils gleiche Abstände haben (sog. „Ratingsskalenmodell“).

Ordinales Rasch-Modell

Ausschluss von Items mit nicht erwartungskonformen Funktionsverläufen

Items mit gewünschten Schwellenabständen auswählen

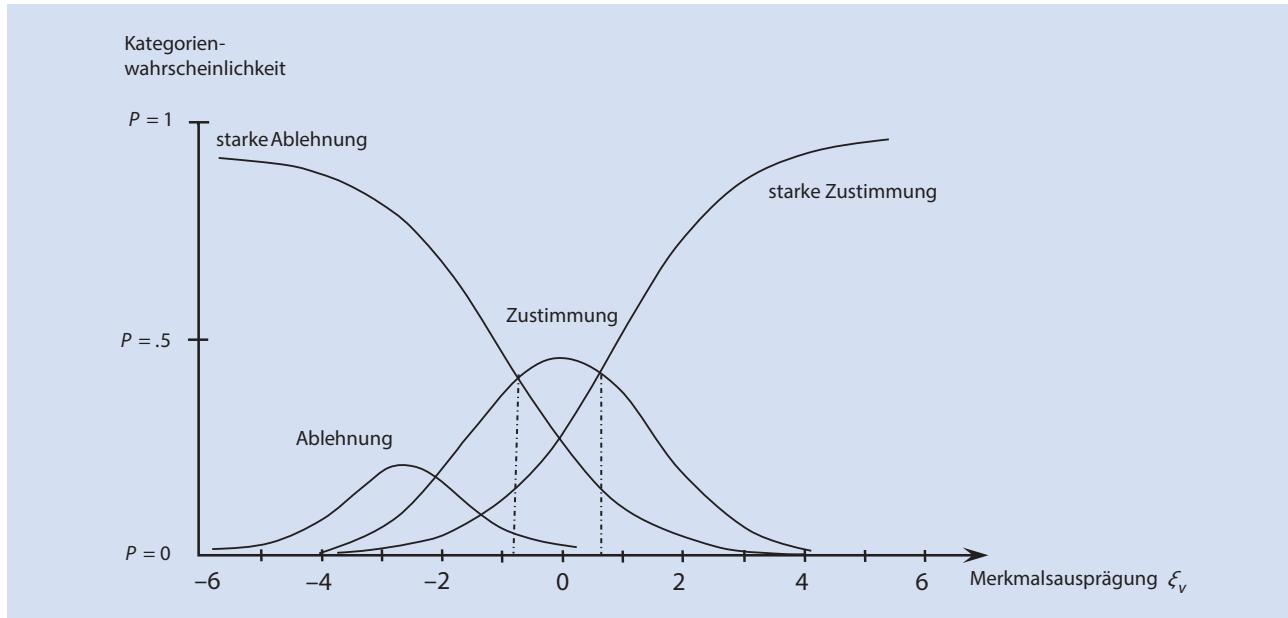
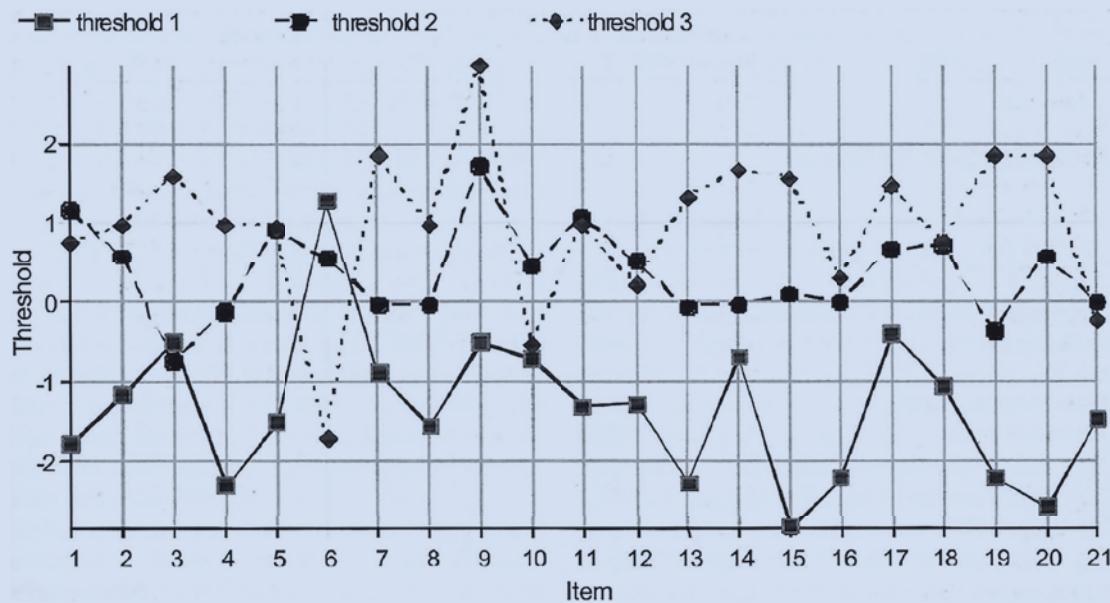


Abb. 2.33 Kategorienwahrscheinlichkeiten eines Items mit ungeordneten Schwellen. (Angelehnt an Rost 2004, S. 205, © Hogrefe)

Schwellenparameter im Beck Depressions-Inventar, Revision (BDI-II)

Das BDI-II (Hautzinger et al. 2006) ist ein Test zur Messung der Depressivität. Er verwendet ein etwas unorthodoxes Antwortformat: Jede Abstufung der Antwortskala ist dem jeweiligen Iteminhalt entsprechend ausformuliert (z. B. für Traurigkeit: 0=„Ich bin nicht traurig“, 1=„Ich bin oft traurig“, 2=„Ich bin ständig traurig“ usw.). Allerdings zeigt sich durch Betrachtung der Schwellenparameter, dass für manche Items bei höherer Depressivitätsausprägung niedrigere Kategorien gewählt werden. So ist in Item 6 eine höhere Depressivitätsausprägung nötig, um die Schwelle zwischen Kategorie 1 und 2 zu „überspringen“ als für das „Überwinden“ der Schwelle zwischen Kategorie 3 und 4. Eine Analyse des Verlaufs der Schwellenparameter über alle Items eines Tests hinweg kann helfen, solche Items zu identifizieren und auszusortieren.



Verlauf der Schwellenparameter des BDI-II. (Aus Hautzinger et al. 2006, S. 23, © Pearson). Threshold bezeichnet hier den Schwellenparameter – meint also den Punkt, an dem sich die Kategorienwahrscheinlichkeiten schneiden (in Abb. 2.33 sind die Schwellen durch gestrichelte Linien dargestellt).

Itemselektion bei der Testentwicklung

Weiterführende Literatur

Ausführlichere Informationen zur Itemselektion nach Probabilistischen Testtheorien finden sich bei Rost (2004) sowie bei Bühner (2021).

Kurzversionen von Fragebögen

Es ist an dieser Stelle wichtig, zu betonen, dass bei einer Itemselektion im Zuge der Testentwicklung immer mehrere Merkmale eines Items gleichzeitig beachtet werden müssen. Trägt das Item dazu bei, den theoretisch relevanten Merkmalsbereich abzudecken? Trägt das Item zu einer Differenzierung im oberen (oder unteren) Bereich der Merkmalsausprägung bei? Leidet die Reliabilität der Skala, wenn das Item entfernt wird (► Abschn. 2.6.2)? Trägt das Item dazu bei, dass der Test höher mit intendierten Kriterien korreliert (► Abschn. 2.6.3)? Diese und weitere Fragen müssen Testentwickelnde berücksichtigen, wenn sie Items aus einem Testentwurf entfernen. Eine reine Maximierung eines Itemkennwertes ist nicht ratsam. Beispielsweise würde die unkritische Maximierung von Cronbachs Alpha (► Abschn. 2.6.2.1) dazu führen, dass vorwiegend hoch miteinander korrelierende Items im Testentwurf verblieben. Diese wären wahrscheinlich alle sehr ähnlich formuliert und würden den gleichen spezifischen Aspekt des intendierten Messanspruchs abdecken – aber nicht in angemessener Breite. Daher basieren Entscheidungen bei der Itemselektion am besten auf einer Kombination verschiedener Kennwerte und werden mit Bedacht vorgenommen.

Gleichwohl ist es manchmal nötig, Items zu selektieren, auch wenn diese nicht eindeutig negativ auffallen. Das ist dann der Fall, wenn zu entwickelnde Instrumente besonders kurz, also zeitsparend auszufüllen sein sollen. Mittlerweile liegen für viele Langformen von Fragebögen auch Kurzversionen vor. Eine sehr umfangreiche Sammlung verfügbarer Kurzversionen von Fragebögen findet sich bei Kemper et al. (2014). Das GESIS – Leibniz-Institut für Sozialwissenschaften in Mannheim hat ebenfalls einige Kurzversionen von Fragebögen zur Messung gängiger psychologischer Merkmale entwickelt, dokumentiert und zum Download zur Verfügung gestellt. Dies ist unter ► <https://www.gesis.org/kurzskalen-psychologischer-merkmale/home/> zu finden.

Dabei ist zu beachten, dass kürzere Testversionen häufig mit zumindest geringen Einbußen der Reliabilität verbunden sind und zu diesen selten eigene Validierungen, separat von der Validierung der Langversion, vorliegen (Kruyken et al. 2013). Je nach Einsatzweck ist eine etwas geringere Reliabilität zu verschmerzen – beispielsweise wenn ein Test für Screeningzwecke (d. h. für eine initiale „grobe“ Diagnostik, der weitere diagnostische Erhebungen nachfolgen) oder nicht in der Individualdiagnostik, sondern in der Forschung eingesetzt wird. Oder man greift doch besser auf die Langversion zurück. Eine Diskussion zum angemessenen Umgang mit und zum Einsatz von Kurzskalen findet sich bei Ziegler et al. (2014).

Optimierung der Itemselektion: Ant Colony Optimization

Eine spannende Methode zur Itemselektion wurde durch die Biologie inspiriert. Dort hat man festgestellt, dass Ameisen auf ihrer Suche nach Futter Pheromone hinterlassen und somit anderen Ameisen signalisieren, auf welchem Weg sie zu Futter gelangt sind. Während die ersten Ameisen zufällig auf ihrem Weg hin und her laufen, folgen später eintreffende Ameisen mehr und mehr der deutlichsten Pheromonspur. Das Entscheidende: Der kürzeste Weg zum Futter wird mit der Zeit die deutlichste Pheromonspur aufweisen, da die vergleichsweise kürzere Strecke in der gleichen Zeit von mehr Ameisen zurückgelegt wurde (für eine ausführliche Darstellung s. Schultze 2017). Nach einem ähnlichen Prinzip kann man bei der Itemselektion vorgehen. Man lässt ein Programm zunächst zufällig Items selektieren und kurze Testversionen erstellen. Dabei werden die Items, die zu psychometrisch guten Kurzformen führen, mit „Pheromonen“ versehen. Diese Items werden dann in den nächsten Runden der Kurzformgenerierung bevorzugt eingesetzt und (manche davon) wiederum mit „Pheromonen“ versehen (nach Olaru et al. 2015). Ein großer Vorteil dieser Methode ist, dass mit ihr Items ausgewählt werden können, die mehrere Kennwerte (z. B. Reliabilität und Korrelation mit Kriterien) eines Fragebogens berücksichtigen, die gleichzeitig optimiert werden. Olaru et al. (2015) konnten zeigen, dass diese Form der Kurzformgenerierung anderen Methoden überlegen ist.

Itemselektion am Vorbild der Natur



Tests „bauen“ nach dem Vorbild der Ameisen. (© Andrey Burmakin/stock.adobe.com)

2.6 Testgütekriterien

Analyse auf Testebene

Sind alle Analysen auf Itemebene abgeschlossen (► Abschn. 2.5), so lohnt sich die Zusammenstellung eines ersten Testentwurfs. Nach erneuter Datenerhebung kann dann die Güte des Testentwurfs geprüft werden – also eine Analyse auf Test- und nicht auf Itemebene erfolgen.

Gütekriterien zur Auswahl von Tests

Vorab sei bereits betont: Gütekriterien von Tests sind nicht nur im Rahmen der Testkonstruktion bedeutsam. Dort geht es für Testentwicklerinnen und -entwickler darum, eine Testversion zu erstellen, die gängigen Gütekriterien entspricht, und deren Ermittlung zu dokumentieren. Aber auch für Anwenderinnen und Anwender von Tests sind die daraus resultierenden Angaben in Testmanualen für die Auswahl und richtige Handhabung der jeweiligen Tests wichtig.

Haupt- und Nebengütekriterien

Eine gängige Gliederung der Gütekriterien besteht in der Unterscheidung von sog. „Haupt- und Nebengütekriterien“. Während die Gruppe der Hauptgütekriterien allgemein akzeptierte „Kandidaten“ enthält, variieren die Nebengütekriterien je nach Lehrbuch und Priorität der Testautorinnen und -autoren.

Haupt- und Nebengütekriterien

- Hauptgütekriterien:
 - Objektivität
 - Reliabilität
 - Validität
- Nebengütekriterien:
 - Normierung
 - Skalierung
 - Ökonomie
 - Nützlichkeit
 - Unverfälschbarkeit
 - Zumutbarkeit/Akzeptanz
 - Fairness

Nachfolgend werden wir die Hauptgütekriterien sowie das Nebengütekriterium der Normierung etwas ausführlicher behandeln.

Beurteilung der Testgüte

Die Güte eines Tests zu bewerten kann Testanwenderinnen und -anwendern schwerfallen. Sie finden jedoch für einige der gängigen Testverfahren Rezensionen, die nach einem vom Diagnostik- und Testkuratorium entwickelten, standardisierten Prinzip erstellt wurden. Die Rezensionen sowie eine genaue Erläuterung des Vorgehens bei der Rezension sind unter dem folgenden Link zu finden: ► <https://www.psypindex.de/tests/testkuratorium/>. Ein besonderer Vorteil dieser Rezensionen ist die summarische Gesamtbeurteilung eines Tests anhand eines einfachen Beurteilungsschemas. Die Beurteilungen stellen ein Konsensurteil von in der Regel 2 Rezessenten bzw. Rezessentinnen dar. Es bewertet Tests hinsichtlich ihrer Hauptgütekriterien und der Verfügbarkeit und Stringenz der Testdokumentation. Eine solche Beurteilung des Adaptiven Intelligenz Diagnostikum 3 (AID 3; Kubinger und Holocher-Ertl 2014) durch Ziegler und Reichert (2017, S. 239, mit freundlicher Genehmigung des Hogrefe Verlages) ist in folgender Tabelle dargestellt:

AID 3	Die TBS-TK-Anforderungen sind erfüllt			
	voll	weitgehend	teilweise	nicht
Allgemeine Informationen und diagnostische Zielsetzung	X			
Objektivität	X			
Zuverlässigkeit	X			
Validität			X	

Das Diagnostik- und Testkuratorium hat für Autorinnen und Autoren solcher Rezensionen klar spezifiziert, anhand welcher Kriterien die Qualität der Testkonstruktion zu prüfen ist. Darauf basierende Checklisten können Testanwenderinnen und -anwender nutzen, um die Qualität eines Tests zu beurteilen. Wir zeigen in den nachfolgenden Abschnitten Auszüge dieser Checklisten, sofern sie sich auf die jeweiligen Schritte der Testkonstruktion bzw. auf die jeweiligen Testgütekriterien beziehen. Eine umfangreiche Checkliste zur Beurteilung der Qualität von Testverfahren, die die Grundlage der Kriterien des Diagnostik- und Testkuratoriums bildet, steht in Form des *DIN Screen* (Kersting 2006) zur Verfügung (s. auch Diagnostik- und Testkuratorium 2018a).

Testrezensionen des Diagnostik- und Testkuratoriums

Checklisten zur Beurteilung der Qualität von Tests

2.6.1 Objektivität

Definition

Objektivität bedeutet, dass die Ergebnisse eines diagnostischen Verfahrens unabhängig davon zustande kommen, wer die Untersuchung, die Auswertung und die Interpretation durchführt.

In der Definition sind 3 Phasen der Testanwendung angegeben, in denen die Objektivität leiden kann: bei der Durchführung, der Auswertung und der Interpretation. Man unterscheidet dementsprechend 3 „Unterformen“ der Objektivität. Von Ausnahmefällen abgesehen wird die Objektivität nicht numerisch bestimmt. Stattdessen werden Maßnahmen zur Standardisierung von Durchführung, Auswertung und Interpretation genannt, die die Objektivität

3 Phasen der Testanwendung, in denen die Objektivität leiden kann

Maximale Standardisierung der Durchführung

gewährleisten sollen. Diese Maßnahmen sind quasi Bestandteil des diagnostischen Verfahrens und müssen im Manual zum Verfahren dokumentiert sein. Eine Aussage über die Objektivität beruht zumeist auf der Bewertung der genannten Maßnahmen. Nachfolgend betrachten wir die 3 Formen der Objektivität etwas näher.

■ Durchführungsobjektivität

Diese ist dann gegeben, wenn ein Verfahren immer auf die gleiche Weise durchgeführt wird. Dazu ist es zunächst erforderlich, dass das Testmaterial (Testhefte, Antwortbögen etc.) mitgeliefert wird. Wenn weitere Materialien wie eine Stoppuhr, Bleistifte, Kugelschreiber oder Notizzettel benötigt werden, sind diese genau zu benennen. Die Durchführungsbedingungen müssen genau spezifiziert werden. Ist eine Einzeluntersuchung vorgeschrieben, oder können mehrere (wie viele?) Personen gleichzeitig untersucht werden? Wie muss der Arbeitsplatz beschaffen sein? Beispielsweise können gute Lichtverhältnisse oder eine feste Schreibunterlage gefordert werden. Für die Untersuchungsleiterin bzw. den Untersuchungsleiter sind Anweisungen nötig, damit sie bzw. er sich möglichst immer auf die gleiche Weise verhält. Wichtige Bestandteile dieser Anweisungen sind eine Instruktion, die entweder wörtlich oder sinngemäß vorzutragen oder von den Testpersonen selbst zu lesen ist, eindeutige Zeitvorgaben sowie Regeln für den Umgang mit Fragen oder Störungen. Ziel ist insgesamt eine maximale Standardisierung der Durchführung.

Unzulängliche Standardisierung

Bei dem Test „Familie in Tieren“ (Brem-Gräser 2001) soll das Kind seine Familie als Tiere zeichnen. Diese Zeichnung soll über „die Qualität der Beziehung des Kindes zu seinen Eltern im Hinblick auf Geborgenheits-, Macht- und Kontakterleben“ und „über die teils unbewusste Struktur und Dynamik innerhalb der Familie, so wie sie das Kind erlebt“ Aufschluss geben (Baumgärtel und Thomas-Langel 2015, S. 453). In der Testanweisung wird nicht spezifiziert, wie diese Anweisung genau vorzutragen ist, welches Papier (Größe, Qualität) und welche Stifte (Bleistift, Buntstifte, bunte Faserstifte ...?) zu verwenden sind. Daher fassen Baumgärtel und Thomas-Langel (2015, S. 453) zusammen:

» „Andere Angaben, die für eine objektive Durchführung notwendig wären, wie Härtegrad des Zeichenstiftes sowie Qualität und Format des Papiers, fehlen. Das ist angesichts der geforderten detaillierten Analyse der grafischen Merkmale der Zeichnung, wie Strichbreite, Strichführung, Schattierung und Schwärzung, ein bemerkenswerter Mangel. Die für Auswertung und Interpretation angegebenen Vorgehensweisen der Deutung und intuitiven Einfühlung können zu subjektiver Gewissheit des Diagnostikers, nicht aber zu objektiv begründbaren diagnostischen Aussagen führen.“

Vollständige Standardisierung ist nicht möglich

Eine völlige Standardisierung ist jedoch nicht möglich und manchmal sogar nicht einmal wünschenswert. Geschlecht, Alter, Aussehen, Kleidung etc. des Untersuchungsleiters bzw. der Untersuchungsleiterin variieren naturgemäß. Bei Testungen von Kindern wird manchmal bewusst auf eine wörtlich vorzutragende Instruktion verzichtet. Stattdessen wird angegeben, was sinngemäß gesagt werden soll. Damit versucht man, zu gewährleisten, dass die Anweisungen alters- und kindgerecht vorgetragen werden. Die angestrebte Standardisierung besteht darin, dass alle Testpersonen den Auftrag gleich verstehen,

und nicht darin, dass der Auftrag mit den gleichen Worten vorgetragen wird. Dass die Durchführungsobjektivität damit eventuell leicht eingeschränkt wird, muss hingenommen werden. Bei diagnostischen Interviews ist es ohnehin nur schwer möglich, das Verhalten der Interviewerin bzw. des Interviewers völlig zu vereinheitlichen. Ihr Verhalten wird nicht nur von habituellen Merkmalen wie Akzent, Aussprache oder Körpersprache mit beeinflusst, sondern auch vom Verhalten der interviewten Personen. Diese fragen nach, schweigen, schweifen vielleicht vom Thema ab und zwingen damit Interviewerinnen und Interviewer, vom Leitfaden abzuweichen.

- !** Die Durchführungsobjektivität darf als hoch oder „gegeben“ angesehen werden, wenn alle Bedingungen festgelegt sind, die sich erfahrungsgemäß auf das Testverhalten auswirken können.

■ Auswertungsobjektivität

Sie gibt das Ausmaß an, in dem Antworten der Testperson unabhängig von der Person, die den Test auswertet, zu den gleichen Ergebnissen führen. Gleicher Verhalten der Testpersonen wird in einem objektiv auswertbaren Test stets nach exakt denselben Regeln abgebildet.

In Tests liegt das Antwortverhalten zumeist in Form von Kreuzchen vor, die die Testperson beispielsweise bei „Ja“ oder „Nein“ oder – bei Multiple-Choice-Items – bei den Antworten a, b, c, d gesetzt haben. Die Auswertung besteht darin, eine Antwort als richtig oder falsch bzw. bei Fragebögen als Grad der Merkmalsausprägung zu klassifizieren. Bei mehrstufigen Antwortskalen muss der angekreuzten Stufe eine Zahl zugewiesen werden. Dazu dienen in der Regel Schablonen, die auf das Testformular oder einen Antwortbogen aufgelegt werden. Die „richtigen“ Antworten werden dann gezählt. Bei mehrstufigen Antwortskalen sind die ermittelten Zahlen zu addieren. Auf diese Weise erhält man einen Rohwert.

Bei einigen Leistungstests, etwa dem Wechsler-Intelligenztest für Erwachsene (WAIS-IV; Petermann 2014) wird auch eine freie Beantwortung von Fragen verlangt. Testleiterinnen und Testleiter müssen die Antworten unter Umständen sogar sofort bewerten, weil der Test nach einer bestimmten Anzahl von Falsch-Antworten abgebrochen wird. Das Manual muss genaue Angaben enthalten, wann eine Antwort als richtig oder falsch zu bewerten ist. Beispiele für richtige und falsche Antworten sind dabei hilfreich.

Für die Auswertungsobjektivität ist entscheidend, dass das Vorgehen im Manual mit klaren und unmissverständlichen Anweisungen beschrieben wird. Dazu gehören auch Anweisungen, wie mit Auslassungen, Korrekturen und Doppelankreuzungen („richtig“ und „falsch“ angekreuzt) zu verfahren ist. Manchmal liegen mehrere Schablonen vor. Dies ist der Fall bei mehrseitigen Antwortbögen und bei mehrdimensionalen Fragebögen, die pro Skala eine Schablone benötigen. Um Verwechslungen auszuschließen, müssen die Schablonen gut sichtbar gekennzeichnet sein. Anstelle von Schablonen finden manchmal Auswertungsprogramme Verwendung, die von den Testverlagen angeboten werden. Das Verrechnen der Antworten entfällt damit; die Antworten müssen dennoch abgelesen und in das Programm eingegeben werden. Testverlage bieten für einige Tests auch eine maschinelle Auswertung an. In idealer Weise wird die Auswertungsobjektivität bei computerbasierten Tests gewährleistet.

- !** Bei klaren Anweisungen zum Vorgehen bei der Auswertung und wenn angemessene Hilfsmittel zur Verfügung gestellt werden, kann die Auswertungsobjektivität als gegeben gelten.

Tests stets nach exakt denselben Regeln auswerten

Nutzung von Schablonen

Freie Antwortformate erfordern detaillierte Auswertungsregeln

Maschinelle Auswertung ideal

Übereinstimmung der Auswerterinnen bzw. Auswerter als Maß der Auswertungsobjektivität

Klären, welche Schlussfolgerungen zulässig sind

Verständliche Interpretationshinweise im Manual erforderlich

Normtabellen ermöglichen Einordnung der Testwerte

Die Auswertungsobjektivität kann auch empirisch ermittelt und quantitativ bestimmt werden. Dazu wird eine größere Anzahl von Testprotokollen von mindestens 2 Personen ausgewertet. Die Übereinstimmung der Auswerterinnen bzw. Auswerter wird als Intraklassenkorrelation berechnet (vgl. Yoder und Symons 2010). Solche Berechnungen sind vor allem für Auswertungen von Interviews und Verhaltensbeobachtungen (z. B. im Rahmen von Assessment-Centern) üblich und sinnvoll.

■ Interpretationsobjektivität

Für die Sicherstellung der Interpretationsobjektivität sollte in Tests klar benannt werden, welche Aussagen zulässig und welche unzulässig sind. Sind anhand eines Tests zum Entwicklungsstand eines Kindes auch Aussagen über dessen Intelligenz zulässig? Solche Forderungen sind keineswegs trivial. Der Test d2-R (Brickenkamp et al. 2010), ein weitverbreiteter Test, wird auf dem Deckblatt des Manuals als „Aufmerksamkeits- und Konzentrationstest“ beschrieben. Ist das Ergebnis nun als Aussage über die Aufmerksamkeit einer Person zu interpretieren? Oder als Aussage über deren Konzentration? Oder über beides? Die Vorgängerversion trug noch den Namen „Aufmerksamkeits-Belastungs-Test“, und es kam tatsächlich vor, dass in einem studentischen Gutachten zu lesen war, die Testperson könne sich gut konzentrieren, habe eine hohe Aufmerksamkeit und sei zudem sehr belastbar.

Antworten auf die Frage, was mit dem Test gemessen wird, sollten nicht den Testanwenderinnen und -anwendern überlassen werden, da dadurch die Interpretationsobjektivität verletzt würde. Im Kapitel „Interpretation“ des Manuals zum Test d2-R wird deshalb genau erklärt, welche Merkmale mit den Kennwerten des Tests erfasst werden. In jedem Testmanual sollten exakte Hinweise vorliegen, wie die erfassten Merkmale zu benennen sind. Weder der Name des Tests noch die Ausführungen zum theoretischen Hintergrund oder zur Validität sind diesbezüglich eindeutig. Testautorinnen und -autoren sollten im Manual in einem Kapitel „Anwendung“ und dort in einem Unterkapitel „Interpretation“ gut nachvollziehbare und für Adressatinnen und Adressaten von Gutachten oder Ergebnisberichten verständliche Formulierungen vorschlagen.

Die Auswertung eines Tests liefert Rohwerte. Ein Proband hat beispielsweise 210 Aufgaben richtig gelöst. „Interpretieren“ bedeutet, diesem Wert eine Bedeutung zu geben. Interpretationsobjektivität ist dann gegeben, wenn alle Testanwenderinnen und -anwender diesen Rohwert in die gleiche Aussage über die Testperson transformieren. Dazu sollten Testautorinnen und -autoren im Manual klare Angaben machen. Zumeist dienen Normtabellen (► Abschn. 2.6.4) der Einordnung und Interpretation des Testergebnisses. Die Normtabellen zeigen die Position der Testperson – je nach gewählter Normtabelle – im Vergleich z. B. zu Gleichaltrigen, anderen Männern/Frauen oder Personen mit einem bestimmten Schulabschluss (z. B. Abitur). Da sich Normwerte nicht von selbst erklären, ist eine Übersetzungshilfe vorteilhaft. So sollten Testautorinnen und -autoren Angaben machen, welche Werte als „durchschnittlich“ oder „überdurchschnittlich“ etc. zu bezeichnen sind. Alternativ ist es auch möglich, in Testmanualen zu erläutern, welche inhaltlichen Schlüsse aus niedrigen, mittleren oder hohen Testwerten zu ziehen sind.

! Interpretationsobjektivität kann als gegeben angesehen werden, wenn klare Aussagen über zulässige Interpretationen sowie Hilfsmittel für die Einordnung von Ergebnissen (z. B. in Form von Normtabellen) vorliegen.

Empfehlungen des Diagnostik- und Testkuratoriums

Das Diagnostik- und Testkuratorium empfiehlt zur Beurteilung der Objektivität folgende Prüfungen (zitiert nach Diagnostik- und Testkuratorium 2018b, S. 113 f., © Hogrefe; Gender-Formulierungen durch Autoren dieses Lehrbuchs angepasst):

- Ist der Test so weit wie möglich standardisiert?
- Sind die Instruktionen für die Testleiterinnen und Testleiter möglichst wörtlich vorgeschrieben bzw. ist klar, was die Testleiterinnen und Testleiter sagen sollen und was nicht?
- Ist genau angeben, welche Handlungen die Testleiterinnen und Testleiter konkret zu verrichten haben (z. B. das Testmaterial in einer bestimmten Art ordnen)?
- Ist genau ausgeführt, wie auf Fragen der Teilnehmerinnen und Teilnehmer eingegangen werden muss?
- Enthalten die Instruktionen für die getesteten Personen Beispiel- und Übungselemente sowie Informationen über die Art, wie die Reaktionen (Antworten) zu geben sind?
- Falls Auswertungsschablonen gebraucht werden: Ist genau angegeben, wie diese auf die Antwortformulare zu legen sind?
- Falls Auswertungsschablonen benutzt werden: Ist auf den Schablonen angegeben, zu welcher Version des Tests sie gehören?
- Ist angegeben, welcher Testwert für ein nicht bearbeitetes Item gegeben werden soll bzw. wie mit nicht bearbeiteten Items umzugehen ist?
- Ist angegeben, bis zu welcher Anzahl von nicht bearbeiteten Items das Testergebnis noch interpretiert werden darf?
- Falls der Test den Einsatz mehrerer Beurteilerinnen und Beurteiler bzw. Beobachterinnen und Beobachter erfordert: Ist angegeben, wie mit unterschiedlichen Urteilen/Beobachtungen umzugehen ist?
- Bei Tests, die am Computer durchgeführt und ausgewertet werden: Können Anwenderinnen und Anwender die Auswertung vom Prinzip her nachvollziehen?
- Wurden einzelne Fallbeschreibungen in die Verfahrenshinweise (das Testmanual) aufgenommen?
- Wurden, sofern unterschiedliche Normgruppen für die Interpretation angeboten werden, Hinweise gegeben, wie die Entscheidung, welche Normgruppe in welchem Fall heranzuziehen ist, zu treffen ist?
- Wird bei der beispielhaften Interpretation von Testergebnissen darauf eingegangen, welchen Einfluss bestimmte Hintergrundvariablen und die (Test-)Erfahrung auf die Testwerte haben können bzw. wie mit möglichen Messfehlern umzugehen ist (z. B. Konfidenzintervalle oder kritische Differenzen)?
- Wird das Ausmaß an Sachkunde angegeben, das nötig ist, um den Test zu interpretieren?

2.6.2 Reliabilität

Reliabilität = Anteil der Varianz der wahren Werte an der Varianz der beobachteten Werte

Vier gebräuchliche Methoden der Reliabilitätsschätzung

Reliabilitätsschätzung anhand von 2 identischen Stichproben

Personen zu einem späteren Zeitpunkt erneut untersuchen

Die Reliabilität einer Messung wurde als zentrales Konzept der Klassischen Testtheorie bereits in ▶ Abschn. 2.2.2 definiert als der Anteil der Varianz der wahren Werte an der Varianz der beobachteten Werte. Sie bezeichnet also die Genauigkeit einer Messung, und zwar losgelöst davon – in Abgrenzung zur Validität –, ob mit der Messung das intendierte Merkmal erfasst wird. Nachdem das Konzept der Reliabilität bereits eingeführt wurde, soll an dieser Stelle auf Schätzmethoden und die Relevanz der Reliabilität bei verschiedenen diagnostischen Zielen eingegangen werden.

2.6.2.1 Reliabilitätsschätzung

Wir haben Reliabilität als die Korrelation eines Tests „mit sich selbst“ beschrieben (▶ Abschn. 2.2.2). Somit stellt sich als Nächstes die Frage, wie eine Untersuchung geplant werden soll, sodass man die Korrelation eines Tests mit sich selbst ermitteln kann. Dazu sind 4 Methoden gebräuchlich.

Methoden der Reliabilitätsschätzung

1. Retest-Methode: Der gleiche Test wird 2 × dargeboten.
2. Paralleltest-Methode: Der Test und eine parallele Version desselben werden verwendet.
3. Split-Half-Methode bzw. Testhalbierungsmethode: Ein Test wird in 2 Teile „zerlegt“; es wird die Korrelation der beiden Testteile geprüft.
4. Interne Konsistenz: Jedes Item wird als Testteil betrachtet.

■ Schätzung über die Retest-Methode

Hierbei wird der gleiche Test derselben Stichprobe 2 × dargeboten. Die Korrelation zwischen den Ergebnissen der beiden Messungen wird als Reliabilitätschätzung interpretiert. Bei Testung und Retestung muss es sich um mindestens essenziell parallele Messungen handeln.

Die Retest-Methode setzt voraus, dass man die gleichen Personen zu einem späteren Zeitpunkt erneut untersuchen kann und die dann durchgeführte Messung als mindestens essenziell parallel (s. nachfolgender Abschnitt zur Äquivalenz von Messungen) zur 1. Messung gelten kann. Eine damit verbundene Schwierigkeit der Retest-Methode besteht darin, das Intervall zwischen den beiden Testungen sinnvoll zu wählen. Einerseits gilt es, Erinnerungs- und Übungseffekte zu vermeiden. So kann die Reliabilitätsschätzung dadurch zu hoch ausfallen, dass Testpersonen bei der 2. Messung absichtlich ähnlich antworten wie zum Zeitpunkt der 1. Messung. Testpersonen denken vielleicht, ein Fragebogen würde 2 × durchgeführt, um zu kontrollieren, ob sie den Bogen zuvor zuverlässig bearbeitet haben. In solchen Fällen sind eher lange Zeitabstände in der Größenordnung von mehreren Wochen oder gar Monaten sinnvoll. Andererseits soll sich das Merkmal, das der Test zu messen beansprucht, zwischen den beiden Messungen nicht verändern. Dieses Argument spricht für eine baldige Testwiederholung. Das Dilemma kann nur durch pragmatische Überlegungen gelöst werden: Wenn die Forschung

gezeigt hat, dass ein Merkmal sehr stabil ist (Beispiel: Intelligenz), sind lange Retest-Intervalle (z. B. 1 Jahr) anzustreben. Bei stark variierenden Merkmalen wie Emotionen oder Stimmungen kann sich die Ausprägung bereits nach wenigen Minuten deutlich verändert haben. Wie stark die 2. Messung durch Erinnerungs- und Übungseffekte, aber auch durch Ermüdung, Veränderung der Motivation zur ernsthaften Bearbeitung des Tests und andere Faktoren belastet wird, hängt stark vom jeweiligen Test und auch von den Probandinnen und Probanden ab. Bei einem langen Test mit vielen Aufgaben werden Erinnerungseffekte nach einem kurzen Zeitintervall eher wenig stören. Ermüdung und Mitarbeitbereitschaft können dagegen ein ernsthaftes Problem darstellen. Kinder und ältere Leute werden eher unter einer Wiederholung nach nur kurzer Pause leiden als junge Erwachsene.

! Vorsicht

Die Retest-Reliabilität wird nicht durch Merkmalsveränderungen beeinflusst, die alle Personen gleichermaßen betreffen. Mittelwertunterschiede zwischen der 1. und 2. Messung (die bei essenziell parallelen Messungen vorkommen können) haben keine Auswirkung auf die Höhe der Korrelation (sofern dadurch keine Decken- oder Bodeneffekte auftreten). Wenn beispielsweise alle Testpersonen bei der 2. Testdurchführung einheitlich 10 Punkte mehr erreichen, weil sie sich an einige Lösungen erinnern konnten, wird die Reliabilität dadurch nicht gemindert. Die Retest-Reliabilität verringert sich nur, wenn die Effekte interindividuell unterschiedlich groß ausfallen, da z. B. einige Personen große Erinnerungseffekte zeigen und andere kleine.

Merkmalsveränderung beachten!

Bei der Interpretation von Retest-Reliabilitäts-Koeffizienten ist das Zeitintervall zwischen beiden Messungen zu beachten. Große Zeittabstände führen tendenziell zu niedrigeren Werten. Dabei ist die Stabilität des Merkmals relevant. Je stärker das Merkmal über die Zeit variiert, desto stärker vermindert sich die Retest-Reliabilität durch lange Zeitintervalle. Niedrige Koeffizienten sind deshalb unter Umständen nicht dem Test anzulasten, sondern der unsystematischen Veränderung des Merkmals. Das bedeutet schlicht: Bei nicht essenziell parallelen Messungen (z. B. aufgrund instabiler Merkmale) ist die Korrelation von Test und Retest nicht geeignet, die Reliabilität eines Tests zu schätzen.

Zeitintervall beachten

Die Retest-Reliabilität der 1. Fußball-Bundesliga



(© sidorovstock/stock.adobe.com)

Jedes Jahr fieberten Millionen Fans über 34 Spieltage mit „ihren“ Vereinen mit und hofften auf ein möglichst gutes Abschneiden, d. h. einen guten Tabellenplatz am Saisonende. Schaut man sich die Abschlusstabellen der Saisons 2016/2017 und 2017/2018 an – quasi eine Retest-Betrachtung nach 1 Jahr (wenngleich vermutlich keine essentiell parallele Messung) – und ermittelt die Korrelation der beiden Abschlusstabellen, so liegt diese bei $r_{tt}=.52$. Bei einer echten Testwiederholung wäre dieser Wert als zu niedrig anzusehen und der fragliche Test müsste überarbeitet werden. Anders gesagt: Es verändert sich einiges in den Rangreihen zwischen der 1. und 2. Messung. Es macht also – zumindest aus einer Stabilitäts- bzw. Retest-Reliabilitäts-Perspektive – Sinn für Fußballfans, jedes Jahr mit „ihren“ Vereinen mitzufiebern, denn das Vorjahresergebnis korrespondiert nur bedingt mit dem Ergebnis des aktuellen Jahres.

Abschlusstabellen der 1. Fußball-Bundesliga

Verein	Platzierung 2016/2017	Platzierung 2017/2018
FC Bayern München	1	1
RB Leipzig	2	6
Borussia Dortmund	3	4
TSG 1899 Hoffenheim	4	3
1. FC Köln	5	18
Hertha BSC	6	10
SC Freiburg	7	15
SV Werder Bremen	8	11
Borussia Mönchengladbach	9	9
FC Schalke 04	10	2
Eintracht Frankfurt	11	8
Bayer 04 Leverkusen	12	5
FC Augsburg	13	12
Hamburger SV	14	17
1. FSV Mainz 05	15	14
VFL Wolfsburg	16	16
FC Ingolstadt 04	17	– (abgestiegen)
SV Darmstadt 98	18	– (abgestiegen)

■ Schätzung über die Paralleltest-Methode

Hierbei werden (mindestens essenziell) parallele Versionen eines Tests von einer und derselben Gruppe von Personen bearbeitet. Die Reliabilitätsschätzung ergibt sich aus der Korrelation der beiden parallelen Tests. Anders als bei der Retest-Methode ist der 2. Test jedoch nicht identisch mit dem 1., sondern nur inhaltlich äquivalent. Somit spielen Übungs- und Erinnerungseffekte kaum eine Rolle. Es kann daher ein relativ kurzes Zeitintervall zwischen der Durchführung beider Tests gewählt werden. Folglich spielen Veränderungen des zu messenden Merkmals bei dieser Methode der Reliabilitätsschätzung keine Rolle. Aus diesen Ausführungen wird deutlich, warum die Paralleltest-Methode als „Königsweg“ der Reliabilitätsbestimmung gilt. Der wesentliche Nachteil der Paralleltest-Methode besteht allerdings darin, dass für die allermeisten Tests keine parallele Version vorliegt. Das liegt u. a. daran, dass es sehr aufwendig ist, eine „echte“ Parallelversion eines Tests zu erstellen.

Parallele Versionen eines Tests von denselben Personen bearbeiten lassen

Äquivalenz von Messungen

Da wir hier von Parallelversionen eines Tests sprechen, soll an dieser Stelle erläutert werden, dass es verschiedene Abstufungen der Äquivalenz von Messungen gibt. Es werden 5 Formen der Äquivalenz von Messungen unterschieden, die unterschiedlich „strenge“ Annahmen machen:

- Parallele Messungen
- Essenziell parallele Messungen
- Tau-äquivalente Messungen
- Essenziell tau-äquivalente Messungen
- Tau-kongenerische Messungen

Die fünf unterschiedlich strengen Annahmen zur Äquivalenz von Messungen setzen allesamt voraus: Die Messungen müssen eindimensional sein, also nur ein Merkmal messen – die Messfehler der beider Messungen sind unkorreliert. Sofern von parallelen Messungen im strengen Sinne ausgegangen werden soll, müssen die wahren Werte identisch und die Messungen gleich reliabel sein (gleicher Messfehlereinfluss; in nachfolgender Tabelle ist zusammengefasst, welche Annahmen von den 5 Formen der Äquivalenz von Messungen gemacht werden).

Sind beide Messungen gleich reliabel, aber die wahren Werte von Personen in einer der beiden Messungen lediglich um eine Konstante verschoben, so spricht man von essenziell parallelen Messungen.

Die verbleibenden 3 Formen der Äquivalenz verlangen nicht, dass Messungen gleich reliabel sein sollen. Werden dennoch identische wahre Werte gemessen, spricht man von tau-äquivalenten Messungen. Sind die Messungen nicht gleich reliabel, aber die wahren Werte bis auf eine Verschiebung um eine Konstante identisch, spricht man von essenziell tau-äquivalenten Messungen. Sind die wahren Werte nicht identisch, aber durch eine lineare Transformation ineinander überführbar, spricht man von tau-kongenerischen Messungen (vgl. Bühner 2021; Eid und Schmidt 2014).

	Gleiche latente Variable wird gemessen?	Messfehlereinfluss auf Messungen gleich?	Vergleich der wahren Werte der Messungen
Parallele Messungen	ja	ja	exakt gleich
Essenziell parallele Messungen	ja	ja	wahre Werte um additive Konstante verschoben
Tau-äquivalente Messungen	ja	nein	exakt gleich
Essenziell tau-äquivalente Messungen	ja	nein	wahre Werte um additive Konstante verschoben
Tau-kongenerische Messungen	ja	nein	wahre Werte durch lineare Transformation ineinander überführbar

Eid et al. (2017) weisen darauf hin, dass zur Reliabilitätsschätzung eines Tests nur dann dessen Korrelation mit einer Parallelversion (Paralleltestreliabilität) oder die Korrelation von Testteilen (Split-Half-Reliabilität) herangezogen werden kann, wenn die jeweils korrelierten Tests bzw. Testteile wenn mindestens essenziell parallel sind.

Paralleltestversionen nicht nur für Reliabilitätsschätzung nützlich

Paralleltestversionen werden übrigens nicht entwickelt, um die Reliabilität optimal zu schätzen (dafür wäre der Aufwand zu groß), sondern sie erweisen sich in der Praxis als nützlich. Erstens erlauben sie Gruppentestungen, ohne dass die Gefahr besteht, dass Probandinnen und Probanden voneinander abschreiben (dafür reicht es aus, einen „Pseudoparalleltest“ zu erstellen, bei dem lediglich die Abfolge der Items verändert wird). Zweitens kann man sie zur Veränderungsmessung einsetzen, um die Wirksamkeit einer Interventionsmaßnahme (z. B. eines Therapieverfahrens) zu überprüfen. Sie sind dafür besser geeignet als identische Testformen, da bei der 2. Erhebung keine Erinnerungseffekte auftreten.

Die Retest- und die Paralleltest-Methode sind sehr aufwendig. Die Reliabilität eines Tests kann auch ohne Testwiederholung und ohne Konstruktion einer parallelen Form geschätzt werden. Das Prinzip besteht darin, zu prüfen, ob der Test in sich konsistent ist. Für die Schätzung der Reliabilität benötigt man Informationen auf Itemebene, also die Antworten der Testperson auf jedes Item.

■ Schätzung über die Split-Half- bzw. Testhalbierungsmethode

Bei dieser Methode wird das Testergebnis auf Basis von 2 äquivalenten (d. h. mindestens essenziell parallelen) Testhälften berechnet. So erhält man für jede Probandin und jeden Probanden 2 Testwerte. Der ganze Test wird von den Testpersonen zunächst normal bearbeitet; die Aufteilung in Hälften erfolgt erst nach Vorliegen der Ergebnisse. Die Korrelation der Werte aus beiden Testhälften wird – nach einer Korrektur (s. u.) – als Schätzung der Reliabilität verwendet.

Test auf äquivalente Testhälften aufteilen

Vorgehen bei der Testhalbierung

Für die Testhalbierung kommen mehrere Techniken in Betracht:

- *Aufteilung nach ungerader und gerader Nummer der Items:* Diese Methode wird auch als Odd–Even-Methode bezeichnet. Die Items mit den Nummern 1, 3, 5, 7 etc. bilden die eine Testhälfte und die mit den Nummern 2, 4, 6, 8 etc. die andere. (Bei ungerader Itemzahl muss eine entsprechende Korrektur erfolgen.) Diese Aufteilung bietet sich an, wenn die Items im Test nach ihrer Schwierigkeit geordnet sind oder, wie es oft bei Persönlichkeitsfragebögen der Fall ist, überhaupt keine Ordnung aufweisen.
- *Aufteilung in die 1. und 2. Testhälfte:* Besteht der Test aus 40 Items, bilden die Items 1 bis 20 die eine und die Items 21 bis 40 die andere Hälfte. Diese Halbierungsmethode darf nicht angewendet werden, wenn der Test zeitbegrenzt ist (und die 2. Hälfte daher meist nicht vollständig bearbeitet wird) oder wenn die Items nach Schwierigkeit geordnet sind. Beide Hälften wären nicht vergleichbar.
- *Halbierung auf Basis von Itemkennwerten:* Dazu werden für alle Items zunächst die Schwierigkeit und die Trennschärfe ermittelt. Unter Berücksichtigung beider Kennwerte werden möglichst ähnliche Itempaare gebildet.

Zur Schätzung der Reliabilität über die Korrelation von 2 Testteilen muss ebenfalls gelten, dass diese Testteile mindestens essenziell parallele Messungen darstellen.

Die Korrelation der beiden Testhälften unterschätzt allerdings die Reliabilität des Gesamttests. Es wurde ja nicht der (gesamte) Test mit sich selbst, sondern nur eine Testhälfte mit sich selbst (streng genommen mit einer als parallel erachteten anderen Testhälfte) korreliert. Wenn also ein Test aus 40 Items besteht, würde er bei der Retest- oder der Paralleltest-Methode mit einem 40 Items umfassenden Test korreliert. Bei der Testhalbierungsmethode korreliert man dagegen 2 Tests mit 20 Items miteinander. Mithilfe der Spearman-Brown-Formel kann man für diese künstliche Testhalbierung korrigieren und schätzen, wie hoch die Reliabilität des Tests mit der gesamten Itemzahl (im Beispiel 40 Items) wäre. Voraussetzung für die Anwendung der Spearman-Brown-Formel ist jedoch, dass es sich bei den Testhälften um mindestens essenziell parallele Versionen handelt (Cho 2016).

Korrektur der Testhälftenkorrelation mit der Spearman-Brown-Formel

Spearman-Brown-Formel bei Testhalbierung

$$r_{tt'} = \frac{2 \times r_{1,2}}{1 + r_{1,2}}$$

(Formel angelehnt an Bühner 2011, S. 162)

Für ein Beispiel nehmen wir an, dass die Korrelation beider Testhälften $r_{1,2} = .70$ betrage. Daraus errechnet sich ein Wert von $r_{tt'} = .82$ als Schätzung der Split-Half-Reliabilität des Tests.

■ Schätzung über die interne Konsistenz

Die meisten Schätzungen der Reliabilität von Tests erfolgen über eine Ermittlung der internen Konsistenz. Hogan et al. (2000) berichten, dass dies für 75 % der Reliabilitätschätzungen gilt. Mit interner Konsistenz bezeichnen wir, analog zu Bentler (2009), den Anteil der gemeinsamen Varianz an der Gesamtvarianz der jeweils relevanten Items. Die bekannteste Form

Cronbachs Alpha = Koeffizient
Alpha = KR-20

dieser Reliabilitätsschätzung stammt bereits aus dem Jahre 1937; sie wurde von Kuder und Richardson (1937) entwickelt. In einem historischen Abriss beschreibt Cho (2016), dass die damals mit KR-20 bezeichnete Formel (es war schlicht die 20. Formel im Manuskript von Kuder und Richardson) von Cronbach (1951) aus Marketinggründen in „Koeffizient Alpha“ umgetauft wurde. Aus „Koeffizient Alpha“ wurde schnell „Cronbachs Alpha“. Diese Bezeichnung ist zwar insofern irreführend, als dass Lee Cronbach nicht der Urheber dieses Koeffizienten war, sie soll aber in diesem Buch beibehalten werden, vor allem da sie von vielen Testautorinnen und -autoren ebenfalls verwendet wird. Um auf alternative Bezeichnungen hinzuweisen, verwenden wir jedoch wechselweise auch „Koeffizient Alpha“, „Cronbachs α “ oder auch kurz „Alpha“.

Koeffizient Alpha bzw. Cronbachs Alpha

$$\alpha = \frac{k}{k-1} \times \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_t^2} \right)$$

Nach Eid und Schmidt (2014) lässt sich jedoch auch folgende Formel verwenden:

$$\alpha = \frac{k}{k-1} \times \frac{\sum_{i \neq j} cov(x_i x_j)}{s_t^2}$$

k = Zahl der Testitems

s_i^2 = Varianz eines Testitems i

s_t^2 = Varianz des Tests

$cov(x_i x_j)$ = Kovarianz der Testitems i und j

(Formeln aus Eid und Schmidt 2014, S. 287 f.)

Insbesondere die 2. hier vorgestellte Formel macht deutlich, dass Cronbachs Alpha umso höher ist, je mehr ähnliche, d. h. hoch miteinander korrelierende Items ein Test enthält (da dies den Zähler des 2. Bruchs vergrößert). Cronbach (1951) zeigte zudem, dass Alpha dem Durchschnitt der Split-Half-Reliabilitätskoeffizienten aller möglichen Testhälften entspricht.

In Anlehnung an Streiner (2003) können wichtige Hinweise zur Bewertung von Alpha-Koeffizienten gegeben werden:

Was man über Cronbachs Alpha wissen sollte

- Je höher die Items interkorrelieren, desto höher fällt Alpha aus.
- Alpha hängt jedoch nicht nur von der Interkorrelation der Items, sondern auch von weiteren Faktoren ab (s. u.).
- Nicht immer sollte eine hohe Iteminterkorrelation angestrebt werden. Heterogene Konstrukte (also solche, die sich durch relativ unabhängige Facetten oder Komponenten auszeichnen) verlangen zwangsläufig nach einer Operationalisierung durch entsprechend niedrig korrelierende Items.
- Weist ein kurzer Test ein sehr hohes Alpha auf, sind die Items eventuell redundant. Beispielsweise wird die gleiche Frage in unterschiedlichen Varianten immer wieder gestellt.
- Alpha ist (wie alle Reliabilitätskoeffizienten) stichprobenabhängig. In heterogenen Stichproben fällt die Varianz der Testwerte höher aus, was wiederum zu höheren Werten für Alpha führt.

Es muss jedoch beachtet werden, dass Alpha nur dann eine angemessene Schätzung der Reliabilität gewährleistet, wenn die Testitems mindestens tau-äquivalente Messungen darstellen (Eid und Schmidt 2014). Sind Messungen durch Testitems nur tau-kongenerisch, ermöglicht Cronbachs Alpha keine angemessene Schätzung der Reliabilität. Da tau-Äquivalenz der Items selten angenommen werden kann, raten mittlerweile viele Forscherinnen und Forscher von der Nutzung von Cronbachs Alpha ab. In diesen Fällen empfiehlt sich eine Schätzung der Reliabilität anhand von McDonalds Omega (McDonald 1999).

Koeffizient Alpha verlangt, dass Testitems tau-äquivalente Messungen darstellen

McDonalds Omega

Für eine eindimensionale Messung, bei der die Varianz der latenten Variablen auf 1 fixiert wurde, kann folgende Formel verwendet werden:

$$\omega = \frac{\left(\sum_{i=1}^k \lambda_i\right)^2}{\left(\sum_{i=1}^k \lambda_i\right)^2 + \sum_{i=1}^k \varepsilon_i}$$

k = Zahl der Testitems

λ_i = Unstandardisierte Ladung eines Items auf dem gemeinsamen Faktor aller Testitems

ε_i = Fehlervarianz eines Testitems (Anteil des Items, der nicht durch den gemeinsamen Faktor aller Testitems beschrieben wird)

(Formel aus Eid und Schmidt 2014, S. 322)

Vorteile von McDonalds Omega

Dunn et al. (2014) betonen den wesentlichen Vorteil von McDonalds Omega im Vergleich zu Cronbachs Alpha: Es müssen weniger Voraussetzungen erfüllt sein, um eine angemessene Reliabilitätsschätzung zu erhalten. Sie führen weiterhin aus, dass selbst bei Erfüllung der Voraussetzungen, die Cronbachs Alpha impliziert, McDonalds Omega eine robustere Reliabilitätsschätzung ermöglicht. Die Voraussetzung der Eindimensionalität, die für beide Methoden gleichermaßen gilt, sollte jedoch stets erfüllt sein oder es sollten pro Dimension Subskalen gebildet und deren Reliabilität geschätzt werden.

Voraussetzung beachten

„Reliabilitäts-Wirrwarr“

Aus den bisherigen Ausführungen wurde deutlich, dass für die Aufteilung von Tests in 2 Teile (Split-Half-Reliabilität) und die Berechnung interner Konsistenzen Methoden der Reliabilitätsbestimmung existieren (Alpha, Omega), die jeweils unterschiedliche Voraussetzungen fordern. Sind die jeweiligen Voraussetzungen nicht erfüllt oder können sogar strengere Voraussetzung als erfüllt angesehen werden, so können weitere Methoden der Reliabilitätsbestimmung herangezogen werden. Diese firmieren unter diversen, nicht besonders intuitiven Namen – beispielsweise als „Raju-Koeffizient“ (Raju 1970). Cho (2016) schlägt daher vor, Methoden der Reliabilitätsbestimmung nach deren Voraussetzungen zu benennen. Dies hat den Vorteil einer klaren Logik der Bezeichnungen und macht allen Anwenderinnen und Anwendern unmittelbar klar, welche Voraussetzungen erfüllt sein müssen. Die folgende Tabelle nennt einige der vorgeschlagenen Bezeichnungen (vgl. Cho 2016, S. 659 f.):

Voraussetzung	Split-Half-Methoden	Konsistenzmethoden
Parallele Messungen	Parallele Split-Half-Reliabilität (Spearman-Brown Formel)	Parallele Reliabilität (standardisiertes Alpha)
Tau-äquivalente Messungen	Tau-äquivalente Split-Half-Reliabilität (Flanagan-Rulon-Formel, Guttmans λ_4)	Tau-äquivalente Reliabilität (Cronbachs Alpha)
Kongenerische Messungen	Kongenerische Split-Half-Reliabilität (Raju-Koeffizient, Anoff-Feldt-Koeffizient)	Kongenerische Reliabilität (Omega)

Losgelöst von der Bezeichnung der Methoden zur Reliabilitätsbestimmung ist es wichtig, zu wissen, welche Voraussetzungen erfüllt sein müssen. Andernfalls erfolgt eine Über- oder Unterschätzung der Reliabilität.

Reliabilität im Testmanual muss nicht der Reliabilität einer bestimmten Messung entsprechen

Vereinfachend haben wir in den Ausführungen zur Reliabilität mitunter von *der* Reliabilität eines Tests gesprochen. Dies entspricht dem üblichen Sprachgebrauch von Testautorinnen und Testautoren. Nimmt man jedoch ernst, dass jeder unsystematische Einfluss die Reliabilität einer Messung mindert, so wird deutlich, dass dies auch Faktoren sein können, die außerhalb des Tests liegen. Beispielsweise können bei einer Durchführung und Auswertung von Tests unsystematische Faktoren auf das Testergebnis einwirken. Das können Lichtverhältnisse, Geräusche, Luftqualität, Raumtemperatur, Sitzkomfort, Art und Anzahl der anderen Testteilnehmer/-innen sowie die Verwendung von Schablonen bei der Auswertung oder Ungenauigkeiten beim Zusammenzählen von Punkten sein. Diese Fehlerquellen können aber bei der nächsten Durchführung und Auswertung nicht mehr existent sein. Es muss also einerseits gelten: Die Objektivität der Durchführung, Auswertung und Interpretation von Testergebnissen ist unbedingt sicherzustellen. Andererseits wird dies nicht immer perfekt gelingen. Daher muss auch festgestellt werden: Die Reliabilität einer Messung kann im konkreten Anwendungsfall anders ausfallen, als die von Testautorinnen und -autoren im Manual berichtete Schätzung.

Empfehlungen des Diagnostik- und Testkuratoriums

Das Diagnostik- und Testkuratorium empfiehlt die Prüfung, ob im Manual folgende Fragen hinreichend beantwortet werden (zitiert nach Diagnostik- und Testkuratorium 2018b, S. 114 f., © Hogrefe):

- Wurden die jeweiligen Reliabilitätskennwerte für alle (Sub-)Populationen aus einer Stichprobenerhebung geschätzt, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll?
- Sind die jeweiligen Reliabilitätskennwertschätzungen inhaltlich angemessen?
- Sofern mit dem Verfahren Eignungsmerkmale erfasst werden, für die eine zumindest relative Zeit- und Situationsstabilität angenommen wird: Wurde die Zuverlässigkeit (auch) über die Retest-Methode bestimmt oder die Retest-Reliabilität durch einen geeigneten Untersuchungsplan geschätzt?
- Im Fall von Retest-Reliabilitäten: Ist das Intervall zwischen Test und Retest angemessen?
- Ist eine sehr hohe interne Konsistenz auf nahezu identisch gestaltete Items zurückzuführen (was negativ zu bewerten ist)?
- Bei Tests mit einer Speedkomponente, bei denen also nicht alle Testpersonen auch zur Bearbeitung der letzten Items kommen: Wurden andere Reliabilitätsschätzungen als die der internen Konsistenz verwendet?

- ⚠ Erweist sich ein in der Entwicklung befindlicher Test für seine Einsatzzwecke als nicht hinreichend reliabel, müssen Testentwicklerinnen und -entwickler in eine frühere Phase der Testentwicklung (☞ Abb. 2.22) zurückgehen und die danach folgenden Schritte wiederholen.

2.6.2.2 Bedeutung der Reliabilität in der Einzelfalldiagnostik

In vielen Anwendungskontexten der Psychologischen Diagnostik hat man es mit einzelnen Klientinnen und Klienten zu tun, für die es gilt, eine Fragestellung zu beantworten. Sofern Diagnostikerinnen und Diagnostiker dabei auf Verfahren zurückgreifen, für die eine Reliabilitätsschätzung bekannt ist, stellt sich die Frage: Wie soll diese Information im Rahmen der Einzelfalldiagnostik berücksichtigt werden? Auf diese Frage gibt es mehrere Antworten:

1. Identifikation der passenden Reliabilitätsangaben Wir haben bereits festgestellt, dass es nicht *die* Reliabilität eines Tests gibt. Daher sollten Diagnostikerinnen und Diagnostiker zuerst prüfen, welche der im Testmanual berichteten Reliabilitätsangaben unter Randbedingungen und anhand von Stichproben ermittelt wurden, die möglichst gut auf den aktuellen „Fall“ übertragbar sind (z. B. im Rahmen von Einzeltestungen bei gesunden Probandinnen und Probanden). Idealerweise werden in den folgenden Schritten nur Reliabilitätsangaben berücksichtigt, die unter vergleichbaren Bedingungen ermittelt wurden. Bei Tests, die einen großen Altersbereich abdecken und die Altersnormen haben, sollte die Reliabilität in der jeweiligen Altersgruppe und nicht die der gesamten Normstichprobe verwendet werden. Liegen keine Reliabilitätsangaben vor, die auf den aktuellen Fall übertragbar sind, sollte man sich fragen, warum dies so ist. Möglicherweise ist der Test nicht für den vorliegenden Fall geeignet.

2. Nutzung zur Auswahl der Tests Wenn mehrere Tests zur Anwendung bei dem vorliegenden Fall infrage kommen, kann u. a. anhand der Reliabilitätsangaben eine Auswahl der infrage kommenden Tests vorgenommen werden. Selbstverständlich sollte die Reliabilität nicht das alleinige Auswahlkriterium sein. Eine zu geringe Reliabilität kann und sollte jedoch auch losgelöst von anderen Kriterien zum Ausschluss von Testverfahren führen.

3. Feststellung der Genauigkeit, mit der ein Individualergebnis ermittelt wurde Der Anteil der Fehlervarianz an der Gesamtvarianz von Testwerten (Unreliabilität) sollte sinnvollerweise auch bei einzelnen Messungen berücksichtigt werden. Man stelle sich dazu vor, dass die Klientin oder der Klient den Test nicht nur einmal absolviert, sondern sehr oft. Die beobachteten Testwerte würden dann – gemäß den Annahmen der Klassischen Testtheorie – um den wahren Wert schwanken. Da man allerdings nur einmal misst, kann man sich nicht sicher sein, wie viel „schwankungsbedingte Abweichung“ in dem aktuell gemessenen Wert enthalten ist. Man berichtet daher statt eines einzelnen Ergebnisses ein Intervall, dass die zu erwartende Schwankung reflektiert – ein Konfidenzintervall.

Die nachfolgend erläuterten Konfidenzintervalle dienen der Eingrenzung des Bereichs, in dem der wahre Wert einer Person mit hinreichend großer Wahrscheinlichkeit zu vermuten ist. Dabei beziehen wir uns nur auf Konfidenzintervalle, die nach der sog. „Regressionsmethode“ berechnet werden. Nur für solche gilt die vorherige Aussage (Eingrenzung des Bereichs, in dem der wahre Wert einer Person mit einer bestimmten Wahrscheinlichkeit zu vermuten ist; vgl. Eid und Schmidt 2014). Für Konfidenzintervalle, die nach der Äquivalenzmethode berechnet wurden, ist dies nicht zulässig (Hoekstra et al. 2014). Daher wird deren Berechnung nachfolgend nur aus Gründen der Vollständigkeit berichtet.

Bereich, in dem der „wahre Wert“ zu vermuten ist

Berechnung von Konfidenzintervallen nach der Regressionsmethode

Wie der Name „Regressionsmethode“ besagt, wird zunächst der wahre Wert einer Person regressionsanalytisch geschätzt. In die dazu vorzunehmende Berechnung gehen der beobachtete Wert X_i , die Reliabilität r_{tt} und der Mittelwert der Skala M ein:

$$X_{i'} = r_{tt} \times X_i + M \times (1 - r_{tt})$$

(Formel angelehnt an Bühner 2021, S. 683)

In □ Abb. 2.34 sind die hiernach berechneten wahren Werte für einen IQ-Wertebereich von 60 bis 140 (Mittelwert = 100) dargestellt. Dies wurde für 4 unterschiedlich reliable (fiktive) Tests gemacht ($r_{tt} = .9/.8/.7/.6$). Man sieht, dass extreme Werte stets zur Mitte des Wertebereichs korrigiert wurden, und dies umso stärker, je unreliabler der Test ist. So resultiert bei einer Reliabilität von .6 aus einem gemessenen IQ-Normwert von 130 (x-Achse) ein korrigierter IQ-Normwert von 118. Bei einer Reliabilität von .9 fällt diese Korrektur deutlich milder aus (aus IQ-Normwert von 130 wird 124). Erläuterungen zu IQ-Normwerten und anderen Normwerten finden sich in ▶ Abschn. 2.6.4.

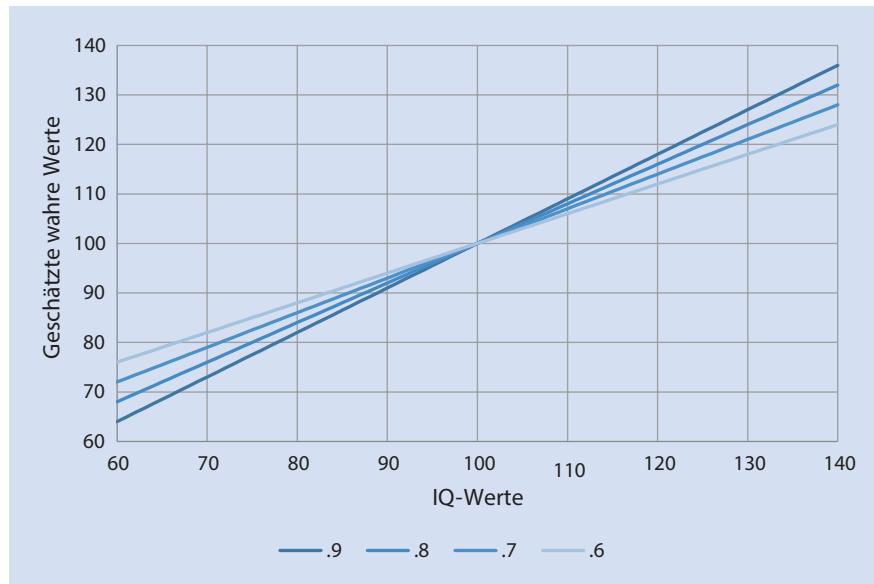


Abb. 2.34 Schätzung der wahren Werte nach der Regressionsmethode bei 4 unterschiedlich reliablen Tests

Um den so geschätzten wahren Wert wird ein Konfidenzintervall gelegt. Zu dessen Berechnung wird der Standardschätzfehler s_{ET} herangezogen. Dieser ergibt sich aus der Standardabweichung der beobachteten Werte s_X und der Reliabilität r_{tt} des Tests.

Konfidenzintervall um geschätzten wahren Wert

$$s_{ET} = s_X \times \sqrt{r_{tt} \times (1 - r_{tt})}$$

(Formel angelehnt an Bühner 2021, S. 684)

Unter der Annahme der Normalverteilung der Schätzfehler beschreibt das Intervall von -1 Standardschätzfehler bis $+1$ Standardschätzfehler einen Bereich in dem 68 % aller geschätzten wahren Werte liegen. Würde man also ein Konfidenzintervall um den geschätzten wahren Wert $X_{i'}$ einer Person legen, so könnte man sagen, dass mit 68 %iger Wahrscheinlichkeit der wahre Wert der Person in diesem Intervall liegt. Es ist jedoch üblich, eine 95 %ige Wahrscheinlichkeit anzunehmen. Dazu multipliziert man den Standardschätzfehler mit dem z -Wert, innerhalb dessen negativer und positiver Ausprägung 95 % der Fläche der Normalverteilung liegen. Üblicherweise wird dieser Wert als $z_{1-\alpha/2}$ beschrieben, wobei α die Irrtumswahrscheinlichkeit (üblicherweise 0,05) ist. $z_{1-\alpha/2}$ beschreibt daher die Fläche unter der Normalverteilung, wenn man an den Rändern der Verteilung insgesamt 5 % „abschneidet“.

Das Konfidenzintervall um den geschätzten wahren Wert $X_{i'}$ einer Person ist also

$$X_{i'} \pm z_{1-\alpha/2} \times s_X \times \sqrt{r_{tt} \times (1 - r_{tt})}.$$

- ! Es muss beachtet werden, dass diese Formel nur bis zu einer Reliabilität von .50 sinnvolle Konfidenzintervall „produziert“. Bei Reliabilitätswerten $< .50$ entstehen keine sinnvollen Ergebnisse, da dann mit weiter fallender Reliabilität immer kleinere Konfidenzintervalle resultieren.

Beispielhafte Konfidenzintervallberechnung

Nehmen wir an, Frau S. habe in einem Intelligenztest einen IQ-Normwert von 118 erzielt. (Die IQ-Normwertska hat einen Mittelwert von 100 und eine Standardabweichung von 15). Nehmen wir zudem an, der bearbeitete Intelligenztest habe eine Reliabilität von .89. Das nach der Regressionshypothese geschätzte 95 %ige Konfidenzintervall errechnet sich wie folgt:

Schritt 1: Geschätzten wahren Wert berechnen:

$$X_{i'} = r_{tt} \times X_i + M \times (1 - r_{tt}) \\ X_{i'} = 0,89 \times 118 + 100 \times (1 - 0,89) = 116,02$$

Schritt 2: Konfidenzintervall um den geschätzten wahren Wert $X_{i'}$ berechnen:

$$X_{i'} \pm z_{1-\alpha/2} \times s_X \times \sqrt{r_{tt} \times (1 - r_{tt})} \\ 116,02 \pm 1,96 \times 15 \times \sqrt{0,89 \times (1 - 0,89)} = 116,02 \pm 9,199$$

Das Konfidenzintervall reicht in diesem Fall also von 106,82 (untere Konfidenzintervallgrenze) bis 125,22 (obere Konfidenzintervallgrenze). Es lässt sich also sagen, dass das Ergebnis von Frau S. in dem verwendeten Intelligenztest unter Berücksichtigung von dessen Messgenauigkeit (also der Reliabilität) mit 95 %iger Wahrscheinlichkeit in einem Bereich zwischen 106,82 und 125,22 liegt. Hätte Frau S. das gleiche Ergebnis (IQ-Normwert = 118) in einem weniger reliablen Test erzielt, würde das Konfidenzintervall größer. Beispielsweise ergeben sich bei einer Reliabilität von .75 Konfidenzintervallgrenzen von 100,77 und 126,23. Für den Fall, dass man eine andere Sicherheitswahrscheinlichkeit als 95 % wünscht, braucht man lediglich den z -Wert von 1,96 durch 2,58 (99 %) oder etwa 1,64 (90 %) zu ersetzen. Bei einseitiger Fragestellung lauten die z -Werte 1,65 (95 %), 2,33 (99 %) und 1,28 (90 %).

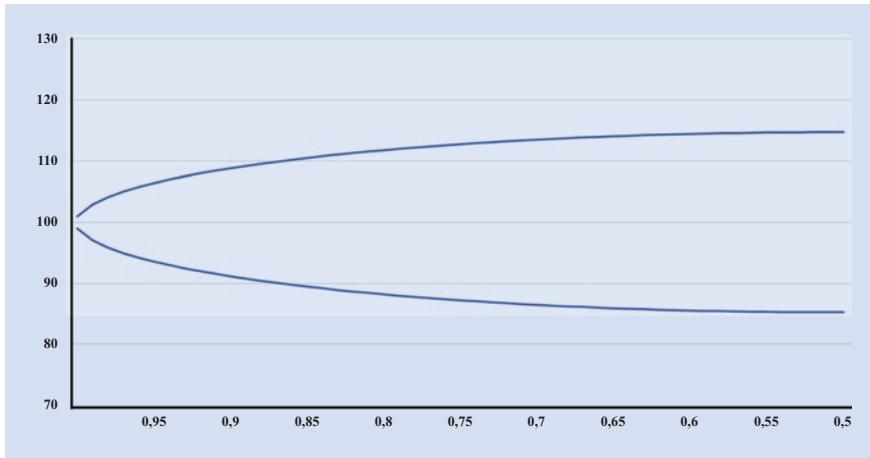
z -Werte zur Berechnung von Konfidenzintervallen:

Fragestellung	Sicherheitswahrscheinlichkeit		
	90 %	95 %	99 %
Einseitig	1,282	1,645	2,326
Zweiseitig	1,645	1,96	2,576

Wer Konfidenzintervalle nicht selbst berechnen will, kann den Kalkulator unter ► <https://www.psychometrica.de/normwertrechner.html> nutzen.

Je unreliabler die Messung, desto unpräziser die Aussage

Es wird deutlich, dass die Reliabilität maßgeblich die Breite des Konfidenzintervalls beeinflusst. Das heißt, Aussagen über eine Person werden umso unpräziser, je unreliabler die vorgenommene Messung war. □ Abb. 2.35 zeigt den Verlauf der Konfidenzintervallbreite in Abhängigkeit von der Reliabilität am Beispiel einer IQ-Werte-Skala (Mittelwert = 100, Standardabweichung = 15). Man sieht, dass bei einer Reliabilität von $r_{tt} = .60$ das 95 %ige Konfidenzintervall um den geschätzten wahren Wert fast so breit wird wie 2 Standardabweichungen (= 30 IQ-Wertpunkte). Damit wird eine Aussage für Individuen (mit der gewünschten Wahrscheinlichkeit von 95 %) ziemlich bedeutungslos. Man stelle sich dazu einfach vor, man habe die betreffende Person 1 h lang mit der Testung beschäftigt, nur um am Ende bei einem solch breiten Konfidenzintervall sagen zu müssen, dass ihr Ergebnis in einem Bereich liegt, der so breit ist, dass 68 % aller Menschen „darunter fallen“ (unter Annahme der Normalverteilung und im mittleren Bereich der Verteilung).



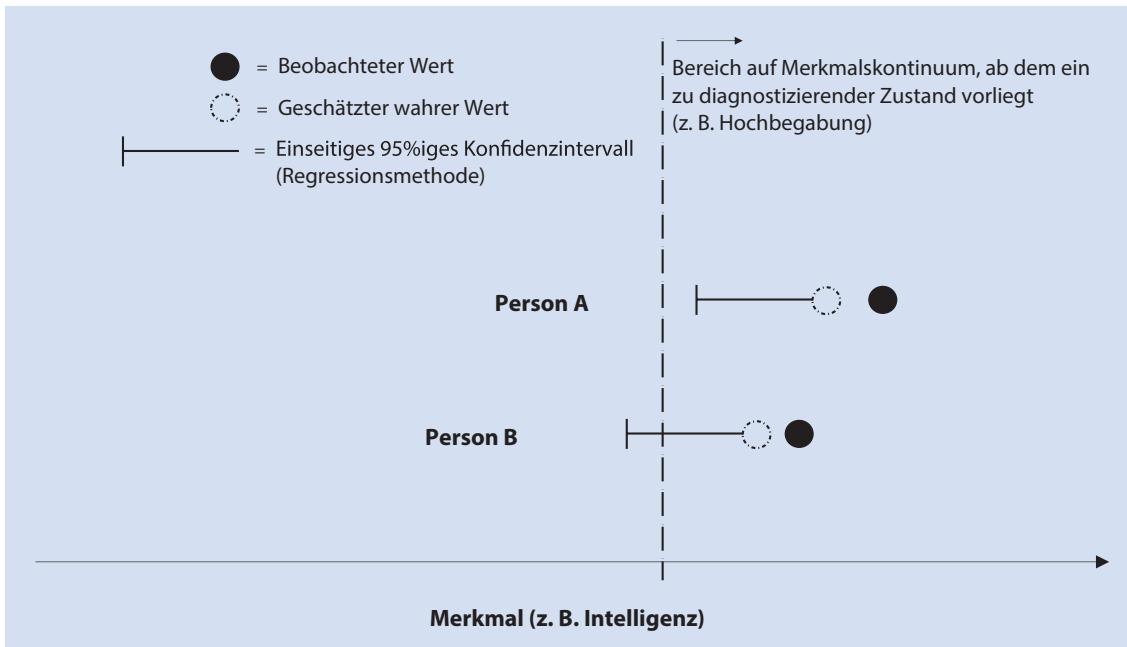
■ Abb. 2.35 Konfidenzintervallgrenzen in Abhängigkeit der Reliabilität. Auf der x-Achse ist die Reliabilität, auf der y-Achse die Grenzen des Konfidenzintervalls (zweiseitig, 95 %ige Sicherheitswahrscheinlichkeit) in IQ-Werten (Standardabweichung von 15) angegeben. Die obere Linie stellt die obere Konfidenzintervallgrenze dar, die untere Linie die untere Konfidenzintervallgrenze – jeweils um einen gemessenen Wert von 100

Selbstverständlich muss ein Konfidenzintervall nicht notwendigerweise zu beiden Seiten um den geschätzten wahren Wert einer Person gelegt werden. Manchmal ist nur eine Grenze des Konfidenzintervalls wichtig. Beispielsweise ist bei der Abklärung einer möglichen Hochbegabung die obere Grenze des Konfidenzintervalls nicht von Interesse; wichtig ist nur, ob die untere Grenze über einem kritischen Wert (üblicherweise einem $\text{IQ} \geq 130$) liegt. Bei einseitig berechneten Konfidenzintervallen ist ein z -Wert von $z_{1-\alpha}$ zu verwenden.

■ Abb. 2.36 veranschaulicht dieses Prinzip. Es wird deutlich, dass man für Person A mit 95 %iger Wahrscheinlichkeit sagen kann, dass ihre Merkmalsausprägung auch unter Berücksichtigung der Messgenauigkeit (also der

Einseitige vs. zweiseitige Konfidenzintervalle

Konfidenzintervalle an Cut-off-Werten



■ Abb. 2.36 Beurteilung von Testwerten anhand von einseitigen Konfidenzintervallen

Reliabilität) über der geforderten Grenze liegt. Für Person B lässt sich das nicht mit 95 %iger Wahrscheinlichkeit sagen. Für Person B könnte anhand der obigen Formel die Wahrscheinlichkeit berechnet werden, mit der das einseitige Konfidenzintervall so klein wird, dass es die geforderte Grenze nicht unterschreitet. Dann könnte man sagen, dass die Merkmalsausprägung von Person B auch unter Berücksichtigung der Messgenauigkeit über der geforderten Grenze liegt, allerdings nur mit einer Wahrscheinlichkeit von beispielsweise 60 %.

Äquivalenzmethode: Beobachteter Wert wird als wahrer Wert angenommen

Berechnung von Konfidenzintervallen nach der Äquivalenzmethode

Bei der Äquivalenzmethode wird der wahre Wert nicht mithilfe der linearen Regression geschätzt. Stattdessen wird davon ausgegangen, dass der beobachtete Wert eine gute Schätzung des wahren Wertes ist – beide Werte werden als äquivalent angenommen. Dementsprechend basiert das Konfidenzintervall nicht auf dem Standardschätz-, sondern dem Standardmessfehler (SE):

$$s_E = s_X \times \sqrt{1 - r_{tt}}$$

Das Konfidenzintervall um den beobachteten Wert einer Person ist dann:

$$X_i \pm z_{1-\alpha/2} \times s_E \times \sqrt{1 - r_{tt}}$$

Allerdings ist bei diesem Vorgehen (und einem α von .05) nur folgende Aussage zulässig: Bei 95 % der nach dieser Methode berechneten Konfidenzintervalle liegt der wahre Wert innerhalb der jeweils resultierenden Konfidenzintervallgrenzen (Hoekstra et al. 2014). Eine Aussage über die Wahrscheinlichkeit der Werte einer Person ist nicht zulässig. Da man aber genau dies in der Regel möchte, sollte auf die – zudem exaktere – Regressionsmethode zurückgegriffen werden.

Reliabilität beim Vergleich von 2 Testwerten berücksichtigen

Die Reliabilität von Messungen ist auch dann zentral, wenn man 2 Testwerte einer Probandin bzw. eines Probanden vergleichen möchte – beispielsweise um zu beurteilen, ob eine Verbesserung infolge einer Behandlung eingetreten ist (vgl. auch ► Abschn. 8.6.2). Aufgrund unsystematischer Schwankungen (Messfehler) könnten sich die beiden Werte zufällig voneinander unterscheiden, auch wenn keine tatsächliche Verbesserung des Merkmals eingetreten ist. Daher muss man wissen, wie groß eine Differenz zwischen 2 Werten (des gleichen Tests, gemessen bei der gleichen Person) sein muss, sodass sie wahrscheinlich nicht alleine durch den Messfehler erklärt werden kann. Dies bezeichnet man als kritische Differenz.

Berechnung von kritischen Differenzen einer Person bei Wiederholung des gleichen Tests

Sofern die Messungen mit einem Test zu 2 Messzeitpunkten gleich reliabel sind, gilt:

$$(x_1 - x_2)_{\text{krit}} = z_{1-\alpha/2} \times s_x \times \sqrt{2 \times (1 - r_{tt})}$$

$z_{1-\alpha/2}$ = z-Wert bei einer Irrtumswahrscheinlichkeit von α

s_x = Standardabweichung der Testwerte

r_{tt} = Reliabilität der beiden Messungen

(Lienert und Raatz 1998, S. 369)

Diese Formel wird auch genutzt, um den sog. „Reliable Change Index“ (RCI) zu berechnen (vgl. Jacobson und Truax 1991, S. 14). Vereinfachend geht man dabei von $\alpha=.05$ und einem entsprechenden z-Wert von $\approx 1,96$ aus.

$$RCI = \frac{(x_1 - x_2)}{s_x \times \sqrt{2 \times (1 - r_{tt})}} = 1,96$$

Übersteigt der Bruch den kritischen Wert von 1,96, geht man von einer „reliablen“ Veränderung aus. Der RCI ist vor allem in der klinisch-psychologischen Diagnostik verbreitet. Mit seiner Hilfe können Veränderungen zwischen den zu Behandlungsbeginn und zu Behandlungsende erfolgten Messungen dagehend beurteilt werden, ob sie wahrscheinlich eher messfehlerbedingt sind oder eine tatsächliche Veränderung der Symptomatik darstellen.

Es ist klar, dass mit abnehmender Reliabilität der Messungen die Differenz zunimmt, ab der man einen Unterschied zwischen der Vor- und der Nachtestung als signifikant bezeichnen könnte. Bereits bei einer Reliabilität der beiden Messungen von $r_{tt}=.86$ müssen sich die beiden Werte um eine Standardabweichung (also 15 Punkte auf der IQ-Werte-Skala) unterscheiden, um als kritische Differenz beurteilt zu werden. Erwartet man eine Verbesserung des Merkmals im Zuge einer Behandlung, so kann die kritische Differenz mit $z_{1-\alpha}$ statt mit $Z_{1-\alpha/2}$ berechnet werden (einseitige Berechnung der kritischen Differenz). Bei einer Reliabilität der Messungen von $r_{tt}=.86$ würde dann die kritische Differenz auf ≈ 13 IQ-Wertpunkte sinken.

Je geringer die Reliabilität, desto größer die Differenz, ab der tatsächlich ein Unterschied vorliegt

Testwertveränderung ohne Merkmalsveränderung

Wenn die Messwerte zum 2. Messzeitpunkt so weit auseinanderliegen, dass sie die kritische Differenz übersteigen, kann dies auf eine tatsächliche Veränderung der gemessenen Merkmale zurückzuführen sein. Um tatsächliche Veränderungen abzubilden, sind änderungssensitive Verfahren erforderlich. Liegen entsprechende Validitätsbelege vor, sind diese für die Auswahl eines Verfahrens hilfreich. Bei Fragebögen erkennt man eventuell an den Items, ob sie zur Veränderungsmessung geeignet sind. Formulierungen wie „In meinem Leben habe ich ...“ oder „Normalerweise bin ich ...“ beziehen sich offensichtlich auf lange Zeiträume. Aussagen wie „Ich bin momentan mit meinem Leben zufrieden“ betreffen dagegen den momentanen Zustand. Manchmal wird in der Instruktion explizit festgelegt, dass sich die Aussagen auf einen bestimmten Zeitraum (z. B. die letzten 2 Wochen) beziehen sollen.

Allerdings gibt es auch Faktoren, die zu einer kritischen Differenz zweier Messwerte führen können, ohne dass eine Merkmalsveränderung stattgefunden hat. Diese sind

- Übungsgewinne,
- Veränderung von anderen Merkmalen der Person (z. B. Motivation), die sich auf die Messung des eigentlichen Merkmals auswirken,
- Regression zur Mitte.

Alleine durch die Wiederholung eines Tests kann die Testleistung steigen, ohne dass sich das gemessene Merkmal verändert hat. Wenn eine Messwiederholung geplant ist, um Änderungen zu evaluieren, sollten Verfahren eingesetzt werden, die wenig anfällig für Übungseffekte sind. Sofern ein echter Paralleltest (also nicht nur eine veränderte Itemabfolge) zur Verfügung steht, sollte dieser verwendet werden; allerdings sind auch dann noch Übungs- oder Transfereffekte möglich. Besteht der Verdacht, dass die Testleistung ansteigen könnte, weil ein Teil der Testpersonen bei der 1. Durchführung die Instruktion nicht gut genug versteht, kann die Instruktion durch zusätzliche Erläuterungen und Übungsaufgaben optimiert werden. Besonders bei der Untersuchung leistungsschwacher Personen können damit Übungseffekte verringert werden. Allerdings weicht man damit oft von der Standardinstruktion ab, sodass die Normtabellen ihre Gültigkeit verlieren. Deshalb ist diese Strategie primär für Forschungsfragen geeignet. In der Forschung kann zudem eine Kontrollgruppe ohne Intervention realisiert werden.

Schließlich kann sich ein Testwert verändern, weil sich die Testperson verändert hat – aber nicht in Bezug auf das gemessene, sondern ein anderes Merkmal, das sich jedoch auf den Testwert auswirkt. Beispielsweise kann sich die Motivation, ein gutes Ergebnis zu erzielen oder als gesund oder krank zu gelten, ändern.

Eine Regression zur Mitte ist bei extremen Werten, deren wiederholtes Auftreten in dieser Extremität unwahrscheinlich ist, zu beachten. Wir gehen in ▶ Abschn. 7.1.1 näher darauf ein.

Möchte man die Werte einer Person aus 2 unterschiedlichen Tests (mit unterschiedlichen Reliabilitäten) vergleichen, so ändert sich die Vorgehensweise zur Berechnung der kritischen Differenz wie folgt:

Berechnung von kritischen Differenzen einer Person bei 2 verschiedenen Tests

Aufgrund der unterschiedlichen Reliabilitäten der Messungen muss zunächst eine tau-Normierung der beobachteten Werte einer Person in beiden Tests vorgenommen werden. Nach Bühner (2021, S. 690 f.) kann dies mit folgender Formel erfolgen:

$$\text{(für Test 1)} \quad Y_{\tau,1} = \frac{X_1}{\sqrt{r_{tt1}}} + M \times \left(1 - \frac{1}{\sqrt{r_{tt1}}} \right)$$

$$(für Test 2) \quad Y_{\tau,2} = \frac{X_2}{\sqrt{r_{tt2}}} + M \times \left(1 - \frac{1}{\sqrt{r_{tt2}}} \right)$$

Die somit tau-normierten Werte können dann eingesetzt werden in:

$$z_{Diff(Y_{\tau,1}), (Y_{\tau,2})} = \frac{Y_{\tau,1} - Y_{\tau,2}}{s_x \sqrt{\frac{1-r_{tt1}}{r_{tt1}} + \frac{1-r_{tt2}}{r_{tt2}}}}$$

$z_{Diff(Y_{\tau,1}), (Y_{\tau,2})}$ = z -Wert der kritischen Differenz einer Person in Tests 1 und 2

$Y_{\tau,1}$ = tau-normierter Messwert einer Person in Test 1

$Y_{\tau,2}$ = tau-normierter Messwert einer Person in Test 2

X_1 = Wert einer Person in Test 1

X_2 = Wert einer Person in Test 2

s_x = Standardabweichung der Testwerte

r_{tt1} = Reliabilität von Test 1

r_{tt2} = Reliabilität von Test 2

M = Populationsmittelwert

2.6.2.3 Bedeutung der Reliabilität bei der Evaluation

Die Reliabilität von Messungen ist nicht nur bei der Individualdiagnostik ein zentraler Aspekt. Auch bei der Betrachtung von vielen Messungen ist sie relevant. Wenn das Vorgehen im Rahmen der Psychologischen Diagnostik quantitativ evaluiert werden soll, stellt sich häufig die Frage nach dem Zusammenhang zwischen 2 Messungen – z. B. zwischen einem Eignungstest und Leistungen in Beruf oder Ausbildung. Dabei gilt: Wenn Messwerte fehlerbehaftet sind, wirkt sich dies mindernd auf die Höhe der Korrelation mit einer anderen Variablen aus. Man stelle sich vor, dass 2 Tests das gleiche Merkmal erfassen sollen, aber so schlecht konstruiert wurden, dass sie nur aus Messfehlern bestehen. Die beobachteten Werte aus diesen Tests werden nicht miteinander korrelieren. Die Begründung ist einfach: die Messfehler zweier Tests korrelieren nicht miteinander (gemäß den Annahmen der Klassischen Testtheorie). Grundsätzlich gilt: Je größer der Anteil der Messfehler an den beobachteten Werten ist oder, mit anderen Worten, je niedriger die Reliabilität der Tests ist, desto weniger systematische Varianzanteile bleiben übrig, die miteinander korrelieren können. Anders ausgedrückt bedeutet dies, dass die (Un-)Reliabilität von Tests deren maximal mögliche Korrelation bestimmt (☞ Abb. 2.37).

Aus den Grundannahmen der Klassischen Testtheorie lässt sich eine Formel herleiten, die als doppelte Minderungskorrektur bezeichnet wird. „Doppelt“ bedeutet hier, dass die Reliabilitäten beider Tests berücksichtigt werden.

Fehleranteil der Messung wirkt sich mindernd auf die Korrelation aus

Doppelte Minderungskorrektur

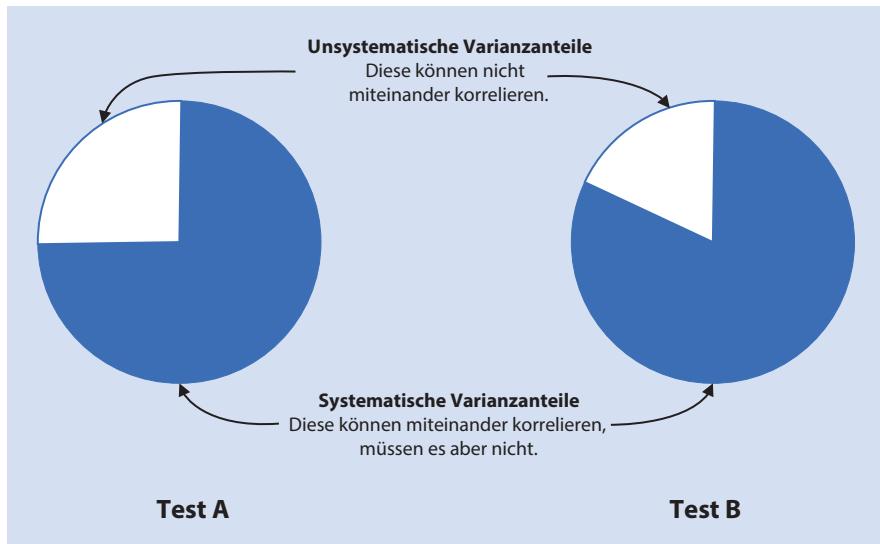


Abb. 2.37 (Un-)Reliabilität von Tests begrenzt deren maximal mögliche Korrelation

Doppelte Minderungskorrektur

$$r_{corr1,2} = \frac{r_{1,2}}{\sqrt{r_{tt1} \times r_{tt2}}}$$

$r_{corr1,2}$ = um die Unreliabilität der Tests 1 und 2 korrigierte Korrelation

$r_{1,2}$ = unkorrigierte Korrelation zwischen den Tests 1 und 2

r_{tt1}, r_{tt2} = Reliabilitäten der Tests 1 und 2

(vgl. Lienert und Raatz 1998, S. 258)

Die Formel gibt an, wie hoch die Korrelation $r_{corr1,2}$ zwischen den wahren Werten von Test (oder Messung) 1 und 2 ausfallen würde. Dazu müssen die Korrelation der beobachteten Werte $r_{1,2}$ beider Tests sowie die Reliabilitäten r_{tt1} und r_{tt2} der Tests bekannt sein.

Ein Beispiel zeigt, dass diese Korrektur beträchtlich ausfallen kann, wenn die Reliabilitäten der miteinander korrelierten Tests gering sind. Nehmen wir an, für beide Tests ist die Reliabilität mit .70 angegeben. Die beobachtete Korrelation sei .30. Dann läge die korrigierte Korrelation bei

$$r_{corr1,2} = \frac{.30}{\sqrt{.70 \times .70}} = .429.$$

Die Formel kann auch für die sog. „einfache Minderungskorrektur“ verwendet werden. Hierbei korrigiert man nur für die (Un-)Reliabilität eines der beiden Tests. Für den anderen Test bzw. die andere Messung geht man daher von einer perfekten Reliabilität aus (man setzt also für r_{tt2} eine 1 ein).

$$r_{corr1,2} = \frac{r_{1,2}}{\sqrt{r_{tt1}}}$$

(vgl. Lienert und Raatz 1998, S. 257)

Korrekturen können deutlich ausfallen

Obergrenze der Korrelation

Die Formel zur Minderungskorrektur zeigt, dass die Korrelation zweier Messungen nicht größer ausfallen kann als die Wurzel aus dem Produkt der beiden Reliabilitätskoeffizienten dieser Messungen. Im Falle des obigen Beispiels

liegt die Obergrenze bei $r = .70$. Betragen beide Reliabilitäten nur .50, liegt die Obergrenze bei .50.

Praktisch bedeutsam werden solche Korrekturen, wenn man beispielsweise Validitätskoeffizienten (Korrelation des Tests mit einem relevanten Kriterium) vergleichen möchte (► Abschn. 2.6.3.4). Solche Koeffizienten sind kaum vergleichbar, wenn sie sich auf Kriterien beziehen, die unterschiedlich genau messbar sind. Man stelle sich 2 Intelligenztests vor, die beide anhand des Schulerfolgs validiert wurden. Test 1 konnte mit der Abiturnote korreliert werden, Test 2 nur mit einer Leistungsbeurteilung durch die Klassenlehrerinnen bzw. -lehrer. Die Abiturnote stellt ein Aggregat mehrerer Einzelnoten dar und wird als solches relativ reliabel sein. Die Reliabilität des Lehrerinnen- bzw. Lehrerurteils wird darunter leiden, dass es sich um eine einzelne Messung handelt und zudem um ein subjektives Urteil. Test 2 wird daher ohne Minderungskorrektur scheinbar weniger valide sein als Test 1. Eine Korrektur der Validitätskoeffizienten um die Reliabilität des Kriteriums schafft einen fairen Vergleich.

Korrektur von
Korrelationskoeffizienten schafft
Vergleichbarkeit

2.6.3 Validität

Die Auffassungen, was unter Validität zu verstehen ist, haben sich in den vergangenen Jahrzehnten immer wieder verändert. Die folgende Definition bezieht sich auf die Standards for Educational and Psychological Testing aus dem Jahr 2014, die gemeinsam von der American Educational Research Association, der American Psychological Association und dem National Council on Measurement in Education herausgegeben werden (AERA et al. 2014).

Definition

Validität bezeichnet das Ausmaß, in dem Evidenz und Theorie die Interpretation von Testwerten rechtfertigen (AERA et al. 2014, S. 11, Übersetzung der Autoren).

Validität wichtigstes
Hauptgütekriterium

Es ist das Ziel jeder Testanwendung, eine sinnvolle Interpretation anzustellen, die zur Beantwortung einer diagnostischen Fragestellung (► Abschn. 1.5) beiträgt. Somit ist Validität als das wichtigste der 3 Hauptgütekriterien (Objektivität, Reliabilität, Validität) anzusehen. Es muss folglich das vorrangige Ziel für Testentwicklerinnen und -entwickler sein, substantielle Evidenz und schlüssige Theorie(n) zu präsentieren, um die intendierten Interpretationen der Testwerte zu untermauern. Es muss umgekehrt die wichtigste Aufgabe von Testanwenderinnen und -anwendern sein, im Testmanual eben solche Evidenz und Theorie zu sichten und zu bewerten (oder sich dabei von unabhängigen Expertinnen und Experten helfen zu lassen).

⚠ Ein Test oder ein anderes diagnostisches Verfahren kann eine sehr hohe Objektivität und eine sehr hohe Reliabilität aufweisen und dennoch für die diagnostische Praxis unbrauchbar sein. Eine hohe Objektivität und Reliabilität sind günstige Voraussetzungen für die Validität von Testwertinterpretationen, aber keine Garantie. Absolut notwendig ist der Nachweis über die Zulässigkeit der anhand des Verfahrens vorgenommenen Interpretationen. Ein vielversprechender Testname oder Aufdrucke wie „klinisch geprüft“, „in der Anwendung erprobt“ oder andere Werbeaussagen können kein Ersatz für den Nachweis von Validität sein.

Validität als Aussage über die
Rechtmäßigkeit der Interpretation
von Testwerten

Die Definition der Validität macht deutlich, dass sie keine Eigenschaft des diagnostischen Instruments (z. B. eines Tests) ist. Aussagen wie „Test A ist valide“ sind demzufolge nicht zulässig. Denn laut Definition ist Validität eine

Eigenschaft der Interpretation von Testwerten. Für viele diagnostische Instrumente ist mehr als nur eine einzige Interpretation der Testwerte denkbar. Beispielsweise können Ergebnisse eines Tests als Aussage über die Ausprägung eines Merkmals interpretiert, zusätzlich aber auch als Eignungsindikator für ein Studium herangezogen werden. Für beide Interpretationen sollte hinreichend Evidenz und Theorie vorliegen. Es kann durchaus sein, dass nur für eine der beiden Interpretationen schlüssige Evidenz vorliegt. Beispielsweise könnte ein kognitiver Test nachweislich die Studienleistung vorhersagen, aber die im Manual berichteten Korrelationen lassen keine eindeutige Zuordnung des Tests zu einer Intelligenzkomponente zu. Somit bleibt unklar, was gemessen wird. Dennoch ist die Interpretation der Ergebnisse im Sinne einer Studieneignung möglich.

! Würde man Tests pauschal als valide bezeichnen, so wäre das vergleichbar mit einem Medikament, dass man pauschal als wirksam bezeichnet. Niemand wird erwarten, dass ein Medikament gegen alle Krankheiten wirkt. Genauso kann ein Test nicht generell valide sein, sondern nur valide für bestimmte Interpretationen.

Unterteilung in 3 Validitätsarten
gebräuchlich

Frühere Auffassung einer Dreiteilung der Validitätsarten in Inhalts-, Konstrukt- und Kriteriumsvalidität wurden durch die genannte, einheitliche Validaitsauffassung abgelöst. An der Annahme von 3 distinkten Validitätsarten wurde vor allem kritisiert, dass diese nicht konsequent voneinander trennbar sind (vgl. Guion 1980; Landy 1986). Trotz der Ablösung dieser Dreiteilung ist sie noch in vielen Testmanualen enthalten und auch für die Beurteilung des Tests nützlich. Daher ist es für Testnutzerinnen und -nutzer nach wie vor wichtig, die 3 historischen Validitätsformen unterscheiden zu können. Wir werden diese auch bei den Testbeschreibungen in ▶ Kap. 3 dieses Buchs verwenden, da sie nach wie vor gebräuchlich sind und in vielen Testmanualen verwendet werden.

Historische Unterscheidung in 3 Validitätsarten

1. Unter *Inhaltsvalidität* verstand man, wie repräsentativ die Items eines Tests für das zu messende Merkmal sind.
2. Unter *Konstruktvalidität* verstand man das Ausmaß, in dem ein Test das Konstrukt erfasst, das er erfassen soll.
3. Unter *Kriteriumsvalidität* verstand man das Ausmaß, in dem das Testergebnis mit konkreten Leistungen oder Verhaltensweisen außerhalb der Testsituation korrespondiert. Das Kriterium muss für den vorgesehenen Einsatzbereich des Tests relevant sein.

Die Unterscheidung zwischen Konstrukt- und Kriteriumsvalidität ist keineswegs so offenkundig, wie es vielleicht zunächst den Anschein haben mag. Ein Intelligenztest, dessen Einsatzzweck nicht die Prognose der Schulleistung ist, mag dennoch hoch mit Schulnoten korrelieren. Wenngleich einige Autorinnen und Autoren dies als Beleg für dessen Kriteriumsvalidität einordnen würden, plädieren wir dafür, dies als Beleg der Konstruktvalidität anzusehen. Die folgenden beiden Gründe sprechen dafür:

1. Die Schulleistung ist kein dem Einsatzbereich des Tests entsprechendes Kriterium.
2. Man weiß, dass Schulleistungen in hohem Maße mit der Intelligenz korrelieren – tut der vorliegende Test dies auch, ist dies ein (wenngleich nicht hinreichender) Beleg, dass er ebenfalls Intelligenz misst.

Es ist zudem wichtig, „Validierung“ von „Validität“ abzugrenzen.

Definition

Validierung beschreibt den Prozess, mit dem Evidenz zur Interpretation von Testwerten generiert wird.

Die aktuellen Standards for Educational and Psychological Testing nennen 4 Quellen, aus denen Belege für die Validität von Testwertinterpretationen generiert werden können. Es muss betont werden, dass diese nicht als distinkte Formen der Validität, sondern als grundsätzliche Möglichkeiten der Validierung aufzufassen sind.

Vier Quellen für Validitätsbelege

Belege für die Validität von Testwertinterpretationen

Belege für die Validität von Testwertinterpretationen können generiert werden anhand

1. des Testinhalts,
2. von Antwortprozessen,
3. der Struktur des Tests,
4. des Zusammenhangs zu anderen Variablen.

2.6.3.1 Analyse des Testinhalts

Testitems sollen den Messgegenstand des Tests repräsentieren. Mit dem früher gebräuchlichen Begriff „Inhaltsvalidität“ wurde die gleiche Forderung aufgestellt. Zur Veranschaulichung bedienen wir uns des Begriffs „Itemuniversum“. Ein Itemuniversum gibt es (vielleicht von wenigen Ausnahmen abgesehen) nicht wirklich. Es handelt sich vielmehr um eine gedankliche Hilfskonstruktion. Wir stellen uns vor, man könne alle Items finden, die zur Messung von Intelligenz, Neurotizismus, Schulangst, Konzentrationsfähigkeit oder eines anderen Merkmals grundsätzlich geeignet sind. Dazu sind genaue Kenntnisse des Merkmals erforderlich. Beispielsweise müsste man wissen, welche Art von Verhaltensweisen, Gedanken, Gefühlen etc. für Narzissten typisch sind. Sinnvollerweise besteht ein Test aus einer repräsentativen Auswahl von Items aus dem Itemuniversum.

Sind die Items repräsentativ für das Merkmal?

Bei einigen Tests ist es besonders wichtig, den Testinhalt systematisch zu evaluieren. Wir haben dies bereits in ► Abschn. 2.4.2.3 für Tests, die nach der kriteriumsorientierten Methode konstruiert werden, betont. Beispielsweise dienen Schulleistungstests dazu, den Wissensstand in einem bestimmten Unterrichtsfach zu erfassen. Was Schülerinnen und Schüler beispielsweise im Fach Mathematik am Ende des 7. Schuljahrs wissen sollen, ist dem Lehrplan oder einem Lehrbuch, das verpflichtend im Unterricht eingesetzt wird, zu entnehmen (vgl. ► Abschn. 2.4.2.3). Die Testautorinnen und -autoren können im Testmanual beschreiben, wie sie vorgegangen sind, um eine repräsentative Auswahl von Unterrichtsinhalten zu finden, zu denen sie dann Items formuliert haben. Dieses Vorgehen und die im Test enthaltenen Items können dann geprüft und der Testinhalt beurteilt werden.

Wenn mit einem Fragebogen das Vorliegen einer bestimmten psychischen Störung, beispielsweise einer Depression, geklärt werden soll, kann die Analyse des Testinhalts wichtig sein. Bei psychischen Störungen wurde mit den verbreiteten Diagnosesystemen DSM-5 (American Psychiatric Association 2015) und ICD-11 (WHO 2018) ein Konsens herbeigeführt, welche Symptome vorliegen müssen, um eine Störung diagnostizieren zu können. Damit wird eine Inspektion des Testinhalts relativ einfach realisierbar: Man prüft,

ob in dem Fragebogen alle für die Störung relevanten Symptome enthalten sind.

Ähnliches gilt übrigens für strukturierte Interviews, wenn sie eingesetzt werden, um die Eignung für einen Beruf zu prüfen (► Abschn. 3.7.1). Bei der Analyse der Fragenauswahl prüft man, ob zuvor eine Anforderungsanalyse durchgeführt wurde und ob zu allen berufsrelevanten Anforderungen Interviewfragen entworfen wurden.

Quantitative Inhaltsanalyse

Colquitt et al. (2019) beschreiben 2 Ansätze zur quantitativen Inhaltsanalyse, die auf Hinkin und Tracey (1999) sowie auf Anderson und Gerbing (1991) zurückgehen. Beide Ansätze gehen davon aus, dass sowohl *definitional correspondence* als auch *definitional distinctiveness* wichtige Aspekte einer Inhaltsanalyse sind. Ersteres meint das Ausmaß, in dem die in einem Fragebogen enthaltenen Items mit der Konstruktdefinition übereinstimmen. Letzteres bezeichnet das Ausmaß, in dem ein Item mehr dem intendierten Konstrukt als einem verwandten Konstrukt zuzuordnen ist. Umgesetzt wird dies, indem Beurteilende gebeten werden, die Zugehörigkeit eines Items zu dem intendierten sowie zu verwandten Konstrukten zu kennzeichnen. Nach Hinkin und Tracey (1999) erfolgt dies auf einer Beurteilungsskala. Anderson und Gerbing (1991) hingegen lassen „nur“ eine dichotome Zuordnung von Items zu Konstrukten vornehmen. In beiden Fällen gilt: Je eher ein Item von Beurteilenden dem intendierten Konstrukt und je weniger es einem verwandten Konstrukt zugeordnet wird, umso besser ist es geeignet. Konkret bedeutet dies, dass Items eines Narzissmusfragebogens nicht nur von (fast) allen Beurteilenden dem Konstrukt Narzissmus zugeordnet werden; idealerweise werden diese Items ebenso von (fast) allen Beurteilenden als nicht zugehörig zu anderen Konstrukten (z. B. Extraversion) bewertet.

Erfolgt eine dichotome Zuordnung (Anderson und Gerbing 1991), können folgende Formeln verwendet werden (zitiert nach Colquitt et al. 2019, S. 1244):

$$\text{definitional correspondence} = \frac{n_c}{N}$$

$$\text{definitional distinctiveness} = \frac{n_c - n_0}{N}$$

n_c = Anzahl richtiger Zuordnungen

n_0 = Höchste Zahl falscher Zuordnungen zu einem der verwandten Konstrukte

N = Anzahl beurteilender Personen

Für eine Zuordnung anhand von Beurteilungsskalen, stehen folgende Formeln zur Verfügung (zitiert nach Colquitt et al. 2019, S. 1249):

$$\text{definitional correspondence} = \frac{\bar{i}}{a}$$

$$\text{definitional distinctiveness} = \frac{\bar{i} - \bar{v}}{a - 1}$$

\bar{i} = Mittlere Zugehörigkeitsbewertung zu intendiertem Konstrukt

\bar{v} = Mittlere Zugehörigkeitsbewertung zu verwandten Konstrukten

a = Höchste Skalenstufe der verwendeten Beurteilungsskala

Colquitt et al. (2019) haben anhand von publizierten Skalen eine Verteilung der zuvor genannten Koeffizienten erstellt und daraus Cut-off-Werte abgeleitet. Demnach kann ein dichotome Zuordnung mit einem Distinktheitskoeffizienten $<.60$ als gut gelten; dasselbe gilt für einen Distinktheitskoeffizienten $>.26$ bei Zuordnung anhand von Beurteilungsskalen (beides ohne Berücksichtigung der Korrelationen zwischen intendiertem und verwandten Konstrukt – hierfür stellen Colquitt et al. je nach mittlerer Korrelationshöhe weitere Tabellen zur Verfügung).

Diese Art der Inhaltsanalyse eignet sich zur Quantifizierung der Übereinstimmung von Items mit dem zu messenden Konstrukt. Das heißt, sie liefert Erkenntnisse, ob oder wie gut die vorhandenen Items zu dem Konstrukt passen. Sie ist aber nicht dazu gedacht, das Fehlen von wichtigen Items zu entdecken. Besteht das Ziel eines Fragebogens nicht darin, ein klar definiertes Konstrukt abzubilden, kann der hier dargestellte Ansatz der definitional correspondence und definitional distinctiveness nicht ohne Weiteres verwendet werden.

- ! In der Regel erfolgt eine Analyse des Testinhalts durch Sichtung des Vorgehens, mit dem Items in den Test aufgenommen wurden, sowie durch Analyse der finalen Itemauswahl. Dies sollte durch Expertinnen und Experten erfolgen.

2.6.3.2 Analyse von Antwortprozessen

Manche Messungen enthalten dezidierte Annahmen über den psychologischen Prozess, der das Antwortverhalten ursächlich beeinflussen soll. So wird beispielsweise in projektiven Verfahren (► Abschn. 3.5) zur Motivmessung angenommen, dass das darin enthaltene Bildmaterial Motive von Testpersonen anregt und deren Antwortverhalten beeinflusst. Eine Analyse des Antwortprozesses könnte darin bestehen, zu prüfen, ob Antworten tatsächlich auf durch das Testmaterial angeregte Motive zurückzuführen sind oder anders zustande kamen.

Psychologischer Antwortprozess

Motivmessung ohne Motiv?

Krumm et al. (2016) haben sich der Frage gewidmet, ob Motivmessungen tatsächlich auf durch Bildmaterial angeregte Motive zustande kommen. Dazu applizierten sie das Multi-Motiv-Gitter (Schmalt et al. 2000) entweder mit oder ohne die zur Motivanregung vorgesehenen Bilder. Testpersonen kreuzten auf den vorgesehenen Antwortvorlagen dieses Tests an, inwiefern motivbezogene Aussagen für sie zutreffen.

Ursprüngliches Testitem



ja nein

Hier kann das eigene Ansehen verloren gehen

Bei diesen Aufgaben an mangelnde spezielle Fähigkeiten denken

Die Macht anderer befürchten

Selber Einfluss haben wollen

Modifizierte Version

Bitte stellen Sie sich eine soziale
Situation aus Ihrem Alltag vor.

ja nein

Hier kann das eigene Ansehen verloren gehen

Bei diesen Aufgaben an mangelnde spezielle Fähigkeiten denken

Die Macht anderer befürchten

Selber Einfluss haben wollen

Das Multi-Motiv-Gitter für Anschluss, Leistung und Macht (MMG) von Heinz-Dieter Schmalt, Kurt Sokolowski und Thomas Langens. ©1999 Swets Test Services GmbH, Frankfurt am Main.

Die Ergebnisse dieser Studie zeigten, dass in 4 der 6 Dimensionen des Multi-Motiv-Gitters keine höhere Motivanregung durch Bilder im Vergleich zur allgemeinen Aufforderung („Bitte stellen Sie sich eine soziale Situation aus Ihrem Alltag vor“) gab. Somit kann der für diesen Test angenommene Antwortprozess infrage gestellt werden.

Neben der experimentellen Manipulation des Testmaterials, das für einen angenommenen Antwortprozess essenziell sein sollte, gibt es weitere Wege, Antwortprozesse zu identifizieren und damit Validitätsbelege zu generieren. Eine Möglichkeit besteht schlicht darin, Testpersonen zu bitten, ihre Kognitionen beim Beantworten eines Tests zu schildern. Dies wird als sog. „Think-aloud-Technik“ beschrieben.

Möglichkeiten, Antwortprozesse zu identifizieren

Think-aloud-Technik

Die Think-aloud-Technik wurde ausführlich von Ericsson und Simon (1984) beschrieben, allerdings bereits früher verwendet, so z. B. von Sargent (1940). Letzterer benutzte diese Technik um kognitive Prozesse beim Lösen von Anagrammen zu identifizieren. Wenn Personen also die Buchstabenfolge E S H C R A P sehen und das daraus zu bildende Wort identifizieren sollten, könnten sie auf die Aufforderung: „Bitte verbalisieren Sie alles, was Ihnen beim Lösen durch den Kopf geht?“ Folgendes antworten:
„P-R-A-S... hm, nein. Vielleicht S-C-H-A-R... auch nicht. Mal sehen, S und P lassen sich kombinieren, ebenso C und H, aha ... S-P-R-A-C-H-E.“
Somit lassen sich Strategien wie „Ausprobieren verschiedener Kombinationen“ oder „Nutzung wahrscheinlicher Buchstabenkombinationen“ identifizieren.

Bei den bereits erwähnten Intelligenztests, deren Distraktoren wertvolle Hinweise auf die richtige Lösung geben (s. „Die verflixten Distraktoren“ in ▶ Abschn. 2.4.2.6), könnte die Think-aloud-Technik darüber Aufschluss geben, ob Testpersonen – wie vom Test intendiert – Aufgaben analysieren, die der Aufgabe inhärente Regel ableiten und auf die Distraktoren anwenden oder schlicht die Distraktoren anhand der in ▶ Abschn. 2.4.2.6 genannten Strategie durcharbeiten. Neben der Think-aloud-Technik könnten auch Analysen der Blickbewegungen helfen, den Antwortprozess zu identifizieren. Würden Testpersonen fast ausschließlich die Distraktoren inspizieren, müsste der eigentlich intendierte Antwortprozess verworfen werden.

Analyse der Blickbewegungen bei Intelligenztests

Diese Quelle für Validitätsbelege wird von Testautorinnen und -autoren leider selten genutzt. Verschiedene Studien zeigen jedoch, dass sich Testpersonen nicht automatisch so verhalten, wie Testautorinnen und -autoren es gerne möchten. So zeigen Studien aus dem Bereich des Leseverständnisses, dass man manche Leseverständnisaufgaben (z. B. aus PISA, s. nachfolgendes Beispiel) auch lösen kann, ohne den eigentlichen Text gelesen zu haben (Sparfeldt et al. 2012). Für Situational-Judgment-Tests (▶ Abschn. 6.2.1.2) ist mittlerweile bekannt, dass gute Testergebnisse auch ohne Kenntnis der infrage stehenden Situation zu erzielen sind (Krumm et al. 2015). Die jeweils intendierten Prozesse sind also nicht essenziell für eine erfolgreiche Bearbeitung der Tests. Verschiedene Autoren fordern einen stärkeren Fokus auf diese Quelle für Validitätsbelege und beschreiben nützliche Vorgehensweisen (z. B. Bornstein 2011).

Testpersonen verhalten sich nicht automatisch wie intendiert

Beispielitem aus dem PISA-Leseverständnistests

Wissenschaftliche Waffen der Polizei

Ein Mord wurde begangen, aber der Verdächtige streitet alles ab. Er behauptet, das Opfer nicht zu kennen. Er sagt, er habe ihn nie gekannt, sei nie in seiner Nähe gewesen, hätte ihn nie angerührt ... Polizei und Justiz sind überzeugt, dass er nicht die Wahrheit sagt. Aber wie ist es zu beweisen?

Am Tatort haben die Ermittlungsbeamten jede noch so kleine denkbare Spur und mögliche Beweistücke zusammengetragen: Gewebefasern, Haare, Fingerabdrücke, Zigarettenstummel ... Die wenigen auf dem Jackett des Opfers gefundenen Haare sind rot. Und sie sehen denen des Verdächtigen merkwürdig ähnlich. Wenn es bewiesen werden könnte, dass diese Haare tatsächlich von ihm stammen, wäre das ein Beweis, dass er dem Opfer doch begegnet ist.

Jedes Individuum ist einzigartig

Die Spezialisten gehen an die Arbeit. Sie untersuchen einige Zellen an der Haarwurzel und ein paar Blutzellen des Verdächtigen. Im Kern jeder Zelle unseres Körpers befindet sich DNS. Was ist das? Die DNS ist wie eine Kette aus zwei umeinander gedrehten Perlenketten. Stelle dir vor, dass

diese Perlen in vier verschiedenen Farben vorkommen und tausende von Perlen (aus denen ein Gen besteht) in einer ganz bestimmten Reihenfolge aufgezogen sind. Bei jedem einzelnen Individuum ist diese Reihenfolge in allen Zellen des Körpers gleich: die von den Haarwurzeln genauso wie die vom großen Zell-, von der Leber sowie der Magens oder des Blutes. Aber die Reihenfolge der Perlen ist bei jedem Menschen anders. Die Wahrscheinlichkeit, dass zwei Menschen die gleiche DNS haben, ist angesichts der Anzahl derart aufgezogener Perlen sehr gering, mit Ausnahme von einigen Zwillingen. Einzigartig für jedes Individuum, ist die DNS damit eine Art genetischer Personalausweis.

Die Genetiker können deshalb den (in seinem Blut festgelegten) genetischen Personalausweis des Verdächtigen mit dem der rothaarigen Person vergleichen.

Wenn der genetische Personalausweis derselbe ist, wissen sie, dass der Verdächtige doch in der Nähe des Opfers war, dem er angeblich nie begegnet ist.

Nur ein Beweisstück

Immer häufiger lässt die Polizei bei sexuellen Vergehen, Mord, Diebstahl oder anderen Verbrechen genetische Analysen durchführen. Warum? Um zu versuchen, Beweise dafür zu finden, dass zwei Menschen, zwei Gegenstände oder ein Mensch und ein Gegenstand miteinander in Berührung gekommen sind. Der Nachweis eines solchen Kontakts ist für die Ermittlungen oft sehr nützlich. Er liefert aber nicht unbedingt den Beweis für ein Verbrechen. Er ist nur ein Beweisstück unter vielen anderen.

Anne Versailles

Genetischer WAS?

Die DNS besteht aus mehreren Genen, von denen jedes aus Tausenden von „Perlen“ gebildet wird. Zusammen bilden diese Gene den genetischen Personalausweis eines Menschen.

Wie findet man den genetischen Personalausweis?

Der Genetiker nimmt die wenigen Zellen von den Wurzeln der Haare, die bei dem Opfer gefunden wurden, oder aus dem Speichel, der an einem Zigarettenstummel haftet. Er taucht sie in eine Substanz, die alles zerstört, was sich um die DNS dieser Zellen herum befindet. Dasselbe macht er dann mit einigen Zellen aus dem Blut des Verdächtigen. Die DNS wird dann speziell für die Analyse vorbereitet. Danach kommt sie in ein spezielles Gel, und durch das Gel wird elektrischer Strom geleitet. Nach ein paar Stunden entstehen dadurch Streifen, ähnlich wie bei einem Strichcode (wie auf Waren, die wir kaufen), die unter einer speziellen Lampe sichtbar werden. Den Strichcode der DNS des Verdächtigen vergleicht man dann mit dem der Haare, die bei dem Opfer gefunden wurden.



Mikroskop in einem Polizeilabor

Frage: Um die Struktur der DNS zu erklären, spricht der Autor von einer Perlenkette. Wodurch unterscheiden sich diese Perlenketten bei verschiedenen Menschen?

- Sie sind von unterschiedlicher Länge.
- Die Reihenfolge der Perlen ist unterschiedlich.
- Die Anzahl der Ketten ist unterschiedlich.
- Die Farbe der Perlen ist unterschiedlich.

(Aus OECD 2000; ► <https://www.oecd.org/berlin/39803735.pdf>, © OECD)

2.6.3.3 Analyse der Struktur**Faktorenanalyse als Validitätsbeleg**

Eine Analyse der inneren Struktur kann helfen, zu prüfen, ob ein Test die theoretische angenommene Grundlage sinnvoll abbildet. Wenn ein Test ein eindimensionales Konstrukt erfassen soll, so sollten die darin enthaltenen Items zueinander so viel Ähnlichkeit aufweisen, dass sie durch nur einen Faktor zu beschreiben sind. Geht die Theorie hingegen von 2 korrelierten Dimensionen aus, so sollte sich ebendies auch empirisch zeigen lassen. In Testmanualen werden solche Prüfungen häufig unter der Überschrift „faktorielle Validität“ behandelt und als ein Beitrag zur Konstruktvalidität angesehen.

Wurde zur Analyse der Struktur eine explorative Faktorenanalyse verwendet (zum Prinzip der Faktorenanalyse s. ► Abschn. 2.5.4), so finden sich Belege für die intendierte Interpretation von Testwerten in der „Natur“ und in der Zahl der extrahierten Faktoren. Die „Natur“ der extrahierten Faktoren lässt sich durch Inspektion der Items, die hoch auf den Faktoren laden, identifizieren. Die

Explorative Faktorenanalyse

Zahl der zu extrahierenden Faktoren wird durch sog. „Extraktionskriterien“ indiziert.

Im Rahmen von konfirmatorischen Faktorenanalysen kann die durch die Theorie implizierte Struktur eines Tests a priori spezifiziert werden. Es kann dann geprüft werden, ob die tatsächlichen Korrelationen der Testitems zu den durch die implizierte Struktur erwarteten Korrelationen passt. Ist dies der Fall, kann das als ein Beleg für die intendierte Interpretation der Testergebnisse gewertet werden.

2.6.3.4 Analyse der Zusammenhänge zu anderen Variablen

Validitätsbelege lassen sich auch dadurch generieren, dass Zusammenhänge eines Tests zu anderen Variablen inspiziert werden. Mit „andere Variablen“ können andere Tests sowie Gegebenheiten des realen Lebens gemeint sein. Soll der infrage stehende Test das Konstrukt Narzissmus erfassen, so ist eine naheliegende Forderung, dass er mit anderen, ggf. bereits etablierten Narzissmusfragebögen hoch korreliert. Ist das Ziel die Vorhersage einer Gegebenheit im Alltag (z. B. Erfolg als Führungskraft oder Teilnahme an einer „Narzissmustherapie“), so sollten die Ergebnisse des infrage stehenden Tests diese Gegebenheiten statistisch vorhersagen können. Auf beide Forderungen gehen wir in der Folge näher ein.

In einer klassischen Arbeit haben Cronbach und Meehl (1955) argumentiert, dass Konstrukte (z. B. Gewissenhaftigkeit, Narzissmus, Intelligenz) in ein nomologisches Netzwerk integriert werden müssen. Darunter verstehen sie Annahmen über die Beziehung des Zielkonstrukts (z. B. Narzissmus) zu anderen Konstrukten (z. B. Extraversion, Verträglichkeit, Offenheit für Erfahrungen). So könnten Testentwicklerinnen und -entwickler annehmen, dass Narzissmus einen engen (negativen) Zusammenhang mit Verträglichkeit, einen moderaten Zusammenhang mit Extraversion und keinen Zusammenhang mit Offenheit für Erfahrungen aufweist. Sofern die entsprechenden Messungen (durch einen Narzissmus-, einen Extraversions-, einen Verträglichkeits- und einen Offenheitsfragebogen) diese Konstrukte adäquat abbilden, sollten die beobachteten Testwerte in einem ähnlichen Zusammenhang stehen, wie er für die Konstrukte angenommen wurde. Lassen sich für einen neu entwickelten Test die so abgeleiteten Zusammenhänge nicht zeigen, misst der Test entweder das Konstrukt nicht (oder nur teilweise) oder die angenommenen Zusammenhänge waren theoretisch nicht gut abgeleitet.

Es wird deutlich, dass Testmanuale eine umfangreiche Einordnung des zu messenden Konstrukts in ein breiteres Gefüge anderer, mehr oder weniger inhaltlich naher Konstrukte enthalten müssen. Nur so kann empirische Evidenz in Form von Korrelationen mit anderen Tests sinnvoll beurteilt werden. Die Wahl der theoretisch nahen und weniger nahen Konstrukte muss theoretisch begründet sein – Testentwicklerinnen und -entwickler benötigen also ein gutes Verständnis der Zusammenhänge zwischen relevanten Konstrukten.

Es wird außerdem deutlich, dass hohe Korrelationen weder immer gut sind noch grundsätzlich als Beleg für die Validität der intendierten Testwertinterpretation gelten können. Zwar erwartet man hohe Korrelationen zwischen dem zu validierenden Test und einem anderen Test mit gleichem Messanspruch (also dem neuen Narzissmustest und einem bereits etablierten Narzissmustest) – hierbei spricht man auch von konvergenten Validitätsbelegen. Das nomologische Netzwerk enthält jedoch auch Konstrukte, die dem Zielkonstrukt inhaltlich wenig nahe sind (z. B. Narzissmus und Offenheit für Erfahrungen). Sinnvollerweise werden zwischen den entsprechenden Messungen niedrige Korrelationen erwartet. Hierbei spricht man auch von diskriminanten Validitätsbelegen. Für eine gelungene Testvalidierung ist zu erwarten,

Konfirmatorische Faktorenanalyse

Neues Instrument sollte mit etablierten Instrumenten gleichen Messanspruchs hoch korrelieren

Nomologisches Netzwerk

Verständnis der Zusammenhänge zwischen relevanten Konstrukten notwendig

Konvergente und diskriminante Validitätsbelege

dass die Zusammenhänge zwischen als „konvergent“ angenommenen Messungen deutlich höher ausfallen als die als „diskriminant“ erachteten.

2

Wozu diskriminante Validitätsbelege?

Man mag sich fragen, warum es nicht ausreicht, Zusammenhänge zu Verfahren mit gleichem Messanspruch zu prüfen und als Validitätsbeleg zu nutzen? Die Bedeutung diskriminanter Validitätsbelege lässt sich anhand der folgenden fiktiven Korrelationsmatrix erläutern. Wir nehmen an, für einen Narzissmusfragebogen ist folgendes Ergebnis einer Korrelationsstudie verfügbar:

	1	2	3	4
1. Narzissmusfragebogen (neu entwickelt)				
2. Narzissmusfragebogen (etabliert)	.50			
3. Verträglichkeitsfragebogen (etabliert)	-.48	-.23		
4. Extraversionsfragebogen (etabliert)	.14	.05	.08	

Würde nur die Korrelation zwischen den beiden Narzissmusfragebögen zur Verfügung stehen ($r=.50$), würde man diese möglicherweise als Beleg für eine gelungene Validierung interpretieren. Die Hinzunahme des Verträglichkeitsfragebogens macht jedoch deutlich, dass nicht klar zu sagen ist, ob der neu entwickelte Fragebogen nun Narzissmus oder (Un-)Verträglichkeit misst. Die absolute Höhe der Korrelation mit $r=|-.48|$ liegt kaum unter der zwischen den beiden Narzissmusfragebögen ($r=.50$). Es wird zudem offenkundig, dass der bereits etablierte Narzissmusfragebogen mit $r=|-.23|$ deutlich besser von Verträglichkeit abgrenzen ist.

Unterschiedliche Methoden verringern in der Regel die Korrelationen

Gleiche Methoden erhöhen in der Regel die Korrelationen

Konstrukte und Messmethode beachten

Es ist zu beachten, dass Messungen nicht nur divergieren können, weil ihnen unterschiedliche Konstrukte zugrunde liegen (also z. B. Narzissmus vs. Offenheit für Erfahrungen). Messungen können auch divergieren, weil sie anhand unterschiedlicher Methoden vorgenommen wurden (z. B. Fragebogen vs. Interview).

Analog dazu können eigentlich als divergent erachtete Messungen künstlich erhöhte Zusammenhänge aufweisen, weil ihnen die gleiche Messmethode zugrunde lag (also beide Merkmale etwa per Interview erfasst wurden). Sofern – wie im gerade präsentierten Beispiel der Korrelationsmatrix – nicht die gleiche Methode über alle Messungen verwendet wurde, sind Methodeneffekte bei der Interpretation der Höhe der Korrelationen unbedingt zu berücksichtigen. Man spricht dann von „gemeinsamer Methodenvarianz“, die für eine unerwartet hohe Korrelation verantwortlich gemacht wird.

In dem in Abb. 2.38 skizzierten nomologischen Netz ist die zu validierende Messung durch einen dick umrahmten Kasten gekennzeichnet. Ihr Messanspruch ist „Narzissmus“, als Methode wurde ein Fragebogen entwickelt. Es kann sein, dass dieser Fragebogen zu der Messung mit gleichem Messanspruch oder zu der Messung mit ähnlichem Messanspruch „Verträglichkeit“ nur moderate oder geringe Korrelationen zeigt – diese also anders ausfallen als auf Konstruktebene erwartet wurde. Dies könnte daran liegen, dass die beiden letztgenannten Messungen mit einer anderen Methode, einem Interview, durchgeführt wurden. Zudem kann es sein, dass der zu validierende Fragebogen zu den Messungen mit dem Messanspruch „Offenheit“ und „Gewissenhaftigkeit“ eine höhere Korrelation aufweist als theoretisch erwartet, da alle Messungen durch Fragebögen vorgenommen wurden. Es ist

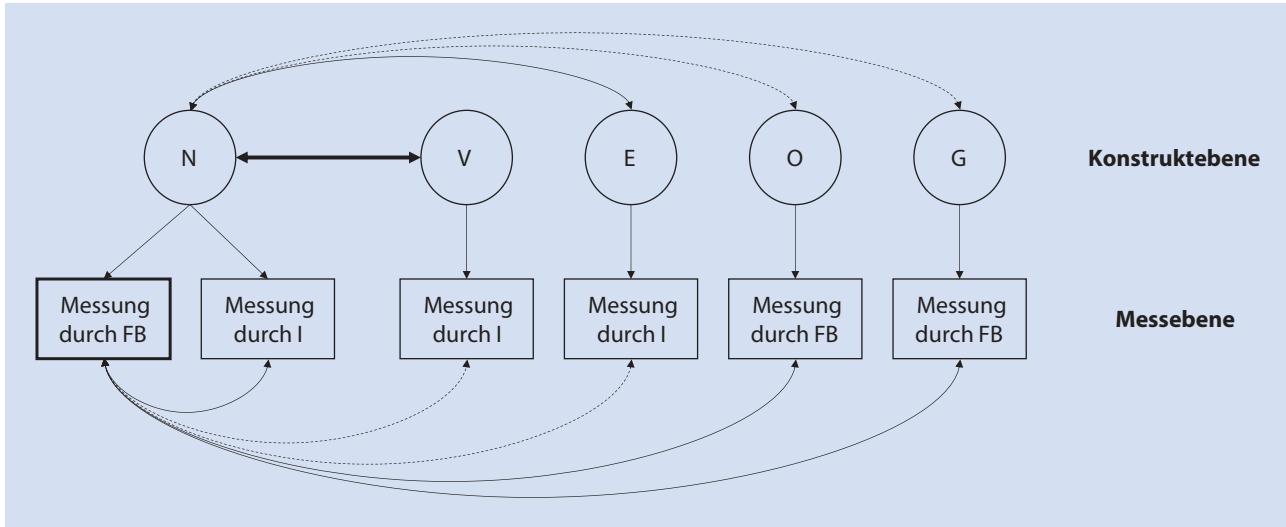


Abb. 2.38 Beispielhaftes nomologisches Netz. Die Dicke und die Art der Striche geben die Höhe der theoretisch erwarteten (Konstruktebene) und die Höhe der tatsächlich beobachteten Korrelationen (Messebene) an. FB = Fragebogen, I = Interview, N = Narzissmus, V = Verträglichkeit, E = Extraversion, O = Offenheit für Erfahrungen, G = Gewissenhaftigkeit

also zentral, Korrelationen vor dem Hintergrund der Konstrukte *und* der Messmethoden zu interpretieren.

Eine einfache Möglichkeit, Methodeneffekte im Rahmen der Testvalidierung zu berücksichtigen, bietet der sog. „Multitrait-Multimethod-Ansatz“ (MTMM-Ansatz; Campbell und Fiske 1959). Er sieht vor, dass alle in einer Validierungsstudie berücksichtigten Konstrukte möglichst mit mehreren Methoden erfasst werden. Die daraus resultierenden Korrelationen werden dann nach einer einfachen Systematik sortiert und bewertet: Statt nur nach „konvergent“ und „diskriminant“ zu unterscheiden, werden Korrelationen zusätzlich nach „gleiche Methoden“ und „ungleiche Methoden“ sortiert. Aus der Kombination dieser Unterscheidungen entstehen die 4 Kategorien der MTMM-Analyse (► Tab. 2.17).

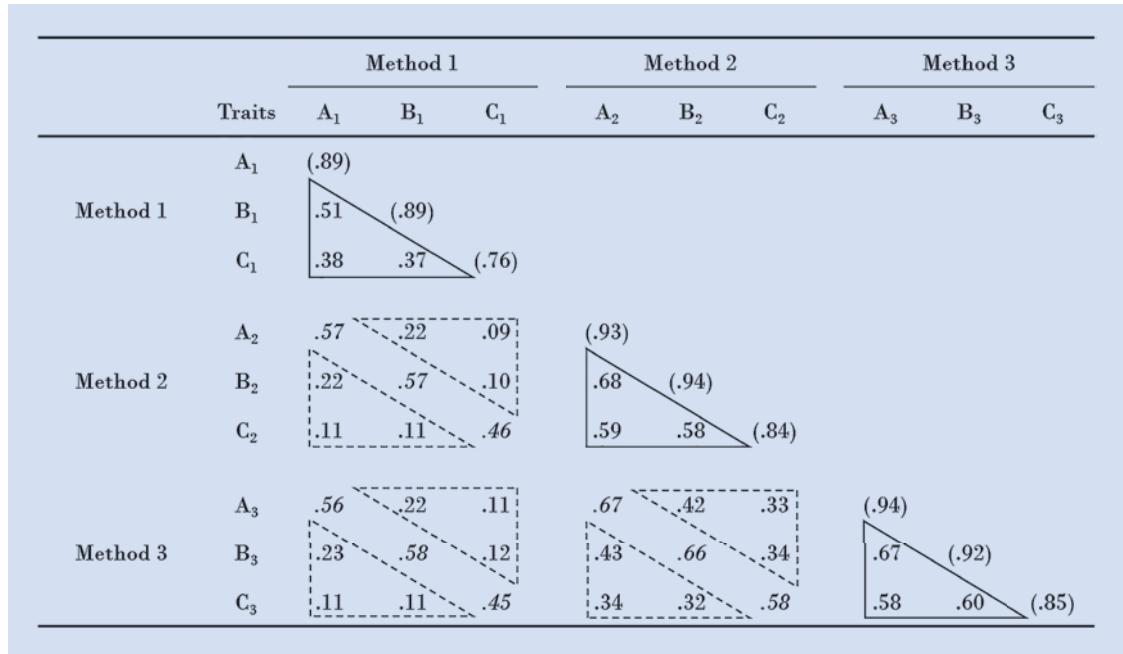
Im Rahmen einer Korrelationsmatrix können dann die entsprechenden Korrelationen systematisch verglichen werden. ► Abb. 2.39 ist der Originalpublikation von Campbell und Fiske (1959) entnommen. Darin sind Korrelationen zwischen 3 verschiedenen Traits und 3 verschiedenen Methoden dargestellt. „Echte“ Monotrait-Monomethod-Korrelationen gibt es darin nicht, da keiner der Traits 2× mit der gleichen Methode erfasst wurde. Da wir Reliabilität bereits als Korrelation eines Tests mit sich selbst kennengelernt haben (► Abschn. 2.2.2), sind die Werte in die Reliabilitätsdiagonale in ► Abb. 2.39 (Werte in Klammern) anstelle echter Monotrait-Monomethod-Werte eingetragen. Die mit durchgezogener Linie umrandeten Dreiecke direkt unter der Reliabilitätsdiagonalen kennzeichnen die Heterotrait-Monomethod-Werte. Innerhalb der „Heteromethod-Blöcke“ (z. B. alle Korrelationen zwischen Me-

Multitrait-Multimethod-Ansatz

Korrelationsmatrix zum
systematischen Vergleich

Tab. 2.17 Bezeichnungen der Korrelationen im Rahmen der MTMM-Analyse

	Gleiche Messmethode	Ungleiche Messmethode
Konvergent	Monotrait-Monomethod-Korrelationen	Monotrait-Heteromethod-Korrelationen
Diskriminant	Heterotrait-Monomethod-Korrelationen	Heterotrait-Heteromethod-Korrelationen



■ Abb. 2.39 MTMM-Matrix. (Aus Campbell und Fiske 1959, S. 82)

thode 1 und Methode 2) befinden sich ebenfalls Diagonalen (kursiv gedruckte Werte) – dies sind Monotrait-Heteromethod-Werte. Die Diagonale wird auch als „Validitätsdiagonale“ bezeichnet, da sie im Sinne konvergenter Validitätsbelege zu interpretieren ist. Die Validitätsdiagonalen befinden sich jeweils zwischen zwei gestrichelt umrandeten Dreiecken. Die gestrichelten Dreiecke kennzeichnen Heterotrait-Heteromethod-Werte, sie markieren – ebenso wie die Heterotrait-Monomethod-Werte – divergente Validitätsbelege.

Belege für eine „gelungene“ Validierung definieren Campbell und Fiske (1959, S. 82–83) wie folgt:

1. Werte in den Validitätsdiagonalen sollten signifikant von 0 verschieden sein.
2. Werte in den Validitätsdiagonalen sollten größer sein als alle Werte in der gleichen Zeile und Spalte, die in den Heterotrait-Heteromethod-Dreiecken liegen.
3. Eine Variable sollte mit Messungen des gleichen Traits, für die aber eine andere Methode verwendet wurde (Monotrait-Heteromethod-Werte), höher korrelieren als mit Messungen eines anderen Traits, für den die gleiche Methode verwendet wurde (Heterotrait-Monomethod-Werte).
4. Die Zusammenhänge zwischen unterschiedlichen Traits sollten das gleiche Muster der Zusammenhänge zeigen, und zwar sowohl in den Monomeethod- als auch in den Heteromethod-Blöcken.

Einfluss des Merkmals soll Einfluss der Methode übertreffen

Zusammenfassend lässt sich sagen, dass sich gelungene Messungen dadurch auszeichnen, dass sie hauptsächlich die intendierten Merkmale (Traits) erfassen und der Einfluss der Methode zwar vorhanden, aber gering oder zumindest geringer als der Einfluss der Merkmale ist. Somit gilt für die 4 Kategorien der MTMM-Matrix, dass sich Korrelationen in ihrer Höhe wie folgt sortieren lassen sollten: Monotrait-Monomethod-Korrelationen > Monotrait-Heteromethod-Korrelationen > Heterotrait-Monomethod-Korrelationen > Heterotrait-Heteromethod-Korrelationen.

Validierung eines Angstfragebogens nach der MTMM-Systematik

Folgende Matrix zeigt Korrelationen eines neu entwickelten Angstfragebogens mit anderen Fragebögen und Interviews. Die jeweiligen Kategorien der MTMM-Systematik sind durch unterschiedliche Grautöne gekennzeichnet:

	1	2	3	4	5	6	7	8
1. Angstfragebogen neu	(.78)							
2. Angstfragebogen (etabliert)	.60	(.79)						
3. Depressivitätsfragebogen	.35	.28	(.78)					
4. Belastbarkeitsfragebogen	.25	.32	.30	(.80)				
5. Angstinterview A	.30	.35	.18	.15	(.81)			
6. Angstinterview B	.50	.45	.22	.25	.60	(.78)		
7. Depressivitätsinterview	.30	.10	.38	.17	.35	.25	(.77)	
8. Belastbarkeitsinterview	.24	.16	.23	.42	.32	.28	.30	(.80)



Monotrait-Monomethod
 Heterotrait-Monomethod
 Monotrait-Heteromethod
 Heterotrait-Heteromethod

Wie man sieht, müssen sich Korrelationen im Rahmen einer MTMM-Matrix nicht immer in Form von Dreiecken anordnen. Dadurch, dass wir hier „echte“ Monotrait-Monomethod-Korrelationen inkludiert haben, entstehen keine Dreiecke und Validitätsdiagonalen mehr. Dennoch ist das Prinzip der MTMM-Matrix beibehalten worden und anzuwenden.

Korrelationen, die zwischen gleichen Konstrukten – gemessen mit gleichen Methoden – ermittelt wurden (Monotrait-Monomethod-Werte) fallen mit $r=.60$ hoch aus. Eine mittlere Korrelation von $r=.40$ ist zwischen unterschiedlichen Methoden (Fragebogen vs. Interview) mit gleichem Messanspruch (Monotrait-Heteromethod-Werte) zu beobachten. Diese ist höher als die mittlere Korrelation von $r=.30$ zwischen unterschiedlichen Methoden mit unterschiedlichem Messanspruch (Heterotrait-Monomethod-Werte). Im Detail fällt jedoch auf, dass der neu entwickelte Angstfragebogen mit dem Depressivitätsfragebogen höher korreliert als mit dem Angstinterview (A). Korrelationen zwischen unterschiedlichen Konstrukten – gemessen mit unterschiedlichen Methoden – fallen im Mittel mit $r=.20$ gering aus. Im Großen und Ganzen liefert diese Validierung auf Basis von Zusammenhängen mit anderen Variablen recht überzeugende Belege für die intendierte Interpretation der Testwerte des neuen Angstfragebogens.

Selbstverständlich stehen auch elaboriertere Methoden zur Separierung von Merkmals- und Methodeineinflüssen zur Verfügung. Interessierte Leserinnen und Leser seien hierzu auf Eid et al. (2008) verwiesen.

Tests werden auch zu dem Zweck konstruiert, bestimmte Leistungen im Alltag vorherzusagen. So werden beispielsweise Intelligenztests zur Vorhersage von Schul- oder Ausbildungserfolg verwendet. In diesem Fall bezeichnet man den Schul- bzw. Ausbildungserfolg als Kriterium. Als Kriterium wird überwiegend etwas Konkretes, direkt Messbares und Relevantes aus dem Alltag verstanden – etwa Prüfungsleistungen, erzielter Umsatz,

■ Tab. 2.18 Beispiele für Kriterien zur Validierung von Tests

Diagnostisches Verfahren (Verwendungszweck)	Mögliche Kriterien	Begründung
Allgemeine Psychopathologie	Dauer des Aufenthaltes in einer psychiatrischen Klinik	Je ausgeprägter die Psychopathologie ist, desto länger sollte die Behandlung im Krankenhaus dauern
Intelligenztest (soll Schulerfolg vorhersagen)	Abiturnote mehrere Jahre nach Testdurchführung	Die Abiturnote ist ein anerkanntes Maß für Schulerfolg; da prognostische Validität angestrebt wird, muss das Kriterium deutlich später erhoben werden
Aufmerksamkeitstest (soll Fahreignung erfassen)	Fehler in einer standardisierten Fahrprobe	Aufmerksamkeitsdefizite sollten sich in bestimmten Fehlern wie Übersehen von Verkehrszeichen, Gefahren oder der Geschwindigkeitsanzeige im Auto niederschlagen. Das Verhalten sollte im Straßenverkehr erfasst werden, weil der Test für diesen Bereich eingesetzt wird

Vorgesetztenbeurteilungen oder auch Verhaltensweisen wie Studienabbruch, Reduktion der Nahrungsaufnahme, Alkoholkonsum, Begehen einer Straftat. Was als Kriterium infrage kommt, ergibt sich aus der diagnostischen Zielsetzung des Tests. Wenn Testautorinnen und -autoren angeben, ihr Fragebogen solle Alkoholismus messen, so kann das Verfahren am tatsächlichen Alkoholkonsum als Kriterium validiert werden. ■ Tab. 2.18 enthält einige Beispiele für mögliche Kriterien. Kriterien dürfen keinesfalls beliebig sein; die Testautorinnen und -autoren sollen begründen, warum sie ein bestimmtes Kriterium gewählt haben.

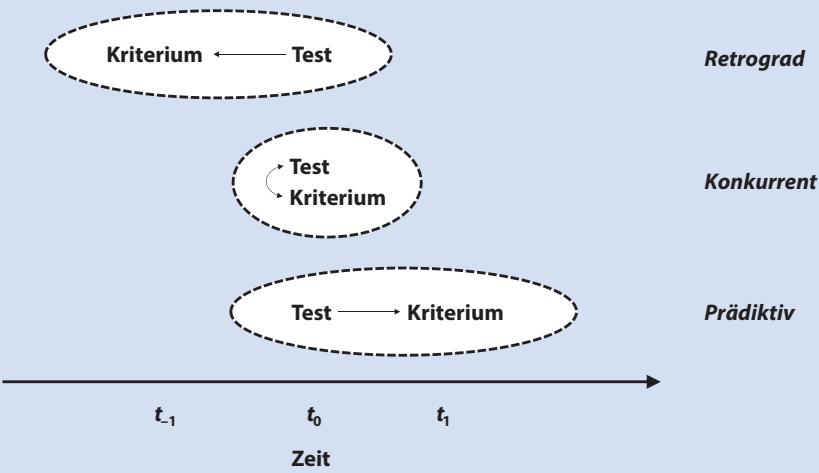
Prädiktive Validitätsbelege

Die Erhebung des Kriteriums kann im gleichen Zeitraum erfolgen wie die Testdurchführung, aber auch deutlich später. Man spricht in diesem Zusammenhang von *konkurrenten bzw. prädiktiven Validitätsbelegen* (s. nachfolgenden Kasten: Zeitpunkt des Entstehens der Kriteriumsdaten). Wenn ein Test dafür geeignet sein soll, spätere Leistungen vorherzusagen, beispielsweise den Berufserfolg einige Jahre nach der Einstellung, so sind prädiktive (auch manchmal bezeichnet als prognostische) Validitätsbelege erforderlich. Würde der Berufserfolg zeitgleich mit dem Test erhoben, so wäre dies ein schlechter Ersatz für eine prognostische Studie und damit ein schwacher Validitätsbeleg. Bei der prädiktiven Form der Validierung ist zu beachten, dass eine längere Prognosedauer in der Regel mit einer geringeren Prognosegüte einhergeht (engl. validity degradation; Dahlke et al. 2018). Manchmal sind Kriterien bereits vor der eigentlichen Testentwicklung und damit natürlich auch vor der durchgeführten Validierungsstudie entstanden. Bei Validierungen von Intelligenztests ist dies häufig der Fall – ein gängiges Kriterium sind Schulnoten, die in Zeugnissen dokumentiert sind. Sie sind vor der Validierungsstudie entstanden. Man spricht hierbei von *retrograden Validitätsbelegen*.

Zeitpunkt des Entstehens der Kriteriumsdaten

Validitätsbelege anhand von Zusammenhängen zu Kriteriumsdaten können, je nach Zeitpunkt der Entstehung der Kriteriumsdaten, bezeichnet werden als

- retrograd,
- konkurrent,
- prädiktiv.



Die Wahl der jeweiligen Alternative muss vor dem Hintergrund des Messanspruchs und pragmatischer Überlegungen getroffen werden. So impliziert eine prädiktive Validierung, dass man warten muss, bis die entsprechenden Daten entstehen.

Aus einer pragmatischen Perspektive haben Testanwenderinnen und -anwender ein starkes Interesse daran, das interessierende Kriterium möglichst umfassend aufzuklären. Meist kommen dazu mehrere diagnostische Verfahren zum Einsatz, die sich in der Vorhersage des Kriteriums ergänzen. Entscheidend ist dabei, welcher Zuwachs an Prognosekraft mit einem weiteren Verfahren erzielt wird. Ist dieser Zuwachs substanziell, so kann dies als *inkrementeller Validitätsbeleg* interpretiert werden. Das zusätzliche Verfahren erfasst in diesem Fall einen bisher noch nicht berücksichtigten Aspekt des Kriteriums. Inkrementelle Validitätsbelege liegen auch vor, wenn ein neues Verfahren gegenüber seinem Vorgänger oder einem konkurrierenden Verfahren mit gleichem Messanspruch zusätzliche Kriteriumsvarianz aufklärt.

Inkrementeller Validitätsbeleg

Inkrementelle Validitätsbelege

Intelligenztests korrelieren in etwa zu $r=.51$ mit Berufserfolg (nach Korrektur für Varianzeinschränkung in Intelligenztests sowie nach Korrektur für Varianzeinschränkung und Unreliabilität im Kriterium Berufserfolg). Das wissen wir aus umfangreichen Metaanalysen (Schmidt und Hunter 1998). Strukturierte Interviews sind ein ebenso guter Prädiktor des Berufserfolgs ($r=.51$; Schmidt und Hunter 1998). Es wäre aber naiv zu glauben, dass sich die Korrelationen einfach addieren lassen. Im Gegenteil: Würden Eignungsinterviews nichts anderes als Intelligenz messen, könnten sie die mit Intelligenztests erzielte Voraussage überhaupt nicht verbessern. Anders gesagt: Je deutlicher ein Prädiktor mit einem anderen überlappt (korreliert), desto weniger wirkt er als inkrementeller Prädiktoren über den anderen hinaus. Tatsächlich leistet eine Kombination von Intelligenztest und strukturiertem Eignungsinterview eine Varianzaufklärung für Berufserfolg von $R=.63$ (Schmidt und Hunter 1998; R steht für eine multiple Korrelation). Der inkrementelle Zuwachs des Interviews ist mit $\Delta R=.12$ also beträchtlich, entspricht aber nicht der Addition beider Prädiktoren.

Wenn ein Test nicht oder nur in geringem Ausmaß mit dem intendierten Kriterium (z. B. Studienerfolg) korrespondiert, muss geschlussfolgert werden, dass entsprechende Interpretationen des Testergebnisses (z. B. Eignung für ein Studium) nicht zulässig sind.

Die Varianzaufklärung durch ein diagnostisches Verfahren mag oftmals gering erscheinen. Ein Vergleich mit Korrelationen aus anderen Disziplinen ist sehr aufschlussreich. Meyer et al. (2001) haben Korrelationen und andere Effektstärken (die sie zur Vergleichbarkeit in Korrelationen umgerechnet haben) aus anderen Disziplinen gesichtet und sie Korrelationen aus dem Bereich der Psychologischen Diagnostik gegenübergestellt. Sie konnten sich dabei meist auf große Metaanalysen beziehen. Die Ergebnisse sind zum Teil verblüffend. Beispielsweise nehmen Millionen von Menschen Azetylsalizylsäuretabletten (z. B. Aspirin) ein, um ihr Blut zu verdünnen und damit einem Herzinfarkt vorzubeugen. Das Risiko, an einem Herzinfarkt zu sterben, wird durch Azetylsalizylsäure nur wenig gemindert; die Korrelation beträgt gerade einmal .02. Alkohol macht aggressiv – das klingt stimmig. Der Effekt entspricht aber nur einer Korrelation von .23. Gemessen an diesen und weiteren Korrelationen kann sich die Psychologische Diagnostik sehen lassen. Ausgewählte Ergebnisse dieser Studie sind in □ Tab. 2.19 aufgeführt. Die Korrelationen sind nach ihrer Größe geordnet; Ergebnisse aus der Psychologischen Diagnostik wurden kursiv hervorgehoben. Es ist unschwer zu erkennen, dass sich scheinbar niedrige Korrelationskoeffizienten im Vergleich mit anderen als beachtliche Größen entpuppen.

Natürlich kann man Effektstärken oder Korrelationen nicht mit Nutzen gleichsetzen. Wenn man durch Aspirin ein Menschenleben retten kann, ist dies höher zu bewerten, als wenn man 100 Bewerberinnen und Bewerber entdecken kann, die ein Studium voraussichtlich mit guten Noten abschließen werden. Der Nutzen Psychologischer Diagnostik lässt sich manchmal in Geldeinheiten messen. Durch Einsatz eines validen Verfahrens gegenüber dem üblichen Standard kann ein großes Unternehmen einen wirtschaftlichen Nutzen in Höhe von mehreren Millionen Euro erzielen (► Abschn. 5.2.3). In anderen Fällen, etwa wenn es gelingt, gefährliche Straftäterinnen und Straftäter zu erkennen und damit eine vorzeitige Entlassung zu verhindern, sollten ähnliche Maßstäbe angelegt werden wie beim Erkennen von schweren behandelbaren Krankheiten oder der Verminderung von Sterblichkeitsraten.

Vergleich von Korrelationen aus der Psychologie mit anderen Disziplinen

Korrelation ist nicht gleich Nutzen

Tab. 2.19 Höhe von Korrelationen in Psychologischer Diagnostik und anderen Bereichen

Untersuchungen/Metaanalysen	r	N (k)
Azetylsalizylsäure (z. B. Aspirin) und reduziertes Sterberisiko durch Herzinfarkt	.02	22.071
Effekt von Alkohol auf aggressives Verhalten	.23	k = 47
<i>Wert in Hare Psychopathy Checklist und Rückfall bei entlassenen Straftätern</i>	.28	1605
Schlaftabletten (Benzodiazepine oder Zolpidem) und kurzzeitige Verbesserung des Schlafs bei chronischen Schlafstörungen	.30	680
Sildenafil (z. B. Viagra) und verbesserte sexuelle Funktion bei Männern	.38	779
Rorschach Testergebnis und Ergebnis einer Psychotherapie	.44	783
Körpergröße und Gewicht bei Erwachsenen in den USA	.44	16.948
<i>Intelligenztestleistung und erreichtes Bildungsniveau</i>	.44	k = 9
Magnetresonanztomografiebefunde und Differenzierung zwischen Demenzpatienten und Kontrollprobanden	.57	374
Nähe zum Äquator und Tagestemperatur in den USA	.60	k = 19.724
Geschlecht und Körpergröße bei Erwachsenen in den USA (Männer sind im Durchschnitt größer als Frauen)	.67	16.962
<i>Neuropsychologische Tests und Differenzierung zwischen Demenzpatienten und Kontrollprobanden</i>	.68	k = 94
Immunglobulin-G-Test und Entdecken von rheumatoider Arthritis	.68	2541
<i>MMPI-Validitätsskalen und Entdecken der Simulation von Psychopathologie</i>	.74	11.204

Quelle (soweit nicht anders vermerkt): Meyer et al. (2001). N = Anzahl der untersuchten Personen, k = Anzahl der Korrelationskoeffizienten, die gemittelt wurden. Zahlreiche weitere Effekte sind der Originalpublikation zu entnehmen.

Bei der Beurteilung der Höhe der Korrelationen ist zudem zu beachten, unter welchen Randbedingungen sie zustande kamen. Möglicherweise sind die intendierten Interpretationen auf Basis des infrage stehenden Tests zulässig – die Art und Weise, wie Evidenz dafür generiert wurde, war jedoch suboptimal. Es kann auch vorkommen, dass die Art und Weise, wie die Evidenz zustande kam, besonders „hilfreich“ für hohe Korrelationen war, d. h., eine Überschätzung des eigentlichen Zusammenhangs vorliegt. Solche Einflussfaktoren werden nachfolgend näher beschrieben.

Einflussfaktoren auf die Höhe von Korrelationen

Einfluss auf die Höhe von Korrelationen haben folgende Faktoren:

- Asymmetrie der Messungen
- Drittvariablen
- Stichprobe
- Reliabilität der Messungen (► Abschn. 2.6.2.3)
- Unterschiedliche Methoden zwischen Test und Validierungsverfahren (vgl. MTMM-Systematik, s. o.)

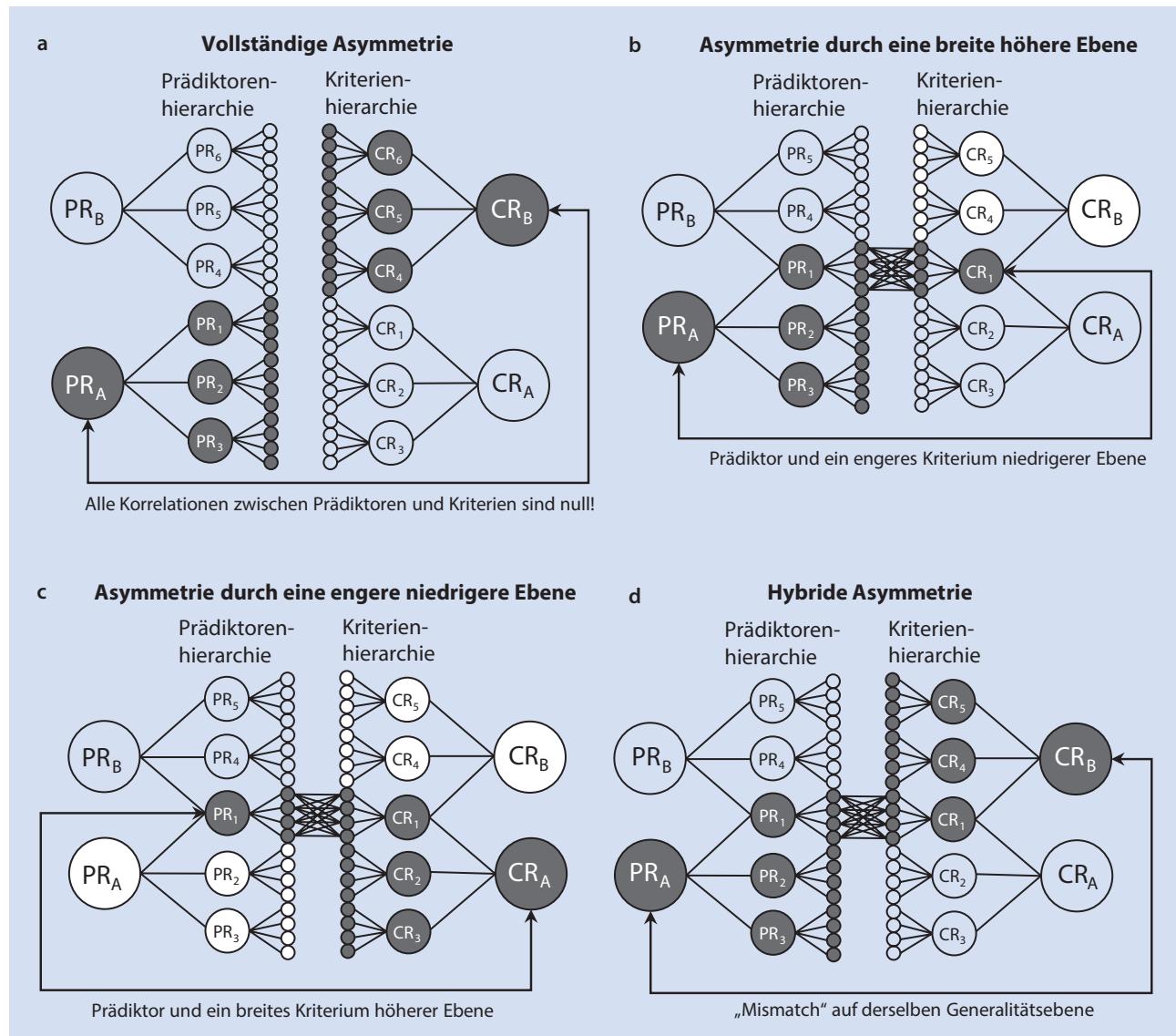
Prädiktor und Kriterium können in Inhalt und Generalitätsniveau disparat sein

Vollständige Asymmetrie = der Prädiktor erfasst etwas völlig anderes als das Kriterium

■ Asymmetrie

Das Problem der Asymmetrie von Messungen geht auf Brunswik (1952) zurück und wurde von Wittmann (1988) weiterentwickelt. Asymmetrie beschreibt dabei das Problem, dass Prädiktor (z. B. ein neu entwickelter und zu validierender Test) und Kriterium in Bezug auf den Inhalt und das Generalitätsniveau nicht korrespondieren. In diesen Fällen ist die zu erwartende Korrelation gemindert. Wittmann (1988) beschreibt 4 Formen der Asymmetrie, die in Abb. 2.40 veranschaulicht werden.

Im 1. Fall – als *vollständige Asymmetrie* beschrieben (Abb. 2.40a) – erfasst der Prädiktor konzeptuell etwas völlig anderes als das Kriterium. Beispielsweise wird der Fragebogen zur Gewissenhaftigkeit mit Berufserfolg korreliert – auf den ersten Blick ein plausibles Vorgehen, weil ein solcher Zusammenhang gut belegt ist (vgl. Barrick et al. 2001). Eine vollständige



■ Abb. 2.40 Vier Formen der Asymmetrie nach Wittmann (2012, S. 363): **a** vollständige Asymmetrie, **b** Asymmetrie durch eine breite höhere Ebene, **c** Asymmetrie durch eine engere niedrigere Ebene, **d** hybride Asymmetrie. PR = Prädiktor, CR = Kriterium. (Used with permission from Antisocial Behavior & Crime by Thomas Bliesener, Andreas Beelmann, Mark Stemmler, ISBN 9780889374249 © 2012 by Hogrefe Publishing, ► www.hogrefe.com)

Asymmetrie entsteht jedoch dann, wenn Berufserfolg über die Vorgesetztenbeurteilung von Belastbarkeit und Teamfähigkeit operationalisiert wurde und der Prädiktor Gewissenhaftigkeit über die Fragebogenskalen „Ordnungsliebe“ und „Leistungsstreben“ – also die Komponenten der Gewissenhaftigkeit in Prädiktor und Kriterium vollständig andere sind. Prädiktor und Kriterium weisen also keine inhaltlichen Gemeinsamkeiten auf. Wenn Prädiktor und Kriterium teilweise überlappen, teilweise aber auch nicht, entsteht eine Form der Asymmetrie, die als *hybrid* beschrieben wird (Abb. 2.40d). In unserem Beispiel könnte Berufserfolg über „Belastbarkeit“ und „Teamfähigkeit“ erfasst worden sein, der zu validierende Fragebogen zwar auch „Teamfähigkeit“ enthalten, aber zusätzlich „Leistungsstreben“ und nicht „Belastbarkeit“.

In den bisher diskutierten Fällen der Asymmetrie wurden der Prädiktor und das Kriterium gleich breit (mit je 2 Komponenten) erfasst. Asymmetrie entsteht auch, wenn sich der Prädiktor und das Kriterium in Bezug auf das Generalitätsniveau unterscheiden. Wenn also der Fragebogen „Belastbarkeit“, „Teamfähigkeit“ und „Leistungsstreben“ als Komponenten der Gewissenhaftigkeit inkludiert, die Vorgesetztenbeurteilung jedoch nur „Leistungsstreben“ berücksichtigt, so liegt für den Fragebogen ein höheres Generalitätsniveau als für das Kriterium vor. Dieser Fall, dass relevante Komponenten des Zielkonstrukt nicht im Kriterium enthalten sind, wird auch als *Kriteriumsdefizienz* beschrieben. In Abb. 2.40b wäre das Kriterium um „CR2“ und „CR3“ defizient. Im umgekehrten Fall – Berufserfolg wird breit erfasst, der Fragebogen fokussiert aber auf eine Komponente der Gewissenhaftigkeit – liegt für den Fragebogen ein niedrigeres Generalitätsniveau vor. Wenn es tatsächlich beabsichtigt war, Gewissenhaftigkeit eng zu erfassen, so würde die ungerechtfertigte Breite des Kriteriums auch als *Kriteriumskontamination* beschrieben werden. In Abb. 2.40c wäre das Kriterium um „CR2“ und „CR3“ kontaminiert. Natürlich können Kriterien durch vieles kontaminiert sein, also auch durch vollständig irrelevante Aspekte wie die Attraktivität der zu Beurteilenden.

Nicht immer ist die Zusammensetzung von Prädiktor und Kriterium so offensichtlich wie in den genannten Beispielen. Prädiktor und Kriterium können asymmetrisch sein, ohne dass man dies den Instrumenten auf den ersten Blick ansieht (vgl. Schulze et al. 2020). Eine bestehende Asymmetrie wird verschleiert, wenn die Skalen von Prädiktor und Kriterium nominell zwar gleich sind, sich hinter den gleichen Namen aber inhaltlich unterschiedliche Konzepte verbergen. Die Definition der „Teamfähigkeit“ in Instrument A ist eventuell nicht die gleiche wie die in Instrument B. Entdecken kann man solche Diskrepanzen, indem man den theoretischen Hintergrund und die Konstruktions schritte der Verfahren genau studiert. Es ist auch zu beachten, dass Verfahren mehrdimensional sein können, ohne dass dies explizit angegeben wird. Vielleicht wurde auf eine Prüfung der faktoriellen Struktur verzichtet oder diese wurde nicht sachgerecht durchgeführt. Je gründlicher ein Verfahren erforscht ist, desto geringer ist die Gefahr, dass die tatsächliche von der angenommenen Struktur abweicht.

Generalitätsniveau beachten

Asymmetrie nicht immer leicht zu erkennen

Jingle-jangle fallacy (Jingle-Jangle-Irrtum)

Als jingle-jangle fallacy (vgl. z. B. Block 1995) beschreibt man 2 Phänomene, die in der Psychologischen Diagnostik wahrscheinlich häufig auftreten. Wenn 2 Instrumente etwas ganz Unterschiedliches erfassen, aber den gleichen Namen bzw. Messanspruch haben, handelt es sich um eine jingle fallacy. Wenn also auf 2 Tests „Konzentrationstest“ steht, aber beide etwas völlig Verschiedenes messen, so wäre dies ein solcher Irrtum. Als jangle fallacy wird der Fall beschrieben, bei dem 2 Instrumente das Gleiche messen, aber unterschiedliche Namen tragen – etwa ein Test als Konzentrationstest, ein anderer als Test zur mentalen Geschwindigkeit bezeichnet wird und die Testwerte fast perfekt korrelieren.

Konfundierung durch Drittvariablen

■ Drittvariablen

Es gibt eine Reihe von Variablen, die quasi unbemerkt in Test und Validierungsinstrumente bzw. Kriterien einfließen können und damit die Korrelation zwischen ihnen künstlich erhöhen:

Anstrengungsbereitschaft (Motivation) Freiwillige Testpersonen sind unterschiedlich stark motiviert, gute Testergebnisse zu erzielen. Bei Leistungstests arbeiten sie unterschiedlich schnell und strengen sich z. B. bei schweren Aufgaben unterschiedlich stark an.

Soziale Erwünschtheit Bei Fragebögen tendieren Probandinnen und Probanden unterschiedlich stark dazu, sich sozial erwünscht darzustellen. Die Korrelation zwischen einer Skala zur Aggressivität und einer Skala zu körperlichen Beschwerden (beides sozial unerwünschte Merkmale) kann sich dadurch erhöhen, dass manche Probandinnen und Probanden ihre Antworten gar nicht, andere etwas und wieder andere stark in Richtung sozialer Erwünschtheit verändern.

Antwortstile Bei der Bearbeitung von Fragebögen tendieren manche Probandinnen und Probanden dazu, die Antwortskala auf eine individuelle Art und Weise zu nutzen – und zwar unabhängig vom Inhalt eines Items. Einige bevorzugen es, eher eine niedrige Ausprägung anzukreuzen, andere hingegen kreuzen eher hohe Ausprägungen an, und wiederum andere tendieren zur Mitte. Auch eine Tendenz, extreme Werte anzukreuzen, ist bekannt. Weijters et al. (2010) berechneten Indikatoren für diese 4 Antwortstile und wandten sie auf die Daten einer Umfrage an. Alle Fragen waren auf 7-stufigen Beurteilungsskalen zu beantworten. Die Befragung fand in 2 Wellen mit einem Abstand von 12 Monaten statt. Die Fragen an beiden Zeitpunkten waren inhaltlich völlig unabhängig. Die 4 Antwortstile erwiesen sich erstens als zeitlich relativ stabil, und zweitens konnte die Varianz dieser stabilen Anteile durch biografische Merkmale der Befragten teilweise erklärt werden. Die Varianzaufklärung variierte zwischen 1 % (Präferenz für niedrige Skalenwerte) und 8 % (Präferenz für Extremwerte). Je älter die Befragten und je niedriger ihre Bildung waren, desto stärker tendierten sie zu hohen, zu extremen oder zu mittleren Skalenwerten. Solche Antwortstile treten bei der Messung vieler Merkmale auf. Dies führt dazu, dass der tatsächliche Zusammenhang zweier Merkmale überschätzt wird. Durch das Auftreten von Antwortstilen kann sich auch die Retest-Reliabilität künstlich erhöhen.

Konfundierung mit anderen Merkmalen Test und Validierungsinstrumente bzw. Kriterien können zudem ein weiteres Konstrukt messen, ohne dass dies beabsichtigt oder bekannt ist. An einem Beispiel lässt sich das Problem leicht erklären: Assessment-Center-Ergebnisse werden anhand beruflicher Leistungsbeurteilungen (Vorgesetztenurteile) validiert. Der daraus resultierende empirische Zusammenhang könnte hoch ausfallen, da sowohl in das Assessment-Center-Urteil als auch in die Vorgesetztenbeurteilung die physische Attraktivität der Kandidatinnen und Kandidaten einfließt.

■ Stichprobe

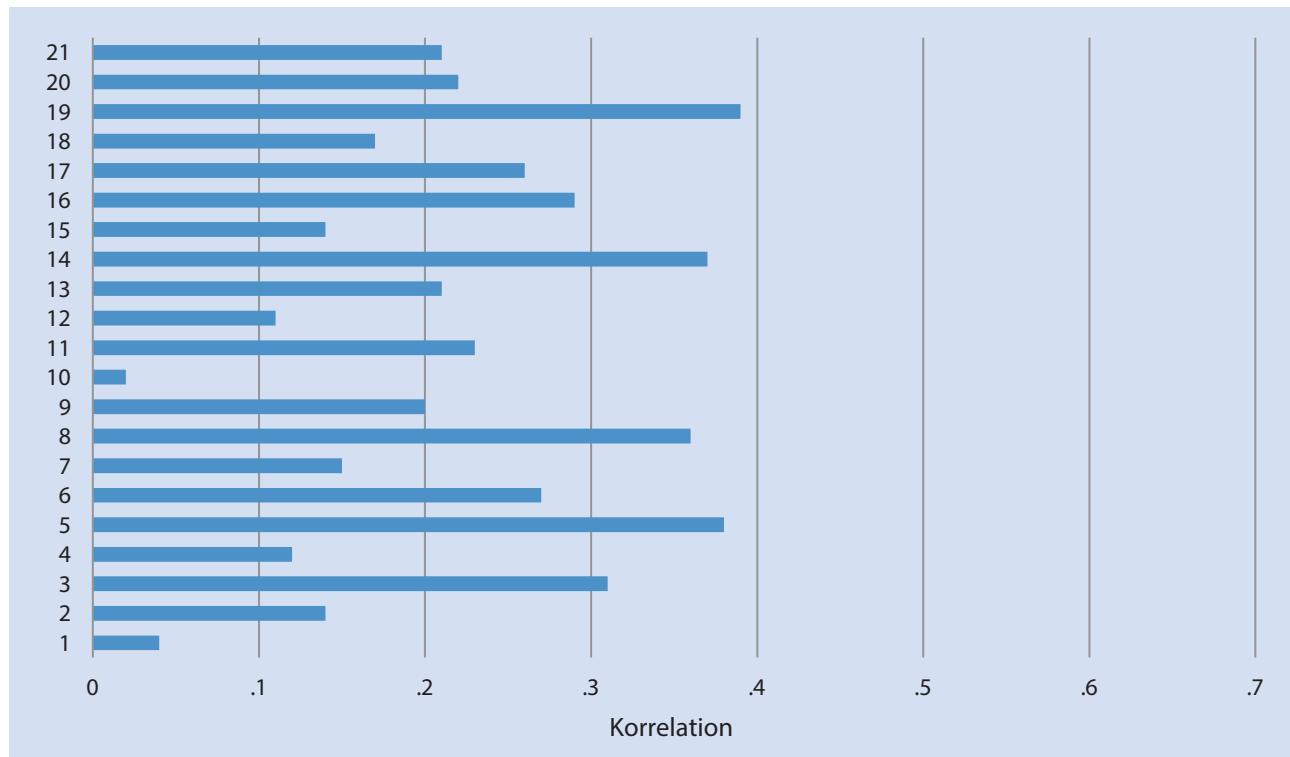
Mehrere Merkmale der Stichprobe nehmen auf die Höhe der Korrelationen Einfluss. Dies inkludiert den Umfang der Stichprobe. Stichproben, an denen die Korrelationen eines Tests mit anderen Variablen ermittelt werden, sind manchmal relativ klein. Die Validität der Testwertinterpretationen kann dadurch falsch eingeschätzt werden. Je kleiner die Stichprobe ist, desto stärker kann die beobachtete Korrelation nach oben oder unten von dem tatsächlichen Wert in der Population abweichen. Diesen Fehler nennt man auch *Stichprobenfehler*.

Schmidt (1992) hat den Effekt kleiner Stichproben auf die Höhe von Korrelationen anschaulich demonstriert. In einer Studie ($N=1428$) betrug die Korrelation $r=.22$. Er unterteilte nun diesen Datensatz nach Zufall in 21 Teilstichproben auf, deren Größe ($n=68$) für die Fachliteratur zur Validität eignungsdiagnostischer Verfahren typisch war, und berechnete für jede Teilstichprobe die Korrelation. □ Abb. 2.41 zeigt das Ergebnis: Die beobachteten Korrelationen reichten von $r=.02$ bis $.39$. Schönbrodt und Perugini (2013) berichten auf Basis von Simulationsstudien, dass für typische Anwendungsfälle eine Stichprobengröße von annähernd 250 erforderlich ist, um stabile Korrelationsschätzungen zu erhalten.

Auch die Zusammensetzung der Untersuchungsstichprobe kann sich auf die Höhe der Korrelation wie folgt auswirken: Erstens kommt es vor, dass Korrelationen an einer Personengruppe ermittelt wurde und nun auf eine andere übertragen werden sollen. Dieser Schritt ist eventuell unzulässig. Intelligenztests werden häufig anhand von Korrelationen zwischen Testergebnissen und späterem Berufserfolg validiert. Nehmen wir an, dass die Untersuchung an einer Gruppe leitender Angestellter erfolgte. Der Test soll nun zur Auswahl von Auszubildenden aus der gleichen Branche eingesetzt werden. Es ist

Merkmale der Stichprobe beachten

Übertragbarkeit auf andere Stichproben prüfen



□ Abb. 2.41 Beobachtete Validitätskoeffizienten nach zufälliger Aufteilung einer Stichprobe von $N=1428$ in 21 Teilstichproben mit jeweils $n=68$. (Nach Schmidt 1992, Tab. 1, mit freundlicher Genehmigung der American Psychological Association)

Streuung der Testwerte beachten

bekannt, dass die Korrelation zwischen Intelligenz und Berufserfolg bei einfachen Berufen niedriger ist als bei komplexen (Salgado et al. 2003a; Schmidt und Hunter 1998). Folglich wird das Ergebnis der Validierung für leitende Angestellte nicht auf Bürogehilfinnen und -gehilfen übertragbar sein. Gegen den Fehler, einen Validitätsbefund ungerechtfertigt auf eine andere Personengruppe zu übertragen, kann man sich schützen: Die Stichprobe, an der die Validität ermittelt wurde, sollte der Personengruppe, bei welcher der Test eingesetzt wird, möglichst ähnlich sein.

Zweitens unter- oder überschätzt man die Validität der Testwertinterpretationen leicht, wenn man die Streuung der Testwerte ignoriert. Die Problematik lässt sich gut am Beispiel der Personalauswahl erklären. Zur Besetzung von 100 Stellen werden 500 Bewerberinnen und Bewerber untersucht. Als Prädiktor des Berufserfolgs dient ein Eignungstest. Das Unternehmen stellt nur Bewerberinnen und Bewerber mit den besten Testergebnissen ein. Während die Streuung der Testwerte in der Bevölkerung relativ groß ist, findet durch die Selektion der Bewerberinnen und Bewerber eine Einengung der Variabilität („Varianzeinschränkung“) statt. Für die Validitätsprüfung steht nicht mehr die gesamte Streubreite der Messwerte zur Verfügung. Dies hat eine Minderung der Korrelation zur Folge. Es sind deshalb Formeln entwickelt worden, um den bei eingeschränkter Streuung im Prädiktor ermittelten Korrelationskoeffizienten auf repräsentative Breite aufzuwerten.

Aufwertung von Korrelationen bei Streuungseinschränkung

$$r'_{tc} = \frac{\frac{r_{tc} \times s_x}{s_x}}{\sqrt{1 - r_{tc}^2 + \frac{r_{tc}^2 \times s_x^2}{s_x^2}}}$$

r'_{tc} = aufgewertete Korrelation zwischen Test t und Kriterium c

r_{tc} = beobachtete Korrelation zwischen Test t und Kriterium c

s_x = angenommene Messwertestreuung (wenn keine Streuungseinschränkung vorläge)

s_x = vorliegende (eingeschränkte) Messwertestreuung

(Lienert und Raatz 1998, S. 266)

Beispiel: Ein neu entwickelter Test hat in der Normierungsstichprobe eine Streuung von $S_x = 10$. In der Validierungsgruppe ist die Streuung mit $s_x = 5$ deutlich kleiner; die beobachtete Korrelation von $r_{tc} = .50$ stellt daher eine Unterschätzung des wahren Zusammenhangs dar. Wie hoch wäre der Zusammenhang rechnerisch, wenn die uneingeschränkte Messwertestreuung vorgelegen hätte?

$$r'_{tc} = \frac{\frac{.50 \times 10}{5}}{\sqrt{1 - .50^2 + \frac{.50^2 \times 10^2}{5^2}}} = .76$$

Nach Korrektur für Varianzeinschränkung ergibt sich eine Korrelation von $r'_{tc} = .76$.

Wie Höhe der Korrelationen beurteilen?

Nachdem wir verschiedene Einflussfaktoren auf die Höhe von Korrelationen thematisiert haben, stellt sich nun mehr die Frage, wie hoch Korrelationen ausfallen sollten, um als Beleg für die intendierte Interpretation von Testwerten zu gelten. In der Tat fällt es vielen Testnutzerinnen und Testnutzern

schwer, die Höhe von Korrelationskoeffizienten angemessen zu beurteilen. Wenig brauchbar sind Faustregeln zur Höhe von niedrigen, mittleren und hohen Korrelationen bzw. Effektstärken. Cohen (1988), der in diesem Zusammenhang oft zitiert wird, betont, dass solche Anhaltspunkte nur gelten, wenn keine Vergleichswerte vorliegen. Glücklicherweise liegen aus vielen Forschungsarbeiten eben solche Vergleichswerte vor. Oftmals sind die Befunde so zahlreich, dass nicht nur eine, sondern inzwischen mehrere Metaanalysen dazu durchgeführt wurden.

Hemphill (2003) hat aus Metaanalysen, die Meyer et al. (2001) zusammengestellt hatten (Tab. 2.19), diejenigen 78 ausgewählt, welche sich explizit auf psychologische Kriterien beziehen. Es muss betont werden, dass keine Korrelationen zwischen Fragebögen oder Leistungstests berichtet worden waren, sondern meist zwischen Testwerten und dazu passenden realen Kriterien wie Studiennoten, Schwere einer Kopfverletzung oder Berufserfolg. Die Koeffizienten stehen in der Regel für beobachtete, nicht korrigierte Korrelationen. Hemphill hat die Korrelationskoeffizienten in eine Rangreihe gebracht und die Koeffizienten danach in ein unteres, mittleres und oberes Drittel eingeteilt. Im mittleren Bereich finden sich Koeffizienten zwischen $r = .21$ und $.33$. Der untere Bereich ist durch Werte zwischen $r = .02$ und $.21$, der obere durch Werte zwischen $r = .35$ bis $.78$ gekennzeichnet. Wenn man also überhaupt keinen spezifischen Vergleichswert findet (dazu unten gleich mehr), kann man dieses Raster zur Beurteilung von Korrelationskoeffizienten heranziehen.

Metaanalysen zur Orientierung nutzen

- ! Bei Verwendung eines Vergleichsmaßstabs, der auf Metaanalysen basiert, ist unbedingt zu beachten, dass dort in der Regel Korrekturen für Varianzeinschränkung, Reliabilität des Kriteriums und manchmal auch des Prädiktors vorgenommen wurden. Deshalb ist es unbedingt erforderlich, die empirisch ermittelten unkorrigierten Validitätskoeffizienten durch die entsprechenden Korrekturen aufzuwerten. Nur so ist eine faire Beurteilung möglich.

Zur Validität der Testwertinterpretation bei Persönlichkeitsfragebögen, die den Big-Five-Dimensionen zuzuordnen sind, finden sich in Tab. 2.20 umfangreiche Vergleichswerte. Bei Verwendung der Tab. 2.20 ist zu beachten, dass die Korrelationen teilweise korrigiert wurden (Angaben dazu bei den Anmerkungen). Ferner ist zu beachten, dass hier Mittelwerte berichtet werden; diese Werte können also als „durchschnittlich“ gelten. Für detaillierte Angaben sei auf die zitierten Quellen verwiesen. Dort finden sich meist auch Angaben zur Standardabweichung der Korrelationskoeffizienten. Beispielsweise wird für Berufserfolg und Gewissenhaftigkeit eine mittlere Korrelation von $.27$ (korrigierter Wert) berichtet; die Standardabweichung beträgt $.05$ (vgl. Barrick et al. 2001). Ein korrigierter Korrelationskoeffizient von $.37$ wäre demnach sehr hoch, liegt er doch 2 Standardabweichungen über dem mittleren Wert. Will man also genauer wissen, wie weit ein Korrelationskoeffizient über einem Vergleichswert aus der Tab. 2.20 liegt, muss man sich in den Publikationen über die Streuung der Korrelationskoeffizienten informieren. Für eine grobe Orientierung reichen die in der Tab. 2.20 aufgeführten mittleren Korrelationskoeffizienten jedoch aus.

Metaanalytische Befunde für Intelligenztests

Zu Intelligenztests wurden in Tab. 2.21 metaanalytische Ergebnisse zusammengestellt. Den höchsten Korrelationskoeffizienten findet man für Maße des Schulerfolgs. Man beachte, dass hier ein beobachteter (unkorrigierter) Wert berichtet wird. Beim Kriterium Berufserfolg wurden mittelkomplexe Berufe berücksichtigt. Die Korrelation fällt für komplexere, anspruchsvollere Berufen höher aus als die hier angegebenen Werte. Bei weniger komplexen,

■ Tab. 2.20 Zusammenhang zwischen Big-Five-Persönlichkeitsfragebögen und Kriteriumsdaten

Persönlichkeitsmerkmal (Fragebogen)	Kriterium					
	Verhalten im Alltag ^a	Fremdbeurteilung ^b	Schulerfolg ^c	Ausbildungserfolg ^d	Studienerfolg ^e	Berufserfolg ^f
Neurotizismus	.53	(.51)	(.20)	.05 (.09)	(.01)	.06 (.13)
Extraversion	.42	(.62)	(.18)	.13 (.28)	(-.01)	.06 (.15)
Verträglichkeit	.55	(.46)	(.30)	.07 (.14)	(.06)	.06 (.13)
Gewissenhaftigkeit	.48	(.56)	(.28)	.13 (.27)	(.23)	.12 (.27)
Offenheit für Erfahrungen	.56	(.59)	(.24)	.14 (.33)	(.07)	.03 (.07)

Durchschnittliche beobachtete Korrelationen (in Klammern korrigierte Werte, s. u.)

^aQuelle: Fleeson und Gallagher (2009, Tab. 4); Mittelwert von vielen Angaben zum momentanen persönlichkeitsbezogenen Verhalten im Alltag (z. B. „Wie hart haben Sie in der letzten halben Stunde gearbeitet?“), $N=495$, über 21.000 Messgelegenheiten

^b Quelle: Connolly et al. (2007, Tab. 2); Korrelation zwischen Selbstbeschreibung in Persönlichkeitsfragebögen und Fremdbeurteilung; Metaanalyse über 38–55 Studien, Korrelationen korrigiert für Reliabilität von Prädiktor und Kriterium; $N=5.333\text{--}8.000$

^c Quelle: Poropat (2009, Tab. 2); Kriterium ist die Leistung in der Grundschule (Noten); Metaanalyse über 8 unabhängige Studien, $N=3869$. Korrelationen korrigiert für Reliabilität von Prädiktor und Kriterium

^d Quelle: Barrick et al. (2001, Tab. 1–5); Zusammenfassung von 2 Metaanalysen; Kriterium explizit Trainingserfolg; Korrelationen in Klammern korrigiert für Varianzeinschränkung und Reliabilität von Prädiktor und Kriterium; 18–25 unabhängige Studien, $N=3177\text{--}4100$

^e Quelle: Poropat (2009, Tab. 2); Kriterium ist der Studienerfolg (Noten); Metaanalyse über 75–92 unabhängige Studien ($N=27.944\text{--}32.887$). Korrelationen korrigiert für Reliabilität von Prädiktor und Kriterium

^f Quelle: Barrick et al. (2001, Tab. 1–5); Zusammenfassung mehrerer unabhängiger Metaanalysen; Korrelationen in Klammern korrigiert für Varianzeinschränkung und Reliabilität von Prädiktor und Kriterium; 143–239 unabhängige Studien, $N=23.225\text{--}48.100$

einfachen Berufen ist sie dagegen niedriger (Salgado et al. 2003b; Schmidt und Hunter 1998). Die Werte in ■ Tab. 2.21 dienen als Referenz zur Beurteilung der Korrelation von Intelligenztests mit realen Leistungsdaten. Wurde etwa ein Test in Deutschland am Ausbildungserfolg validiert, so kann der Wert von .59 als Vergleichswert herangezogen werden. Der beobachtete bzw. in einem Testmanual berichtete Korrelationskoeffizient muss dafür jedoch den gleichen Korrekturen unterworfen werden wie der Referenzwert. Für eine genaue Beurteilung ist zudem die Streuung der Korrelationskoeffizienten in den Metaanalysen informativ. Kramer (2009) schätzt die Standardabweichung der korrigierten Korrelationskoeffizienten auf .17 (Kriterium: Ausbildungserfolg). Erzielte man einen (korrigierten) Korrelationskoeffizienten von .40, läge dieser mehr als eine Standardabweichung unter dem in ■ Tab. 2.21 berichteten Wert und dürfte damit als eher niedrig gelten.

Tab. 2.21 Zusammenhang zwischen Intelligenztestleistungen und realen Leistungsdaten

Herkunft der Studien	Leistungsindikator			
	Berufserfolg	Ausbildungserfolg	Bildungsniveau	Schulerfolg
International ^{a, d}	(.51)	(.56)	.46 (.56)	
Europa ^{b, e}	.27 (.53)	.29 (.53)		.69
Deutschland ^c	.33 (.62)	.37 (.59)		

Tests zur allgemeinen Intelligenz. Berufserfolg wurde in allen Studien zumeist durch Vorgesetztenbeurteilung erfasst, Ausbildungserfolg zumeist durch Prüfungsergebnisse

^aQuelle für Berufs- und Ausbildungserfolg: Schmidt und Hunter (1998, Tab. 1 und 2); Test: General Aptitude Test Battery; $N=32.000$

Berufstätige in den USA, Berufe mittlerer Komplexität, bei Ausbildungserfolg keine Angabe zu N . Korrelationen in Klammern sind für Varianzeinschränkung sowie für die Reliabilität von Test und Kriterium korrigiert

^bQuelle: Salgado et al. (2003a, Tab. 6); diverse Tests; bei Berufserfolg 43 Studien ($N=4744$), bei Ausbildungserfolg 35 Studien ($N=4304$); jeweils nur Berufe mittlerer Komplexität. Korrelationen in Klammern sind für Varianzeinschränkung sowie für die Reliabilität von Test und Kriterium korrigiert

^cQuelle: Kramer (2009, Tab. 3 und 4); diverse Intelligenztests; 18 Studien ($N=2739$) zur Arbeitsleistung; zu Lernleistung (meist Ausbildungserfolg) 210 Studien ($N=30.451$). Korrelationen in Klammern sind für Varianzeinschränkung sowie für die Reliabilität von Test und Kriterium korrigiert

^dQuelle: Strenze (2007, Tab. 1); Bildungsniveau definiert als Gesamtdauer der Bildung in Jahren oder höchster erreichter Bildungsabschluss; Metaanalyse über 59 Längsschnittstudien ($N=84.828$); Korrelationen in Klammern korrigiert für Reliabilität von Prädiktor und Kriterium

^eQuelle: Deary et al. (2007, Tab. 2); bei englischen Schülerinnen und Schülern wurde im Alter von 11 Jahren die Intelligenz mit einem Test zum schlussfolgernden Denken (Cognitive Abilities Test, CAT) gemessen und mit Prüfungsergebnissen in der Schule im Alter von 16 Jahren korreliert ($N=70.530$). Auch Angaben zu einzelnen Schulfächern; die höchste Korrelation fand sich mit $r=.77$ für das Kriterium Mathematiknote ($N=68.125$)

Empfehlungen des Diagnostik- und Testkuratoriums

In seinem Testbeurteilungssystem empfiehlt des Diagnostik- und Testkuratorium die Prüfung, ob im Manual folgende Fragen hinreichend beantwortet werden (zitiert nach Diagnostik- und Testkuratorium 2018b, S. 115 f., © Hohlfeld):

- Wurden die Validitätskoeffizienten für alle (Sub-)Populationen aus einer Stichprobenerhebung geschätzt, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll?
- Ist eine hypothesenleitete Prüfung von Validitätsbelegen zur Stützung des Testeinsatzes gemäß der diagnostischen Zielsetzung erfolgt?
- Wurden die Validitätsuntersuchungen hypothesen- bzw. theoriegeleitet entwickelt (statt nur im Nachhinein signifikante Korrelationen als Validitätsbeleg anzuführen)?
- Ist die inhaltliche und psychometrische Qualität der zur Validierung herangezogenen Maße (z. B. andere Tests zur Konstruktvalidität; Kriteriumsmaße) angemessen?
- Fand die Untersuchung zur Übereinstimmung mit einem Kriterium unter solchen Testbedingungen statt, wie sie den Bedingungen bei der Nutzung des Tests in der Praxis weitgehend entsprechen?

! Erweist sich ein in der Entwicklung befindlicher Test für seine Einsatzzwecke und die intendierte Interpretation als nicht hinreichend valide, müssen Testentwicklerinnen und -entwickler in eine frühere Phase der Testentwicklung (Abb. 2.22) zurückgehen und die danach folgenden Schritte wiederholen.

2.6.4 Nebengütekriterium: Normierung

Normen dienen der Einordnung von Testrohwerten

Die Normierung eines Tests liefert ein Bezugssystem, um die individuellen Testwerte (Rohwerte) im Vergleich zu denen einer größeren und meist repräsentativen Stichprobe von Testteilnehmerinnen und -teilnehmern einordnen zu können. Normen sind wichtig, wenn ein Test zur Individualdiagnostik eingesetzt wird. Um in der Einzelfalldiagnostik beispielsweise beurteilen zu können, was 15 richtige Lösungen in einem Leistungstest bedeuten, muss man wissen, wie viele Aufgaben andere Testteilnehmerinnen und -teilnehmer lösen. Die Normierung stellt den diesbezüglich erforderlichen Bezugsrahmen zur Verfügung und sagt uns, was die Rohpunktswerte „bedeuten“. Zu diesem Zweck werden die Rohwerte in transformierte Werte überführt.

Arten von Normen

- Äquivalentnormen
- Variabilitäts- oder Abweichungsnormen
- Prozentrangnormen

Zuordnung der Rohwerte zu bestimmten Referenzgruppen

Bei der Bildung von *Äquivalentnormen* erfolgt eine Zuordnung der Rohwerte zu bestimmten Referenzgruppen. Ein typisches Beispiel sind Altersgruppen. Ein 10-jähriges Kind habe in einem Test 15 richtige Lösungen erzielt. Zum Vergleich dient eine Tabelle, in der die durchschnittlichen Leistungen von Kindern unterschiedlicher Altersgruppen aufgeführt sind. Diese Tabelle zeigt, dass 15 richtige Lösungen der durchschnittlichen Leistung von 9-jährigen Kindern entspricht. Nun weiß man, dass das untersuchte Kind bezüglich des untersuchten Merkmals etwas „rückständig“ ist: Es hat einen Leistungsstand, der eigentlich für Kinder typisch ist, die 1 Jahr jünger sind. Generell kann man auch vom Entwicklungsalter sprechen: Das untersuchte Kind weist bei dem Merkmal ein Entwicklungsalter von 9 Jahren auf. Das früher gebräuchliche „Intelligenzalter“ stellt eine spezielle Variante des Entwicklungsalters dar. Es besagt, für welche Altersgruppe ein Intelligenztestwert typisch ist.

Variabilitäts- oder Abweichungsnormen setzen voraus, dass die Messwerte eines Tests normalverteilt sind und mindestens Intervallskalenniveau haben (also Interpretationen der Abstände zwischen Testwerten zulässig sind). Der Normwert gibt dann an, wie weit eine Person mit ihrer Testleistung unter oder über dem Mittelwert einer Vergleichsgruppe liegt. Die Abweichung jedes einzelnen Messwertes X vom Mittelwert M der Normgruppe wird dabei in Einheiten der Streuung der Normgruppe ausgedrückt. Als Vergleichsgruppe können eine bevölkerungsrepräsentative Gesamtstichprobe dienen

Abweichung vom Mittelwert in Streuungseinheiten

Tab. 2.22 Roh- und Normwerte von 5 Personen

Person	Rohwert	Abweichung vom Mittelwert der Referenzgruppe in Standardabweichungseinheiten
1	12	-2
2	16	-1
3	20	0
4	24	1
5	28	2

oder auch nur Personen gleichen Alters (Altersnormen), gleichen Geschlechts (Geschlechtsnormen) oder etwa gleicher Bildung (schul- oder bildungsspezifische Normen) herangezogen werden.

Nehmen wir zur Erläuterung des Prinzips von Variabilitäts- bzw. Abweichungsnormen an, für einen fiktiven Test liegt eine bevölkerungsrepräsentative Vergleichsgruppe vor. Die Werte der Vergleichsgruppe in dem fiktiven Test sind normalverteilt mit einem Mittelwert von 20 und einer Standardabweichung von 4. Nun bearbeiten 5 Testpersonen den Test und schneiden wie folgt ab (Tab. 2.22).

Die Darstellung der Testergebnisse als Abweichung vom Mittelwert der Referenzgruppe in Einheiten der Standardabweichung der Referenzgruppe hat den Vorteil, dass diese leicht zu interpretieren sind. Ein Wert, der 2 Standardabweichungen unter dem Mittelwert einer normalverteilten Vergleichsgruppe liegt, ist unmittelbar als gering einzustufen: Person 1 ist besser oder gleich gut wie nur ca. 2 % der Vergleichsgruppe (Abb. 2.42). Ein Wert der 2 Standardabweichungen über dem Mittelwert dieser Vergleichsgruppe liegt, ist als hoch zu beurteilen: Person 5 ist besser oder gleich gut wie ca. 98 % der Vergleichsgruppe.

Die Transformation der Rohwerte in eine Skala mit einem Mittelwert von 0 und einer Standardabweichung von 1 – diese Skala haben wir soeben in unserem Beispiel benutzt – wird auch z -Transformation genannt. Die entsprechenden Werte werden als z -Werte bezeichnet. Allerdings lässt sich diese Skalierung beliebig verändern. Statt eines Mittelwertes von 0 könnte man auch 100 annehmen und statt einer Standardabweichung von 1 auch 15. Damit wäre man bei sog. „IQ-Werten“. Daneben sind T-Werte, Standardwerte und Stanine-Werte gebräuchlich. Abb. 2.42 zeigt die einzelnen Normen und ihren Bezug zur Normalverteilung (für Stanine-Werte Abb. 2.43). Diese Normwerte können bei Bedarf ineinander überführt werden. Sie verhalten sich wie verschiedene Währungen zueinander, die man nach festen Wechselkursen tauschen kann. Sehr hilfreich ist dabei der frei verfügbare Normwert-Rechner von Psychometrica (<https://www.psychometrica.de/normwertrechner.html>). Wir zeigen nachfolgend beispielhaft das rechnerische Vorgehen, um die Transformation von Normwerten zu erläutern.

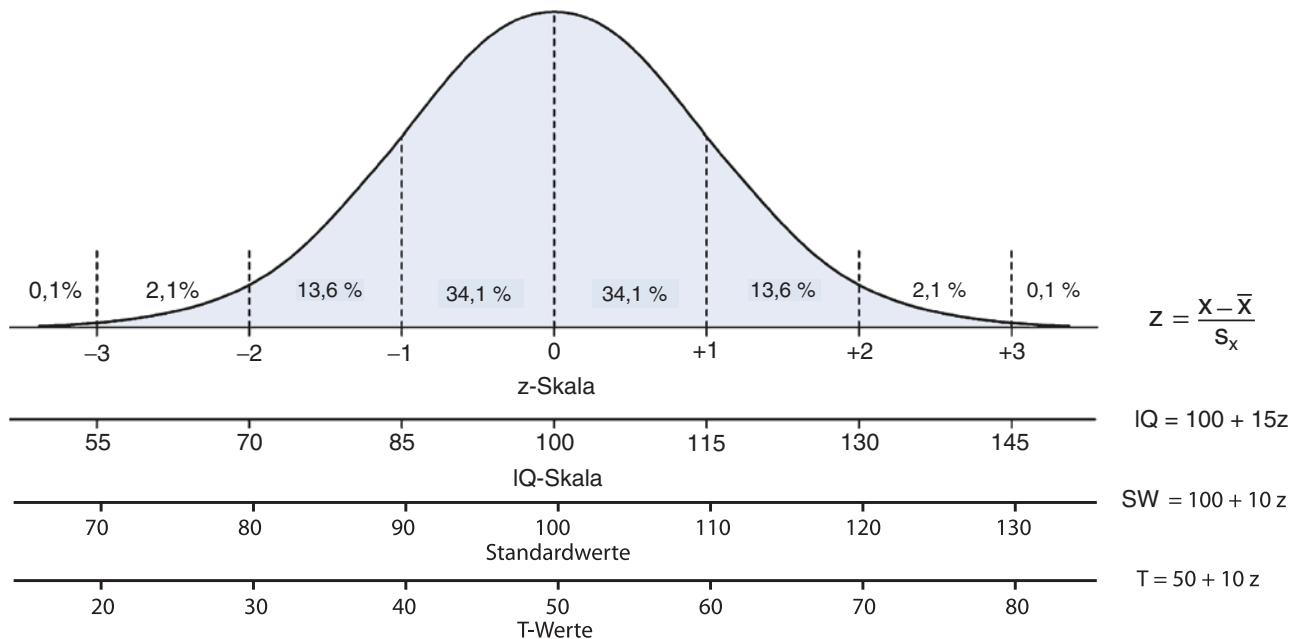


Abb. 2.42 Relative Häufigkeiten von z-, IQ-, Standard- (SW) und T-Werten unter den einzelnen Abschnitten der Normalverteilung

Beispiele für die Transformation von Normwerten

Der IQ-Normwert einer Person kann beispielsweise folgendermaßen in einen T-Normwert der Person umgerechnet werden.

$$X_T = \frac{X_{IQ} - \bar{X}_{IQ}}{s_{IQ}} \times s_T + \bar{X}_T$$

X_T = T-Normwert einer Person

X_{IQ} = IQ-Normwert einer Person

\bar{X}_{IQ} = Mittelwert der IQ-Normwertskala (= 100)

\bar{X}_T = Mittelwert der T-Normwertskala (= 50)

s_{IQ} = Standardabweichung der IQ-Normwertskala (= 15)

s_T = Standardabweichung der T-Normwertskala (= 10)

Die weithin bekannten PISA-Testungen nutzen Werte mit einem Mittelwert von 500 und einer Standardabweichung von 100. Mit dem Wissen um den Mittelwert und die Standardabweichung der Normwertskala lassen sich die Ergebnisse der PISA-Ländervergleiche gut einordnen. So erreichten deutsche Schülerinnen und Schüler 2015 im Mittel einen Wert von 509 in Naturwissenschaften (Platz 16 im Länderranking). Ein skandinavisches Land, das häufig als Vorbild genannt wird, ist Finnland. Finnische Schülerinnen und Schüler erzielten im Mittel einen Wert von 531 in Naturwissenschaften (und damit Platz 5 im Länderranking). Mit der Formel

$$z = \frac{X - \bar{X}}{s_x}$$

lassen sich beide PISA-Werte in z-Normwerte umrechnen. Diese sind:

$$z_{\text{Deutschland}} = \frac{509 - 500}{100} = .09$$

$$z_{\text{Finnland}} = \frac{531 - 500}{100} = .31$$

Da z-Normwerte eine Standardabweichung von 1 haben, lässt sich sagen, dass der Unterschied zwischen Deutschland und Finnland ($z_{\text{Diff}} = .22$) ca. einer fünfstel Standardabweichung entspricht. In IQ-Normwerten wären dies ca. 3 IQ-Punkte (1/5 der Standardabweichung von 15). Der PISA-Wert von 509 entspricht einem IQ-Wert von 101,35, der PISA-Wert von 531 entspricht einem IQ-Wert von 104,65.

Äquivalent- bzw. Abweichungsnormen können nach Grob- und Feinnormen untergliedert werden. IQ-, Standard- und T-Werte ermöglichen feine Unterscheidungen innerhalb der Standardabweichungsintervalle (bei IQ-Werten existieren innerhalb einer Standardabweichung 15 unterscheidbare Einheiten). Sie werden als „Feinnormen“ bezeichnet. Erlauben die verwendeten diagnostischen Instrumente eine entsprechend feine Unterscheidung, sind solche Normen zu präferieren. Sind derart feine Unterscheidungen nicht möglich (beispielsweise weil ein Test nur über eine mäßige Reliabilität verfügt, so suggerieren sie jedoch eine Differenzierungsfähigkeit des Tests, die in der Realität gar nicht vorhanden ist. Dann sind sog. „Grobnormen“ (Stanine-Werte und C-Werte) zu bevorzugen.

Grob- und Feinnormen

2.6.4.1 Stanine-Werte

Neben den bereits eingeführten z-, IQ-, Standard-, T- und PISA-Werten sind auch sog. „Stanine-Werte“ und „Sten-Werte“ gebräuchlich. Diese unterscheiden sich von den anderen Normwertskalen dadurch, dass nur ganzzahlige Werte vorgesehen sind, mit denen ganze Bereiche einer Normverteilung bezeichnet werden.

Die Stanine-Skala (Abkürzung für **standard nine**) reicht von 1 bis 9. Sie hat einen Mittelwert von 5 und eine Standardabweichung von ungefähr 2. Entwickelt wurde sie im Zweiten Weltkrieg von der amerikanischen Luftwaffe – und zwar aus einem sehr pragmatischen Grund: Die 9 Ziffern konnten auf den damals üblichen Lochkarten (eine frühe Form eines Datenträgers, der per Computer verarbeitet werden konnte) in einer einzigen Spalte und damit platzsparend eingestanzt werden (Kaplan und Saccuzzo 2005).

Stanine-Skala

Die Umrechnung von Rohwerten in Stanine-Werte ist sehr einfach: Die Testwerte werden in eine aufsteigende Reihenfolge gebracht, die ersten 4 % der Rohwerte erhalten den Stanine-Wert 1, die nächsten 7 % den Wert 2 etc. (Tab. 2.23). Stanine-Werte können somit auch mit nicht normalverteilten Rohwerten berechnet werden (Abb. 2.43).

Die Prozentränge (4, 11, 23 etc.) entsprechen in der Normalverteilung bestimmten z-Werten, die in der untersten Zeile von Tab. 2.23 aufgeführt sind. Ein Stanine-Wert steht immer für einen Bereich von z-Werten. So wird die Mitte der Skala durch die z-Werte –0,25 bis +0,25 markiert. In der Normalverteilung ist sie 0,5 Standardabweichung breit und schließt 20 % aller Personen ein. Die sich anschließenden Stanine-Werte sind ebenfalls eine halbe Standardabweichung breit. Nur für die Stanine-Werte 1 und 9 ist die Grenze zum Rand der Verteilung hin offen; theoretisch reicht sie bis minus bzw. plus unendlich.

Tab. 2.23 Stanine-Werte

Stanine-Wert	1	2	3	4	5	6	7	8	9
Relative Häufigkeit (in %)	4	7	12	17	20	17	12	7	4
z -Wert-Bereiche	< -1,75	-1,251 bis -1,75	-0,751 bis -1,25	-0,251 bis -0,75	-0,25 bis 0,25	0,251 bis 0,75	0,751 bis 1,25	1,251 bis 1,75	> 1,75
$M(z)$	-2,0	-1,5	-1	-0,5	0	0,5	1	1,5	2,0

Relative Häufigkeitsangaben nach Gregory (2004, S. 68). Erläuterungen für die zu Stanine 1 und 9 gehörigen z -Werte im Text

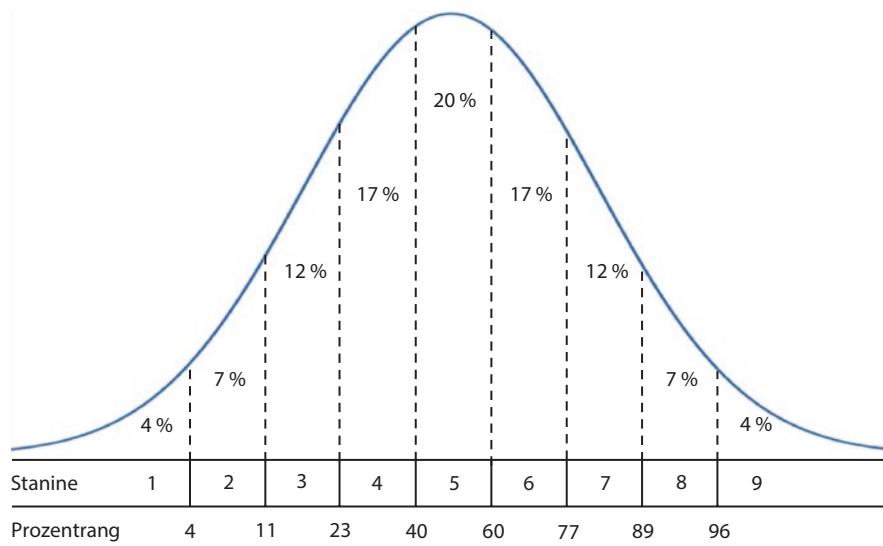


Abb. 2.43 Stanine-Werte

Konfidenzintervalle für Stanine-Werte bestimmen

Die Formel zur Schätzung des wahren Wertes nach der Regressionshypothese und die Berechnung von Konfidenzintervallen (► Abschn. 2.6.2) sind im Prinzip auch auf Stanine-Werte anwendbar. Wir sagen „im Prinzip“, weil jeder Stanine-Wert für einen Bereich steht. Die Randbereiche mit Stanine 1 und 9 sind aber als nach unten bzw. nach oben offen definiert ($z < -1,75$ bzw. $z > 1,75$; vgl. Tab. 4.3). Wir nehmen daher einfach an, dass der Stanine-Wert von 1 einem z -Wert von -2 und der Stanine-Wert von 9 einem z -Wert von +2 entspricht. Die Stanine-Skala wird mit dieser Vereinfachung über die ganze Breite von 1 bis 9 durch den Mittelwert 5 und die Standardabweichung 2 definiert.

Zu den gleichen Ergebnissen wie bei der Berechnung der Konfidenzintervalle über die Formeln in ► Abschn. 2.6.2.2 gelangt man durch Anwendung des Normwertrechners (► <https://www.psychometrica.de/normwertrechner.html>).

- Das Konfidenzintervall wird für den z -Wert bestimmt, der nach ► Tab. 2.23 der Intervallmitte des beobachteten Stanine-Wertes entspricht. Beispiel: Bei Stanine 3 wird für $z = -1$ das Konfidenzintervall bestimmt. Es beträgt bei $r_{tt} = .90$, Urteilsicherheit 90 % und zweiseitiger Fragestellung $z = -1,39$ bis $-0,41$. Für Stanine 1 verwenden wir $z = -2$ und ermitteln als Intervall von $-2,29$ bis $-1,31$ (z -Werte).

2. Mithilfe von □ Tab. 2.23 wird geprüft, bis in welche Stanine-Bereiche das Konfidenzintervall reicht. Im Beispiel Stanine 3 reicht der z -Wert $-1,39$ bis in den Stanine-Bereich 2 und für $z = -0,41$ bis in den Bereich von Stanine 4. Für Stanine 1 reicht das Konfidenzintervall mit $z = -1,31$ in den Stanine-Bereich 2. Der untere Wert für Stanine 1 kann nur 1 betragen, weil dies der kleinstmögliche Wert auf der Stanine-Skala ist – egal wie weit der z -Wert in den negativen Bereich weist. Analog dazu kann das Konfidenzintervall für Stanine 9 nach oben nicht weiter als 9 reichen.

2.6.4.2 Sten-Werte

Sten-Werte (Abkürzung für standard scale of ten units) haben einen Mittelwert von 5,5 und eine Standardabweichung von 2. Sie boten bei ihrer Einführung gegenüber den Stanine-Werten für manche Zwecke den Vorteil, dass man eine Stichprobe damit exakt halbieren kann, d. h., die Sten-Werte 1 bis 5 beschreiben die eine Hälfte, die Sten-Werte 6 bis 10 die andere Hälfte der Verteilung (Canfield 1951). Wie die Stanine-Werte stellen sie Wertebereiche dar, die – mit Ausnahme der Randbereiche 1 und 10 – in einer Normalverteilung jeweils 0,5 Standardabweichungen breit sind (□ Tab. 2.24). Die am Rand der Verteilung liegenden Werte 1 und 10 sind nach unten bzw. oben offen.

Sten-Werte

Normwerte bei Bedarf selbst rechnen

Wenn für einen Test keine Normtabellen vorliegen, kann man dennoch durch eine einfache Berechnungsmethode jeden beliebigen Rohwert in einen Normwert transformieren. Voraussetzung dafür ist, dass die Rohwerte normalverteilt und ihr Mittelwert und ihre Standardabweichung bekannt sind. Ferner sollten die Testwerte von einer passenden Personengruppe stammen.

Beispielsweise liegen von 200 Studierenden Testwerte vor. Der Mittelwert M der Testwerte betrage 65, die Standardabweichung s ist 12. Eine Person habe in dem Test einen Rohwert X von 77 erzielt. Die Frage ist, wie viele Standardabweichungen der Testwert von 77 über oder unter dem Mittelwert der Vergleichsgruppe liegt. Unter Verwendung der Formel für z -Werte (□ Abb. 2.42) erhalten wir einen z -Wert von 1. Der Testwert liegt also genau eine Standardabweichung über dem Mittelwert von 65. In □ Abb. 2.42 lässt sich ablesen, dass diesem Mittelwert ein Standardwert von 110 entspricht. Das gleiche Ergebnis erhalten wir, indem wir den z -Wert mit 10 multiplizieren und zu 100 addieren (vgl. die Umrechnungsformel in □ Abb. 2.42).

□ Tab. 2.24 Sten-Werte

Sten-Wert	1	2	3	4	5	6	7	8	9	10
Relative Häufigkeit (in %)	2,3	4,4	9,2	15,0	19,2	19,2	15,0	9,2	4,4	2,3
z -Wert	< -2,001	-1,501 bis -2,0	-1,01 bis -1,5	-0,501 bis -1,0	-0,001 bis -0,5	0,001 bis 0,5	0,501 bis 1,0	1,01 bis 1,5	1,501 bis 2,0	> 2,001

Summe der Prozentwerte rundungsbedingt nicht exakt 100.

Prozentrang = relativer Anteil von Personen, die einen schlechteren oder gleich guten Wert erzielen

2.6.4.3 Prozentränge

Prozentränge beschreiben den relativen Anteil von Personen in der Vergleichsgruppe, der einen schlechteren oder gleich guten Wert wie eine Testperson erzielt hat. □ Tab. 2.25 illustriert das Vorgehen dabei.

Wie man sieht, wird für jeden Rohwert berechnet, wie viel Prozent der Personen einer Vergleichsgruppe diesen Wert erzielt haben. Danach werden diese Prozentanteile vom niedrigsten bis zum höchsten Rohwert kumuliert. Erzielt eine Person nun einen Rohwert von 9, so erhält sie im vorliegenden Beispiel den Prozentrang 70. Dies bedeutet, dass sie besser oder gleich gut ist wie 70 % der Vergleichsgruppe. Man könnte auch sagen: Nur 30 % der Vergleichsgruppe erzielten einen höheren Wert.

! Vorsicht

Bei Prozenträngen sind keinerlei Annahmen über die Verteilung der Testwerte nötig. Es muss lediglich ein Ordinalskalenniveau vorliegen. Die Transformation besteht darin, dass dem Testwert die relative Position auf der nach Größe ranggereichten Messwerteskala der Bezugsguppe zugeordnet wird.

Ordinalskalenniveau beachten

Prozentränge sind sehr anschaulich und einfach zu verstehen. Sie haben aber auch einen Nachteil. In Rohwertbereichen, die von vielen Personen erzielt werden, steigen die Prozentränge stark an. In Rohwertbereichen, die wenige Personen erreichen, verändern sich die Prozentränge kaum. Im Beispiel in □ Tab. 2.25 „springt“ der Prozentrang von 35 auf 70, wenn man statt eines Rohwerts von 8 einen Wert von 9 erzielt. Es entsteht oberflächlich das Bild eines großen Unterschieds zwischen solchen Personen. Ein Rohwert von 10 oder 11 hingegen führt nur zu kleinen Prozentrangunterschieden (90 vs. 100). Es muss also stets bedacht werden, dass Prozentränge nur über ein Ordinalskalenniveau verfügen und darüber hinausgehende Interpretationen von Prozentrangunterschieden nicht zulässig sind.

2.6.4.4 Anforderungen an Normen

Die Normierung gilt als Gütekriterium eines Tests. Konkret ist zu fordern, dass die Normierungs- oder Eichstichprobe

1. repräsentativ für die intendierte Population der Testteilnehmenden ist,
2. eine Differenzierung nach Personenmerkmalen, die mit der Testleistung korrelieren, vornimmt,
3. hinreichend groß ist und
4. die Erhebung der Daten möglichst aktuell ist.

Repräsentative Zusammenstellung der Normstichprobe

Eine repräsentative Zusammenstellung der Eichstichprobe ist unerlässlich; nur dann ergibt es Sinn, einzelne Personen mit ihrem Punktewert auf den durch die Population definierten Hintergrund zu beziehen. Soll ein Test beispielsweise nur bei Schülerinnen und Schülern der 4. Klasse in einem

□ Tab. 2.25 Beispiel für die Berechnung von Prozenträngen

Rohwert	Anzahl der Personen in der Vergleichsgruppe mit diesem Rohwert	Anteil der Personen in der Vergleichsgruppe mit diesem Rohwert (in %)	Kumulierter Anteil der Personen in der Vergleichsgruppe mit diesem Rohwert (in %)
7	20	10	10
8	50	25	35
9	70	35	70
10	40	20	90
11	20	10	100

bestimmten Bundesland eingesetzt werden, muss die Eichstichprobe repräsentativ für Schülerinnen und Schülern der 4. Klasse in diesem Bundesland sein. Anders verhält es sich bei einem Intelligenztest, der bundesweit eingesetzt werden soll: Da die Intelligenz bildungskorreliert ist, muss die Eichstichprobe entsprechend der Bildung in der Gesamtbevölkerung zusammengesetzt sein. Zudem wurden über Regionen in Deutschland hinweg unterschiedliche mittlere Intelligenztestleistungen gefunden (Ebenrett et al. 2003). Deshalb ist es wichtig, dass die Normierung nicht in einer Region durchgeführt wird, sondern an verschiedenen Orten in Deutschland.

Manchmal hängt die Testleistung mit weiteren Personenmerkmalen zusammen. Beispielsweise ist die kognitive Leistungsfähigkeit mit dem Bildungsgrad und dem Alter von Personen korreliert. Möchte man dies bei der Einordnung der Testwerte einer Person berücksichtigen, sollten nach Alter und Bildungsgrad differenzierte Normen vorliegen. Das bedeutet, dass eine 60-jährige Person nicht mit der Gesamtnormgruppe verglichen wird, sondern nur mit Personen, die in etwa das gleiche Alter haben. Da üblicherweise nicht genügend viele Personen in der Normgruppe sind, die exakt 60 Jahre alt sind, bildet man üblicherweise Subgruppen und fasst beispielsweise alle 55- bis 65-jährigen Personen zusammen. Das Ergebnis könnte sein, dass diese Person eine für 55- bis 65-jährige Personen durchschnittliche Testleistung erbracht hat, die aber verglichen mit allen Personen eher unterdurchschnittlich bewertet würde.

Differenzierung durch
Subgruppenbildung

Kontinuierliche Testnormen

Die vorherigen Ausführungen zur Normdifferenzierung durch Subgruppenbildung offenbart eine Schwäche: Die Subgruppe müssen breit genug gewählt werden, sodass ihr ausreichend viele Personen aus der Normgruppe zugeordnet werden. Das bedeutet im Falle der zuvor erwähnten Alterskategorien, dass eine 55-jährige Testperson ebenso mit allen 55- bis 65-jährigen Personen verglichen wird wie eine 65-jährige Testperson – obwohl beide 10 Jahre auseinanderliegen. Eine 66-jährige Testperson hingegen würde mit der nächsthöheren Altersgruppe (z. B. 66- bis 80-jährige Personen) verglichen – auch wenn sie de facto vielleicht nur ein paar Tage älter ist als die 65-jährige Testperson. Kontinuierliche Testnormen berechnen auf Basis der Gesamtnorm eine altersspezifische Verteilung der Testleistung – und dies für jedes Alter. Die individuelle Leistung kann dann anhand dieser errechneten Verteilung eingeordnet werden. Allgemeiner gesagt wird also bei kontinuierlichen Normen der Verlauf der Mittelwerte und Streuungen über die gesamte Stichprobe als mathematische Funktion abgebildet. Das hat zudem den Vorteil, dass kleine Abweichungen der empirisch vorliegenden Daten von diesen Kurven geglättet werden (s. z. B. Lenhard et al. 2018).

Die Größe der Eichstichprobe richtet sich danach, wie stark die Normen nach Alter, Bildung und/oder Geschlecht differenziert werden. Jeder einzelnen Normtabelle sollten möglichst mehrere Hundert Personen zugrunde liegen. Sinnvoll und aussagekräftig sind neben einer Aufgliederung in Altersgruppen bei Leistungstests auch gesonderte Normen für verschiedene Schultypen. Bei Fragebögen zu bestimmten Merkmalen (z. B. Aggressivität) kann es erforderlich sein, innerhalb der Altersgruppen auch noch nach Geschlecht zu differenzieren.

Ausreichender Umfang der
Normgruppen

Normen sollten aktuell sein

Eine weitere Forderung besteht darin, dass die Normdaten aktueller Herkunft sein sollen. Die Literatur ist voller Beispiele über markante Leistungszuwächse im Laufe der Zeit, teils als Folge allgemein verbesserter Anregungs- und Schulungsbedingungen, teils als Folge spezifischer Ereignisse in Technik, Sport oder Wissenschaft. Diese führen dazu, dass ein und derselbe individuelle Punktwert immer leichter zu erzielen ist. Vor diesem sich ändernden Hintergrund müssten die Verfahren laufend „nachnormiert“ werden. Zumindest sollte überprüft werden, ob die Normen noch aktuell sind. Dazu reicht es aus, eine ausreichend große neue Stichprobe zu erheben und deren Testwerte mit denen der Eichstichprobe zu vergleichen. Die aktuelle Stichprobe ist selbstverständlich nach den gleichen Kriterien auszuwählen wie die Eichstichprobe. Finden sich keine bedeutsamen Unterschiede zwischen den Mittelwerten und den Streuungen, können die „alten“ Normen weiterverwendet werden.

Empfehlungen des Diagnostik- und Testkuratoriums

Das Diagnostik- und Testkuratorium empfiehlt die Prüfung, ob im Manual folgende Fragen hinreichend beantwortet werden (zitiert nach Diagnostik- und Testkuratorium 2018b, S. 114, © Hogrefe):

- Stehen (sofern nötig) für jedes genannte diagnostische Ziel Normen zur Verfügung?
- Ist die Eichstichprobe repräsentativ?
- Im Falle altersspezifischer oder in anderer Hinsicht spezifischer Normen: Sind die Altersintervallbreite und die betreffende Größe der jeweiligen Eichstichprobe angemessen?
- Entspricht die verwendete Normwertskala (z. B. z -Werte) in ihrer Differenziertheit dem im Testmanual formulierten Anspruch zur Differenzierungsfähigkeit des Tests?

2.6.5 Weitere Nebengütekriterien

2.6.5.1 Skalierung

Ist die Verrechnung der Itemergebnisse zu Testwert angemessen?

Das Gütekriterium der Skalierung ist erfüllt, wenn die gewählte Form der Verrechnung der Testitems zu einem Gesamtwert so vorgenommen wird, dass dies der tatsächlichen Merkmalsausprägung gerecht wird. (Bei den meisten Tests werden die Antworten einfach zu einem Gesamtwert addiert). Dies wäre beispielsweise nicht gegeben, wenn man auch Antworten zu Items addieren würde, die eigentlich nichts mit dem intendierten Merkmal zu tun haben. McNeish und Wolf (2020) betonen zudem, dass die Verwendung eines Summen- oder Mittelwertes über alle Testitems impliziert, dass alle Items mit dem gleichen Gewicht in den Testwert eingehen. Faktorenanalysen könnten jedoch ergeben, dass Items unterschiedlich hoch auf dem zugrunde liegenden Faktor laden. Dann erlauben der Summen- und der Mittelwert keine angemessenen Verrechnungen zu einem Testwert. Im Rahmen Probabilistischer Testtheorien kann geprüft werden, ob die vorgenommene Verrechnung das zu messende Merkmal adäquat reflektiert. Beispielsweise nimmt das einparametrische dichotome Rasch-Modell an, dass der Summenwert eine erschöpfende Aussage über die Merkmalsausprägung macht. Lässt sich empirisch zeigen, dass die gewonnenen Testdaten das einparametrische dichotome Rasch-Modell bestätigen, so kann davon ausgegangen werden, dass Summenwerte eine angemessene Form der Verrechnung darstellen.

2.6.5.2 Zumutbarkeit

Die Durchführung eines Tests kann für Testpersonen belastend sein: Teilnehmerinnen und Teilnehmer sind nach der Testdurchführung erschöpft, die Durchführung kostet viel Zeit, bestimmte Fragen empfinden sie ggf. als zu persönlich, und die Beantwortung ist ihnen unangenehm. Wird etwa bei einem Reaktionszeittest verlangt, sehr oft so schnell wie möglich eine Taste zu drücken, so kann dies auch körperlich anstrengend sein. Die Zumutbarkeit ist kein festes Merkmal eines Tests: Ob ein Test zumutbar ist oder nicht, hängt immer auch von der Testperson und den Untersuchungsumständen ab. Es ist abzuwägen, ob der Nutzen durch die Testanwendung in einem angemessenen Verhältnis zu den Belastungen steht, die Probandinnen und Probanden zugezumutet werden.

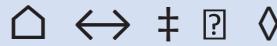
Ist die Belastung durch den Test angemessen?

Zumutbarkeit als „Teil des Messanspruchs“

Manche Tests stellen den getesteten Personen Aufgaben, die bewusst an die Grenze der Zumutbarkeit gehen. Folgende Abbildung zeigt eine hypothetische Vigilanzaufgabe, die häufig Teil von Aufmerksamkeitstestbatterien sind. Die Aufgabe ist es, eine einfache Reizabfolge ($2 \times$ der gleiche Reiz nacheinander) zu entdecken. Der Test dauert bewusst sehr lange (z. B. 30 min) und die Reizabfolgen, auf die zu reagieren ist, sind sehr selten – diese Operationalisierung entspricht der Definition von Vigilanz. Ermüdung und Monotonie werden hierbei also bewusst hervorgerufen.

Hypothetische Vigilanzaufgabe

Nachfolgend erscheint immer eines der folgenden Symbole auf dem Bildschirm.



Drücken Sie die Reaktionstaste so schnell wie möglich, sobald das gleiche Symbol $2 \times$ hintereinander erscheint.

Testdauer: 30 Minuten

2.6.5.3 Akzeptanz

Die Akzeptanz kann als Teilaспект der Zumutbarkeit gesehen werden. Erwachsene können sich daran stören, dass die Aufgaben sie zu sehr an die Schule erinnern oder dass sich Fragen anscheinend nicht auf das Problem beziehen, mit dem sie eine Psychologin oder einen Psychologen aufgesucht haben. Es ist zu bedenken, dass sich Probleme der Zumutbarkeit negativ auf die Akzeptanz eines Tests oder sogar den gesamten diagnostischen Prozess auswirken können. Die Ernsthaftigkeit der Testbearbeitung kann darunter

Stören sich Teilnehmende am Test oder an dessen Komponenten?

leiden. Manche Autorinnen und Autoren verwenden für die Akzeptanz auch den Begriff „soziale Validität“ (Schuler und Stehle 1983). Dennoch handelt es sich bei Zumutbarkeit und Akzeptanz um Nebengütekriterien und nicht um spezielle Formen der Validität. Weitere Ausführungen zur Akzeptanz finden sich in ▶ Abschn. 6.2.1.4.

2.6.5.4 Unverfälschbarkeit

Ist das Testergebnis verfälschbar?

Testpersonen haben manchmal ein Interesse an einem bestimmten Ergebnis. Das können sowohl hohe als auch niedrige Testwerte sein. Beispielsweise hat jemand großen Leidensdruck und möchte daher unbedingt, dass die Krankenkasse eine Psychotherapie bezahlt. Nach einem nicht selbst verschuldeten Unfall will eine Person vielleicht Schmerzensgeld erhalten oder sogar frühzeitig berentet werden. In diesen Fällen besteht ein Interesse an „schlechten“ Testergebnissen, die als Beleg für eine schwere Störung oder Beeinträchtigung gelten. Man kann sich sowohl in Persönlichkeitsfragebögen als auch in Leistungstests (also beispielsweise Intelligenz- oder Konzentrationstests) absichtlich schlecht darstellen (in der Literatur als „faking bad“ bezeichnet). Manchmal ist Testpersonen daran gelegen, sich besonders gut darzustellen. Wenn jemand eine ausgeschriebene berufliche Stelle bekommen möchte, sind „gute“ Testergebnisse in der Eignungsuntersuchung erstrebenswert. Da mit Leistungstests bereits die bestmögliche Leistungsfähigkeit geprüft wird, ist hier in der Regel keine Verfälschung im positiven Sinne möglich – man soll bereits die bestmögliche Leistung erbringen, die sich in der Regel nicht weiter steigern lässt. Dies wäre nur denkbar, wenn man beispielsweise über irreguläre Testkenntnisse verfügen würde, zwischendurch „spicken“ könnte oder mehr Zeit als zulässig bekäme. Bei vertraulicher Handhabung des Testmaterials und ordnungsgemäßer Testdurchführung sind solche Einflüsse jedoch nicht zu erwarten. „Faking good“ ist daher vorwiegend bei Fragebögen sowie in Interviews und Assessment-Centern ein Problem (für eine Übersicht s. Ziegler et al. 2011). Dabei verfälschen Testpersonen nicht immer absichtlich ihre Testergebnisse, auch eine unabsichtliche Selbstdarstellung im positiven Sinne ist als Phänomen bekannt (Paulhus 1984).

Was tun gegen Verfälschung?

Durch Faking-Studien (z. B. Ziegler et al. 2015) können Testautorinnen und -autoren zunächst einmal herausfinden, wie anfällig ihr Instrument für eine absichtliche Verfälschung ist. In der Regel sieht man neben einer „ehrlichen“ Gruppe von Testpersonen 1 oder 2 weitere Gruppen vor, die gebeten werden, sich besonders gut oder besonders schlecht in dem vorliegenden Testverfahren darzustellen. Um ein realistisches Verfälschungsverhalten zu evozieren, gibt man oft als Instruktion „Bearbeiten Sie den Test so, dass es für eine Textexpertin oder einen -experten realistisch aussieht“. Man kann zudem konkrete Stellenausschreibungen beifügen, um zu prüfen, ob gezielte Verfälschungen im Sinne der Stellenanforderungen möglich sind.

Sollte man anhand dieser Studien herausfinden, dass ein Instrument prinzipiell verfälscht werden kann, ist es nicht so einfach, einen Mechanismus gegen die Verfälschung zu implementieren oder einen „Detektor“ für eine Verfälschung einzubauen. Bei Fragebögen scheint zur Verhinderung von Verfälschung lediglich die Verwendung von Forced-Choice-Antwortformaten zu helfen (Brown und Maydeu-Olivares 2013). Verfälschung nach unten („faking bad“) in Leistungstests kann man manchmal an unrealistischen Fehlerhäufungen erkennen.

So konnten Schmidt-Atzert et al. (2004) zeigen, dass bei ernsthaftem Bearbeiten des Tests d2 (Brickenkamp 2002) fasst nie sog. „Buchstabenfehler“ vorkommen (beispielsweise wird kaum „p“ statt „d“ angekreuzt). Da dies simulierte Personen nicht wissen, bauen sie auch solche sehr seltenen Fehler ein – und werden deshalb entdeckt. Schmidt-Atzert et al. (2004) konnten anhand des Kriteriums „2 oder mehr Buchstabenfehler“ 63 % der Simulantinnen und Simulanten entdecken.

2.6.5.5 Fairness

Ein Test gilt als fair, wenn er Personengruppen nicht systematisch benachteiligt. Denkbar ist etwa eine Diskriminierung von Personen aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit. Die Gefahr besteht besonders bei Tests, die explizit Wissen erfassen oder Wissen für die Lösung einer Aufgabe voraussetzen. Zum Beispiel enthält der Test zur Praktischen Alltagsintelligenz (PAI 30; Mariacher und Neubauer 2005) eine Aufgabe, bei der Wege aufgezeigt werden sollen, wie ein Auto in eine eigentlich zu kleine Garage gelangen kann (► Abb. 2.44). Personen aus Großstädten oder mit niedrigen Einkommen könnten durch eine solche und

Besteht keine Benachteiligung von Personengruppen?

Als Sie eines Tages für einen Ausflug mit Ihrer fünfköpfigen Familie Ihr Auto aus der Garage holen wollen, klemmt das elektrische Garagentor in einer Höhe, sodass Sie mit Ihrem Wagen gerade nicht darunter durchfahren können (siehe Foto). Um das Tor reparieren zu können, müssen Sie jedoch erst das Fahrzeug entfernen. Wie bewerkstelligen Sie das, ohne Ihr Auto oder das Garagentor zu beschädigen?

Betrachten Sie bitte genau das Foto und beschreiben Sie in Stichworten, wie Sie Ihr Fahrzeug aus der Garage bringen, um das Garagentor reparieren zu können!



► Abb. 2.44 Item aus dem PAI 30. (Aus Mariacher und Neubauer 2005, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

Bei Fairnessbeurteilung die vom Test intendierte Zielgruppe beachten

Subgruppenunterschiede müssen einen Test nicht automatisch unfair machen

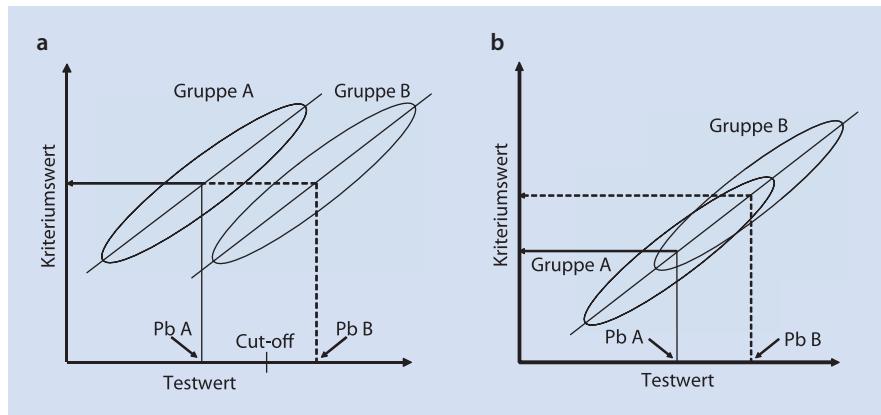
Fairnessmodell nach Cleary

ähnliche Aufgaben benachteiligt sein, etwa wenn ein Großteil dieser Personengruppen kein Auto besitzt.

Ein Test ist aber nicht an sich fair oder unfair: Eine Unfairness ergibt sich erst, wenn der Test in einer Population eingesetzt wird, die zum Teil aus benachteiligten Personen besteht. Dies wäre etwa der Fall, wenn ein sprachloser Intelligenztest als Auswahlinstrument dient und ein Teil der Bewerberinnen und Bewerber die deutsche Sprache nicht fließend beherrscht. Der gleiche Test kann fair sein, wenn in der Bewerberpopulation niemand Schwierigkeiten mit der deutschen Sprache hat.

Mangelnde Fairness eines Tests lässt sich nicht automatisch daran erkennen, dass eine bestimmte Gruppe von Menschen niedrigere Werte erzielt als andere. Beispielsweise könnten Hauptschülerinnen und -schüler im Durchschnitt niedrigere Werte in einem Allgemeinwissenstest als Abiturientinnen und Abiturienten haben. Das könnte sowohl auf eine mangelnde Testfairness als auch auf ein unterschiedlich hohes Allgemeinwissen zurückzuführen sein. □ Abb. 2.45 zeigt die Streudiagramme zweier Gruppen, die sich deutlich in ihren Testwerten unterscheiden. Die Korrelation zwischen Test und Kriterium ist in beiden Gruppen gleich hoch. Gruppe A könnten die Hauptschülerinnen und -schüler und Gruppe B die Abiturientinnen und Abiturienten sein. In □ Abb. 2.45a ist ein, in Bezug auf das vorherzusagende Kriterium, unfairer Test dargestellt: Obwohl sich beide Gruppen in Bezug auf ihre Testwerte unterscheiden, besteht kein Unterschied des mittleren Kriteriumswertes. Nehmen wir an, dass nur Bewerberinnen und Bewerber eingestellt werden, deren Testwert über dem Cut-off-Wert liegen. Probandin A würde abgelehnt, obwohl sie bei der Einstellung eine ebenso gute Kriteriumsleistung zeigen würde wie Probandin B. Fairness ist also nicht gegeben, wenn niedrigere Werte in einem Test nicht auch mit entsprechend niedrigeren Werten im relevanten Kriterium einhergehen, das mit dem Test vorhergesagt werden soll. □ Abb. 2.45b zeigt dagegen einen Test, der nach manchen Fairnessauffassung als fair gelten kann: Unterschiedlich hohen Testwerten entsprechen auch unterschiedliche hohe Kriteriumswerte.

Eine ebensolche Fairnessauffassung vertritt Cleary (1968); ihr zufolge ist ein Test dann fair, wenn bei seiner Anwendung für keine der miteinander verglichenen Gruppen eine systematische Über- oder Unterschätzung der Kriteriumswerte entsteht. Diese Forderung ist Cleary zufolge dann erfüllt, wenn die zur Vorhersage des Kriteriums verwendeten gruppenspezifischen Regressionsgeraden miteinander identisch sind, d. h. gleiche Steigungen aufweisen



□ Abb. 2.45 Testfairness. a Mitglieder der Gruppe A erhalten bei gleichen Kriteriumswerten systematisch niedrigere Testwerte als Mitglieder der Gruppe B. b Unterschiedliche Testwerte gehen mit entsprechenden Kriteriumswerten einher

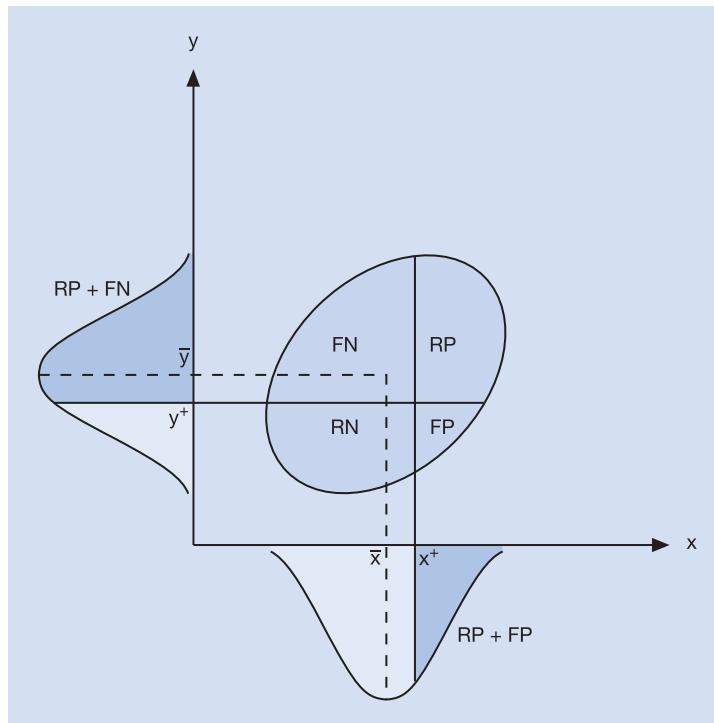
und an derselben Stelle die Ordinate schneiden (► Abb. 2.45b). Für die Prädiktion des Kriteriums aus den Testwerten kann deshalb in einem solchen Fall für alle Messwertträgerinnen und -träger (ohne Berücksichtigung ihrer Gruppenzugehörigkeit) eine gemeinsame Regressionsgerade angenommen werden, ohne dass dadurch einzelne Personen aufgrund ihrer Gruppenzugehörigkeit systematisch bevorzugt oder benachteiligt würden.

Thorndike (1971) diskutiert ein alternatives Fairnesskriterium, bei dem das Verhältnis der durch die Testung ausgewählten Bewerberinnen und Bewerber an allen Bewerberinnen und Bewerbern in den miteinander verglichenen Gruppen gleich oder konstant ist (engl. constant ratio model). Gemäß der schematischen Darstellung in ► Abb. 2.46 entspricht das der Identität der Proportionen ($RP + FP$): $(RP + FP + RN + FN)$. Beispielsweise wären von 100 Bewerberinnen 50 für eine Ausbildung zugelassen worden, so müssten nach dieser Fairnessauffassung von 50 Bewerbern 25 aufgenommen werden. Dadurch läge die Quote der aufgenommenen Frauen und Männer bei jeweils 50 % und entspräche damit der hier skizzierten Quotendefinition von Fairness (identische Selektionsraten über Subgruppen).

Die in den USA viel beachteten Uniform Guidelines for Employee Selection Procedures (► <https://www.uniformguidelines.com/>) sprechen von diskriminierenden bzw. unfairen Auswahlverfahren, wenn die Selektionsrate (also der Anteil aller Ausgewählten an allen Bewerberinnen und Bewerbern) in einer Subgruppe unterhalb von 80 % der Gruppe mit der höchsten Selektionsrate liegt. Im Sinne von ► Abb. 2.46 entspricht die Selektionsrate ebenfalls: $(RP + FP)$: $(RP + FP + RN + FN)$. Würde man also feststellen, dass ein Auswahlverfahren dazu führt, dass 8 von 10 Bewerbern eingestellt würden, aber

Fairnessmodell nach Thorndike

80 %-Regel der Uniform Guidelines for Employee Selection Procedures



► Abb. 2.46 Die 4 Ergebnisse eines Selektionsverfahrens. RP = richtige Positive (Anzahl ausgewählter Personen, die auch tatsächlich erfolgreich sind), FP = falsche Positive (Anzahl ausgewählter Personen, die tatsächlich nicht erfolgreich sind), RN = richtige Negative (Anzahl zurückgewiesener Personen, die tatsächlich auch nicht erfolgreich sind), FN = falsche Negative (Anzahl zurückgewiesener Personen, die tatsächlich erfolgreich sind). x^+ = Cut-off im Zulassungstest, y^+ = Cut-off im Kriterium (schlechterster Kriteriumswert, der noch als Erfolg gilt)

nur 5 von 10 Bewerberinnen, so wäre dies im Sinne der Uniform Guidelines unfair, da die Selektionsrate der Bewerberinnen nicht mindestens 80 % der Selektionsrate der Bewerber entspricht.

Zur Handhabung von Testverfahren für Angehörige unterschiedlicher Personengruppen gibt die International Test Commission konkrete Empfehlungen (ITC 2013).

Beachtung von Fragen der Fairness bei der Testanwendung gemäß der International Test Commission (ITC 2013; Auszüge aus der deutschen Fassung: ZPID 2001)

Wenn Tests mit Angehörigen verschiedener Gruppen durchgeführt werden sollen (z. B. Gruppen, die sich hinsichtlich des Geschlechts, des kulturellen Hintergrunds, der Ausbildung, der ethnischen Abstammung oder des Alters unterscheiden), bemühen sich fachkompetente Testanwendende in jeder Weise, um sicherzustellen, dass ...

- Die Tests für die verschiedenen zu testenden Gruppen unvoreingenommen und angemessen sind.
- Die zu erfassenden Konstrukte in jeder der repräsentierten Gruppen von Bedeutung sind.
- Nachweise über mögliche Gruppenunterschiede in der Testleistung vorliegen.
- Gegebenenfalls Nachweise über Benachteiligungen durch spezifische Items hinsichtlich der ethnischen Zugehörigkeit oder des Geschlechts („Differential Item Functioning“) vorliegen.
- Belege über die Validität eines Tests vorliegen, die die Anwendung für die verschiedenen Gruppen unterstützen.
- Die Auswirkungen von Gruppenunterschieden, die im Hinblick auf den vorrangigen Testzweck nicht bedeutsam sind (z. B. Unterschiede in der Antwortmotivation oder Lesefähigkeit) minimiert werden.
- In allen Fällen die mit der fairen Anwendung von Tests zusammenhängenden Richtlinien im Kontext der örtlich geltenden Grundsätze und Rechtslage interpretiert werden.

Wenn ein Test in mehr als einer Sprache durchgeführt wird (sei es innerhalb eines oder in verschiedenen Ländern), bemühen sich fachkompetente Testanwendende in jeder Weise, um sicherzustellen, dass ...

- Jede Sprach- oder Dialektversion unter Verwendung einer strikten Methodik entwickelt wurde, die den Anforderungen fachkompetenten Vorgehens genügt.
- Aspekte des Inhalts, der Kultur und Sprache von den Testentwicklerinnen und Testentwicklern in sensibler Weise berücksichtigt wurden.
- Die Testleiterinnen bzw. Testleiter in der Lage sind, in unmissverständlicher Form in der vorgesehenen Testsprache zu kommunizieren.
- Die Sprachfähigkeiten des Probanden in der vorgesehenen Testsprache in systematischer Form erfasst werden und die ihm gemäße sprachliche Version vorgegeben oder gegebenenfalls in bilingualer Form getestet wird.

(ZPID 2001, S. 16; gendergerechte Formulierungen wurden durch die Autoren dieses Lehrbuchs vorgenommen)

Die Diskussion über die Fairness von Tests ist weiterhin aktuell, wie die große Zahl von Publikationen zu diesem Thema zeigt. Diagnostische Verfahren, die nicht völlig fair gegenüber Minoritäten sind und dennoch bei diesen eingesetzt werden, sind ethisch und gesellschaftliche sowie – unter Umständen auch – juristisch problematisch. Daher wird weiter nach Methoden zur Verbesserung der Fairness gesucht (für eine Übersicht s. Colella et al. 2017).

2.6.5.6 Ökonomie

Der Einsatz diagnostischer Instrumente kann mit erheblichen „Kosten“ verbunden sein: Die Anschaffung eines Tests kostet Geld; für die Durchführung und die Auswertung muss man Zeit investieren; eventuell müssen auch Ressourcen aufgewendet werden, um sich mit dem Test vertraut zu machen. Die Zeit der Testperson ist ein weiterer Kostenfaktor, auch wenn dieser schwer in Geldeinheiten gemessen werden kann. Ähnliches gilt für Interviews und Verhaltensbeobachtungen: Sie müssen sorgsam geplant werden, Materialen zur Durchführung und Auswertung müssen erstellt werden, das mit der Durchführung und Auswertung betraute Personal muss geschult werden, bei der Durchführung sind oft 2 oder mehr Personen zugegen. Zudem ist die Auswertung häufig aufwendiger als bei Testverfahren. Ein diagnostisches Instrument ist ökonomisch, wenn zum Einsatz nötige Aufwände (Zeit, Geld, Personal) in sinnvoller Relation zum diagnostischen Nutzen stehen. Monetäre Nutzenanalysen (► Abschn. 5.2.3) können helfen, den durch ein diagnostisches Instrument zu erwartenden Mehrwert zu beurteilen.

Ist der Testeinsatz ökonomisch?

2.6.5.7 Nützlichkeit

Ein ökonomischer Test ist nicht unbedingt auch nützlich. Misst der Test ein Merkmal, für das sich niemand interessiert, ist er nicht nützlich – auch wenn er gratis und in 5 min durchzuführen und auszuwerten ist. Entscheidend ist die praktische Relevanz des Merkmals, das gemessen wird: Beispielsweise hilft der Test, eine behandlungsbedürftige psychische Störung zu erkennen oder einen Beruf zu finden, der den Interessen oder den Fähigkeiten der Testperson entspricht. Liegen bereits diagnostische Verfahren vor, die dieses Merkmal messen, so ist dies bei der Beurteilung der Nützlichkeit zu berücksichtigen. Der Test ist nur dann nützlich, wenn er das Merkmal zuverlässiger, valider und/oder ökonomischer erfassen kann als andere Verfahren.

Ist der Test nützlich?

Nützlichkeit neuer Konstrukte

Nicht nur für neue Tests muss ihre Nützlichkeit gegenüber bereits existierenden Tests belegt werden, dasselbe gilt auch für psychologische Konstrukte. Nach Geiger et al. (2018) sollten neue Konstrukte nur als solche aufgefasst und etabliert werden, wenn gezeigt werden kann, dass sie die folgenden 3 Kriterien erfüllen:

1. Sie sollten messtheoretisch fundiert erfassbar sein (z. B. dadurch, dass sich die intendierte Faktorenstruktur zeigt).
2. Sie sollten von bereits etablierten Konstrukten unterscheidbar sein (d. h., es sollte diskriminante Validitätsbelege geben).
3. Sie sollten bedeutsame Gegebenheiten des realen Lebens vorhersagen, und zwar inkrementell zu bereits etablierten Konstrukten (vgl. inkrementelle Validitätsbelege, ► Abschn. 2.6.3.4).

Beispiele für „neue“ Konstrukte, die diesen 3 Kriterien nicht gerecht werden, gibt es viele. So zeigen Geiger et al. (2018), dass das Konstrukt „Nachsichtigkeit mit sich selbst“ (engl. self-compassion) eher als Facette von Neurotizismus denn als eigenes Konstrukt aufzufassen ist. Ähnliche Erkenntnisse liegen für das Konstrukt „Ausdauer und Leidenschaft im Verfolgen langfristiger Ziele“ (engl. grit) vor, das sich als nicht von Gewissenhaftigkeit unterscheidbar erwiesen hat (Credé et al. 2017).

2.7 Zusammenfassung

Psychologische Tests (und Fragebögen) werden als Messmethode verstanden, bei der Personen auf standardisierte Reizvorlagen wie Aufgaben, Bilder, Fragen, Aussagen oder ganze Texte reagieren. Die Psychologische Diagnostik hat eine Vielzahl unterschiedlicher psychologischer Tests hervorgebracht, die sich u. a. in der Art der Reizvorlage und im Antwortformat unterscheiden. Ziel all dieser Testverfahren ist es, einen wissenschaftlich begründbaren und empirisch belegten Rückschluss auf psychologische Merkmale zu ermöglichen. Wie es gelingen kann, aufgrund von beobachtetem Verhalten in einem Test auf ein latentes Merkmal zu schließen, wird von Testtheorien spezifiziert. Bestehende Testtheorien lassen sich grob unterteilen in die Klassische Testtheorie und Item-Response-Theorien.

Eine Kernannahme der Klassischen Testtheorie ist, dass alle psychologischen Messungen fehlerbehaftet, also nicht perfekt reliabel sind. Sie definiert Reliabilität als den Anteil der Varianz der wahren Werte an der Varianz der beobachteten Werte. Dieses ist recht eingängig, vermutlich findet daher die Klassische Testtheorie in der Praxis der Testentwicklung auch viel Anklang. Allerdings gibt es auch Kritik an der Klassischen Testtheorie.

Je nach Messintention (Quantifizierung oder Klassifikation) und Itemergebnissen (z. B. dichotome oder mehrkategoriale Antworten) können unterschiedliche Item-Response-Theorien herangezogen werden. Allen Item-Response-Theorien gemeinsam ist, dass sie das Antwortverhalten als Wahrscheinlichkeitsfunktion in Abhängigkeit von der Ausprägung des zu messenden Merkmals beschreiben. So geht das 1PL-Modell davon aus, dass die Wahrscheinlichkeit dichotomer Antworten (Item gelöst vs. nicht gelöst) alleine durch die Schwierigkeit der Items und die zu messende Fähigkeit der Testperson beschrieben werden kann. Das 2PL-Modell geht davon aus, dass ergänzend dazu die „Diskriminationsfähigkeit“ der Items zu beachten ist, die von Item zu Item variieren kann. Im 3PL-Modell wird zusätzlich dem Umstand Rechnung getragen, dass manche Items durch Raten der Testpersonen gelöst werden.

Ordinale Rasch-Modelle können auf Fragebogenitems, die aus Beurteilungsskalen mit mehreren Antwortkategorien bestehen, angewendet werden. Sie ermöglichen eine Aussage darüber, wie wahrscheinlich die Wahl einer Antwortkategorie in Abhängigkeit von der Merkmalsausprägung der Person und den Schwellenlokationen des Items ist.

Sofern nicht die Ausprägung von Merkmalen, sondern die Klassenzugehörigkeit von Personen das interessierende Ergebnis eines Tests ist, kann die Latent-Class-Analyse herangezogen werden. Sie ermöglicht Aussagen über die wahrscheinliche Klassenzugehörigkeit von Personen in Abhängigkeit von deren Antwortmuster (d. h. deren Antworten auf alle Items eines Tests).

Bei der Konstruktion von Tests ist ein schrittweises Vorgehen zu empfehlen, das mit der Definition des Messgegenstands beginnt und mit einer hinreichenden Validierung des Tests bzw. des Fragebogens endet. Je nach Ergebnis der Item- und Testanalysen müssen Teile des Tests überarbeitet und erneut geprüft werden – Testentwicklung kann also ein iteratives Vorgehen erfordern.

Es werden 4 unterschiedliche Itementwicklungsstrategien unterschieden, die deduktive, die induktive, die kriteriumsorientierte und die exterale Methode. Losgelöst davon muss eine Entscheidung über das passende Antwortformat (von vielen möglichen Antwortformaten) der Items getroffen und deren Vor- und Nachteile bedacht werden.

Itemanalysen nach Klassischer Testtheorie umfassen in der Regel: Itemschwierigkeit, Itemstreuung, Trennschärfe, Itemladungen auf Faktoren und ggf. die Itemvalidität. Item-Response-Theorien sind in ihren Annahmen detaillierter als die Klassische Testtheorie – ihre Annahmen sind auf Ebene der Items formuliert. Daher bestehen Itemanalysen vorwiegend aus einer Prüfung, welche Items zu den in Item-Response-Theorien inhärenten Annahmen passen und welche nicht.

Testentwürfe werden anhand von Testgütekriterien beurteilt. Wir unterscheiden Haupt- und Nebengütekriterien. Hauptgütekriterien sind die Objektivität, die Reliabilität und die Validität. Objektivität bedeutet, dass die Ergebnisse eines diagnostischen Verfahrens unabhängig davon zustande kommen, wer die Untersuchung, die Auswertung und die Interpretation durchführt. Reliabilität wurde bereits im Rahmen der Klassischen Testtheorie als zentrales Konzept erwähnt. Ihr kommt auch für die Einzelfalldiagnostik eine besondere Bedeutung zu, da von ihr wesentlich die Breite der Konfidenzintervalle, die zur Interpretation von Testergebnissen verwendet werden, abhängt. Validität bezeichnet das Ausmaß, in dem Evidenz und Theorie die Interpretation von Testwerten rechtfertigen (AERA et al. 2014, S. 11). Interpretationen von Testwerten lassen sich z. B. dadurch rechtfertigen, dass ein Test mit anderen Tests, die den gleichen Messanspruch haben, hoch korreliert. Sorgfältige Prüfungen dieses Umstands sind durch MTMM-Analysen möglich. Zudem sind bei der Beurteilung der Höhe solcher Korrelationen beeinflussende Randbedingungen und die Symmetrie der Messungen zu beachten.

Die Güte der Normierung stellt ein wichtiges Nebengütekriterium dar. Bei der Erstellung der Normen ist auf eine Repräsentativität der Stichprobe und eine angemessene Normdifferenzierung (z. B. nach Geschlecht) zu achten. Es werden Äquivalentnormen, Variabilitäts- oder Abweichungsnormen und Prozentrangnormen besprochen. Für die Interpretation von den gängigen Variabilitäts- bzw. Abweichungsnormen sollte Anwenderinnen und Anwender deren Logik bekannt sein.

Weiterführende Literatur und Internetressourcen

Zu dem Themenkomplex Testtheorie und Testkonstruktion liegen hervorragende, didaktisch jeweils etwas anders gestaltete Bücher von Bühner (2021), Eid und Schmidt (2014), Moosbrugger und Kelava (2020) und Rost (2004) vor. Alle Bücher mit Ausnahme von Rost (2004) gehen auch auf die Methode der Faktorenanalyse ein, die hier nur sehr knapp skizziert wurde. Weiterführende Informationen zum Thema Validität und Validierung sowie weiteren Testgütekriterien finden sich zudem in den Standards for Educational and Psychological Assessment (AERA et al. 2014).

Eine hilfreiche Onlinequelle für die Berechnung von Normwerten und Konfidenzintervallen ist: ► <https://www.psychometrica.de/index.html>.

?

Übungsfragen

— Abschn. 2.1:

- Was sind zentrale Elemente der Definition von Psychologischen Tests?
- Was versteht man unter einer Testtheorie?
- Was versteht man unter reflexiven Messungen?

— Abschn. 2.2:

- Was sind zentrale Annahmen der Klassischen Testtheorie?
- Was versteht man in der Klassischen Testtheorie unter dem „wahren Wert“?

- Wie lässt sich (zur Reliabilitätsschätzung) bemessen, wie hoch der Anteil der Varianz der wahren Werte und wie hoch der Anteil der messfehlerbedingten Varianz ist?
- Was wird an der Klassischen Testtheorie kritisiert?
- **Abschn. 2.3:**
 - Nennen Sie verschiedene Item-Response-Modelle und die Antwortformate, für die diese verwendet werden können!
 - Was versteht man unter einer itemcharakteristischen Kurve?
 - Wie ist die Itemschwierigkeit im einparametrischen dichotomen Rasch-Modell definiert?
 - Was versteht man unter spezifischer Objektivität?
 - Was versteht man unter lokaler stochastischer Unabhängigkeit?
 - Erläutern Sie die Logik des grafischen Modelltests!
 - Welche Erweiterungen des einparametrischen dichotomen Rasch-Modells gibt es und wodurch unterscheiden diese sich von selbigem?
 - Was beschreiben Item- und Testinformationsfunktion?
 - Welche Schwellenabstände impliziert das Ratingskalenmodell?
 - Wozu dient eine Latent-Class-Analyse?
- **Abschn. 2.4:**
 - Nennen Sie wichtige Schritte des Testentwicklungsprozesses!
 - Nennen und beschreiben Sie grundlegende Methoden der Itemgenerierung!
 - Welche Randbedingungen sind bei der Itemformulierung zu beachten?
 - Was sind die Vor- und Nachteile gängiger Antwortformate?
 - Was versteht man unter einer negativen Itempolung?
- **Abschn. 2.5:**
 - Wie ist Itemschwierigkeit in der Klassischen Testtheorie definiert?
 - Was versteht man unter einer part-whole-korrigierten Trennschärfe?
 - Von welchen Faktoren hängt die Trennschärfe ab?
 - Was sind Eigenwerte im Rahmen einer Faktorenanalyse und wie kann man an deren Verlauf die Zahl der zu extrahierenden Faktoren ermitteln?
 - Was beschreibt der Q-Index?
- **Abschn. 2.6:**
 - Nennen Sie die 3 Hauptgütekriterien und wesentliche Nebengütekriterien!
 - Welche Formen der Objektivität unterscheidet man?
 - Was sind Methoden der Reliabilitätsschätzung und was ist bei deren Anwendung jeweils zu beachten?
 - Welche Formen der Äquivalenz von Messungen gibt es und wodurch unterscheiden sich diese?
 - Inwiefern ist McDonalds Omega dem Koeffizient Alpha vorzuziehen?
 - Wie kann die Reliabilität einer Messung in der Einzelfalldiagnostik genutzt werden?
 - Wodurch wird die Breite eines Konfidenzintervalls, das man um einen Testwert legt, beeinflusst?
 - Welche Rolle spielt die Reliabilität der Messung beim Vergleich von 2 Testwerten einer Person (etwa vor und nach einer Behandlung)?
 - Wofür korrigiert eine doppelte Minderungskorrektur?
 - Wie ist Validität definiert?
 - Anhand welcher Testeigenschaften lassen sich Belege für die Validität von Testwertinterpretationen generieren?
 - Was versteht man unter einem nomologischen Netz?

- Wie kann eine Multitrait-Multimethod-Analyse im Rahmen der Testvalidierung genutzt werden?
- Was versteht man unter retrograden, konkurrenten und prädiktiven Validitätsbelegen?
- Wie beeinflusst die Symmetrie/Asymmetrie von 2 Messungen deren Korrelation?
- Was versteht man unter Variabilitäts- bzw. Abweichungsnormen?
- Nennen Sie gängige Normwertskalen sowie deren Mittelwert und Standardabweichung!
- Wann kann ein Test als unfair gegenüber einer oder mehreren Personengruppen bewertet werden?

Literatur

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychiatric Association. (2015). *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5: Deutsche Ausgabe herausgegeben von Peter Falkai und Hans-Ulrich Wittchen*. Göttingen: Hogrefe.
- Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology* 76, 732–740.
- Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology* 105, 1013–1037.
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment* 9, 9–30.
- Baumgärtel, F., & Thomas-Langel, R. (2015). TBS-TK Rezension. *Psychologische Rundschau* 66, 152–154.
- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika* 74, 137–143.
- Bledow, R., & Frese, M. (2005). Situational Judgment Test on personal initiative. ► <https://docplayer.org/34298704-Situational-judgment-test-on-personal-initiative.html> Zugegriffen: 01.09.2021.
- Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology* 62, 229–258.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin* 117, 187–215.
- Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment* 23, 532–544.
- Brem-Gräser, L. (2001). *Familie in Tieren: Die Familiensituation im Spiegel der Kinderzeichnung. Entwicklung eines Testverfahrens* (8. Aufl.). München: Reinhardt.
- Brickenkamp, R. (2002). *Test d2: Aufmerksamkeits-Belastungs-Test* (9. Aufl.). Göttingen: Hogrefe.
- Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *d2-R: Test d2 – Revision. Aufmerksamkeits- und Konzentrationstest*. Göttingen: Hogrefe.
- Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology* 47, 1334–1346.
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods* 18, 36–52.
- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.
- Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4. Aufl.). München: Pearson.

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait-multimethod matrix. *Psychological Bulletin* 56, 81–105.
- Canfield, A. A. (1951). The "sten" scale—a modified C-Scale. *Educational and Psychological Measurement* 11, 295–297.
- Cattell, R. B. (1946). *Description and measurement of personality*. Oxford: World Book Company.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* 1, 245–276.
- Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods* 19, 651–682.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement* 5, 115–124.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, R. J., Swerdlik, M. E., & Sturman, E. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). New York, NY: McGraw-Hill.
- Colella, A., Hebl, M., & King, E. (2017). One hundred years of discrimination research in the Journal of Applied Psychology: A sobering synopsis. *Journal of Applied Psychology* 102, 500–513.
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology* 104, 1243–1265.
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment* 15, 110–117.
- Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology* 113, 492–511.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52, 281–302.
- Dahlke, J. A., Kostal, J. W., Sackett, P. R., & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. *Journal of Applied Psychology* 103, 980–1000.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence* 35, 13–21.
- DeCastellarnau, A. (2017). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity* 52, 1–37.
- Diagnostik- und Testkuratorium (2018a). *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430*. Berlin, Heidelberg: Springer.
- Diagnostik- und Testkuratorium (2018b). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 03. Jan. 2018. *Psychologische Rundschau* 18, 109–116.
- Diedenhofen, B., & Musch, J. (2017). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment* 33, 336–344.
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling* 13, 440–464.
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 399–412.
- Ebenrett, H. J., Hansen, D., & Puzicha, K. J. (2003). Verlust von Humankapital in Regionen mit hoher Arbeitslosigkeit. *Politik und Zeitgeschichte* 6–7, 25–31.
- Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: different models for different types of methods. *Psychological Methods* 13, 230–253.
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (4. Aufl.). Weinheim: Beltz.
- Entertainment and Sports Programming Network (ESPN) (2012). Serena Williams collapses in opener. ESPN.com news services vom 29. Mai 2012. Zugegriffen: 06. Juni 2020.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Bradford Books/The MIT Press.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4, 272–299.

- Fahrenberg, J., Hampel, R., & Selg, H. (2010). *FPI-R: Freiburger Persönlichkeitsinventar* (8. Aufl.). Göttingen: Hogrefe.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin* 51, 327–358.
- Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology* 97, 1097–1114.
- Frank, F., & Kanning, U. P. (2014). Lücken im Lebenslauf. *Zeitschrift für Arbeits- und Organisationspsychologie* 58, 155–162.
- Freudenstein, J.-P., Strauch, C., Mussel, P., & Ziegler, M. (2019). Four personality types may be neither robust nor exhaustive. *Nature Human Behaviour* 3, 1045–1046.
- Geiger, M., Pfattheicher, S., Hartung, J., Weiss, S., Schindler, S., & Wilhelm, O. (2018). Self-compassion as a facet of neuroticism? A reply to the comments of Neff, Tóth-Király, and Colosimo (2018). *European Journal of Personality* 32, 393–404.
- Gerlach, M., Farb, B., Revelle, W., & Amaral, L. A. N. (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature Human Behaviour* 2, 735–742.
- Golden, J. P., Bents, R., & Blank, R. (2004). *Golden Profiler of Personality (GPOP). Deutsche Adaptation des Golden Personality Type Profiler von John P. Golden*. Bern: Huber.
- Gough, H. G., & Heilbrun, A. B. (1980). *The adjective check list manual*. Palo Alto, CA: Consulting Psychologists Press.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Boston: Pearson.
- Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology* 11, 385–398.
- Gulliksen, H. (1950). *Theory of mental tests*. Hoboken, NJ: John Wiley & Sons Inc.
- Hartmann, P., Reuter, M., & Nyborg, H. (2006). The relationship between date of birth and individual differences in personality and general intelligence: A large-scale study. *Personality and Individual Differences* 40, 1349–1362.
- Hathaway, S. R., McKinley, J. C., & Engel, R. R. (2000). *MMPI-2: Minnesota Multiphasic Personality Inventory-2 (Manual)*. Bern: Huber.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology* 92, 373–385.
- Hautzinger, M., Keller, F., & Kübler, C. (2006). *BDI-II: Beck Depressions-Inventar, Revision*. Frankfurt am Main: Harcourt Test Services.
- Heidenreich, K. (1993). Die Verwendung standardisierter Tests. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (3. Aufl., S. 389–406). München: Oldenbourg.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist* 58, 78–80.
- Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods* 1, 104–121.
- Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods* 2, 175–186.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21, 1157–1164.
- Höft, S., & Muck, P. M. (2009). TBS-TK Rezension: "Golden Profiler of Personality (GPOP). Deutsche Adaptation des Golden Personality Type Profiler von John P. Golden". *Report Psychologie* 34, 322–323.
- Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement* 60, 523–531.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185.
- Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York, NY: Guilford Press.
- Hoyle, R. H., & Duval, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 302–317). Thousand Oaks, CA: Sage Publications.
- International Test Commission (ITC). (2013). ITC Guidelines on Test Use. 8th October, 2013, Version 1.2. Final Version. Document reference: ITC-G-TU-20131008. ► https://www.intestcom.org/files/guideline_test_use.pdf. Zugriffen: 26. März 2020.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59, 12–19.
- Kanning, U. P. (2012). Diagnostik zwischen Inkompétence und Scharlatanerie: Phänomen, Ursachen, Perspektiven. *Report Psychologie* 37, 100–113.

- Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing – Principles, applications, and issues*. Belmont, CA: Thomson Wadsworth.
- Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJT). *European Journal of Psychological Assessment* 32, 230–240.
- Kelava, A., & Moosbrugger, H. (2020a). Deskriptivstatistische Itemanalyse und Testwertbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 143–158). Berlin, Heidelberg: Springer.
- Kelava, A., & Moosbrugger, H. (2020b). Einführung in die Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 369–410). Berlin, Heidelberg: Springer.
- Kemper, C., Brähler, E., & Zenger, M. (2014). *Psychologische und sozialwissenschaftliche Kurzskalen für Wissenschaft und Praxis – Eine Einführung*. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.
- Kersting, M. (2006). "DIN Screen": Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. Lengerich: Pabst.
- Kramer, J. (2009). Allgemeine Intelligenz und beruflicher Erfolg in Deutschland: Vertiefende und weiterführende Metaanalysen. *Psychologische Rundschau* 60, 82–98.
- Krauth, J. (1996). Klassische Testtheorie. In K. Pawlik (Hrsg.), *Grundlagen und Methoden der Differentiellen Psychologie* (Enzyklopädie der Psychologie, Serie Differentielle Psychologie und Persönlichkeitsforschung, Bd. 1, S. 647–671). Göttingen: Hogrefe.
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippel, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). Hoboken, NJ: John Wiley & Sons Inc.
- Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in an situational judgment test? *Journal of Applied Psychology* 100, 399–416.
- Krumm, S., Schäpers, P., & Göbel, A. (2016). Motive arousal without pictures? An experimental validation of a hybrid implicit motive test. *Journal of Personality Assessment* 98, 514–522.
- Kruyken, P. M., Emons, W. H., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing* 13, 223–248.
- Kubinger, K. D., & Holocher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum 3*. Göttingen: Beltz.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160.
- Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist* 41, 1183–1192.
- Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID). (2001). Internationale Richtlinien für die Testanwendung, Version 2000. Deutsche Fassung. ► https://www.zpid.de/pub/tests/itec_richtlinien.pdf. Zugriffen: 15. Apr. 2020.
- Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment* 25, 112–125.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R: Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mariacher, H., & Neubauer, A. (2005). *PAI 30: Test zur Praktischen Alltagsintelligenz*. Göttingen: Hogrefe.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology* 52, 81–90.
- McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Laurence Erlbaum Associates.
- McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–181). New York, NY: Guilford Press.
- McNeish, D., & Wolf, M. (2020). Thinking twice about sum scores. *Behavior Research Methods* 52, 2287–2305.
- Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin* 115, 300–307.
- Menold, N., & Bogner, K. (2015). *Gestaltung von Ratingskalen in Fragebögen*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften.

- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Read, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist* 56, 128–165.
- Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement* 25, 707–726.
- Mittring, G., & Rost, D. H. (2008). Die verfixten Distraktoren. Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests (für besser Begabte und Hochbegabte). *Diagnostica* 54, 193–201.
- Moosbrugger, H. (2012a). Item-Response-Theorie (IRT). In L. Schmidt-Atzert, & M. Amelang (Hrsg.), *Psychologische Diagnostik* (2. Aufl., S. 62–83). Berlin, Heidelberg: Springer.
- Moosbrugger, H. (2012b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103–117). Berlin, Heidelberg: Springer.
- Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Berlin, Heidelberg: Springer.
- Moosbrugger, H., & Schermelleh-Engel, K. (2012). Exploratorische (EFA) und konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 325–343). Berlin, Heidelberg: Springer.
- Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.), *PISA 2009 – Bilanz nach einem Jahrzehnt* (S. 23–71). Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).
- Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality* 59, 56–68.
- Organisation for Economic Co-operation and Development (OECD). (2000). *PISA 2000: Beispielaufgaben aus dem Lesekompetenztest*. ► <https://www.oecd.org/berlin/39803735.pdf>. Zugriffen: 02. April 2020.
- Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung*. Göttingen: Hogrefe.
- Oswald, W. D., & Roth, E. (1997). *Der Zahlen-Verbindungs-Test*. Göttingen: Hogrefe.
- Pargent, F., Hilbert, S., Eichhorn, K., & Bühner, M. (2019). Can't make it better nor worse. *European Journal of Psychological Assessment* 35, 891–899.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology* 46, 598–609.
- Petermann, F. (2014). *WAIS-IV: Wechsler Adult Intelligence Scale. Deutschsprachige Adaptation nach David Wechsler*. Frankfurt am Main: Pearson Assessment.
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin* 135, 322–338.
- Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal lengths. *Proceedings of the Annual Convention of the American Psychological Association* 5(Pt. 1), 143–144.
- Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the Big Five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment* 23, 32–38.
- Raven, J. G. (1965). *Standard Progressive Matrices*. Cambridge: University Press.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin* 130, 261–288.
- Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung: *Zeitschrift für Soziologie* 9, 222–245.
- Rost, J. (1999). Test- und Fragebogenanalysen. In B. Strauß, H. Haag, & M. Kolb (Hrsg.), *Datenanalyse in der Sportwissenschaft* (S. 455–480). Schorndorf: Hofmann.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003a). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology* 88, 1068–1081.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003b). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology* 56, 573–605.
- Sargent, S. S. (1940). Thinking processes at various levels of difficulty. *Archives of Psychology* 249, 5–58.

- Schaarschmidt, U., & Fischer, A. W. (2008). *AVEM – Arbeitsbezogenes Verhaltens- und Erlebensmuster* (3. Aufl.). Göttingen: Hogrefe.
- Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research* 25, 341–383.
- Schmalt, H. D., Sokolowski, K., & Langens, T. A. (2000). *Das Multi-Motiv-Gitter für Anschluss, Leistung und Macht MMG*. Frankfurt am Main: Swets.
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist* 47, 1173–1181.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* 124, 262–274.
- Schmidt-Atzert, L. (2007). *Objektiver Leistungsmotivationstest OLMT – Software und Manual* (2. Aufl., unter Mitarbeit von M. Sommer, M. Bühner & A. Jurecka). Mödling: Schuhfried.
- Schmidt-Atzert, L., Hommers, W., & Hess, M. (1995). Der I-S-T 70: Eine Analyse und Neubewertung. *Diagnostica* 41, 108–130.
- Schmidt-Atzert, L., Bühner, M., Rischen, S., & Warkentin, V. (2004). Erkennen von Simulation und Dissimulation im Test d2. *Diagnostica* 50, 124–133.
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality* 47, 609–612.
- Schuler, H., & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. *Zeitschrift für Arbeits- und Organisationspsychologie* 27, 33–44.
- Schultze, M. (2017). Constructing subtests using ant colony optimization [Dissertation]. Berlin: Freie Universität Berlin.
- Schulze, J., West, S. G., Freudenstein, J.-P., Schäpers, P., Mussel, P., Eid, M., & Krumm, S. (2020). Hidden framings and hidden asymmetries in the measurement of personality – A combined lens-model and frame-of-reference perspective. *Journal of Personality*, 89, 357–375.
- Sparfeldt, J. R., Kimmel, R., Lowenkamp, L., Steingraber, A., & Rost, D. H. (2012). Not read, but nevertheless solved? Three experiments on PIRLS multiple choice reading comprehension test items. *Educational Assessment* 17, 214–232.
- Staatsinstitut für Schulqualität und Bildungsforschung (ISB). (2004). Jahrgangsstufe 7: Mathematik. Genehmigter Lehrplan. ► <https://www.isb-gym8-lehrplan.de/contentserv/3.1.neu/g8.de/index.php?StoryID=26298>. Zugegriffen: 20. April 2020.
- Stemmler, G., Hagemann, D., Amelang, M., Spinath, F. M., Hasselhorn, M., Kunde, W., & Schneider, S. (2016). *Differentielle Psychologie und Persönlichkeitsforschung*. Stuttgart: Kohlhammer.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment* 80, 99–103.
- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence* 35, 401–426.
- Thielsch, M. T., Lenzner, T., & Melles, T. (2012). Wie gestalte ich gute Items und Interviewfragen? *Praxis der Wirtschaftspsychologie II: Themen und Fallbeispiele für Studium und Praxis*, 221–240.
- Thissen, A., Koch, M., Becker, N., & Spinath, F. M. (2016). Construct your own response: The cube construction task as a novel format for the assessment of spatial ability. *European Journal of Psychological Assessment* 34, 304–311.
- Thissen, A., Spinath, F. M., & Becker, N. (2019). The cube construction task allows for a better manipulation of item difficulties than current cube rotation tasks *European Journal of Psychological Assessment*, e-pub ahead of print. doi: ► <https://dx.doi.org/10.1027/1015-5759/a000534>.
- Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement* 8, 63–70.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. New York, NY: Routledge.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods* 15, 96–110.
- Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. *European Journal of Psychological Assessment* 34, 1–5.
- Wirtz, M. A. (2013). *Dorsch – Lexikon der Psychologie* (16. Aufl.). Bern: Huber.

- Wittchen, H.-U., Zaudig, M., & Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen/Achse II: Persönlichkeitsstörungen*. Göttingen: Hogrefe.
- Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade, & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum Press.
- Wittmann, W. W. (2012). Principles of symmetry in evaluation research with implications for offender treatment. In T. Bliesener, A. Beelmann, & M. Stemmler (Eds.), *Antisocial behavior and crime: Contributions of developmental and evaluation research to prevention and intervention* (pp. 357–368). Cambridge: Hogrefe.
- World Health Organization (WHO) (2018). *International Statistical Classification of Diseases and Related Health Problems (ICD-11)*. Genf: World Health Organization.
- Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer.
- Ziegler, M. (2014). Stop and state your intentions! *European Journal of Psychological Assessment* 30, 239–242.
- Ziegler, M., & Reichert, A. (2017). TBS-TK Rezension: „Adaptives Intelligenz Diagnostikum 3 (AID 3)“. *Psychologische Rundschau* 68, 237–239.
- Ziegler, M., Kemper, C. J., & Kruyken, P. (2014). Short scales—Five misunderstandings and ways to overcome them. *Journal of Individual Differences* 34, 185–189.
- Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods* 18, 679–703.
- Ziegler, M., MacCann, C., & Roberts, R. (2011). *New perspectives on faking in personality assessment*. Oxford: Oxford University Press.



Diagnostische Verfahren

Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang

Inhaltsverzeichnis

- 3.1 Einleitung – 211**
- 3.2 Leistungstests – 216**
 - 3.2.1 Allgemeines zu Leistungstests – 216
 - 3.2.2 Aufmerksamkeits- und Konzentrationstests – 225
 - 3.2.3 Intelligenztests – 253
 - 3.2.4 Spezielle Fähigkeitstests – 284
 - 3.2.5 Entwicklungstests – 288
 - 3.2.6 Schultests – 301
- 3.3 Persönlichkeitsfragebögen – 303**
 - 3.3.1 Persönlichkeitsmerkmale und ihre Messung – 303
 - 3.3.2 Allgemeine Vor- und Nachteile von Persönlichkeitsfragebögen – 312
 - 3.3.3 Persönlichkeitstestsysteme – 322
 - 3.3.4 Verfahren zur Erfassung aktueller Zustände – 361
- 3.4 Objektive Persönlichkeitstests – 367**
 - 3.4.1 Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität – 369
 - 3.4.2 Objektiver Leistungsmotivations-Test (OLMT) – 371
 - 3.4.3 Implizite Assoziationstests (IAT) – 375
 - 3.4.4 Weitere Forschung zu objektiven Persönlichkeitstests und impliziten Assoziationstests – 376
 - 3.4.5 Weitere digitale Ansätze – Machine Learning und künstliche Intelligenz – 377
 - 3.4.6 Objektive sprachbasierte Eignungsdiagnostik – 381
- 3.5 Projektive Verfahren – 382**
 - 3.5.1 Klassische projektive Tests – 384
 - 3.5.2 Abgeleitete Testprinzipien und semiprojektive Tests – 393
 - 3.5.3 Zeichnerische und Gestaltungsverfahren – 398

- 3.6 Verhaltensbeobachtung und -beurteilung – 400**
 - 3.6.1 Arten der Verhaltensbeobachtung – 401
 - 3.6.2 Systematische Verhaltensbeobachtung – 407
 - 3.6.3 Verhaltensbeurteilung – 415
 - 3.6.4 Gütekriterien von Beobachtungs- und Beurteilungsverfahren – 419
- 3.7 Diagnostisches Interview – 426**
 - 3.7.1 Standardisierte strukturierte Interviews – 432
 - 3.7.2 Interviews selbst konstruieren – 443
 - 3.7.3 Techniken der Gesprächsführung – 451
- 3.8 Zusammenfassung – 460**
- Literatur – 461**

3.1 Einleitung

Dieses Kapitel befasst sich mit unterschiedlichen diagnostischen Verfahren. Nicht nur Laien denken bei „diagnostischen Verfahren“ zuerst an Tests. Fragt man aber Praktikerinnen und Praktiker, wie häufig sie bestimmte Verfahren anwenden, steht das *diagnostische Interview* („exploratives Gespräch“) an erster Stelle, gefolgt von *Verhaltensbeobachtung*; diese Reihenfolge gilt für ältere wie für jüngere Psychologinnen und Psychologen. In Skaleneinheiten ausgedrückt werden Interviews „sehr häufig/immer“ und Verhaltensbeobachtungen „oft“ durchgeführt (Roth und Herzberg 2008). Testverfahren wurden in der Befragung von Roth und Herzberg (2008) in 10 Kategorien unterteilt; die Häufigkeitsangaben zu den einzelnen Kategorien fielen zwangsläufig niedriger aus, als wenn nach einer Gesamtkategorie „Tests“ gefragt worden wäre. Die „Hitliste“ der Tests wird von den Persönlichkeitstests angeführt, dicht gefolgt von den Intelligenztests.

Interview und
Verhaltensbeobachtung werden
häufig angewendet

In diesem Kapitel nehmen Tests im weiteren Sinne, also neben Leistungs- tests auch Persönlichkeitsfragebögen, objektive Persönlichkeitstests, projek- tive Verfahren etc., den größten Raum ein. Aus der Sicht von praktisch täti- gen Diagnostikerinnen und Diagnostikern stellen sie wichtige Hilfsmittel dar: Sie sind sozusagen gebrauchsfertig, enthalten alle Materialien zur Durchfüh- rung und Auswertung sowie Benutzeranweisungen. Da sie in der Regel auch normiert sind (s. dazu ▶ Abschn. 2.6.4), gestatten sie Aussagen über die Aus- prägung der gemessenen Merkmale im Vergleich zu anderen Personen. Das kann aber nur funktionieren, wenn die Tests nicht in die breite Öffentlichkeit gelangen.

Verschiedene Arten von Tests

Testschutz

Psychologische Tests sind in aller Regel urheberrechtlich geschützt. Ihre Nut- zung, Vervielfältigung und Weitergabe sowie das Offenlegen von Testbestand- teilen (z. B. der Testitems oder des Auswertungsschlüssels) bedarf der Zustim- mung der jeweiligen Rechteinhaberinnen bzw. -haber.

Bei kommerziell vertriebenen Tests erlangt man die Zustimmung zur (sach- und instruktionsgemäßen) Nutzung durch den rechtmäßigen Erwerb der Test- materialien. Dabei ist zu beachten, dass der einmalige Erwerb eines Testsets (meist bestehend aus Testmanual, Testheften und Testauswertebögen) nicht zur unbegrenzten oder anderweitigen Nutzung berechtigt. So kann beispielsweise ein als Papier-und-Bleistift-Version erworbener Test nicht einfach von Testnut- zenden über eine eigene Webseite administriert werden. Es ist ebenfalls unzu- lässig, das Testmaterial zu vervielfältigen und dann für die jeweiligen diagno- tischen Zwecke einzusetzen.

Testmaterialien nicht verbreiten

Die ungeschützte Verbreitung von Tests verbietet sich nicht nur aus Sicht des Urheberrechts. Ein angemessener „Testschutz“ dient auch dazu, dass Test- verfahren über längere Zeit eingesetzt und ihrem Messanspruch gerecht wer- den können. Sollten beispielsweise Aufgaben eines Wissenstests bekannt sein, wäre es sehr leicht, in diesem Test gut abzuschneiden, indem man die Antwor- ten vorab recherchiert und memoriert. Damit würde der Wissenstest weniger das Wissen als vielmehr die Vorbereitung und Erinnerungsfähigkeit der Teil- nehmenden messen. Insbesondere wenn Testverfahren in großen Stil zum Ein- satz kommen (z. B. für die Auswahl aller Medizinstudierenden in Deutschland) gilt es, angemessene Maßnahmen des Testschutzes zu ergreifen und zu verhin- dern, dass Testaufgaben über einschlägige Internetportale verbreitet werden.

Dem berechtigten Wunsch von Testautorinnen und -autoren sowie von Testverlagen, ihre Testmaterialien zu schützen, steht die ebenso berechtigte Forderung nach Transparenz hinsichtlich der Testentwicklung und der psychometrischen Qualität entgegen. Dementsprechend müssen Testverfahren für Expertinnen bzw. Experten einsehbar sowie zum Einsatz in der universitären Lehre nutzbar sein. Dies befreit die daran beteiligten Personen jedoch nicht von ihrer Pflicht zum angemessenen und vertraulichen Umgang mit den Testmaterialien. Das Diagnostik- und Testkuratorium (DTK) hat in einer Broschüre wichtige „Dos und Don'ts“ beim Einsatz von Tests in Forschung und Lehre zusammengestellt. Diese ist unter ► https://www.dgps.de/fileadmin/user_upload/foederation/dtk_tests_in_lehre_und_forschung.pdf abrufbar.

Prototypen von Tests gesucht

Bei der Auswahl von Tests, die ausführlicher vorgestellt werden, spielen die folgenden beiden Grundsätze eine große Rolle: Erstens sollten die Leserinnen und Leser mit verschiedenen Typen oder Arten von Test vertraut werden. Sie sollten also etwas über Leistungstests und über Persönlichkeitstests erfahren. Innerhalb dieser Kategorien lassen sich die Verfahren thematisch unterschiedlichen Konstruktbereichen zuordnen (bei Leistungstests also Intelligenz, Konzentration etc.). Die Verfahren unterscheiden sich ferner in Bezug auf die Art, wie sie etwas messen. Beispielsweise werden Persönlichkeitsmerkmale nicht nur mit Fragebögen erfasst, sondern auch mit sog. „objektiven Persönlichkeitstests“ und manchmal sogar durch die Analyse von Sprachproben, die unter standardisierten Bedingungen erhoben werden. Diese thematische und messtechnische Vielfalt sollte in diesem Kapitel abgebildet werden. Damit verbindet sich die Erwartung, dass das Kennenlernen von Prototypen einen Transfereffekt hat: Ein neu auf den Markt kommender Test lässt sich leichter einem Inhaltsbereich und einem Messprinzip zuordnen, wenn man prototypische Vertreter derselben kennt: „Aha, das ist also ein klassischer Fragebogen wie das Freiburger Persönlichkeitsinventar – revidierte Fassung (FPI-R), der einige aus pragmatischen Gründen ausgewählte Persönlichkeitmerkmale erfassen soll.“ Zweitens wird es für die Berufspraxis nützlich sein, die meisten der weitverbreiteten Verfahren zu kennen. Deshalb wurde bei der Auswahl von Prototypen möglichst auf bekannte und in der Praxis häufig eingesetzte Verfahren zurückgegriffen.

Prototypen repräsentieren nicht die ganze Vielfalt der Verfahren. Deshalb haben wir entschieden, im Anschluss an ein ausführlich dargestelltes Verfahren auch auf Alternativen hinzuweisen und dabei sowohl Gemeinsamkeiten als auch Unterschiede zum Prototyp herauszuarbeiten. Damit soll der erwähnte Transfereffekt auf die Berufspraxis verstärkt werden.

Für Informationen über weitere Tests sowie auch ggf. über Neuauflagen hier vorgestellter Tests oder aktuelle Rezensionen stehen verschiedene Quellen zur Verfügung. Diese sind zusammen mit deren Vor- und Nachteilen in □ Tab. 3.1 dargestellt.

Testverfahren, aber auch einige andere standardisierte diagnostische Verfahren (insbesondere diagnostische Interviews) werden von Testverlagen vertrieben, die sie gemeinsam mit Testautorinnen und -autoren entwickeln und pflegen. Das Interview mit Dr. G.-Jürgen Hogrefe gibt Einblicke in die Arbeit einer der führenden Testverlage in Europa. Der Geschäftsführer des Verlages beantwortet diese und andere Fragen.

Tests sollten aktuell oder verbreitet sein

Die nachfolgend vorgestellten Testverfahren untergliedern sich in Leistungstests, Persönlichkeitsfragebögen, objektive Persönlichkeitstests und projektive Verfahren.

Informationsquelle	Beschreibung	Vorteile	Einschränkungen
Testkompendien ^a	Bücher, in denen viele Tests kurz beschrieben und systematisiert werden	<ul style="list-style-type: none"> – Guter Überblick – Sachlich, neutral 	<ul style="list-style-type: none"> – Oft nicht aktuell – Meist keine Bewertung
Lehrbücher zur Psychologischen Diagnostik, Bücher zu Fachgebieten wie Personalauswahl	Lehrbücher wie das vorliegende informieren auch über diagnostische Verfahren. Anders als in Testkompendien wird keine Vollständigkeit angestrebt.	<ul style="list-style-type: none"> – Sachlich, neutral – Meist ausführliche Informationen – Meist auch Bewertung 	<ul style="list-style-type: none"> – Nur ausgewählte Verfahren – Viele nicht mehr ganz aktuell
Online-Testverzeichnis in PSYNDEx Tests ^b	Vom Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) erstelltes und gepflegtes Testverzeichnis. Ab Erscheinungsjahr 1945 über 7.900 Testnachweise (Stand: Mai 2020; jährlicher Neuzugang ca. 150)	<ul style="list-style-type: none"> – Fundierte Testbeschreibungen – Wird oft aktualisiert – Kostenfrei – Weitere Links – Hinweise auf Rezensionen 	<ul style="list-style-type: none"> – Nicht zu allen Tests auch detaillierte Informationen
Testmanuale	Handbuch zum Test, in dem nicht nur die Durchführung, die Auswertung und die Interpretation erklärt werden, sondern auch die Testkonstruktion beschrieben wird und Angaben zu den Gütekriterien gemacht werden.	<ul style="list-style-type: none"> – In der Regel sehr ausführliche Informationen 	<ul style="list-style-type: none"> – In der Regel nur verfügbar, wenn der Test angeschafft wurde – Betrifft immer nur ein Verfahren – Keine unabhängige Bewertung
Testrezensionen	Beschreibung und Bewertung eines Verfahrens durch unabhängige Expertinnen und Experten; wird meist in Fachzeitschriften publiziert.	<ul style="list-style-type: none"> – Beschreibung des Tests – Unabhängige Bewertung – Manchmal mehrere Rezensionen verfügbar – TBS-TK-Rezensionen^c: – Hoch standardisiert – Von mindestens 2 Expertinnen/Experten erstellt – frei verfügbar 	<ul style="list-style-type: none"> – Fachzeitschrift eventuell schwer zugänglich – Informiert immer nur über einen Test – Nicht zu allen Tests verfügbar
Kataloge und Onlineverzeichnis der Testverlage	Kurze Informationen zu den einzelnen Verfahren dienen der Produktinformation.	<ul style="list-style-type: none"> – Wird ständig aktualisiert – Kurze, verständliche Beschreibung der Verfahren 	<ul style="list-style-type: none"> – Dient dem Verkauf, daher eventuell einseitig – Beschränkung auf Angebot des Verlages – Keine detaillierten Informationen.

^a Standardwerk: Brähler et al. (2002)

^b Quelle: ▲ <https://www.psynDEX.de/tests/>

^c TBS-TK = Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen; Testrezensionen nach dem Testbeurteilungssystem des Testkuratoriums unter ▲ <https://www.psynDEX.de/tests/testkuratorium/>

Interview mit Dr. G.-Jürgen Hogrefe



Dr. G.-Jürgen Hogrefe (Psychologe und Verleger, Hogrefe Verlagsgruppe)

Der Hogrefe Verlag wurde 1949 von Ihrem Vater gegründet. Ihr Verlag hat mehrere Schwerpunkte: Bücher, Zeitschriften, Schulungen und eben Tests. Wann kamen die ersten Tests ins Verlagsprogramm?

Zur eigentlichen Gründung des Verlages kam es durch ein Zusammenspiel verschiedener, fast zufälliger Umstände. Einer davon war, dass mein Vater, damals wissenschaftlicher Assistent am Institut für Psychologie der Uni Göttingen, keinen Verlag fand für die Zeitschrift, die er gründen wollte, die *Psychologische Rundschau*. Man kann die Stellung der Psychologie in der damaligen Zeit mit der heutigen Situation nicht vergleichen. Psychologie war eine vergleichsweise junge, kleine und wenig etablierte Wissenschaft. All die Anwendungsfelder, die für uns heute selbstverständlich sind, gab es praktisch noch nicht. Daher das mangelnde Interesse bei den angefragten Verlagen. Als Anekdote darf ich in einem Springer-Buch nicht unerwähnt lassen, dass mein Vater auch dem Springer-Verlag die *Psychologische Rundschau* vergeblich anbot. Man riet ihm – durchaus wohlmeinend – lieber die Finger von seinem

Plan zu lassen. Man könne mit „dieser Psychologie“ eigentlich nur Geld verlieren. In seiner Not, denn die Manuskripte für das 1. Heft warteten schon auf die Veröffentlichung, gründete er seinen eigenen Verlag, zunächst nur für diese eine Zeitschrift. Bekanntlich entwickelte sich das Fach mit großer Dynamik weiter. Mein Vater, zunächst nur Teilzeit-Verleger und als Wissenschaftler an der Universität mitten drin in der Entwicklung, erkannte von daher wohl früher als andere Verlage die Chancen, die darin steckten.

Nun zu Ihrer Frage nach den Tests: Mein Vater entwickelte die Vision, dass wir mit unserer Nähe zum Fach ein qualitativ hochstehendes, wissenschaftlich fundiertes Programm entwickeln sollten, dass die Psychologie mit allen benötigten Publikationen abdeckt, ob es nun Lehrbücher, Monografien, Handbücher, Zeitschriften, für Studium und Wissenschaft, oder eben auch Publikationen für die neu entstehenden Anwendungsfelder seien. Und in letzteren sah er für die Diagnostik eine große Bedeutung. „Tests sind das Handwerkszeug der Psychologie“, war seine Überzeugung. Kurz nachdem das 1. Buch in unserem Verlag erschien, erschien 1953 auch schon der erste Test, der „Intelligenz-Struktur-Test“ (I-S-T) von Rudolf Amtshauer. Ein Verfahren übrigens, das nach vielen Revisionen und Überarbeitungen auch heute noch zu den beliebtesten und weitverbreitetsten Tests zählt. Bald kam es dann auch zur Gründung der sog. „Testzentrale“.

Was hat es mit der Testzentrale auf sich?
Man kann sich das heute auch nicht mehr vorstellen, aber psychologische Testverfahren wurden zu der Zeit ganz frei wie Bücher über Buchhandlungen verkauft. Das heißt, jeder, ob qualifiziert oder nicht, konnte Tests kaufen und anwenden. Mein Vater wollte das für die Hogrefe-Tests auf gar keinen Fall so handhaben. Er gründete eine eigene spezielle Vertriebsorganisation, die „Testzentrale“. In Zusammenarbeit mit

dem Berufsverband Deutscher Psychologen wurden Kriterien für die Bezugsberechtigung von Tests festgelegt. Die Testzentrale verkauft Testverfahren nur an entsprechend qualifizierte Anwender. Im Laufe der Zeit konnte mein Vater auch andere Verlage, die Tests publizierten, von der Idee des kontrollierten Testvertriebs überzeugen und mit ihnen entsprechende Vertriebsverträge machen. Die Testzentrale ist so zur zentralen Bezugsquelle für Testverfahren geworden, nicht nur für Verfahren von Hogrefe, die allerdings heute die überwiegende Mehrheit ausmachen, sondern auch für Verfahren anderer Verlage.

Hogrefe allein hat heute über 2000 Tests im Programm. Hinzu kommen die erwähnten Tests anderer Anbieter. Braucht die Psychologie so viele Tests – und wenn ja, wozu?

Wenn Sie nur den deutschsprachigen Markt betrachten, ist die Zahl zu hoch gegriffen. Etwa 2000 ist die Anzahl aller Hogrefe-Tests in allen Sprachen. Wir sind ja heute in 16 Ländern, in Europa, Nord- und Südamerika mit eigenen Verlagen aktiv und publizieren dort auch Tests, natürlich in den jeweiligen Landessprachen. Aber im deutschen Sprachraum kommen wir auch auf etwa 800 Verfahren. Natürlich ist das viel, aber nicht alles sind große Verfahren mit breitem Anwendungsbereich. Vieles sind auch Verfahren für sehr spezielle diagnostische Fragestellungen.

Es gibt natürlich in unserem Verlagsprogramm auch Verfahren, die sich thematisch überschneiden oder auch gegenseitig „Konkurrenz“ machen. Wir sehen das aber eher als Vorteil. In der diagnostischen Praxis möchte man doch häufig auch auf Alternativen zurückgreifen können, das gleiche Thema mit einem anderen Ansatz angehen etc. Zu einem Thema oder Konstrukt gibt ja immer auch Neuentwicklungen auf dem Hintergrund neuer Methoden und Theorien. Entsprechende Neuentwicklungen möchten wir natürlich anbieten. Andererseits wollen Testanwender bestehende Verfahren weiterverwenden,

mit denen sie vielleicht schon sehr viele Daten erhoben haben, sodass wir diese nicht einfach einstellen können. Aber auch das gibt es, wenn Verfahren nicht mehr nachgefragt werden oder veraltet sind. Letzten Endes entscheidet natürlich der Markt. Wir versuchen ein möglichst breites Angebot zu machen.

Welche Entwicklungen sehen Sie auf dem Testmarkt? Was ändert sich, wo hin führt der Weg?

Wie überall, ist auch hier die Digitalisierung ein ganz entscheidender Treiber. Sie ermöglicht das Entwickeln neuer Testkonzepte, das Umsetzen neuer Methoden, das Erfassen von bisher nicht greifbaren, aber diagnostisch relevanten Aspekten. Neben rein digitalen Verfahren wird es auch weiterhin materialgebundene Verfahren geben, bei denen allerdings auch heute schon erwartet wird, dass die Auswertung digital erfolgen kann. Digital heißt dabei online. Von uns als Verlag wird erwartet, dass wir, unter Berücksichtigung aller datenschutzrechtlichen Aspekte, die anonymisiert anfallenden Daten produktiv zur ständigen Verbesserung der Verfahren verwerten, z. B. zur Überprüfung und Aktualisierung von Normen. Wir können auch online den Anwender gezielter, schneller und effizienter mit neuen Informationen zum Test versorgen, z. B. mit neuen relevanten Forschungsergebnissen, etwa zur Validität. Ein weiterer wichtiger Treiber ist die Globalisierung. Sowohl im klinischen Bereich als auch im berufsbezogenen Eignungsbereich werden Verfahren erwartet, die in den verschiedensten Sprachen quasi als internationale „Goldstandards“ zur Verfügung stehen. Die Rolle des Verlages hat sich sehr geändert und ändert sich weiter. Unsere Aufgabe ist es dabei weiterhin, ein aktiver und gestaltender Partner für Wissenschaft und Anwendung in der psychologischen Diagnostik zu sein. Unser großer Vorteil dabei ist unsere Nähe zum Fach Psychologie. Das Führungsgremium unserer Verlagsgruppe besteht fast ausschließlich aus Psychologinnen und Psychologen, insgesamt beschäftigen wir über 100 Psychologinnen und Psychologen.

Testleistung als Arbeit pro Zeit

Testleistung als Indikator für Fähigkeit, Fertigkeit und Wissen

Maximales Verhalten gesucht

3.2 Leistungstests

3.2.1 Allgemeines zu Leistungstests

„Leistung“ wird in der Physik als Arbeit pro Zeiteinheit definiert. Auch in Leistungstests müssen die Testpersonen arbeiten: Sie rechnen, vergleichen geometrische Figuren miteinander, suchen Fehler in Texten oder bestimmte Figuren unter ähnlichen etc. Gemessen wird, wie viele solcher Aufgaben sie in einer feststehenden Bearbeitungszeit lösen oder wie viel Zeit sie zur Bearbeitung einzelner oder auch aller Aufgaben brauchen. Die geleistete Arbeit ist damit quantifizierbar; das Ergebnis nennen wir *Testleistung*.

Die Testleistung kann als Indikator für eine *Fähigkeit* (z. B. fluide Intelligenz), für eine *Fertigkeit* (z. B. das Beherrschung der Grundrechenarten) oder für *Wissen* verstanden werden. Fähigkeiten werden als das Potenzial zum Fertigkeits- oder Wissenserwerb verstanden.

Die Übergänge zwischen Fähigkeit und Fertigkeit bzw. Wissen sind jedoch fließend. Erstens können auch Fähigkeiten manchmal durch Training verbessert werden. Zweitens stellen auch Fertigkeiten oft eine Voraussetzung zum Erwerb weiterer Fertigkeiten dar bzw. (Vor-)Wissen ist oft für den Aufbau von weiterem Wissen förderlich. Drittens kommt es vor, dass ein und dasselbe Merkmal sowohl als Fähigkeit als auch als Fertigkeit oder Wissen betrachtet wird. So versteht man im Intelligenzbereich unter rechnerischem Denken eher eine Fähigkeit, während in Schultests explizit Rechenfertigkeiten überprüft werden. Wegen der unscharfen konzeptuellen Abgrenzung wird in diesem Buch nicht streng zwischen dem Potenzial und dem Gelernten unterschieden. Die Konstrukte „Aufmerksamkeit“, „Konzentrationsfähigkeit“ und „Intelligenz“ (mit Ausnahme der kristallinen Intelligenz) sind grundsätzlich eher dem Fähigkeitsbereich zuzuordnen, während Schultests eher Fertigkeiten und Wissen erfassen. Entwicklungstests können so konzipiert sein, dass sie beide Aspekte der Leistung messen.

„Gleiche Fähigkeit = gleiche Testleistung?“

Stellen wir uns 10 gleichaltrige Personen vor, die über exakt die gleiche Konzentrationsfähigkeit verfügen (wie auch immer man das feststellen kann). Werden sie in einem Konzentrationstest alle die gleiche Testleistung erzielen? Aus verschiedenen Gründen ist daran zu zweifeln.

- Wegen der begrenzten Messgenauigkeit des Tests wird die gleiche „wahre“ Merkmalsausprägung mit leicht unterschiedlichen Testergebnissen einhergehen (s. dazu „Konfidenzintervalle“ in ▶ Abschn. 2.6.2.2).
- Die Testpersonen haben sich vielleicht unterschiedlich stark angestrengt.
- Sie könnten möglicherweise aufgrund von Testangst ihr Potenzial bei der Testdurchführung nicht gleich gut ausnutzen.
- Sie hatten vielleicht zum Teil schon Erfahrungen mit diesem Test, waren also unterschiedlich geübt.

Es ist also zu beachten, dass es neben der zu messenden Fähigkeit (in diesem Fall der Konzentrationsfähigkeit) auch andere Einflussfaktoren auf die Testleistung gibt. Dies gilt auch für die Diagnostik von anderen Fähigkeiten, Fertigkeiten und Wissen.

Der Umstand, dass sich Testpersonen unterschiedlich stark angestrengt haben können, verdient besondere Beachtung. Bei Leistungstests werden Testpersonen ja explizit angewiesen, ihr Bestes zu geben. Anders als bei

Persönlichkeitsfragebögen, in denen das *typische Verhalten* einer Person interessiert, soll mit Leistungstests das *maximal mögliche Verhalten* erfasst werden. Das gelingt jedoch nur, wenn sich die Testpersonen tatsächlich entsprechend anstrengen. Besonders wenn ein Leistungstest „nur“ zu Forschungszwecken durchgeführt wird, kann die Anstrengung interindividuell stark variieren. Beim Bewerten der Leistung ist daher die Motivation bei der Testbearbeitung zu berücksichtigen. Bei geringer Motivation besteht die Gefahr, mit dem Testwert die gemessene Ausprägung zu unterschätzen.

Doch wie prüft man, wie sehr sich Testpersonen in einem Test angestrengt haben? Zunächst einmal kann man deren grundsätzliche *Leistungsmotivation* erheben. Ein Test zur Messung der Leistungsmotivation ist der Objektive Leistungsmotivations-Test (OLMT; Schmidt-Atzert 2007; ▶ Abschn. 3.4.2). Die Anstrengungsbereitschaft wird in diesem Test über die Leistung bei einer kognitiv nicht anspruchsvollen Aufgabe erfasst. Die Testpersonen müssen mithilfe von 2 Tasten eine gewundene „Straße“ auf dem Bildschirm möglichst schnell nachfahren. Damit soll die aufgabenbezogene Anstrengung erfasst werden. Registriert wird, wie viele Felder sie in 10 Durchgängen von jeweils 1 min Dauer zurücklegen. In einer Studie bearbeiteten 100 studentische Versuchspersonen u. a. auch einen Intelligenztest ohne Zeitbegrenzung, den Standard Progressive Matrices plus (SPM plus). Die Korrelation von $r = .35$ spricht dafür, dass die Intelligenztestleistung möglicherweise von der Leistungsmotivation abhängt.

In anderen Studien wurde die Motivation, bei einem Test gut abzuschneiden, erfasst. Im Unterschied zur allgemeinen, testunspezifischen Leistungsmotivation geht es hier um die aktuelle Motivation, in einem ganz bestimmten Test gute Leistungen zu erzielen. Zur Erfassung der „test-taking motivation“ liegen verschiedene Fragebögen vor. Ein typisches Item lautet: „Ich war extrem motiviert, bei diesen Tests gut abzuschneiden.“ Ein solcher Fragebogen kam etwa in der Studie von Chan et al. (1997) zum Einsatz, in der die Testpersonen einen 2-stündigen kognitiven Leistungstest durchführten. Der Fragebogen war nach dem Test auszufüllen. Danach bearbeiteten die Testpersonen eine Parallelform des Tests. Die Korrelation zwischen der Testmotivation und den Testleistungen betrug $r = .37$ bzw. $.40$.

Während sich die Testpersonen in dieser Studie lediglich vorstellen sollten, dass es sich um eine Eignungsuntersuchung handelt, konnten in anderen Studien die Testergebnisse echter Bewerberinnen und Bewerber mit ihrer Testmotivation korreliert werden. Die eingesetzten Tests dienten dazu, ausbildungs- bzw. berufsrelevantes Wissen zu erfassen. In einer Studie von Clause et al. (2001) mit fast 500 Bewerberinnen und Bewerber für eine Stelle im Polizeidienst korrelierte die vorher erhobene Testmotivation zu $r = .25$ mit der Testleistung. McCarthy et al. (2013, Studie 1) haben bei 1750 Bewerberinnen und Bewerber für einen Medizinstudienplatz zwischen der Leistung in einem fachspezifischen Wissenstest (Chemie, Physik, Mathematik und Biologie) und der danach erfassten Testmotivation eine Korrelation von $r = .19$ ermittelt. Die im Vergleich zu der fiktiven Auswahlsituation (vgl. Chan et al. 1997) niedrigeren Korrelationen in realen Auswahltests könnten damit zusammenhängen, dass die Testmotivation hier vermutlich bei allen Personen hoch oder sehr hoch ist und damit nur wenig variiert.

Leistungstests werden u. a. zur Auswahl von Bewerberinnen und Bewerbern und zur Untersuchung von Personen eingesetzt, die ihren Führerschein verloren haben oder eine Lizenz zur Fahrgastbeförderung anstreben. Die Testpersonen haben ein Interesse daran, gute Ergebnisse zu erreichen. Daher liegt es für sie nahe, sich nicht nur anzustrengen, sondern sich auch gezielt vorzubereiten. Im Internet finden sich zum Teil Informationen über den

Intelligenztestleistung hängt mit Leistungsmotivation zusammen

Leistungstestergebnisse hängen mit der Testmotivation zusammen

Sich fit machen für Testuntersuchung

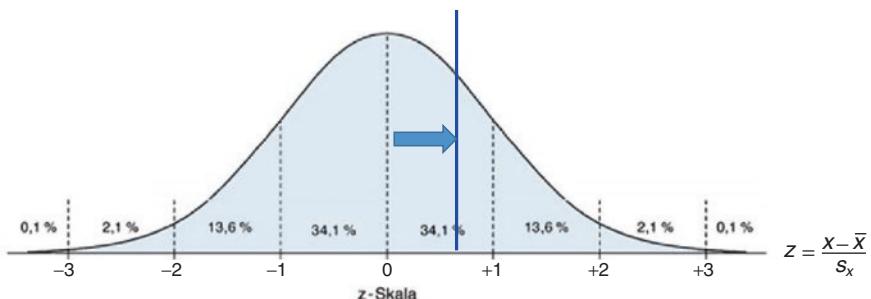
3

Übung und Vorbereitung verbessern die Testleistung

Ablauf und die Inhalte von Eignungsuntersuchungen. Im Buchhandel kann man „Testknacker“ erstehen (z. B. Hesse und Schrader 2015), also Bücher, die über Tests informieren, Trainingsmaterial anbieten und eventuell auf geschickte Strategien hinweisen. Darüber hinaus finden Interessierte kommerzielle Vorbereitungskurse für Führerscheinprüfungen, Auswahltests für ein Medizinstudium oder etwa die Eignungsprüfung zur Auswahl von Pilotinnen und Piloten.

Die gezielte Vorbereitung auf eine Testuntersuchung wird als *Coaching* bezeichnet. In vielen Fällen sammeln Bewerberinnen und Bewerber automatisch Testerfahrung, da sie von mehreren Unternehmen zu Eignungsuntersuchungen eingeladen werden, in denen die gleichen oder ähnlichen Tests eingesetzt werden. Daher stellt sich die Frage: Wie stark wirkt sich Coaching bzw. generell *einschlägige Testerfahrung* auf das Ergebnis in einem Leistungstest aus? Hausknecht et al. (2007) sind in einer Metaanalyse der Frage nachgegangen, wie stark sich unterschiedliche Formen der Testerfahrung auf die mit Tests gemessene kognitive Leistungsfähigkeit auswirken. Sie fanden 50 einschlägige Untersuchungen mit insgesamt rund 130.000 Testpersonen. Der über alle Studien gemittelte Effekt war mit $d=0,24$ nicht sehr groß (wir berichten hier unkorrigierte Effektstärken, da in der diagnostischen Praxis der Messfehler, also mangelnde Reliabilität immer präsent ist). Allerdings konnten die Autorinnen und Autoren Bedingungen entdecken, unter denen der Effekt deutlich größer ist. So kommt es darauf an, ob vorher exakt derselbe Test ($d=0,40$) oder nur ein ähnlicher durchgeführt wurde ($d=0,22$). Die Anzahl der Übungsdurchgänge spielt erwartungsgemäß auch eine Rolle. Nach der 2. Wiederholung ist der Übungsgewinn größer als nach der 1. ($d=0,51$ bzw. $0,24$). Fand ein gezieltes Training (Coaching) statt, verbesserte sich die Leistung stärker als bei reiner Testwiederholung ($d=0,64$ bzw. $0,21$). Außerdem stieg die Testleistung mit der in das Coaching investierten Zeit. Fazit ist, dass Testergebnisse von der Vorerfahrung abhängen. Intensives Coaching und mehrfache Bearbeitung des Tests verbessern die Testleistung deutlich. □ Abb. 3.1 veranschaulicht die Effektstärke $d=0,64$ für Coaching.

In einer neuen Metaanalyse geben Scharfen et al. (2018) einen ausgeweiteten Überblick über den aktuellen Stand der Forschung zur Testwiederholung. Die Autorinnen und der Autor haben einschlägige Literatur zu Retest-Effekten gesichtet, die noch nicht von Hausknecht et al. (2007) erfasst wurde. Coaching-Studien blieben allerdings ausgeschlossen. Sie fanden 122 brauchbare Studien mit Daten von über 150.000 Personen. Eine einmalige Testwiederholung resultierte in einer mittleren Effektstärke von .33, was 3,3 Standardwertpunkten (zu den Normwerten s. ▶ Abschn. 2.6.4 und dort insbesondere □ Abb. 2.42) entspricht. Aufschlussreich sind die Ergebnisse der Moderatoranalysen:



□ Abb. 3.1 Durch Coaching verbessert sich nach Hausknecht et al. (2007) die Testleistung durchschnittlich um 6,4 Standardwertpunkte ($d=z=.64$)

1. Der Effekt nimmt mit der Anzahl der Testwiederholungen zu.
2. Das Retest-Intervall spielt eine, wenn auch kleine, Rolle.
3. Der Retest-Effekt ist kleiner, wenn anstatt des exakt gleichen Tests eine Parallelform verwendet wird.
4. Der Effekt variiert mit dem Messgegenstand. Die größten Effekte wurden bei der Allgemeinen Intelligenz und der Verarbeitungsgeschwindigkeit gefunden.
5. Es fanden sich signifikante Unterschiede zwischen den Testinhalten. Der Retest-Effekt war bei numerischen Aufgaben signifikant kleiner als bei verbalen, figuralen und gemischten Aufgaben.
6. Das Alter der Testpersonen (Bereich von 12 bis 70 Jahren) und deren Intelligenz erwiesen sich nicht als bedeutsame Moderatoren.

Insgesamt konnten damit die Befunde von Hausknecht et al. (2007) repliziert werden, dass der Retest-Effekt mit der Anzahl der Testwiederholungen ansteigt und dass die Wiederholung des gleichen Tests einen stärkeren Effekt hat als die einer Parallelform. Bei Wiederholung des gleichen Tests fällt das Ergebnis um 4 bzw. 3,3 Standardwertpunkte höher aus als bei der 1. Testung. Der etwas höhere Wert bei Hausknecht et al. (2007) ist wohl darauf zurückzuführen, dass auch Coaching-Studien eingeschlossen wurden. Damit lässt sich festhalten, dass sich eine Testperson, die den gleichen Test schon einmal durchgeführt hat, bei der erneuten Testung nicht dramatisch verbessert. Bei Auswahlentscheidungen können 3 Standardwertpunkte mehr allerdings manchmal darüber entscheiden, ob man eine Stelle bekommt oder nicht. Einen deutlichen Vorteil erlangen jedoch Testpersonen, die intensiv auf die Testuntersuchung vorbereitet wurden. Coaching bringt Hausknecht et al. (2007) zufolge einen durchschnittlichen Zugewinn von 6,4 Punkten, bei besonders intensiven Training sogar mehr. Neu ist die Erkenntnis durch Scharfen et al. (2018), dass der Zeitraum zwischen den Messungen nur eine kleine Rolle spielt. Der Retest-Effekt hält also sehr lange an. Scharfen et al. (2018) haben errechnet, dass er erst nach 8,2 Jahren ganz verschwunden ist. Offenbar gibt es Unterschiede zwischen verschiedenen Testarten; bei Tests zur allgemeinen Intelligenz und zur „Verarbeitungsgeschwindigkeit“ (weitgehend identisch mit Konzentration; ► Abschn. 3.2.2.2) scheinen die Wiederholungseffekte am stärksten zu sein.

Wie sich die häufige Wiederholung eines Tests auswirkt, wurde in den oben genannten Metaanalysen nicht systematisch untersucht, weil solche Studien sehr selten sind. Westhoff und Dewald (1990) ließen Versuchspersonen, die sie über Zeitungsinsserate angeworben hatten, im Abstand von jeweils 3–4 Tagen insgesamt 11× den gleichen Durchstreich- oder Rechenkonzentrationstest durchführen. Von der ersten bis zum letzten Messung stieg die Testleistung im Durchschnitt um 62 bzw. 48 % an. Diese Angaben beziehen sich auf Testrohwerte. Wir haben aus den Angaben in der Publikation (Tab. 2 und 4) für jeden der 9 Subtests die Effektstärken berechnet und anschließend über die Subtests gemittelt. Bei den Effektstärken haben wir die Standardabweichungen zum Zeitpunkt 1 als Referenz gewählt, da bei einer Normierung stets die Ergebnisse einer 1. Messung herangezogen werden. Glass Delta beträgt für den Durchstreichtest 2,85 und für den Rechentest 2,57. Dieser Effekt ist extrem groß und bedeutet, dass eine Testperson nach einem Training mit 10 Testdurchgängen mit dem gleichen Test 29 bzw. 26 Standardwertpunkte (!) besser abschneidet als ohne Training.

Eine ähnliche Studie stammt von Albers und Höft (2009). Sie unterscheidet sich aber in 2 wesentlichen Punkten von Westhoff und Dewald (1990). Erstens wurden die Tests, abgesehen von einer 10-minütigen Zwischenpause, ohne Unterbrechung durchgeführt. Zweitens war der Messgegenstand ein anderer. Allerdings weisen die Aufgaben eine starke Ähnlichkeit

Replikation und weitere Erkenntnisse zur Testwiederholung

Sehr starke Übungseffekte bei mehrmals wiederholten Konzentrationstests

mit Konzentrationstestaufgaben auf: Es ging um die Bearbeitung vieler sehr ähnlicher und relativ einfacher Aufgaben (Reaktionszeit ca. 2 s pro Item) unter Zeitdruck. Genauer gesagt analysierten die Autoren die Daten aus einer Eignungsuntersuchung für Nachwuchspilotinnen und -piloten. Der computerbasierte Test bestand aus 10 je 5-minütigen identischen Durchgängen mit einer Pause nach 5 Durchgängen. Der Test soll das räumliche Denken erfassen. Die Testpersonen sehen jeweils die Silhouette eines Flugzeugs und einen Punkt auf dem Bildschirm. Ihre Aufgabe besteht darin, die Position des Punktes aus Sicht der Pilotin/des Piloten zu erkennen. In den 5 min sollten die Testpersonen so viele Aufgaben wie möglich bearbeiten. Vor der Testuntersuchung konnten die Bewerberinnen und Bewerber zu Hause die Tests erproben und vor dem eigentlichen Test mussten sie Übungsaufgaben bearbeiten. Dennoch führte schon die 1. Wiederholung zu einer Leistungssteigerung von umgerechnet 11 Standardwertpunkten ($d=1,09$). Die Leistungskurve stieg weiter steil an, und erreichte mit der 5. Wiederholung einen Zugewinn von 27 Standardwertpunkten gegenüber dem Ausgangswert. Bei der 9. Wiederholung betrug der Zugewinn 29 Standardwertpunkte gegenüber dem Ausgangswert, was für eine starke Abflachung des Leistungszuwachses ab der 5. Testung spricht.

Vorsicht

Bei Tests, in denen kognitiv relativ einfache Aufgaben unter Zeitdruck zu bearbeiten sind, können bei Wiederholung sehr starke Übungseffekte auftreten.

Sehr starke Übungseffekte bei unmittelbarer Testwiederholung

Größe des Übungseffekts selbst berechnen

Größe der Übungseffekte selbst ermitteln

Wie groß der Übungseffekt bei einem Test ist, lässt sich in der Regel einfach aus den Angaben zur Retest-Reliabilität berechnen. Oft findet man dort auch die Mittelwerte und Streuungen für die 1. und die 2. Messung. Mit diesen Angaben lässt sich Glass Delta berechnen, das den durchschnittlichen Übungsgewinn anzeigt. Dazu bildet man die Differenz der Mittelwerte und dividiert sie durch die Standardabweichung der 1. Messung. Im Manual zur elektronischen Version des Aufmerksamkeits- und Konzentrationstests d2-R (Schmidt-Atzert und Brickenkamp 2017; ► Abschn. 3.2.2.2) werden für den Hauptkennwert „Konzentrationsleistung“ (KL) Mittelwerte von 213,9 und 244,6 berichtet. Dividiert man die Differenz von 30,7 Punkten durch die Standardabweichung in der 1. Testung ($SD=38,1$), so erhält man als eine Effektstärke von Glass Delta = 0,81. Bei einer Testwiederholung nach durchschnittlich 16 Tagen ist also mit einer Leistungssteigerung von 8,1 Standardwertpunkten zu rechnen.

Testvorbereitung schwer erkennbar

Beim Einsatz von Tests, deren Ergebnisse mit erheblichen Konsequenzen für die Testperson verbunden sein können (Auswahltest, medizinisch-psychologische Begutachtung im Rahmen der Fahreignungsdiagnostik), ist also mit massiven Verfälschungen der Testleistung zu rechnen, wenn ein „Intensivtraining“ zum verwendeten Test durchgeführt wurde. Leider ist es bislang nicht gelungen, Geübtheit in einem Test anhand von irgendwelchen Kennwerten oder deren Kombination zu erkennen. Angesichts der Fehlentscheidungen, die bei der Untersuchung von hochgeübten Testpersonen möglich sind, ist hier Forschungsbedarf zu erkennen. In der Praxis wird ein erwartungswidrig sehr gutes Testergebnis Zweifel nähren. Im Sinne eines multimethodalen Vorgehens sollten wichtige diagnostische Entscheidungen möglichst durch mehr als ein Verfahren abgesichert werden.

Übungseffekte reduzieren

Vor dem Hintergrund der vorhandenen Übungseffekte auf Leistungstergebnisse, stellt sich die Frage, wie diese minimiert werden können. Zwei

mögliche Maßnahmen sollen hier erwähnt werden. Erstens kann man versuchen, Tests „resistent“ gegen Übungseffekte zu machen. Zur Gestaltung von Tests, die weniger übungsanfällig sind, hat Powers (1986) mit seiner Metaanalyse wichtige Hinweise geliefert: Trainingseffekte werden kleiner, wenn die Testanweisungen einfach, klar und kurz sind sowie wenn ein festes Antwortformat statt freier Beantwortung vorgesehen ist. Zweitens bietet es sich an, in Untersuchungen zu Auswahlzwecken allen Testpersonen schon einige Zeit vor der Untersuchung Beispiel- und Übungsaufgaben zur Verfügung stellen. Diese Maßnahme ist bei der Auswahl von Pilotinnen bzw. Piloten oder von Studienbewerberinnen bzw. -bewerbern verbreitet. Dadurch reduziert sich der „Vorsprung“ von Personen, die sich ein professionelles Testtraining leisten können und wollen.

Bedeutung von Übungseffekten

In der Praxis kommt es oft vor, dass ein Test mehrmals durchgeführt wird. Mit dem 2. Messergebnis wird die Ausprägung der Fähigkeit also mehr oder weniger stark überschätzt. In der Verkehrseignungsdiagnostik wird für Berufskraftfahrerinnen und -kraftfahrer ein Prozentrang von mindestens 16 als Nachweis ihrer Eignung verlangt. Sommer et al. (2017) ließen über 200 Testpersonen 4 in der Verkehrseignungsdiagnostik verwendete Tests wiederholt durchführen und betrachteten die Personen genauer, die bei der 1. Testdurchführung den Mindestwert verfehlt hatten. In den verwendeten Tests waren das zwischen 20 und 41 Personen (je nach Test). Nach Wiederholung der Tests wären nur noch 7–8 Betroffene erneut negativ beurteilt worden; alle anderen hatten nun den kritischen Prozentrang überschritten. Im Ernstfall hätte die einschlägige Testfahrung also in den meisten Fällen für ein positives Eignungsurteil ausgereicht.

Eine andere Situation liegt vor, wenn eine (Behandlungs-)Maßnahme durchgeführt wird, die auf eine Verbesserung der Fähigkeit abzielt. Beispielsweise möchte man prüfen, ob sich die Konzentration oder die Merkfähigkeit nach einer neuropsychologischen Reha-Maßnahme verbessert hat. Der reine Effekt der Maßnahme addiert sich zum Übungseffekt. Beide Effekte können im Einzelfall und in reinen Prä-post-Studien nicht separiert werden. Eine schwer praktizierbare Lösung wäre eine Normierung von Zweitmessungen. Auf jeden Fall sollten Retest-Effekte bei der Interpretation von Testergebnissen bedacht werden. Vielleicht findet man im Testmanual Angaben dazu, wie groß der Übungseffekt ist. Damit kann man zumindest annäherungsweise abschätzen, wie viel der Übungseffekt wohl zu der 2. Testleistung beigetragen hat. Auch die zuvor beschriebenen Befunde aus einschlägigen Metaanalysen bieten eine grobe Orientierung. Bei der Evaluierung von Maßnahmen ist eine unbehandelte Kontrollgruppe, bei der ebenfalls 2 Messungen im gleichen zeitlichen Abstand erfolgen, zwingend erforderlich.

Bedeutung der Übungseffekte für die Evaluation von Maßnahmen

Gesundes Essen

Das folgende, fiktive Beispiel soll zeigen, wie absurd Schlussfolgerungen sein können, wenn man die für manche Tests üblichen Übungseffekte nicht berücksichtigt.

Ein bekannter Fernsehkoch verpflegte eine Gruppe Kinder einige Tage gut mit selbstgekochtem Essen. Vor und nach der Ernährungsmaßnahme führte er einen etablierten Konzentrationstest durch, dessen Ergebnisse tatsächlich bei der 2. Messung besser ausfielen. Er hielt einen Testbogen in die Kamera und meinte, es sei nun wissenschaftlich bewiesen, dass sich eine gesunde Ernährung positiv auf die Konzentrationsfähigkeit auswirke.

Einfluss der Testwiederholung auf die Validität

Übungseffekte besagen zunächst nur, dass sich die Mittelwerte von einer Testung zur anderen verändern. Neben der Frage der Veränderung von Mittelwerten sollte jedoch auch der Frage nachgegangen werden, ob Tests bei wiederholter Durchführung andere Korrelationen zu anderen Tests oder Kriterien aufweisen als beim ersten Mal. Lievens et al. (2007) konnten bei einem Test zur Auswahl von Medizin- und Zahnmedizinstudierenden unter realistischen Bedingungen untersuchen, ob sich die Korrelation des Auswahltests mit einem Kriterium durch Wiederholung verändert. Kriterium waren die späteren Examensleistungen. Zunächst abgelehnte Bewerber durften den Test 1 Monat später erneut durchführen. Bei dieser Bewerbendengruppe konnten das 1. und 2. Testergebnis mit den Examensnoten korreliert werden. Es zeigte sich, dass die Korrelation des 2. Tests mit der Examensleistung geringer ausfiel. Die bereits erwähnte Studie von Albers und Höft (2009) betrachtete ebenfalls die Veränderung solcher Korrelationen bei Testwiederholung. Zur Erinnerung: Es ging in dieser Studie um ein Auswahlverfahren für Nachwuchspilotinnen und -piloten. Die Befunde zur Validität des dabei wiederholt durchgeföhrten Tests zum räumlichen Vorstellungsvermögen sind aufschlussreich. Die Leistungen in der Aufgabe korrelierten über alle 10 Durchgänge hinweg immer höher mit einem an einem anderen Tag einmalig durchgeföhrten Konzentrationstest (r von .16 bis .35 ansteigend). Der Befund spricht dafür, dass sich die Validität der Testung verändert hat. Mit jedem Durchgang war die Testleistung immer mehr von der Konzentrationsfähigkeit abhängig.

Neben den bereits diskutierten Einflüssen der Anstrengung und Übung kann auch die Angst vor dem Test (oder dem Testergebnis) die Testleistung beeinflussen. „Ich hatte so viel Angst, dass ich meine Leistungsfähigkeit gar nicht richtig entfalten konnte.“ So könnte jemand ein vermeintlich schlechtes Testergebnis erklären. Die Forschung hat sich ausführlich mit dem Zusammenhang zwischen Testangst und der Leistung in Tests oder Prüfungen befasst. Dabei muss differenziert werden, ob Testangst als momentaner Zustand oder als habituelles Merkmal verstanden wird. Wir befassen uns hier nur mit der habitualen Testangst. Die Effekte sind aber vergleichbar, wenn man die momentane Testangst mit Testleistungen korreliert.

In einer Metaanalyse hatte Hembree (1988) einen Zusammenhang zwischen Testangst (mit einem Fragebogen gemessen) und Intelligenztestergebnissen von $r = -.23$ gefunden (61 Studien ab Klasse 3; s. Tab. 1, S. 55). Für schulische Leistungstests fiel der Zusammenhang in der gleichen Altersgruppe mit $r = -.29$ (44 Studien, ab Klasse 4) sogar etwas höher aus. Differenziert man bei der Testangst zwischen den Komponenten der kognitiven „Besorgtheit“ (worry) und emotionaler Aufgeregtheit (emotionality), so wird deutlich, dass die Besorgtheit deutlich höher mit schulischen Leistungstests korreliert als die Aufgeregtheit ($r = -.31$ vs. $-.15$). Der Zusammenhang zwischen Testangst und Testleistungen konnte in späteren Metaanalysen bestätigt werden. Für Intelligenztests fanden Ackerman und Heggestad (1997) eine doppelt minderungskorrigierte Korrelation von $r = -.33$. Rechnet man die doppelte Minderungskorrektur heraus (Annahme: Reliabilität Testangstfragebogen .80, Intelligenztest .90), so reduziert sich die Korrelation auf $r = -.28$, was wiederum einer Effektstärke von $d = 0,58$ oder einem „Leistungsverlust“ von 6 Standardwerten entspricht. Weitere Erkenntnisse liegen für Tests im schulischen bzw. akademischen Bereich (Seipp 1991, $r = -.21$) sowie für Studiennoten bzw. Ergebnisse in Zulassungstests vor (Richardson et al. 2012, $r = -.24$ bzw. $-.16$). Fazit ist, dass ein eher schwacher bis moderater negativer Zusammenhang zwischen Testangst und Ergebnissen in Leistungstests gut belegt ist: je größer die Angst, desto schlechter die Leistung. Was hier aber Ursache und was Folge ist, bleibt zunächst unbeantwortet.

Ergebnisse in Leistungstests korrelieren negativ mit Testangst

Der korrelative Zusammenhang zwischen Testangst und Testleistung lässt verschiedene Erklärungen zu. Erstens ist es denkbar, dass Angst bei der Testbearbeitung leistungsmindernd wirkt. Zweitens könnte es sein, dass die Angst nur Ausdruck einer selbst erkannten niedrigen Leistungsfähigkeit (eines „Defizits“) ist: Je stärker Personen zu Recht etwa davon überzeugt sind, dass sie eine niedrige Intelligenz haben, desto mehr Angst werden sie vor Intelligenztests entwickeln. Die Testangst würde demnach lediglich das vorhandene „Defizit“ widerspiegeln. Sommer und Arendasy (2015) haben die 1. Annahme überprüft.

Über 1700 Bewerberinnen und Bewerber für einen Medizinstudienplatz nahmen an einem Auswahltest teil. Sie wurden anschließend gebeten, noch einen Testangstfragebogen auszufüllen. Die Autoren überprüften, ob das 1PL-Rasch-Modell (s. dazu ▶ Abschn. 2.3.1) gut zu den Daten der 4 eingesetzten Leistungstests passt. Die Annahme im 1PL-Modell lautet, dass die Itemantworten in den Leistungstests allein von den Itemschwierigkeiten und der Fähigkeit der Personen abhängt. Sollte sich eine andere Variable, die Testangst, auf die Itemantworten auswirken, wäre ein schlechter Fit des 1PL-Modells zu erwarten. Das war aber nicht der Fall. Die Parameterschätzungen waren offensichtlich invariant gegenüber der Testangst. Zusätzlich wurde mit Strukturgleichungsmodellen überprüft, ob die Leistung im Auswahltest auf eine der beiden Komponenten der Testangst – worry bzw. emotionality (s. o.) – und auf aufgabenirrelevante Gedanken zurückgeführt werden kann. Dies konnte nicht bestätigt werden. Die Ergebnisse sprechen deshalb insgesamt gegen die Annahme, dass sich die Testangst kausal auf die Leistung im Test auswirkt. Sie sind aber mit der Defizithypothese vereinbar.

Weitere Befunde betreffen die kausale Beziehung zwischen Testangst und Testleistung. In der Metaanalyse von Hembree (1988) wurden auch experimentelle Studien betrachtet, in denen die Testangst durch therapeutische Maßnahmen reduziert wurde. Es zeigte sich, dass sich durch die Interventionen auch die Testleistung verbesserte. Betrachten wir nur Interventionsformen, die an mindestens 1000 Personen untersucht wurden (Hembree 1988, Tab. 11, S. 70), so finden sich für systematische Desensibilisierung ($d=0,32$) und für diverse kognitive-behaviorale Ansätze ($d=0,52$) signifikante und zudem deutliche Effekte, während Entspannungstraining nur minimal wirksam war ($d=0,13$). Diese Ergebnisse sprechen dafür, dass eine erhöhte Testangst tatsächlich zu einer Verringerung der Testleistung führen kann.

Leistungstestergebnisse können von den Testpersonen auch „nach unten“ verfälscht werden. Dieses Phänomen wird auch als „faking bad“ bezeichnet. Manchmal haben Testpersonen ein Interesse an schlechten Testergebnissen, etwa bei der Begutachtung nach einem fremdverschuldeten Unfall oder wegen einer beantragten Frühberentung.

Schlechte Testergebnisse können durch absichtliche Fehler oder durch langsameres Arbeiten (bei Tests mit einer Speedkomponente; s. dazu ▶ Abschn. 3.2.3.1) herbeigeführt werden. Für die Minderung der Testleistung gibt es praktisch keine Grenze. Eine Testperson, die ein schlechtes Ergebnis anstrebt, wird aber bemüht sein, nicht mit gänzlich unplausiblen Leistungen aufzufallen und daher eher vorsichtig faking bad betreiben. In einer experimentellen Untersuchung zu faking im Konzentrationstest d2 (Schmidt-Atzert et al. 2004b) sollten studentische Versuchspersonen so vorgehen, dass selbst eine Testexpertin bzw. ein Testexperte nicht merkt, dass sie eigentlich ein anderes Ergebnis bekommen würden. Die Testpersonen arbeiteten langsamer und machten mehr Fehler, wodurch sich ihr Gesamttestwert (KL-Wert) gegenüber einer neutralen Kontrollbedingung um umgerechnet 19,4 Standardwertpunkte verringerte. Dieser Befund konnte mit der Computerversion des d2-R repliziert werden. Bei der Bedingung „faking bad“ fiel der KL-Wert

Auswirkung der Testangst auf die Lösung der Items in einem Auswahltest unwahrscheinlich

Angstreduktion verbessert die Testleistung

Faking bad bei Leistungstests möglich

um umgerechnet 18,1 Standardwerte niedriger aus als in der Kontrollgruppe (Schmidt-Atzert und Brickenkamp 2017). Bei einem multimethodalen Vorgehen können sie durch Diskrepanzen zu anderen diagnostischen Informationen auffallen. In der Neuropsychologie (► Abschn. 9.1) ist die absichtliche Verfälschung von Testergebnissen ein wichtiges Thema. Deshalb kam es gerade in dieser Disziplin zur Entwicklung von Validierungstests, mit deren Hilfe eine absichtliche Verfälschung aufgedeckt werden kann. Diese Tests bestehen aus Aufgaben, die leichter zu lösen sind, als man vermutet. Bei faking bad fallen die Testergebnisse daher auffallend schlecht aus. In den oben zitierten Studien mit dem Test d2 bzw. der elektronischen Version des d2-R konnte faking bad relativ sicher daran erkannt werden, dass bestimmte Fehler extrem häufig aufraten (siehe in ► Abschn. 3.2.2.2 Aufmerksamkeits- und Konzentrationstest d2-R und d2-R elektronisch).

Was verändert die Testwerte?

Im Folgenden werden die im vorangegangenen Abschnitt diskutierten Einflussfaktoren zusammenfassend aufgelistet und die zu erwartenden durchschnittlichen Veränderungen der Testwerte in Standardwerten (SW) angegeben. Außer bei Leistungs- bzw. Testmotivation und Testangst darf ein Ursache-Wirkung-Zusammenhang als gut gesichert gelten:

Hohe Leistungs- oder Testmotivation (Forschungskontext)	+8 SW
Hohe Testmotivation bei Auswahltests	+4 SW
Test früher schon einmal durchgeführt	+3 SW
Intensive Übung (10 Durchgänge) bei Konzentrationstests	+26 SW
Gezielte Testvorbereitung (Coaching)	+6 SW
Hohe Testangst (bei Intelligenztests)	-6 SW
Faking bad (bei Konzentrationstests)	-19 SW

Arten von Leistungstests

Im Folgenden werden ausgewählte Leistungstests vorgestellt. Im deutschen Sprachraum sind heute mehrere Hundert Leistungstests verfügbar, die man verschiedenen Kategorien zuordnen kann. □ Tab. 3.2 führt jene Kategorien auf, die auch dem Testkompendium *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (Brähler et al. 2002) zugrunde liegen. Die angegebenen Zahlen aus dem Jahre 2002 sind für heute sicherlich zu niedrig, da jedes Jahr neue Tests auf den Markt kommen und alte Tests eher selten aus dem Vertrieb genommen werden.

Hitliste der Leistungstests

In der Praxis werden diese Tests unterschiedlich häufig eingesetzt. Roth und Herzberg (2008) erhielten bei einer Befragung per Post von 398 praktisch tätigen Psychologinnen und Psychologen Auskunft über den Testeinsatz: Die meisten (72 %) nannten den klinischen Bereich als ihr Arbeitsgebiet, gefolgt von Pädagogischer sowie Arbeits- und Organisationspsychologie (jeweils 19 %; Mehrfachnennungen waren möglich). Die Ergebnisse finden sich in

□ Tab. 3.2 Leistungstests nach Kategorien. (Nach Brähler et al. 2002)

Testkategorie	Anzahl der aufgeführten Verfahren
Allgemeine Leistungstests	17
Intelligenztests	57
Spezielle Funktions- und Eignungstests	29
Entwicklungstests	18
Schultests	72

■ Tab. 3.3 Die in Deutschland am häufigsten verwendeten Leistungstests

Test	Verwendungshäufigkeit (Nennungen in Prozent)			
	Roth und Herzberg (2008)	Evers et al. (2012)	Steck (1997)	Schorr (1995)
CFT ^a	22	–	14	7
d2 ^a	17	9	32	16
HAWIK/WISC ^a	16	22	29	22
SPM	15	–	18	12
HAWIE/WAIS ^a	13	9	28	18
K-ABC ^a	12	9	4	–
Benton	7	–	19	11
DRT ^a	7	–	8	–
MWT	6	–	8	–
I-S-T ^a	6	10	16	9

Prozentualer Anteil der Befragten, die angeben, den Test zu verwenden. Mehrfachnennungen möglich; bei Schorr: Nennen Sie 5 Tests, die Sie am häufigsten verwenden. $N=398$ (Roth und Herzberg 2008), 187 (Evers et al. 2012), 169 (Steck 1997) und 661 (Schorr 1995). Tests geordnet nach Nennungshäufigkeit bei Roth et al. (2008). CFT = Grundintelligenztest (mehrere Versionen), d2 = Aufmerksamkeits-Belastungs-Test d2, HAWIK/WISC = Hamburg-Wechsler-Intelligenztest für Kinder bzw. Wechsler Intelligence Scale for Children, SPM = Standard Progressive Matrices (Raven 2009), HAWIE/WAIS = Hamburg-Wechsler-Intelligenztest für Erwachsene bzw. Wechsler Adult Intelligence Scale, K-ABC = Kaufman Assessment Battery for Children, Benton = Benton-Test, DRT = Diagnostischer Rechtschreibtest, MWT = Mehrfachwahl-Wortschatz-Test (Lehrhrl 1977), I-S-T = Intelligenz-Struktur-Test. Auf weitere Quellenangaben wird verzichtet, da unterschiedliche Versionen und Auflagen der Tests in Gebrauch waren. Neuere Testversionen sind bei den Testnamen eingeschlossen.

Striche (–) bedeuten, dass zu dem Test keine Informationen vorliegen. Bei Evers et al. (2012) wurden nur zu den 10 am häufigsten genannten Tests Ergebnisse berichtet, darunter befanden sich auch Persönlichkeitsfragebögen.

^aDiese Tests werden in ▶ Abschn. 3.2.2 und 3.2.3 in ihrer aktuellen Form ausführlich behandelt; zum DRT s. ▶ Abschn. 7.4.1.1.

■ Tab. 3.3, in der auch 2 frühere Befragungen mit ähnlicher Methodik und eine neuere Studie mit aufgeführt sind. Die „Hitliste“ enthält 7 Intelligenztests, einen Konzentrationstest (Test d2), mit dem Benton-Test (Benton-Sivan und Spreen 2009) einen neuropsychologischen Gedächtnistest und einen Schultest (DRT; ▶ Abschn. 7.4.1.1). Eine Untersuchung wurde europaweit durchgeführt (Evers et al. 2012). Europaweit sind übrigens die Wechsler-Intelligenztests für Kinder (engl. Abkürzung: WISC) und für Erwachsene (engl. Abkürzung: WAIS) am häufigsten unter den Top Ten der verwendeten Tests vertreten (▶ Abschn. 3.2.3.2).

3.2.2 Aufmerksamkeits- und Konzentrationstests

Viele Leistungen in Schule, Studium, Beruf und Alltag – vom Autofahren bis zum Kochen – verlangen nicht nur ein Mindestmaß an Intelligenz sowie spezielles Wissen oder Fertigkeiten, sondern auch die grundlegende Fähigkeit, sich den Aufgaben effizient zuzuwenden. Man stelle sich vor, diese Grundfähigkeit sei durch aktuellen Drogen- oder Alkoholkonsum stark eingeschränkt: In diesem Zustand wird man viele Tätigkeiten nicht mehr richtig

Eingeschränkte allgemeine Leistungsfähigkeit

3

Aufmerksamkeit und Konzentration im Leben relevant

ausführen können; die Konzentrationsfähigkeit leidet. Die Leistung wird gegenüber dem Normalzustand stark abfallen, obwohl die Intelligenz, das Wissen und die Fertigkeiten vorhanden sind. Während alkohol- und drogenbedingte Leistungseinbußen vorübergehender Art sind, kann es durch Verletzungen oder Erkrankungen des Gehirns zu lange andauernden Einbußen kommen. Manche Menschen haben aber auch ohne eine erkennbare neuropsychologische Beeinträchtigung Schwierigkeiten, sich auf eine Aufgabe zu konzentrieren.

Wer allerdings Berufe wie Flutlotsin/Fluglotse, Pilotin/Pilot, Rennfaherin/Rennfahrer oder Chirurgin/Chirurg ausübt, muss über eine hohe Ausprägung dieser allgemeinen Leistungsfähigkeit verfügen. Mit anderen Worten: Diese Fähigkeit, für die wir nun die Begriffe „Aufmerksamkeit“ und „Konzentrationsfähigkeit“ einführen, variiert auch im Normalbereich. Es ist bezeichnend, dass Lebewesen im Laufe der Evolution Strategien entwickelt haben, um die Aufmerksamkeit von Artgenossen oder anderen Lebewesen zu wecken, also aufzufallen. Beispiele sind akustische, olfaktorische und visuelle Reize wie Balzrufe, Geruchsmarkierungen und bunte Federn oder bei Pflanzen auffällige Blüten, um Insekten zur Bestäubung anzulocken. Auf der anderen Seite kennen wir die Strategie der Tarnung, mit deren Hilfe sich Räuber an ihre Beute annähern können oder sich Beutetiere vor ihren Fressfeinden verstecken. Beide Strategien werden auch von Menschen angewendet. Man denke nur an auffällige Kleidung und an Tarnanzüge. Viele technische Geräte dienen dazu, unsere Aufmerksamkeitsleistung zu verbessern, indem sie Gefahrenhinweise „verstärken“. Beispiele sind Warnwesten, die man bei einer Autopanne tragen muss, Rauchmelder in Wohnungen oder auffällige Kennzeichnungen für Säure, Gift oder brennbare Flüssigkeiten auf Behältnissen.

Aufmerksamkeit und Konzentrationsfähigkeit sind in vielen Anwendungsbereichen relevant: Beeinträchtigungen weisen auf bestimmte psychische Störungen hin (Klinische Psychologie), bei hirnorganischen Störungen sind Aufmerksamkeit und Konzentrationsfähigkeit häufig eingeschränkt (Neuropsychologie), Leistungsprobleme in Schule oder Studium können durch Aufmerksamkeits- oder Konzentrationsprobleme mit bedingt sein (Pädagogische Psychologie), und bestimmte Berufe stellen mehr oder weniger hohe Anforderungen an diese Fähigkeiten (Berufseignungsdiagnostik). Folglich besteht ein hoher Bedarf, Aufmerksamkeit und Konzentrationsfähigkeit zu messen.

Die Konstrukte „Aufmerksamkeit“ und „Konzentration“ sind bislang in der Literatur nicht gut definiert; zumindest haben sich noch keine konsensfähigen Definitionen durchgesetzt. Viele Autorinnen und Autoren vermeiden deshalb eine begriffliche Festlegung und nennen Aufmerksamkeits- und Konzentrationstests in einem Atemzug. In einem einflussreichen Beitrag hatte Bartenwerfer (1964) vorgeschlagen, diese Tests als „allgemeine Leistungstests“ zu bezeichnen. Mit dem Begriff wollte er zum Ausdruck bringen, dass die Tests allgemeine Voraussetzungen für das Erbringen von kognitiven Leistungen erfassen. Eine konzeptuelle Klärung sah er als überflüssig an:

- » Jedoch weiß der unbefangene und fachkundige Leser ungefähr, was gemeint ist, wenn von einem Test für Konzentrationsfähigkeit, Aufmerksamkeit, Willenskraft usw. gesprochen wird. Glücklicherweise ist eine eindeutige sprachlich-definitorische Klarheit über die genannten Bezeichnungen nicht erforderlich, wenn es darum geht, menschliches Verhalten vorherzusagen (Bartenwerfer 1964, S. 387).

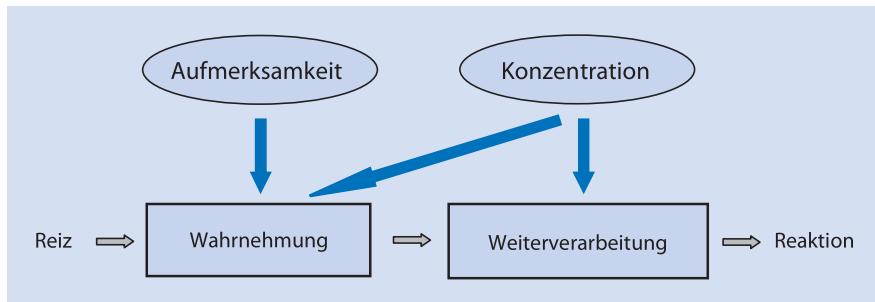
Anwendungsbereiche

„Allgemeine Leistungstests“ als Überbegriff

Der Begriff „allgemeine Leistungstests“ dient auch heute noch als Überbegriff, z. B. im *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (Brähler et al. 2002); die dort aufgeführten Tests werden in der Regel als Aufmerksamkeits- oder Konzentrationstests bezeichnet.

Kontrastierend dazu wird auch die Auffassung vertreten, dass Aufmerksamkeit und Konzentration nicht gleichzusetzen sind. Schmidt-Atzert et al. (2004a) plädieren dafür, Aufmerksamkeit allein mit Wahrnehmung in Verbindung zu bringen und darunter das selektive Beachten relevanter Reize oder Informationen zu verstehen; demgegenüber soll sich der Begriff „Konzentration“ auf alle Stufen der Verarbeitung von Informationen beziehen: von der selektiven Wahrnehmung (= Aufmerksamkeit) über die Ver- oder Bearbeitung bis hin zur Anbahnung einer motorischen Reaktion. □ Abb. 3.2 veranschaulicht diese Trennung und den Überlappungsbereich von Aufmerksamkeit und Konzentration.

Unterscheidung von
Aufmerksamkeit und Konzentration



□ Abb. 3.2 Aufmerksamkeit und Konzentration als unabhängige Konstrukte. Die Aufmerksamkeit hat ausschließlich einen Einfluss auf die Wahrnehmung, die Konzentration wirkt primär auf die Weiterverarbeitung der selegierten Reize, kann aber auch die Wahrnehmung betreffen („konzentrierte Aufmerksamkeit“) und den mentalen Anteil einer Reaktion (Handlungsplanung, Psychomotorik). (Aus Schmidt-Atzert et al. 2004a, S. 11, courtesy of Hogrefe)

Das Validitätsdilemma

Wenn in ▶ Abschn. 3.2.2 Aufmerksamkeits- und Konzentrationstests vorgestellt werden, stellt sich die Frage: Woran lässt sich feststellen, ob ein Test überhaupt Aufmerksamkeit oder Konzentration misst? Um es ganz deutlich zu sagen: Es gibt weder für Aufmerksamkeit noch für Konzentration ein allgemein anerkanntes externes Kriterium. Anders als etwa bei Intelligenz hat man noch keine Leistung im Leben gefunden, die einen deutlichen Zusammenhang mit Konzentrations- oder Aufmerksamkeitstests aufweist. Bei der Kriteriumsvalidität fehlen daher überzeugende Belege für diese Tests. Testentwicklerinnen und Testentwickler machen folglich etwas, was im Grunde unbefriedigend ist: Sie validieren ihre Tests an anderen, die das gleiche Merkmal messen sollen. Weil diese Tests auch nur an anderen Tests validiert wurden, ist das Vorgehen zirkulär.

Wir gehen am Ende von ▶ Abschn. 3.2.2.1 zu den Aufmerksamkeitstests auf die Schwierigkeiten ein, die bei dem Versuch einer „Verankerung“ im Alltagsleben bestehen („Validierung an Leistungen im Alltag“). Am Ende von ▶ Abschn. 3.2.2.2 zu den Konzentrationstests stellen wir einen Ansatz vor, der sich mit der „Anatomie“ von Testaufgaben befasst. Welche psychologischen Prozesse laufen ab, wenn eine Testperson ein Item löst oder nacheinander mehrere Items bearbeitet? Ziel ist es, damit herauszufinden, was das Gemeinsame von Konzentrationstests ist und worin sie sich unterscheiden. Wenn das Gemeinsame konzeptuell als Konzentrationsfähigkeit erklärt werden kann, gibt es auch eine Lösung für das Validierungsproblem.

Die angekündigten Ausführungen sind im Grunde gleichermaßen für Aufmerksamkeits- wie Konzentrationstests relevant, zumal der eingesetzte Test d2-R (▶ Abschn. 3.2.2.2) im Grenzbereich beider Konstrukte zu verorten ist. Für die Beurteilung der konvergenten Validität von Aufmerksamkeits- und Konzentrationstests wurde noch kein Weg gefunden, der allgemein akzeptiert ist – sieht man von der zirkulären Begründung ab, dass die Tests mit anderen Tests korrelieren, die das Gleiche messen wollen.

3.2.2.1 Aufmerksamkeitstests

Aufmerksamkeit wird in der Psychologie schon sehr lange erforscht, und zwar sowohl aus allgemeinpsychologischer, neuropsychologischer, differentialpsychologischer und diagnostischer Perspektive, was eine Verständigung auf eine allgemein akzeptierte Definition erschwert. Hilfreich für ein Verständnis des Konstrukts sind die Überlegungen zu dessen Funktion. Cohen (1993) hat die Funktion der Aufmerksamkeit mit einem Tor verglichen, das den Informationsfluss zum Gehirn beschränkt. Dieses „Tor“ ist genauso eine Metapher wie die Postulierung eines „Filters“ oder etwa eines „Scheinwerfers“, der nur bestimmte Teile der Umwelt erhellt. In der Aufmerksamkeitsforschung haben verschiedene solcher Metaphern Verwendung gefunden, um eine gezielte Selektion von Informationen zu „erklären“ (Moosbrugger und Goldhammer 2006). Die Selektion ist das zentrale Merkmal der Aufmerksamkeit. Dabei kann die Selektion oder Beachtung von Reizen sowohl willentlich als auch unwillentlich geschehen. Bestimmte Reize wie etwa plötzliche laute Geräusche oder Schmerzreize dringen unwillentlich in unser Bewusstsein vor (Eimer et al. 1996). Diese unwillentliche Aufmerksamkeit ist überlebenswichtig, weil so bestimmte Gefahren rechtzeitig entdeckt werden. Diese umfassende Betrachtung des Phänomens wird in einer Arbeitsdefinition der Aufmerksamkeit aufgegriffen:

Definition

Arbeitsdefinition der Aufmerksamkeit

Wir verstehen unter **Aufmerksamkeit** die Fähigkeit, ganz bestimmte Reize/Ereignisse unter vielen willentlich oder nicht willentlich wirksam zu beachten.

Verschiedene Aufmerksamkeitsformen und -funktionen

Der Unterschied zwischen willentlicher und nicht willentlicher Aufmerksamkeit lässt sich durch eine einfache Darstellung veranschaulichen (Abb. 3.3). Jede der beiden Hälften der Abbildung enthält 100 Zeichen, von denen eines anders ist als die restlichen 99 Zeichen. Auf der linken Seite muss man willentlich nach dem „anderen“ Zeichen suchen. Auf der rechten Seite fällt das „andere“ Zeichen sofort auf – ob man will oder nicht.

In der Testdiagnostik wird allerdings nur die willentliche Aufmerksamkeit beachtet. Der „Wille“ ist dabei identisch mit dem Vorsatz der Testperson, die Testinstruktion zu befolgen und auf bestimmte Reize zu achten und zu reagieren. Im Beruf und Alltag können ganz ähnliche Anforderungen vorkommen: Jemand sucht ein bestimmtes Buch im Regal, eine Wohnungsanzeige mit bestimmten Merkmalen im Anzeigenteil einer Zeitung oder eine bestimmte Person unter den austiegenden Fahrgästen am Bahnhof. Eine Hautärztin bzw. ein Hautarzt betrachtet Auffälligkeiten an der Haut und achtet dabei auf bestimmte Merkmale, die für ein Karzinom sprechen können. Oder eine Fluglotsin bzw. ein Fluglotse beobachtet Radar- und Computersysteme, um ggf. bestimmte Gefahrensignale zu entdecken. Tests zu verschiedenen Formen der Aufmerksamkeit haben ein gemeinsames Merkmal: Sie erfassen, wie schnell und genau Testpersonen kritische Reize entdecken. Die Tests unterscheiden sich vor allem darin, welche kritischen Reize verwendet und unter welchen Bedingungen diese dargeboten werden. Die Bedingungen sind ausschlaggebend dafür, welche „Form“ der Aufmerksamkeit gemessen wird. In Tab. 3.4 sind einige häufig genannte Aufmerksamkeitstypen aufgeführt. Eine weitgehend vollständige Auflistung deutschsprachiger Aufmerksamkeitstests mit kurzen Angaben zu jedem Verfahren findet sich bei Schmidt-Atzert et al. (2008). Die Unterscheidung verschiedener Aufmerksamkeitsfunktionen ist für die Einteilung von Tests nützlich; allerdings scheint es mehr Begriffe für Formen der Aufmerksamkeit zu geben, als sich durch die faktorenanalytische Forschung belegen lässt (Schmidt-Atzert et al. 2008). Deshalb sind einige Erläuterungen zu den Aufmerksamkeitsfunktionen angebracht.

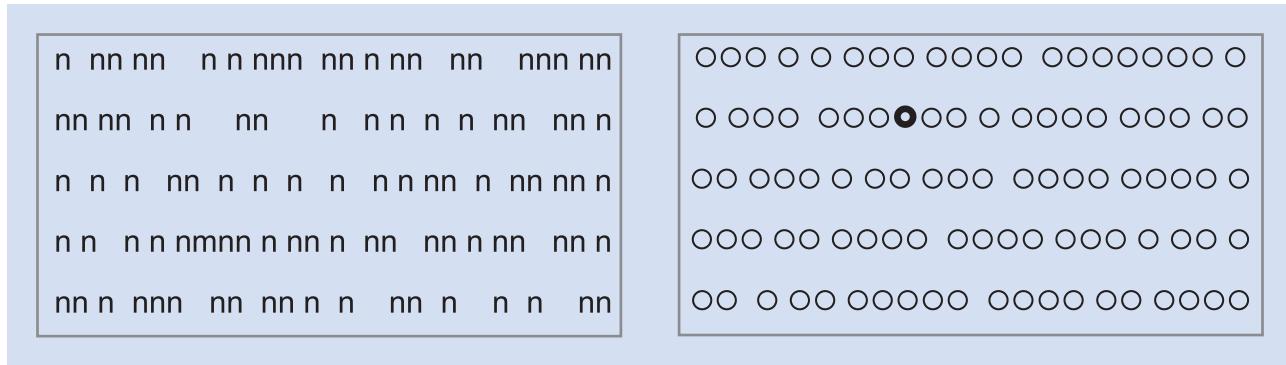


Abb. 3.3 Wirkung willentlicher und nicht willentlicher Aufmerksamkeit. Eines der 100 Zeichen weicht jeweils von den anderen ab. Auf der linken Seite findet man diese Abweichung nur durch willentliche Suche, auf der rechten Seite fällt sie unwillentlich („automatisch“) ins Auge.

Alertness Das Konzept „Alertness“ spielt vor allem in der Neuropsychologie eine Rolle. Darunter wird eine basale Wachheit oder Ansprechbarkeit auf Reize jeder Art verstanden. Eine extreme Verminderung der Alertness ist bei komatösen Patienten zu beobachten. Alertness bezeichnet im Grunde nicht eine bestimmte Form der Aufmerksamkeit, sondern vielmehr eine generelle Voraussetzung für Aufmerksamkeit (im Sinne einer Reizselektion). Ein verbreiteter Test zur Alertness ist der Subtest „Alertness“ der Testbatterie zur Aufmerksamkeitsprüfung (TAP) von Zimmermann und Fimm (1993, 2017). Die Testpersonen sind aufgefordert, beim Erscheinen eines Kreuzes auf dem Bildschirm sofort eine Antworttaste zu drücken. Es gibt keine anderen Reize, die zu ignorieren wären.

Prinzipiell unterscheiden sich Alertness-Tests nicht von Tests zur Messung der Reaktionsschnelligkeit. So wird beim Reaktionstest (Schuhfried 2020b) in der Serie 1 ein gelbes Lichtsignal als einziger Reiz verwendet, der immer so

Ansprechbarkeit auf Reize

Reaktionsschnelligkeit

Tab. 3.4 Aufmerksamkeitsformen und -tests

Aufmerksamkeitsbegriff	Testbedingung (Prinzip)	Testbeispiel und Kurzbeschreibung ^a
Alertness ^b	Einfache Reize schnell und zuverlässig beantworten	TAP Alertness Kreuz auf dem Bildschirm
Fokussierte oder selektive Aufmerksamkeit	Beachtung eines bestimmten Reizes bzw. einiger weniger Reize innerhalb einer Reizklasse	TAP Go/NoGo Einzelabwertung von ähnlichen Mustern, 2 Muster davon sind kritische Reize
Geteilte Aufmerksamkeit	Beachtung von mindestens je einem Reiz aus 2 deutlich verschiedenen Reizklassen	TAP Geteilte Aufmerksamkeit Visuell: wechselnde Kreuze in einer 4×4-Matrix – <i>Quadrat?</i> Akustisch: abwechselnd hoher und tiefer Ton – <i>Unregelmäßigkeit?</i>
Daueraufmerksamkeit	Fokussierte oder geteilte Aufmerksamkeit über längere Zeit	DAUF Reihe von 5 bzw. 7 ständig wechselnden Dreiecken mit Spitze nach oben oder unten – <i>vorher definierte Anzahl von Dreiecken mit Spitze nach unten</i> (Dauer: 20 bzw. 35 min)
Vigilanz	Beachtung seltener Reize über längere Zeit	VIGIL Hell aufleuchtender Punkt springt auf einer Kreisbahn (ähnlich einer Uhr) um einen Schritt – <i>Doppelsprung</i> (Dauer: 30, 35 oder 70 min)

TAP = Testbatterie zur Aufmerksamkeitsprüfung (Zimmermann und Fimm 1993, 2017). DAUF und VIGIL aus dem Wiener Testsystem (Fa. Schuhfried).

^aKritische Reize (kursiv gedruckt) sind mit Tastendruck zu beantworten.

^bZu Alertness s. Erläuterungen im Text

schnell wie möglich zu beantworten ist. Der einzige Unterschied zum TAP-Test Alertness besteht darin, dass in der TAP zusätzlich zwischen tonischer und phasischer Alertness unterschieden wird: Bei der Bedingung „phatische Alertness“ geht dem visuellen Reiz ein akustischer Warnreiz voraus, der die Alertness anheben soll. Bei der Bedingung „tonische Alertness“ gibt es keinen solchen Warnreiz. In einer faktorenanalytischen Untersuchung (Bühner et al. 2001) mit hirngeschädigten Patientinnen und Patienten wiesen die beiden Tests aus der TAP und 2 klassische Reaktionszeittests (Reaktion auf einen Licht- bzw. akustischen Reiz) sehr hohe Ladungen auf einem gemeinsamen Faktor auf.

Selektives Beachten von Reizen

Selektive und geteilte Aufmerksamkeit Die Abgrenzung der geteilten von der selektiven Aufmerksamkeit basiert auf bestimmten Eigenschaften der kritischen Reize: Stammen sie aus 2 unterschiedlichen Reizklassen (meist akustisch und visuell), spricht man von geteilter Aufmerksamkeit. Empirisch ist die Unterscheidung zwischen selektiver und geteilter Aufmerksamkeit schwer zu begründen: Die entsprechenden Tests laden meist auf einem einzigen Faktor (Schmidt-Atzert et al. 2008). Gemeinsam ist diesen Tests, dass sie das selektive Beachten relevanter Reize verlangen. In Sinne der oben vorgeschlagenen Definition handelt es sich also um Aufmerksamkeitstests – nicht mehr und nicht weniger. Die Art und Anzahl der kritischen Reize sowie der Distraktoren (Reize, auf die nicht zu reagieren ist) bestimmen den Schwierigkeitsgrad der Testaufgaben.

Aufrechterhaltung der Aufmerksamkeit

Daueraufmerksamkeit Einige Testautorinnen und -autoren sprechen von Daueraufmerksamkeit, wenn die selektive bzw. geteilte Aufmerksamkeit über einen längeren Zeitraum aufrechterhalten werden muss. Die Aufgaben sind zumeist mit denen bei der Messung der selektiven bzw. geteilten Aufmerksamkeit identisch, die Tests werden lediglich verlängert.

Aufmerksamkeitsleistungen bei selbst bestimmtem Tempo

Konzentrierte Aufmerksamkeit Tests, die von Schmidt-Atzert et al. (2008) der konzentrierten Aufmerksamkeit zugeordnet werden, erfordern Konzentration bei Aufgaben, die Aufmerksamkeit (und nicht etwa Rechenfertigkeit) verlangen. Der entscheidende Unterschied zu Tests zur Daueraufmerksamkeitstests besteht darin, dass die Testpersonen ihr Arbeitstempo selbst bestimmen können. Nicht der Computer gibt vor, wann das nächste Item erscheint, sondern die Testperson bearbeitet alle Aufgaben nacheinander in dem von ihr gewählten Arbeitstempo. Mit diesem auch als „self-paced“ bezeichneten Arbeitsstil ist ein nach Westhoff (1995) zentrales Merkmal von Konzentrationstests erfüllt. Dieser scheinbar kleine Unterschied zwischen computer- und selbstbestimmtem Arbeitstempo hat einen großen Effekt auf die Validität eines Tests, wie Krumm et al. (2008) in einer Studie mit experimentell variierten Testversionen zeigen konnten.

Validierung an Leistungen im Alltag

Möchte man Aufmerksamkeits- (oder Konzentrationstests) validieren, liegt es nahe, Leistungen im Alltag oder im Beruf heranziehen, die Aufmerksamkeit bzw. Konzentration verlangen. Allerdings ist es sehr schwierig, wenn nicht gar unmöglich, Leistungen zu finden, die nur von der Aufmerksamkeit oder der Konzentration abhängen. Das Problem liegt darin, dass vermeintlich einschlägige Leistungen von vielen anderen Faktoren abhängen. Anhand eigener Erfahrungen kann dies konkret erläutert werden.

In 2 Diplomarbeiten (Hof 2012; Nimax 2012) wurde der, wie sich herausstellte, allzu ambitionierte, Versuch unternommen, Aufmerksamkeitsleistungen im Alltagsleben zu erheben. In 3 verschiedenen Alltagssituationen sollten studentische Testpersonen „Reize“ suchen: In einem Supermarkt bestimmte Produkte in einem Regal, auf einem Parkplatz von einem festen Übersichtspunkt aus bestimmte Autos (die zuvor mit einem Smartphone aus der Nähe fotografiert worden waren) und in der Mensa eine ihnen bekannte Person. Die Situationen wurden so gut wie möglich standardisiert, und es gab immer mehrere Durchgänge.

Die Standardisierung soll am Beispiel des „Einkaufens im Supermarkt“ verdeutlicht werden. Selbstverständlich erhielten die Testpersonen eine einheitliche (schriftliche und zudem mündlich erläuterte) Instruktion und absolvierten einen Übungsdurchgang. Vor einem bestimmten Gang im Supermarkt war am Boden ein Startpunkt markiert. Der Testleiter oder die Testleiterin zeigte der Testperson 5 s lang ein Produkt, z. B. eine bestimmte Zahncreme. Diesen Artikel sollte die Testperson auf der rechten oder linken Seite (das wurde mitgeteilt) eines 12 m langen Ganges suchen. Im Regal standen immer mehrere Exemplare des Produkts. Die Testperson startete eine Stoppuhr und machte sich auf die Suche. Sobald sie den Artikel entdeckt hatte, hielt sie die Stoppuhr an. Die Testleiterin oder der Testleiter folgte ihr und überzeugte sich, ob es sich auch um den gesuchten Artikel handelte. Nur 2 der 110 Testpersonen fanden die Zahncreme nicht in den vorgesehenen 90 s und 3 verwechselten sie mit einem ähnlich aussehenden Produkt. Die Aufgabe war also leicht. Damit kam der Suchzeit die größte Bedeutung zu. Die Suche nach der Zahncreme dauerte zwischen 2,6 und 58,6 s ($M=17,4$). Dann begann der nächste Durchgang (Suchobjekt ein bestimmtes Fabrikat eines Gardinen-Weißspülers). Insgesamt gab es 5 Durchgänge mit jeweils anderen Suchobjekten.

Eine Itemanalyse mit der Suchzeit als Itemkennwert offenbarte, dass die Trennschärfen bei allen 3 Aufgaben sehr niedrig waren, jedes Item also überwiegend etwas anderes zu messen schien. Durch eine Nachbefragung und eigene Überlegungen fanden wir Erklärungen dafür. Bei der Artikelsuche kann man vermuten (oder auch nicht), dass ein teures Markenprodukt nicht gerade unten im Regal steht. Bei der Autosuche lernten wir, dass einige Studierende als Radfahrende nicht mit Automarken und -typen vertraut waren; damit fiel für sie eine Strategie wie „ich suche einen blauen Golf, älteres Baujahr“ weg. Generell bergen die verwendeten Suchaufgaben die Gefahr, dass man ein Objekt erst einmal übersieht und dann das Feld erneut absuchen muss. Die Entscheidung, wo man anfängt zu suchen, kann sich als vorteilhaft oder unvorteilhaft erweisen. Es darf deshalb nicht überraschen, dass die Korrelationen mit Aufmerksamkeits- und Konzentrationstests extrem niedrig ausfielen. Der Gesamtwert der Suchzeit über alle 3 Situationen korrelierte zu $r=-.195$ mit dem d2-R (s. u.), zu $r=.03$ mit dem Test zu Wahrnehmungs- und Aufmerksamkeitsfunktionen – selektive Aufmerksamkeit (WAF-S; Sturm 2008) und zu $r=.008$ mit dem Cognitron (ein figuraler Konzentrationstest; Schuhfried 2020a).

Man kann die Messung von Alltagsleistungen auch so weit standardisieren, dass sie (fast) Testcharakter hat. Bei der Validierung des Test d2-R (► Abschn. 3.2.2.2) wird ein solcher Ansatz berichtet. Beispielsweise wurde eine standardisierte Tippfehlersuche durchgeführt. Das zur Validierung herangezogene „Kriterium“ war dabei aber extrem eng gefasst – zumindest im Vergleich zu Schul-, Berufs- oder Ausbildungserfolg als Kriterium für Intelligenztests.

Im Alltag war es also überraschend schwierig, die Aufmerksamkeit zuverlässig zu messen. Unter laborähnlichen Bedingungen war eine zuverlässige Messung möglich, aber der Messgegenstand war dabei sehr „eng“.

Arbeitsdefinition der Konzentrationsfähigkeit

3.2.2.2 Konzentrationstests

Definition

Wir verstehen unter **Konzentration** die Fähigkeit, sich über mehr oder weniger lange Zeit einer Tätigkeit effizient zuwenden zu können. Die Effizienz zeigt sich an der Quantität (bewältigte Aufgabenmenge oder benötigte Zeit) und der Qualität (Güte bzw. niedrige Fehlerquote) der dabei erbrachten Leistung. Die Leistung ist dabei in Bezug auf die spezifischen (für die Tätigkeit benötigten) Fähigkeiten, Fertigkeiten und/oder Kompetenzen zu relativieren.

Zur Erläuterung der Definition betrachten wir eine Konzentrationsleistung im Alltag, nämlich das Korrekturlesen von Texten. Wir nehmen einmal an, dass der Text 2 DIN-A4-Seiten umfasst und ziemlich viele Fehler enthält. Wir könnten die Zeit messen, die jemand für die Bearbeitung benötigt und hätten damit ein Maß für die *Quantität* der Konzentrationsleistung. Je schneller eine Person den Text bearbeitet, desto besser kann sie sich konzentrieren. Wir können auch die Bearbeitungszeit z. B. auf 5 min begrenzen und ermitteln, wie viele Wörter jemand in dieser Zeit bearbeitet hat. Es liegt auf der Hand, dass sich mehrere Personen auch in der *Qualität* ihrer Leistung unterscheiden. Sie machen Fehler, indem sie falsch geschriebene Wörter übersehen und richtig geschriebene Wörter „korrigieren“. Am Rande sei angemerkt, dass wir hier vor dem Problem stehen, Quantität und Qualität irgendwie zu verrechnen. Bei der Darstellung von Konzentrationstests werden Lösungen dafür aufgezeigt.

Was bedeutet nun die Feststellung in der Definition, dass die Leistung „in Bezug auf die spezifischen (für die Tätigkeit benötigten) Fähigkeiten, Fertigkeiten und/oder Kompetenzen zu relativieren“ ist? Die Unterschiede zwischen verschiedenen Personen beim Korrekturlesen könnten auf mindestens 2 Faktoren zurückzuführen sein, die nichts mit ihrer Konzentrationsfähigkeit zu tun haben, nämlich die Rechtsschreibkompetenz und die Schreibgeschwindigkeit. Die Schreibgeschwindigkeit ließe sich als Einflussfaktor eliminieren, indem wir falsch geschriebene Wörter nur markieren und nicht korrigieren lassen. Den Einfluss der Rechtschreibkenntnisse könnten wir nicht ausschalten, sondern nur kontrollieren. So könnten wir zusätzlich einen Rechtschreibtest ohne Zeitdruck durchführen und dessen Ergebnis bei der Interpretation beachten oder mit unserem „Konzentrationstestergebnis“ verrechnen.

Auch bei „richtigen“ Konzentrationstests besteht das Problem, dass die Leistung von anderen Faktoren abhängt. Wir betrachten dazu im Vorgriff auf eine ausführliche Testbeschreibung (s. u.) den Test d2-R.

Die Testpersonen sollen jeweils den Buchstaben d, der mit insgesamt 2 Strichen versehen ist (das sind die „Zielobjekte“), durchstreichen (Abb. 3.4). Alle d's mit mehr oder weniger Strichen sowie jeweils der Buchstabe p, auch wenn er 2 Striche hat, (d. h. alle „Distraktoren“) sind zu ignorieren. In knapp 5 min sind fast 800 Items zu bearbeiten. Die Quantität der Testleistung wird über die Anzahl der in dieser Zeit bearbeiteten Items bzw. wegen der einfacheren Auszählung der Anzahl der bearbeiteten Zielobjekte bestimmt. Die Qualität der Testleistung ergibt sich aus der Anzahl der Fehler, die der Testperson dabei unterlaufen. Fehler sind alle übersehnen



Abb. 3.4 Items aus dem Test d2-R. (Aus Brickenkamp 1994, © Hogrefe)

Zielobjekte und alle fälschlicherweise durchgestrichen Distraktoren. Wie in ▶ Abschn. 3.2.2.2 im Kontext einer Prozessanalyse der Itembearbeitung gezeigt wird, hängt die Testleistung im d2-R auch davon ab, wie schnell jemand visuelle Informationen wahrnehmen kann und wie gut jemand vorausschauend arbeitet.

Konzentrationstests verlangen die schnelle Durchführung von einfachen kognitiven Operationen, also beispielsweise das Entdecken bestimmter Reize oder das Überprüfen einfacher Rechenaufgaben. In der Regel können sukzessiv sehr viele solcher Aufgaben unter Zeitdruck durchgeführt werden. Erstaunlicherweise findet man solche Aufgaben auch in verschiedenen Intelligenztests. Liegt dem Test ein differenziertes Intelligenzmodell zugrunde, werden diese Tests dem Bereich der Be- oder Verarbeitungsgeschwindigkeit zugeordnet und gehen dann außerdem in die Berechnung des Gesamtintelligenzquotienten (Gesamt-IQ) ein. Dies wird etwa im Berliner Intelligenzstruktur-Test (BIS-Test; Jäger et al. 1997) oder in den Wechsler-Intelligenztests wie der WISC-V (▶ Abschn. 3.2.3) so gehandhabt. Im Zahlen-Verbindungs-Test (ZVT; Oswald und Roth 1987; ▶ Abschn. 3.2.2.2) soll die Testperson auf dem Testbogen mit 90 Zahlen fortlaufen eine Zahl mit der jeweils nächsthöheren mit einem Strich verbinden (also 1 – 2 – 3 etc.); die Zahlen sind so auf dem Papier verteilt, dass die nächste Zahl aber immer in der Nähe der vorausgehenden zu finden ist. Der Test wird von den Autoren dennoch als sprachfreier Intelligenztest bezeichnet, der Intelligenz über die Messung der „kognitiven Leistungsgeschwindigkeit“ erfassen soll.

Konzentrationstestaufgaben auch in Intelligenztests

In der einschlägigen Forschung, die übrigens überwiegend der Intelligenzforschung zuzurechnen ist, werden verschiedene Begriffe für die schnelle Verarbeitung von Informationen verwendet. Leider sind diese definitorisch nicht immer klar voneinander abgegrenzt; die Unterschiede sind aber marginal:

- *Mental Speed*: Meist wird darunter die nicht näher spezifizierte Schnelligkeit von Denkprozessen verstanden. Der Begriff ist unter allen hier aufgeführten der allgemeinste.
- Nahezu identisch ist die *allgemeine Verarbeitungsgeschwindigkeit*, d. h. die allgemeine Fähigkeit, verschiedene mentale Operationen, z. B. einfache Rechenaufgaben schnell durchzuführen.
- Das Gleiche gilt für die *Verarbeitungsgeschwindigkeit*. Der Begriff wird z. B. zur Benennung einer Intelligenzkomponente in den Wechsler-Intelligenztests verwendet, die mit Aufgaben wie dem Zahlen-Symbol-Test erfasst wird.
- Die *Bearbeitungsgeschwindigkeit* im Berliner Intelligenzstruktur-Test (Jäger et al. 1997) meint Dasselbe. Sie ist dort eine von 4 basalen mentalen Operationen und wird als „Arbeitstempo, Auffassungsleichtigkeit und Konzentrationskraft beim Lösen einfacher strukturierter Aufgaben von geringem Schwierigkeitsniveau“ (Jäger et al. 1997, S. 6) definiert; eine der Aufgaben ist ein Zahlen-Symbol-Test.
- *Informationsverarbeitungsgeschwindigkeit*: Nimmt man den Begriff wörtlich, so wird hier spezifiziert, was verarbeitet wird, nämlich Informationen. Aber auch in den Tests zur Ver- oder Bearbeitungsgeschwindigkeit werden Informationen verarbeitet.
- *Wahrnehmungsgeschwindigkeit*: Nach Thurstone (1938) stellt sie eine von 7 basalen Faktoren der Intelligenz dar; entsprechende Tests verlangen die schnelle Unterscheidung visuell ähnlicher Reize. Ackerman et al. (2002) unterscheiden 4 Komponenten, darunter die Mustererkennung in Durchstreichtests.

Diese Konzepte passen erstaunlich gut zum Konstrukt „Konzentration“ (effiziente = schnelle und richtige Erledigung von Tätigkeiten – unabhängig von weiteren aufgabenspezifischen Fähigkeiten/Fertigkeiten/Kompetenzen).

Konzentrationsfähigkeit und Verarbeitungsgeschwindigkeit sind weitgehend das Gleiche

Verarbeitungsgeschwindigkeit als Komponente der Intelligenz

Wie ist es möglich, dass ein und dieselbe Art von Aufgaben einmal zur Messung der Konzentrationsfähigkeit und ein andermal der Intelligenz verwendet werden? Diese scheinbare Unstimmigkeit lässt sich erklären. Klassische Konzentrationstests wie etwa der d2 korrelieren nur um $r=.30$ mit breiten Intelligenztests wie etwa dem I-S-T 2000 R (► Abschn. 3.2.3). Dass Konzentrationstestaufgaben eine gemeinsame Varianz mit einem Maß der Allgemeinen Intelligenz oder mit dem Kernbereich des schlussfolgernden Denkens aufweisen, ist also nicht zu bestreiten. Und Konzentrationsfähigkeit und Verarbeitungsgeschwindigkeit wurden und werden mit teilweise identischen Testverfahren operationalisiert. Deshalb kann ein Intelligenztest, dem ein passendes Strukturmodell zugrunde liegt, auch Aufgaben zur „Verarbeitungsgeschwindigkeit“ – und damit zur Konzentrationsfähigkeit – enthalten. Gerade Test zur „Allgemeinen Intelligenz“ sollen Aufgaben enthalten, die unterschiedliche Komponenten der kognitiven Leistungsfähigkeit abdecken. Die Verarbeitungsgeschwindigkeit ist eine dieser Komponenten.

In Intelligenztests dient traditionell die Anzahl der richtigen Antworten als Kennwert für die Testleistung. In Konzentrationstests finden dagegen auch die Fehler Beachtung. Meist wird auch ein Kennwert „Fehlerprozent“ (F%) bestimmt, der besagt, wie viel Prozent der Aufgaben falsch gelöst wurden. Das ist insbesondere bei Durchstreichtests (s. u.) sinnvoll, da sich die Anzahl der richtigen Antworten durch zufälliges Markieren oder Raten leicht in die Höhe treiben lässt. Die Zusatzangabe zur Fehlerrate ist nützlich bei der Interpretation der Trefferzahl. Man erkennt, ob die Testleistung („Richtige“) eher durch fehlerfreies Arbeiten zustande gekommen ist oder ob ungenaues Arbeiten möglicherweise über Zufallstreffer zur Erhöhung der Testleistung beigetragen hat. Auch eine Verrechnung von Fehlern bei der Gesamttestleistung kommt vor.

Es liegt auf der Hand, dass die Leistung in Tests in Bezug auf die Schnelligkeit bei der Bearbeitung kognitiv einfacher Aufgaben nicht nur von der Verarbeitungsgeschwindigkeit/Konzentration abhängt. Jede Testaufgabe verlangt auch bestimmte Fähigkeiten oder Fertigkeiten. Je nach Aufgabe sind dies beispielsweise die Rechenfertigkeit, die Kombinationsfähigkeit, die Merkfähigkeit, die Psychomotorik oder auch die Aufmerksamkeit. Die Anforderungen an diese testspezifischen Fähigkeiten bzw. Fertigkeiten sind eventuell gering, aber völlig ignorieren sollte man sie nicht. Würden sie keine Rolle am Zustandekommen der Testleistung spielen, sollten Konzentrationstests, die verschiedene Fähigkeiten bzw. Fertigkeiten verlangen, sehr hoch miteinander korrelieren. Das ist aber nicht der Fall; die Korrelationen bewegen sich oft im Bereich von $r=.60$. Natürlich mindern nicht nur unterschiedliche kognitive Anforderungen den Zusammenhang, sondern auch Unterschiede im Antwortmodus oder etwa in der Testdauer. Jedenfalls lässt sich festhalten, dass die gemessene Konzentrationsfähigkeit immer mit den Fähigkeiten bzw. Fertigkeiten konfundiert ist, die zur Lösung der Aufgaben benötigt werden. In Faktorenanalysen von verschiedenen Konzentrationstests sowie von Tests zur Erfassung der Zusatzanforderungen (Rechenfertigkeit, Merkfähigkeit etc.) konnte diese Aufspaltung der Konzentrationstestleistung in Konzentrationsfähigkeit und testspezifische Fähigkeiten bzw. Fertigkeiten demonstriert werden (Schmidt-Atzert et al. 2006).

Konzentrationstests lassen sich – im Gegensatz zu beispielsweise Intelligenztests – nicht anhand von theoretischen Modellen unterscheiden, da solche Modelle bisher nicht ausgearbeitet wurden. Für Anwenderinnen und Anwender ist vor allem relevant, welche Art von Aufgaben verwendet wird und für welche Zielgruppe ein Test vorgesehen ist. Sowohl Buchstabendurchstreichtests als auch Konzentrationsrechentests haben eine lange Tradition:

Konzentrationstests haben meist auch Kennwerte für Fehler

Konfundierung von Konzentrationstestleistungen und anderen Fähigkeiten/Fertigkeiten

Keine Strukturmodelle zur Konzentration

Die ersten Verfahren dieser Art wurden bereits in den Jahren 1885 bzw. 1888 entwickelt (Bartenwerfer 1964).

Die Tests können nach den Aufgaben unterteilt werden, bei deren Bearbeitung Konzentration verlangt wird. Die verlangte kognitive Operation dient also der Einteilung. □ Tab. 3.5 zeigt die wichtigsten Arten von Konzentrationstests. Da bei den Suchaufgaben meist die Zielobjekte durchzustreichen sind, wurden diese Tests oft auch „Durchstreichtests“ genannt. Mit dieser Bezeichnung wird aber ein unerheblicher Aspekt der Testbearbeitung hervorgehoben, nämlich die Art der Itembeantwortung. Besser ist eine Benennung nach der geforderten mentalen Operation, also das Suchen definierter Zeichen. Auch die Bezeichnungen „Aufmerksamkeitstests“ oder besser „Tests zur konzentrierten Aufmerksamkeit“ treffen zu, weil das Suchen bzw. das willentliche Beachten von bestimmten Reizen ein wesentliches Merkmal von Aufmerksamkeit ist (s. o.).

Einige Konzentrationstests verlangen gleich mehrere kognitive Operationen. In □ Tab. 3.5 ist der KLT-R aufgeführt, der Rechnen, Merken von Zwischenergebnissen und in der Version für Schülerinnen und Schüler der 6. bis 13. Klasse (nicht aufgelistet) zusätzlich noch die Anwendung einer Regel verlangt (wenn das 1. größer als das 2. Zwischenergebnis ist: Zwischenergebnisse subtrahieren – wenn umgekehrt, beide addieren). Das Inventar komplexer Aufmerksamkeit (INKA) von Heyde (2000, 2004) kombiniert die beiden Aufgabentypen Transformation und Suchen. Vorgegeben sind lange Reihen von Konsonanten (z. B. RFLBPHZM...). Für jede Zeile müssen bestimmte Konsonanten anhand einer Umwandlungstabelle in andere transformiert werden (aus B wird beispielsweise Z). Danach sind die transformierten Konsonanten (also beispielsweise Z) in der Zeile zu suchen. Sie werden jedoch nicht markiert, sondern der davorstehende Konsonant ist am Rand zu notieren (RFLB-PHZM... Antwort: H). Die Aufgabe ist also tatsächlich so komplex, wie der Testname vermuten lässt; daher stellt der Test auch erhebliche Anforderungen an die Merkfähigkeit und die Intelligenz der Testpersonen (□ Tab. 3.6).

Durchstreichtests bzw. Suchen definierter Zeichen

Komplexe Anforderungen

□ Tab. 3.5 Einteilung der Konzentrationstests nach Aufgabentypen

Aufgabe	Testbeispiel	Erläuterung zur Aufgabe
Suchen definierter Zeichen	Aufmerksamkeits- und Konzentrationstest d2-R (Brickenkamp et al. 2010)	Alle d's mit 2 Strichen durchstreichen. Die Zielobjekte verbergen sich unter d's mit einer „falschen“ Strichzahl und p's mit unterschiedlich vielen Strichen (□ Abb. 3.4).
Suchen wechselnder Zeichen	Zahlen-Verbindungs-Test (ZVT) von Oswald und Roth (1987)	Die Zahlen von 1 bis 90 sind auf dem Testbogen verteilt und müssen nacheinander (1 – 2 – 3 etc.) verbunden werden.
Vergleichen	Differentieller Konzentrationstest für Kinder (DKT-K; Funsch und Arias Martin 2017)	Ein Bild (z. B. bunte Zeichnung eines Hauses) ist zu sehen; die Kinder müssen prüfen, ob sich dieses Bild unter 3 Vergleichsbildern befindet.
Rechnen	Konzentrations-Leistungs-Test – revidierte Fassung (KLT-R; Düker et al. 2001)	2 Additionsaufgaben (mit je 3 einstelligen Zahlen) durchführen (Version für Kinder im 4. bis 6. Schuljahr)
Merken	Konzentrations-Leistungs-Test – revidierte Fassung (KLT-R; Düker et al. 2001)	Ergebnisse der Rechenaufgaben (s. Rechnen) merken, dann wieder rechnen: kleinere Zahl von der größeren abziehen
Sortieren (Klassifizieren)	Konzentrations-Verlaufs-Test (KVT; Abels 1974)	60 Kärtchen mit zweistelligen Zahlen durchsehen und auf 4 Stapel sortieren: Kärtchen enthält die Zahl 43, die Zahlen 43 und 63, die Zahl 63, weder 43 noch 63
Transformieren	Zahlen-Symbol-Test des Berliner Intelligenzstruktur-Tests: BIS-Form 4 (Jäger et al. 1997)	Transformation von Zahlen (1 bis 9) in Symbole anhand einer Umwandlungstabelle. Vorgegeben sind Zahlen, unter die jeweils das passende Symbol (z. B. + bei der Zahl 9) einzutragen ist.

■ Tab. 3.6 Zuordnung von Konzentrationstests zu Faktoren

Tests	KON	N	F	INT	GED
BIS-ZS	++	-	-	-	-
ZVT	++	-	+	-	-
Rev. T.	++	++	-	-	-
KLT-R	-	++	-	-	++
FAIR	++	-	++	-	-
Test d2	++	-	++	-	-
INKA	-	-	++	+	+

Quelle: Ergebnisse aus Schmidt-Atzert et al. (2006). BIS-ZS (Zahlen-Symbol-Test des Berliner Intelligenzstruktur-Test), Rev. T. (Revisions-Test) und Test d2 wurden in beiden Untersuchungen eingesetzt. ZVT = Zahlen-Verbindungs-Test, KLT-R = Konzentrations-Leistungs-Test – revidierte Fassung, FAIR = Frankfurter Aufmerksamkeits-Inventar, INKA = Inventar komplexer Aufmerksamkeit; + + symbolisiert eine hohe Ladung auf dem Faktor, + eine moderate und - eine niedrige oder nicht spezifizierte Ladung. Benennung der Faktoren: KON = Konzentration, N = numerische, F = figurale Fähigkeiten (oder Aufmerksamkeit), INT = Intelligenz, GED = Gedächtnis

Fragen an die Forschung

Angesichts der offensichtlichen Unterschiedlichkeit der Tests ergeben sich mindestens 3 wichtige Fragen an die Forschung: Erstens ist zu klären, ob die Konzentrationstests so viel gemeinsame Varianz aufweisen, dass man annehmen darf, dass sie eine gemeinsame Fähigkeit messen. Eine alternative Hypothese ist, dass sich mehrere Formen der Konzentration unterscheiden lassen. Zweitens ist von Interesse, welche Tests als die typischsten Vertreter ihrer Gattung gelten können. Es sollten jene Tests sein, die viel Konzentrationsvarianz und wenig andere Testvarianz aufweisen. Drittens werden insbesondere Testanwenderinnen und -anwender wissen wollen, von welchen anderen Fähigkeiten und Fertigkeiten der Testpersonen die Leistung in einzelnen Konzentrationstests abhängt und wie stark diese Abhängigkeit ist.

Zur Beantwortung dieser Fragen haben Schmidt-Atzert et al. (2006) in 2 Untersuchungen insgesamt 10 bzw. 11 Tests zur Erfassung von Konzentration und verwandten Konstrukten (z. B. Informationsverarbeitungsgeschwindigkeit) sowie weitere Tests zur Validierung der Konzentrationsfaktoren bearbeiten lassen. Als Kennwert wurde immer die Anzahl der richtig bearbeiteten Items bzw. Zielobjekte verwendet. Eine Synopse der Ergebnisse aus beiden Untersuchungen findet sich in ■ Tab. 3.6, in der weitere Subtests aus dem Berliner Intelligenzstruktur-Test (BIS) sowie die Tests zur Validierung aus Platzgründen nicht aufgeführt worden sind. Die Benennung der Faktoren orientiert sich an den Ladungen zusätzlicher Tests (Rechentest, Gedächtnistest, Intelligenztests) auf den jeweiligen Faktoren. Beim Faktor „figurale Fähigkeiten“ handelt es sich möglicherweise auch um einen Aufmerksamkeitsfaktor, da die Tests nicht nur figurales Material verwenden, sondern auch eine Selektion anhand von Reizen verlangen.

In beiden Untersuchungen ließ sich anhand konfirmatorischer Faktorenanalysen ein einziger Konzentrationsfaktor nachweisen, der durch den Zahlen-Symbol-Test (ZS) aus dem BIS markiert wurde. Der Zahlen-Symbol-Test hatte keine nennenswerten Ladungen auf anderen Faktoren; das bedeutet, dass dieser Test die Konzentrationsfähigkeit am besten von allen Tests erfasst. Alternativ lässt sich auch eine Kombination des Test d2 (oder FAIR) und des Revisions-Tests einsetzen; die Leistungen in diesen Tests hängen jedoch auch von figuralen (Test d2, FAIR) bzw. numerischen Fähigkeiten

Faktorenanalytische Untersuchung mit vielen Tests

Ein Konzentrationsfaktor

(Rev. T.) ab. Ein erstaunliches Ergebnis war, dass sich der Zahlen-Verbindungs-Test (ZVT), der zur Messung der Intelligenz entwickelt worden ist, als guter Konzentrationstest erwies, der zudem nur wenig intelligenzabhängig zu sein scheint. Die Faktorenanalysen der Tests zeigten ferner, dass die Rechenkonzentrationstests (Rev. T. und KLT-R) erwartungsgemäß auf dem numerischen Faktor laden, wobei der KLT-R primär nicht Konzentration, sondern Rechenfertigkeit und Merkfähigkeit zu erfassen scheint. Auch der INKA misst offenbar weniger die Konzentration, sondern vielmehr figurale Fähigkeiten (oder Aufmerksamkeit), Intelligenz und Merkfähigkeit.

Im Folgenden wird mit dem d2-R ein ausgewählter Konzentrationstest ausführlich dargestellt. Es handelt sich dabei um den im deutschen Sprachraum am häufigsten eingesetzten Konzentrationstest, der auch international bekannt und in vielen Ländern verfügbar ist. Am Beispiel dieses Tests wird auch gezeigt, welche Herausforderung die Übersetzung eines Konzentrationstests, der zunächst als Papier-und-Bleistift-Test entwickelt wurde, in eine Computerversion darstellt. Anschließend werden wir auf ähnliche Tests der gleichen Kategorie sowie auf ausgewählte Konzentrationstests hinweisen, die auf einem anderen Testprinzip aufgebaut sind. Gemeinsamkeiten mit dem d2-R sowie Unterschiede werden herausgestellt.

Aufmerksamkeits- und Konzentrationstest d2-R und d2-R elektronisch

Steckbrief d2-R: Test d2 – Revision. Aufmerksamkeits- und Konzentrationstest (Brickenkamp et al. 2010) und computerbasierte Version (Schmidt-Atzert und Brickenkamp 2017)

Zielsetzung und Testkonstruktion	
Messgegenstand	Aufmerksamkeit und Konzentration („konzentrierte Aufmerksamkeit“)
Anwendungsbereich	Breiter Anwendungsbereich
Theoretischer Hintergrund	Konzeptuelle Einordnung in den Bereich Verarbeitungsgeschwindigkeit anhand der Testeigenschaften
Testentwicklung	Testverlängerung durch Hinzunahme von Items gegenüber der 9. Auflage; keine Itemselektion
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche und mündliche Instruktion; Computerversion voll standardisiert; beide enthalten Hinweise für den Umgang mit Fragen
Auswertung	Durchschreibbogen und feste Regeln zur Auswertung; Computerversion mit automatischer Auswertung
Interpretation	Vorgaben zur Benennung der erfassten Merkmale und zur Verbalisierung ihrer Ausprägung; konkrete Hinweise auf nicht instruktionsgemäße Testbearbeitung; Computerversion erzeugt Report auf Basis dieser Vorgaben
Reliabilität	
Konsistenz	Hauptkennwert KL: $\alpha = .90$ bis $.95$; Tempo BZO (bearbeitete Zielobjekte): $\alpha = .89$ bis $.95$; Genauigkeit F%: $\alpha = .80$ bis $.91$ (Gesamtstichprobe: $\alpha = .96/.97/.87$); Computerversion: KL: $\alpha = .90$ bis $.98$; BZO: $\alpha = .86$ bis $.97$; F%: $\alpha = .81$ bis $.93$ (Gesamtstichprobe: $\alpha = .98/.97/.92$)
Retest	$r = .94/.85$ (KL), $.91/.92$ (BZO) und $.84/.47$ (F%) (1 Tag/10 Tage, $N = 118/145$); mehrere Retest-Ergebnisse zur Vorgängerverision d2 in ähnlicher Höhe; Computerversion: $r = .93$ (KL), $.90$ (BZO), $.75$ (F%) (16 Tage, $N = 86$)

Validität	
Konstruktvalidität	Im Manual der Papierversion zahlreiche neuere Studien zur Korrelation des d2 mit anderen Konzentrationstests (u. a. d2 KL mit Richtigen in Referenztests BIS-B, BIS-Zahlen-Symbol-Test, FAIR, FAKT, Revisions-Test, ZVT, Cognitron): $r=.43$ bis .81; für F% $r=.08$ bis .40, Korrelation KL mit Intelligenz um $r=.30$; Computerversion: ebenfalls deutlicher Zusammenhang von KL mit vergleichbaren Kennwerten in anderen Konzentrationstests und vergleichsweise geringer Zusammenhang mit Reasoning, einfacher Reaktionszeit, Merkfähigkeit etc.
Kriteriumsvalidität	Zur Papierversion ältere Studien zur Fahreignung; Computerversion: $r(KL/F\%)=.32/.34$ bzw. $.54/.27$ mit standardisierten „Arbeitsproben“ mit Bezug zur Berufseignung (Tippfehler finden, Unternehmensbeteiligungen erkennen)
Normen	
Zusammensetzung	Normen ($N=4024$) für 5 Altersgruppen zwischen 9 und 19 Jahren sowie für 20–39 und 40–60 Jahre; $n=268$ bis 728 pro Altersgruppe; Computerversion: Gewichtung nach Bildungsstand wie in Gesamtbevölkerung für jede Altersgruppe; Normen ($N=3046$) für 16 Altersgruppen von 8 bis 70–79 Jahre; $n=35$ bis 98 für 8, 9, 10, 11 bzw. 70–79 Jahre und $n=105$ bis 939 in den anderen Gruppen; zusätzlich gesamteuropäische Normen ($N=2100$) für den Altersbereich 18–55 Jahre (wegen kontinuierlichen Normierung keine Altersgruppen); Stichproben in den 10 Ländern ($n=126$ bis 277) jeweils repräsentativ für Alter, Bildung und Geschlecht
Erhebungszeitraum	Ende 2007 bis Mitte 2008; Computerversion: 2013 bis 2016 (51 % in 2015)
Sonstiges	
Formen	Papier-und-Bleistift- sowie Computerversion; beide auch in mehreren europäischen Sprachen publiziert
Testrezension	
Quelle	Zur Papier-und-Bleistift-Version: Daseking und Putz (2015)

Beim Test d2-R handelt es sich um die revidierte Version eines Tests, der schon seit 1962 auf dem Markt ist (Brickenkamp 1962). Wir stellen unten auch eine „elektronische Version“ des d2-R vor (die Bezeichnung dient der Abgrenzung zu einer bereits publizierten „Computerversion“). Sie verwendet die gleichen Items und Kennwerte wie die Papierversion, ist aber notwendigerweise etwas anders zu bedienen. An geeigneter Stelle wird auch auf relevante Befunde verwiesen.

Das Testmaterial des d2 wurde bis zur 9. Auflage (Brickenkamp 2002) nicht verändert. Ursprünglich sollte der Test vor allem zur Feststellung der Kraftfahreignung dienen. Der Testautor hatte sich für das bewährte Prinzip der Durchstreichtests entschieden, dabei allerdings versucht, gewisse Unzulänglichkeiten der damals verfügbaren Durchstreichtests zu überwinden. Ziel war ein Test mit einer einfachen und verständlichen Instruktion und einer Aufgabe, die ohne lange Einübung ausgeführt werden kann. Allein die Diskrimination von 2 Buchstaben zu verlangen, erschien angesichts der komplexen Anforderungen an Kraftfahrer als zu einfach. Deshalb führte Brickenkamp Striche unter und über den Buchstaben ein. Zielobjekte (der Buchstabe d mit insgesamt 2 Strichen) und Distraktoren (d mit „falscher“ Strichzahl und p mit manchmal auch 2 Strichen) unterscheiden sich damit auf 2 Ebenen, nämlich in Bezug auf den Buchstaben und die Strichzahl. (► Abb. 3.4 (► Abschn. 3.2.2.2) zeigt die bis heute verwendeten Items.

Im Test selbst stehen die Zeichen dichter nebeneinander als in dem hier gezeigten Ausschnitt. Auf dem Testbogen im DIN-A4-Format befinden sich 14 Zeilen. In jeder Zeile stehen 57 Zeichen (im d2 der 9. Auflage waren es

Test mit langer Tradition

noch 47 – aber es traten Deckeneffekte auf). Der d2-R besteht also insgesamt aus 798 Zeichen. Durchzustreichende Zielobjekte sind alle d's mit 2 Strichen, von denen es 3 Varianten gibt: 1 Strich über und 1 Strich unter dem d, 2 Striche über dem d und 2 Striche unter dem d.

Eingesetzt wird der Test u. a. in der Klinischen Psychologie, der Neuropsychologie, der Arbeits- und Organisationspsychologie (zur beruflichen Eignungsdiagnostik), der Pädagogischen Psychologie, der Sportpsychologie und in der Verkehrpsychologie. Befragungen von Psychologinnen und Psychologen in der Berufspraxis zeigen, dass der Test d2 von allen „allgemeinen Leistungstests“ mit Abstand am häufigsten eingesetzt wird (► Abschn. 3.2.2). Bölte et al. (2000, S. 8), die speziell nach der Verwendung von Tests in der Kinder- und Jugendpsychiatrie gefragt hatten, resümieren: „Bei den Aufmerksamkeitstests zeigt sich der d2 in unserer Erhebung als weitgehend konkurrenzlos.“ Der d2-R wird inzwischen auch in vielen anderen Ländern eingesetzt.

Welches Merkmal wird mit dem d2 erfasst? Da der Test eine Reizselektion verlangt (bestimmte Zeichen sind unter visuell ähnlichen Zeichen herauszusuchen), erfasst er *Aufmerksamkeit*. Diese kann aufgrund der genauen Aufgabenstellung näher als selektive oder fokussierte Aufmerksamkeit bestimmt werden (► Tab. 3.4). Diese Aufmerksamkeitsleistung muss kontinuierlich und dabei sowohl schnell als auch richtig erbracht werden, weshalb Brickenkamp (2002) den Test zutreffend auch als *Konzentrationstest* einordnete.

Die Frage nach dem Messgegenstand wurde im Manual des d2-R sorgfältig erörtert. Die bis zur 9. Auflage verwendete Bezeichnung „Belastungstest“ im Namen des Tests bezog sich darauf, dass die Testleistung unter Zeitdruck zu erbringen ist: Die Testleiterin oder der Testleiter fordert die Testperson alle 20 s dazu auf, die Bearbeitung der aktuellen Zeile abzubrechen und mit der nächsten anzufangen. Die Bezeichnung erwies sich aber als irreführend: Einige Anwenderinnen und Anwender nahmen an, dass der Test auch Belastbarkeit misst. Der Test d2-R trägt deshalb nicht mehr den missverständlichen Zusatz „Belastungstest“. Das mit dem Test erfasste Konstrukt wird nun präziser als „konzentrierte Aufmerksamkeit“ bezeichnet, also eine Überschneidung von Aufmerksamkeit und Konzentration.

Dass der d2 und damit auch der d2-R die Aufmerksamkeit und die Konzentrationsfähigkeit erfasst, begründen die Testautoren so: Die Selektion von bestimmten Reizen (d's mit 2 Strichen) unter ähnlichen Reizen ist eine klassische Aufmerksamkeitsleistung. Da diese Leistung ohne Pause und über viele Items hinweg (bei vollständiger Testbearbeitung immerhin 798×) vollbracht werden muss, benennen die Autoren die Kombination von 2 Fähigkeiten „konzentrierte Aufmerksamkeit“ und in englischsprachigen Publikationen „sustained attention“.

Der Test verlangt also, wie viele andere Konzentrationstests, die schnelle Verarbeitung von Reizen oder Informationen. Dafür finden Begriffe wie Mental Speed, Be- oder Verarbeitungsgeschwindigkeit etc. Verwendung (► Abschn. 3.2.2). Im Manual des d2-R sowie der elektronischen Version des d2-R (s. u.) wird herausgearbeitet, dass der d2-R (ebenso wie sein Vorgänger d2) konzeptuell mühelos in all diese Konzepte eingeordnet werden kann. Beispielsweise lässt sich der d2-R im Berliner Intelligenzstrukturmodell (► Abb. 3.12) in der Facette „Bearbeitungsgeschwindigkeit figural“ verorten. Weiterhin wird betont, dass Tests, welche die oben genannten Geschwindigkeiten erfassen sollen, nur schwach mit Maßen der Allgemeinen Intelligenz korrelieren. Der d2-R misst also die Fähigkeit, (visuelle) Informationen schnell zu verarbeiten. Unter den in ► Abschn. 3.2.2 genannten Begriffen

Verbreiteter Test mit vielen Anwendungsgebieten

Der Test d2 misst Aufmerksamkeit und Konzentration

Konzentrierte Aufmerksamkeit (sustained attention)

Konzentrierte Aufmerksamkeit = Fähigkeit, Informationen schnell zu verarbeiten

20 s Bearbeitungszeit pro Zeile

kommt die *Informationsverarbeitungsgeschwindigkeit* dem am nächsten. Mit *konzentrierter Aufmerksamkeit* ist nichts anderes gemeint. Wir haben hier also verschiedene Namen für das gleiche Konstrukt!

Die Markierung erscheint auf dem Durchschreibbogen

Durchführung Die Instruktion erfolgt durch Vorlesen eines Standardtextes. Den Testpersonen wird eine Kurzanleitung vorgelegt, auf der alle wesentlichen Punkte der mündlichen Instruktion aufgeführt sind. Die Kurzanleitung enthält auch eine Auflistung aller Zielobjekte und Distraktoren sowie 2 Übungszeilen. Wichtig ist die Anweisung am Ende der Instruktion: „Arbeiten Sie so schnell wie möglich – aber möglichst ohne Fehler!“ Die Bearbeitung des Tests erfolgt auf einem separaten Testbogen. Die Testleiterin bzw. der Testleiter fordert mit „Achtung! – Los!“ dazu auf, mit der 1. Zeile zu beginnen. Nach 20 s kommt der Befehl „Halt! Nächste Zeile!“. Die Stoppuhr läuft dabei weiter, und die Anweisung zum Zeilenwechsel wird alle 20 s wiederholt. Die Testdurchführung dauert damit ohne Instruktion genau 4 min und 40 s. Der Test kann einzeln und in Gruppen durchgeführt werden.

Kennwerte: BZO, F% und KL

Auswertung Wird auf dem Testbogen ein Zeichen durchgestrichen, drückt sich der Strich auf den Durchschreibbogen durch. Dort wo auf dem Testbogen Zielobjekte stehen, befinden sich auf dem Durchschreibbogen leere Felder. Diese sind nummeriert, damit man leicht die Anzahl der bearbeiteten Zielobjekte (BZO) ablesen kann. Auslassungsfehler (AF = nicht durchgestrichene Zielobjekte) sind an leeren Feldern und Verwechslungsfehler (VF = durchgestrichene Distraktoren) an Markierungen zwischen den Feldern zu erkennen; diese werden jeweils gezählt. Die Ergebnisse werden auf einem separaten Auswertungsbogen festgehalten. Die 1. Zeile wird nicht ausgewertet, weil die Leistung hier leicht durch zu frühes oder zu spätes Starten beeinflusst werden kann. Die letzte Zeile bleibt ebenfalls unberücksichtigt, da in Gruppenuntersuchungen einzelne Personen unbemerkt nach dem Stoppsignal weiterarbeiten können.

Folgende 3 Kennwerte werden zur Beschreibung der Testleistung bestimmt und anhand einer Normtabelle in Standardwerte transformiert.

- Anzahl der bearbeiteten Zielobjekte (BZO) als Maß für das Tempo bei der Testbearbeitung
- Fehlerprozent (F%) als Maß für die Sorgfalt bei der Testbearbeitung = $(AF + VF) / BZO \times 100$
- Konzentrationsleistung (KL) als Gesamtmaß für die Konzentrationsfähigkeit = $BZO - AF - VF$

Konzentrationsleistungswert (KL) als Hauptkennwert

Der Konzentrationsleistungswert (KL) ist das Maß für die Konzentrationsfähigkeit. Die Kennwerte BZO und F% informieren über das Arbeitsverhalten bei der Testbearbeitung und zeigen, welche Bedeutung das Tempo und die Sorgfalt für die Konzentrationsleistung haben. Beispielsweise kann eine Testperson sehr schnell und zugleich sehr sorgfältig gearbeitet haben oder etwa relativ langsam und dabei sehr genau.

Hinweis auf Simulation

Um eine irreguläre Testbearbeitung zu erkennen, wird eine Fehleranalyse empfohlen: In einer experimentellen Studie fielen die Testpersonen, die eine schlechte Testleistung vortäuschen sollten, durch Buchstabenfehler (p's mit 2 Strichen markiert) auf. Die Studie konnte mit der elektronischen Version des d2-R (s. u.) repliziert werden. Das Auftreten von 2 oder mehr Buchstabenfehlern gilt als starker Warnhinweis, dass hier jemand versucht hat, eine niedrige Konzentrationsfähigkeit vorzutäuschen.

Eine andere Form der irregulären Bearbeitung ist das zufällige Markieren von Items. Die Testperson hat möglicherweise die Instruktion nicht verstanden oder vergessen oder sie boykottiert die Testung. Erkennbar ist zufälliges Markieren u. a. an einem negativen KL-Wert.

Zufälliges Markieren erkennbar

Reliabilität Zur Schätzung der *internen Konsistenz* wurden die 4 völlig identischen Blöcke aus je 3 verschiedenen Testzeilen als Items betrachtet. Die 1. und letzte Zeile werden nicht ausgewertet (s. o.). Cronbachs α variiert leicht mit dem Alter der Testpersonen und darf für die Kennwerte KL und BZO (Tempo) mit Werten zwischen .90 und .95 bzw. .89 bis .95 als hoch bis sehr hoch klassifiziert werden. Für F% (Genauigkeit) liegen die Werte mit .80 bis .91 noch im hohen Bereich. Die Angaben beziehen sich auf die Eichstichprobe.

Die *Retest-Reliabilität* wurde durch Testwiederholung nach einem ($N=118$ Studierende) bzw. nach 10 Tagen ($N=145$ Schülerinnen und Schüler) ermittelt. Für die Konzentrationsleistung (KL) ergaben sich Werte von $r_{tt}=.94$ (1 Tag) bzw. .85 (10 Tage). Der niedrigere Wert nach 10 Tagen ist auf eine geringere Stabilität der Genauigkeit F% ($r_{tt}=.47$ vs. .84 nach 1 Tag) zurückzuführen; beim Tempowert BZO finden sich praktisch keine Unterschiede (.91 bzw. .92). Mit Ausnahme der relativ niedrigen Retest-Reliabilität für den Fehlerwert decken sich die Ergebnisse mit denen früherer Untersuchungen. Da der d2-R als nahezu äquivalent zum d2 gelten kann (auch belegt durch eine Äquivalenzstudie), dürfen wir annehmen, dass der d2-R auch für ein Zeitintervall von 1 Jahr und darüber hinaus noch eine hohe Retest-Reliabilität aufweist. Diese lag beim d2 selbst für den Fehlerprozentwert nach 2 Jahren noch im Bereich von .60.

Hohe interne Konsistenz und Retest-Reliabilität

Validität Die Validierung von Aufmerksamkeits- und Konzentrationstests erfolgt üblicherweise über die Korrelation mit anderen Tests, die das gleiche Konstrukt erfassen (sollen). Dieser Ansatz wird klassischerweise als *Konstruktvalidierung* bezeichnet. Problematisch ist daran, dass die Referenztests ebenfalls nur an anderen Tests validiert wurden.

Konstruktvalidierung an anderen Tests

Die Referenztests, mit denen der d2 korreliert wurde, lassen sich nach dem Aufgabentyp ordnen. Mit dem Frankfurter Aufmerksamkeits-Inventar (FAIR; Moosbrugger und Oehlschlägel 1996) liegt ein figuraler Konzentrationstest vor, der dem d2 ähnlich ist, weil Zeichen zu suchen sind, die sich auf 2 Dimensionen unterscheiden (Kreise vs. Quadrate mit 2 oder 3 Punkten; Abb. 3.5). Die Autoren berichten eine Korrelation von $r=.50$ mit dem d2. Im Manual des d2-R ist eine Studie angegeben, in der die Korrelation mit dem FAIR $r=.55$ beträgt.

Moderater Zusammenhang mit einem anderen figuralen Konzentrationstest

		Gestalt	
		Kreis	Quadrat
Punkte-Anzahl	3		
	2		

Abb. 3.5 Die Zellen zeigen die 4 Itemarten des FAIR, die durch die Variation der 2 Reizdimensionen Gestalt und Punkteanzahl erzeugt werden. Jede Itemart kommt in 2 Varianten vor. Die Anordnung der Punkte ist jedoch für die Aufgabe nicht relevant. (Aus Moosbrugger und Oehlschlägel 1996, © Hogrefe)

Test erfasst auch längerfristige Konzentration

Rechenkonzentrationstests verwenden einen deutlich anderen Aufgabentyp. Steck (1996) ließ seine Testpersonen nacheinander verschiedene lange Version des Pauli-Tests (Arnold 1975) sowie den Test d2 bearbeiten. Der Pauli-Test verlangt das fortwährende Addieren einstelliger Zahlen, die in langen Spalten angeordnet sind (z. B. 2 7 5 ...). Dabei sind stets 2 direkt untereinanderstehende Zahlen zu addieren; die Ergebnisse sind zu notieren. Im Beispiel lauten die Ergebnisse 9 ($2+7$) und 12 ($7+5$); bei zweistelligen Zahlen wird nur die letzte Ziffer, hier also 2, notiert. Der Test d2 korrelierte (zu $r=.52$) mit der Kurzversion (5 min) und zu .48 mit der Langversion (20 min) des Pauli-Tests. Eine weitere Personengruppe bearbeitete eine 30-minütige Version des Pauli-Tests (Christiansen 1983). Die Korrelation mit dem Test d2 betrug $r=.45$. Diese Studie wurde mit der elektronischen Version des d2-R (s. u.) repliziert. Der KL-Wert in d2-R korrelierte mit der 30-minütigen Langform des Pauli-Tests sogar etwas höher ($r=.46$) als mit einer Kurzform (den ersten 6 min; $r=.38$), und das bei nahezu identischer Reliabilität der beiden Pauli-Test-Versionen ($\alpha=.97/.98$).

Moderater Zusammenhang mit Rechenkonzentrationstests

In mehreren Studien kam mit dem Revisions-Test (Marschner 1980) ein anderer Rechentest zum Einsatz. Auch hier sind einstellige Zahlen zu addieren. Allerdings wird stets ein Ergebnis gezeigt (z. B. $3+6=8$) und die Testperson muss richtige Ergebnisse mit einem Häkchen versehen und falsche durchstreichen. Der KL-Wert des d2 korrelierte um $r=.60$ mit dem entsprechenden Kennwert des Revisions-Tests (s. Brickenkamp et al. 2010, Tab. 23).

Moderater Zusammenhang mit weiteren Konzentrationstests

Zwei Tests, die weder als figural noch numerisch klassifiziert werden können, sind der Zahlen-Verbundungs-Test (ZVT; Oswald und Roth 1987) und der Zahlen-Symbol-Test (ZS), der u. a. im BIS-Test (Jäger et al. 1997) verwendet wird. Beim ZVT stehen die Zahlen 1 bis 90 auf dem Testbogen und die Testpersonen müssen, beginnend mit 1, die Zahlen mit der jeweils nächsthöheren mit einem Strich verbinden. Die nächste Zahl befindet sich stets in der Nähe der vorigen. Der Test kann damit als Suchaufgabe klassifiziert werden, bei der – anders als beim d2 – nicht immer die gleichen, sondern wechselnde Zeichen zu suchen sind. Der ZS verlangt eine Transformation („Übersetzung“) von Zeichen mithilfe einer Umwandlungstabelle. Auf dem Testbogen stehen lange Reihen der Ziffern 1 bis 9 und die dazugehörigen Symbole (z. B. + für 9) sind in das Feld darunter einzutragen. Im Manual des d2 werden Korrelationen des KL-Werts von $r=.61$ mit dem ZVT bzw. von .52 mit dem ZS berichtet. Der d2-R bzw. seine elektronische Version (s. u.) wurden ebenfalls am ZS validiert (Schmidt-Atzert und Brickenkamp 2017, Studie 2 mit Studierenden und Studie 3 mit Seniorinnen und Senioren). Die Korrelation mit der Papierform fielen mit $r=.51$ bzw. .69 etwas höher aus als die mit der elektronischen Version ($r=.45$ bzw. .52).

Validität des Fehlerprozentwertes unklar

Für den Fehlerprozentwert liegen die Korrelationen mit den Fehlerwerten anderer Konzentrationstests überwiegend im Bereich von $r=.30$. Der Fehlerprozentwert ist deshalb sehr vorsichtig zu bewerten. Es ist sehr fraglich, ob dieser Kennwert eine generelle Fehlerneigung oder Sorgfalt erfasst – und ob die Annahme eines solchen Konstrukts überhaupt sinnvoll ist. Die Fehlerwerte anderer Konzentrationstests korrelieren trotz zum Teil beachtlicher Reliabilität ebenfalls überwiegend niedrig miteinander, sind also größtenteils testspezifisch (Goeters 1981).

Idealerweise wird ein Test auch an Leistungen im Alltag oder im Beruf validiert, was bei Aufmerksamkeits- und Konzentrationstest aber ausgesprochen schwierig ist (► Abschn. 3.2.2). Eine alternative Strategie besteht darin, alltagsnahe Konzentrationsleistungen unter Laborbedingungen zu erheben. Eine solche Aufgabe ist das Suchen von Rechtschreibfehlern in einem Text. In 2 Studien erhielten die Testpersonen eine 4- bzw. eine 2-seitige Reisebeschreibung, in die zahlreiche Fehler wie Buchstabendreher (z. B. Kindergarten) oder Auslassungen von Buchstaben (z. B. Bleistif) eingebaut waren. Die Aufgabe bestand darin, in begrenzter Zeit möglichst viele Fehler zu finden und zu markieren. Der KL-Wert des „alten“ d2 korrelierte um $r=.43$ mit der Leistung bei der Tippfehlersuche (Wilhelm 2005). In einer Replikation mit der elektronischen Version des d2-R (s. u.) fiel die entsprechende Korrelation mit $r=.32$ etwas niedriger aus (Kunz 2015).

Der d2 soll etwas anderes als Intelligenz messen (diskriminante Validität). Mit Intelligenz fanden sich überwiegend niedrige Korrelationen, die Korrelationen lagen meist unter $r=.30$. Zwischen dem Gesamtwert des I-S-T 70 und dem Tempowert des d2 bestand in einer Stichprobe von Auszubildenden ($N=1.560$), bei allerdings eingeschränkter Intelligenztestvarianz, ein Zusammenhang von $r=.14$. Damit wird unterstrichen, dass die Konzentrationsfähigkeit im Test d2 von Intelligenz abzugrenzen ist. Dass die Korrelationen nicht bei 0 liegen, kann mehrere Ursachen haben. Eine Erklärung ist, dass die Intelligenztestleistung (nicht die Intelligenz!) auch von der Konzentration der Testpersonen abhängt (Oswald und Hagen 1997).

Zur *Kriteriumsvalidität* liegen meist ältere Belege vor. Mit der Eignung zum Führen von Kraftfahrzeugen, operationalisiert über den Erfolg in der Führerscheinprüfung, korrelierte der Tempowert des d2 zu .52. Der d2 diskriminierte erfolgreich zwischen Gesunden und bestimmten psychiatrisch auffälligen Gruppen. Durch neurotoxische Stoffe belastete Personen erreichten niedrigere Werte im d2 als Kontrollpersonen. Einige Befunde sprechen dafür, dass die d2-Leistungen sensitiv für verschiedene Psychopharmaka sind.

Fazit Zusammenfassend lässt sich festhalten, dass für den fehlerkorrigierten Hauptkennwert des d2 (KL-Wert beim d2-R bzw. GZ-F beim d2, wobei GZ für Gesamtzahl und F für Fehler steht) weitgehend unabhängig vom Aufgabentyp der Referenztests überwiegend im Bereich von $r=.50$ bis .60 liegen. Die Fehlerwerte von Konzentrationstests korrelieren generell nur niedrig untereinander. Davon ist auch der d2 nicht ausgenommen. Die beiden Untersuchungen mit unterschiedlich langen Versionen des Pauli-Tests (s. o.) sprechen dafür, dass mit dem d2-R nicht nur die Fähigkeit gemessen wird, sich knapp 5 min lang zu konzentrieren, sondern genauso auch die Fähigkeit zur längfristigen Konzentration.

Normierung Der Test d2-R wurde Ende 2007 bis Mitte 2008 in 6 Bundesländern normiert. Die Gesamtstichprobe umfasst 4024 gültige Fälle. Es liegen Normen für Altersgruppen von 9–10 Jahren bis zu 40–60 Jahren vor, wobei die Altersgruppen der Kinder und Jugendlichen immer 2 Jahre umfassen. Die Altersgruppen bestehen aus 268–728 gültigen Fällen. Eine Überprüfung der Normen ist für das Jahr 2022 geplant.

Finden von Rechtschreibfehlern als Validitätskriterium

Diskriminante Validität: Intelligenz

Kriteriumsvalidität

2007/2008 an über 4000 Personen normiert

Bewährtes, ökonomisches, objektives reliable Verfahren

Umgang mit Übungseffekten und fehlende neue Studien zur Kriteriumsvalidität kritisiert

Frühere Computerversion nicht äquivalent

Anforderungen an eine sehr ähnliche Computerversion

Bewertung Zum Vorgänger, dem Test d2, liegen zahlreiche Rezensionen und wertende Darstellungen in diversen Buchbeiträgen vor, die sich zumeist auf ältere Auflagen beziehen (s. den Eintrag zum Aufmerksamkeits-Belastungs-Test d2 in der Datenbank PSYNDEXplus unter ► <https://www.psyndex.de/>).

Zum d2-R liegt eine Rezension nach dem Testbeurteilungssystem des Diagnistik- und Testkuratoriums vor (Daseking und Putz 2015). Dort ist zu lesen: „Der d2-R ist ein etabliertes und umfassend analysiertes Testverfahren zur Erfassung von Aufmerksamkeitsleistungen, das für Kinder und Erwachsene nutzbar ist“ (Daseking und Putz 2015, S. 324). Positiv werden die zeitlich ökonomische, objektive und reliable Erfassung der Konzentrationsleistung hervorgehoben.

Die Kritik richtet sich vor allem darauf, dass für die bekannten Übungseffekte keine Lösung präsentiert wird (Vorschlag: „Normen für geübte Testteilnehmer oder eine einheitliche Vorübung der Testanden“) und auf fehlende aktuelle Studien zur Kriteriumsvalidität.

Eine neue Computerversion Vom Test d2 existiert eine ältere Computerversion (Brickenkamp et al. 1996), die sich bereits dem Augenschein nach vom Original unterscheidet. Auf dem Bildschirm ist immer nur eine aus 9 Zeichen bestehende Zeile zu sehen. Die Buchstaben d und p sind mit Punkten statt mit Strichen versehen. Die Testleistung korreliert relativ niedrig mit der in der Papierversion. Für den Tempowert werden Korrelationen von $r=.63$ und $.62$ aus 2 Stichproben berichtet und für $F\% .42$ und $.31$. Diese Computerversion kann deshalb nicht als hinreichend äquivalent mit der Papier-und-Bleistift-Version angesehen werden (Merten 2000).

Die Vorteile einer (aktuellen) Computerversion liegen auf der Hand: Der nicht unerhebliche Aufwand für die Auswertung und die Dokumentation der Ergebnisse entfällt und die Auswertung ist nicht mehr fehleranfällig. Für Testentwicklerinnen und Testentwickler stellt die Anpassung an die Papier-und-Bleistift-Version eine große Herausforderung dar. Eine große Ähnlichkeit mit der Papier-und Bleistift Version ist erwünscht, damit man die vielen Befunde zur Validität auf den Computertest übertragen kann. Auch exemplarisch für die Probleme, die bei anderen Tests zu lösen sind, soll kurz das Vorgehen bei der „Übersetzung“ des d2-R in eine elektronische Version (Schmidt-Atzert und Brickenkamp 2017) beschrieben werden. Dabei wurden einige Herausforderungen identifiziert, wobei auf Selbstverständlichkeiten wie etwa die Beibehaltung wesentlicher Merkmale der Instruktion, der Übungsaufgaben und die Verwendung der gleichen Zeichen (auch anteilig gleich) und der gleichen Kennwerte hier nicht eingegangen wird.

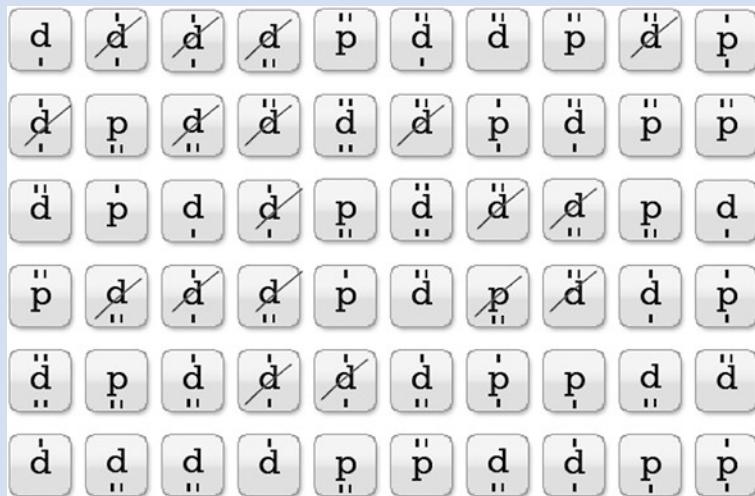
Herausforderungen bei der „Übersetzung“ in Computertests

Darstellung der Items auf dem Bildschirm: Auf dem Bildschirm könnten theoretisch zwar alle Items gleichzeitig dargestellt werden, aber die Anforderungen an die Feinmotorik beim Markieren mit der Maus oder beim gezielten Berühren eines Touchscreens wären immens. Die Lösung für den d2-R bestand darin, auf einer Bildschirmseite etwa so viele Items zu zeigen, wie sich in der Papierversion in einer Zeile befinden. Der Computertest besteht aus Bildschirmseiten anstatt aus Zeilen. Für die Bearbeitung der 60 Items steht so viel Zeit zur Verfügung wie ansonsten für eine Testzeile mit den 57 Items.

Markieren: In der Papierversion sehen die Testpersonen, welche Zeichen sie durchgestrichen haben. Diese fungieren vermutlich auch als Vorbild und helfen, die Regel (Buchstabe d, der 2 Striche hat, markieren) zu memorieren. Deshalb wurden Striche eingescannt, die sich beim Anklicken bzw. Berühren eines Feldes über ein Item legen – und zwar so, dass das Zeichen wie in der Papierversion noch zu erkennen ist. Die Bearbeitung soll analog zur Papierversion oben beginnend zeilenweise von links nach rechts erfolgen.

Durcharbeiten: Zum Durcharbeiten gehört auch, dass man sich seine Kräfte einteilen kann. In der Papierversion sieht die Testperson, wie viele Zeilen noch zu bearbeiten sind. In der neuen Computerversion ist am oberen Bildschirmrand dafür ein Fortschrittsbalken zu sehen. Mit der Bearbeitung einer Seite wird dieser ein Stück länger. Der Zeilenwechsel nach jeweils 20 s durch das Kommando „„nächste Zeile“ wird durch einen leeren Bildschirm für 1 s und anschließender Darbietung einer neuen Testseite ersetzt.

Konkrete Umsetzung in der Computerversion



Testseite des elektronischen d2-R, die bereits bearbeitet wurde (© Hogrefe). Erläuterung: Beim Anklicken (oder Berühren auf einem Touchscreen) eines Feldes wird das Zeichen durchgestrichen. Durch erneutes Anklicken kann der Strich im Falle eines selbst bemerkten Fehlers wieder entfernt werden.

Eigene Normierung erforderlich

Trotz aller Bemühungen um eine große Ähnlichkeit mit der Papierversion kam keine perfekte Äquivalenz zustande. So erweisen sich die Mittelwerte und Streuungen der Testkennwerte als unterschiedlich. Beispielsweise lag der Mittelwerte für KL in der Altersgruppe von 20 bis 39 Jahren bei 158–161 Punkten (= Standardwert 100). In der elektronischen Version erreichten die gleichaltrigen Personen durchschnittlich 184 Punkte. Nach den Normtabellen der Papierversion würde dies einem Standardwert von 106 entsprechen. Dadurch wurde eine eigene Normierung nötig (s. dazu den Steckbrief zum d2-R).

Die elektronische und die Papierversion korrelieren sehr hoch miteinander (Studie 2a im Manual: $r = .83/.84$ für KL, $.78/.79$ für BZO und $.74/.56$ für F%; 1. Wert = elektronische, 2. Wert Papierversion zuerst). Die interne Konsistenz (Cronbachs α) kann für die meisten Altersgruppen wie bei der Papierversion als sehr hoch bezeichnet werden. Die Retest-Reliabilität nach 16 Tagen ist hoch (KL: $r = .93$, BZO: $.90$, F%: $.75$). In mehreren Studien wurde keine bedeutenden Unterschiede der Korrelationen mit konvergenten und diskriminanten Verfahren gefunden. Meist fielen die Korrelationen der Papierversion mit anderen Papierversionen etwas höher aus als die der elektronischen Version. Eine kurze Übersicht zu den Validitätsbefunden gibt der Steckbrief.

Ein Vorteil vieler computerbasierter Tests, so auch des d2-R, ist der automatisch erzeugte Report. Damit wird die Auswertungsobjektivität sichergestellt. Beim d2-R und einigen anderen Tests ist der Report ebenfalls von Vorteil für die Interpretationsobjektivität, weil nach festen Regeln Aussagen über die erfassten Merkmale und deren Ausprägung erzeugt werden.

Weitere Konzentrationstests

■ Tests mit Suchaufgaben

Unterschiedliche Arten von Suchaufgaben

Tests mit Suchaufgaben stellen die größte Untergruppe der Konzentrations- tests dar. Nach dem Prinzip, kritische Reize unter ähnlichen Reizen zu suchen, wurden viele Tests konstruiert. Sie unterscheiden sich vom Test d2-R vor allem darin, welche Art von Zeichen verwendet wird. Exemplarisch werden das Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2) und dessen computerbasierte Variante, der Frankfurter Adaptiver Konzentrationsleistungs-Test II (FAKT-II), sowie der Konzentrationstest für 3. und 4. Klassen – Revision (KT 3–4 R) und der Konzentrationstest für Kinder (KoKi) kurz dargestellt. Unterschiede zum d2-R bestehen vorwiegend hinsichtlich der Zielgruppe, dem Markierungsprinzip, den berechneten Kennwerten und der Testdauer.

FAIR-2: Kreise und Quadrate mit Punkten

Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2) Im FAIR-2 von Moosbrugger und Oehlschlägel (2011) werden als Testitems Kreise und Quadrate, in denen sich jeweils 2 oder 2 Punkte befinden, verwendet. Alle Items sind zusätzlich von einem Kreis umgeben. Insgesamt kommen in dem Test 8 verschiedene Items vor, von denen 2 als Zielobjekte fungieren (Abb. 3.5). In Form A sind alle Items, die in Abb. 3.5 in den Zellen links oben und rechts unten gezeigt werden, zu markieren. Die anderen Items stellen Distraktoren dar. In Form B ist es umgekehrt. Der Testbogen enthält 320 Items, die in 16 Zeilen à 20 Items angeordnet sind. Die reine Testdauer beträgt 6 min. In Testform B dienen alle Quadrate mit 2 Punkten und alle Kreise mit 3 Punkten als Zielitems.

Das FAIR-2 verwendet ein „vollständiges Markierungsprinzip“. Es besteht darin, dass die Testpersonen ihre Urteile Zeile für Zeile von links nach rechts in Gestalt einer durchgehenden Linie abgeben: Bei den Distraktoren ist die Linie nur unter den Zeichen entlangzuführen, bei den Zielitems hingegen ist die Linie zackenförmig hochzuziehen. Der Vorteil dieses Markierungsprinzips besteht darin, dass man bei der Testauswertung genau sieht, bis zu welchem Item die Testpersonen den Test bearbeitet haben. Werden

nur Zielitems durchgestrichen (wie beim d2), so ist bei der Testauswertung nur bekannt, welches Zielitem die Testpersonen zuletzt durchgestrichen haben, nicht aber, ob sie danach bereits weitere Distraktoren angesehen und zu Recht als solche erkannt (d. h. nicht angestrichen) haben.

Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II) Die Items des FAIR bzw. FAIR-2 finden auch in dem computerbasierten adaptiven FAKT-II von Moosbrugger und Goldhammer (2007) Verwendung. Unter Benutzung eines computerbasierten Algorithmus wird das Schwierigkeitsniveau der Items „maßgeschneidert“ an das individuelle Konzentrationsvermögen des Einzelnen dadurch angepasst, dass die Vorgabe der Items umso rascher erfolgt, je höher die Konzentrationsleistung liegt. Intendiert wird damit eine in etwa gleiche Beanspruchung auf den interindividuell unterschiedlichen Leistungsstufen. Als Leistungsscore gilt jene Darbietungsgeschwindigkeit (bzw. die darin erzielte Zahl bearbeiteter Items), bei der in etwa gleich viele richtige wie falsche Lösungen geliefert werden.

Bei der Anwendung kann zwischen 3 Darbietungsvarianten gewählt werden: Einzelitem, 10 Items simultan (Darbietungszeit adaptiv oder nicht). Zur Beurteilung des Leistungsverlaufs kann die Testlänge in 6-min-Schritten bis auf 30 min ausgedehnt werden. In diesem Fall werden Konzentrationsergebnisse für jeden 6-min-Abschnitt berechnet. Trotz Verwendung der gleichen Items wie beim FAIR-2 ist der FAKT-II als eigenständiger Test zu werten. Die Durchführungsbedingungen unterscheiden sich deutlich. So sind weniger Items simultan zu sehen, und beim adaptiven Testen ist die Itemdarbietungszeit so kurz, dass nur jedes 2. Item richtig gelöst werden kann. Im Manual werden Korrelationen von $r = .45$ bis $.55$ mit dem FAIR berichtet.

Sowohl das FAIR-2 wie auch der FAKT-II zeichnen sich wie der Test d2-R durch eine hohe Reliabilität der meisten Kennwerte aus. Als Validitätsbelege werden u. a. überwiegend moderate Korrelationen mit anderen Konzentrationstests und niedrige Korrelationen mit Intelligenztests vorgelegt.

FAKT-II als Computertest mit den Items des FAIR-2

Die Tests sind hoch reliabel

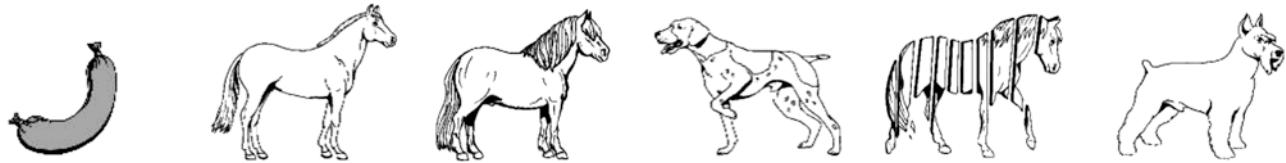
Konzentrationstest für 3. und 4. Klassen – Revision (KT 3–4 R) Der KT 3–4 R von Nell et al. (2004) wurde, wie an der Testbezeichnung schon erkennbar, für Kinder entwickelt und zudem nur für einen engen Altersbereich. Als Items dienen Abbildungen von Würfeln, die zeichnerisch so dargestellt sind, dass man immer 3 Flächen mit jeweils 1–6 Punkten sehen kann. Jedes Item ist mit den 4 Musterwürfeln zu vergleichen. Wenn der Würfel mit einem der 4 Muster identisch ist, wird er durchgestrichen. Die reine Bearbeitungszeit beträgt 20 min. Die Testleiterin oder der Testleiter fordert die Kinder alle 5 min auf, einen Strich als Zeitmarke unter den gerade bearbeiteten Würfel zu setzen. Die Musterwürfel ändern sich auf jeder Testseite, um den Einfluss der Merkfähigkeit auf die Testleistung zu minimieren. Ein ähnliches Aufgabenprinzip ist beim Cognitron (Schuhfried 2020a) realisiert. In diesem computerbasierten Konzentrationstest lautet die Frage, ob die gezeigte Figur mit einer der 4 Vergleichsfiguren identisch ist.

KT 3–4 R: Kindertest mit Würfeln als Items

Konzentrationstest für Kinder (KoKi) Der KoKi von Schmidt-Atzert und Funsch (2021) ist eng an das Testprinzip des d2-R angelehnt. Die Items variieren ebenfalls in Bezug auf 2 Dimensionen: Ein Item ist entweder ein Hund oder ein Pferd; Das Tier schaut entweder in Richtung einer Wurst am Anfang oder Ende der Zeile oder nicht. Zielobjekte sind Hunde, die zur Wurst schauen. Die Items sind zeilenweise angeordnet und werden eines nach dem anderen im selbstbestimmten Tempo bearbeitet, wobei nur die Zielobjekte zu markieren sind. Die Kennwerte entsprechen denen des d2-R.

KoKi: Kindertest mit Hunden und Pferden als Items

Eine Besonderheit stellt ein Subtest mit ablenkenden Reizen dar (Abb. 3.6). Verwendet werden hierzu „verzauberte“ Pferde, also Pferde die



3

Abb. 3.6 Items aus dem KoKi (Auszug). In jeder Zeile befinden sich links oder rechts eine Wurst sowie 10 Hunde und Pferde. Alle Hunde, die in Richtung Wurst schauen, sind durchzustreichen (hier das mittlere Tier). Unter der Bedingung „mit Ablenkung“ sind auch verzauberte Pferde zu sehen, die die Aufmerksamkeit auf sich ziehen und die Testleistung in der Regel mindern. (Aus Schmidt-Atzert und Funsch 2021, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

ungewöhnlich dargestellt sind. Getestete Kinder bekommen dazu die Information, ein Zauberer wolle sie stören und habe dazu „komische Pferde“ gezaubert. Die Kinder sollen sich laut Instruktion davon aber nicht stören lassen. Die verzauberten Pferde sind im Vergleich zu normalen Pferden leichter als Distraktoren zu erkennen, weil sie sich besonders stark von Hunden unterscheiden. Dennoch erschweren sie die Aufgabe, weil sie Aufmerksamkeit auf sich ziehen. Durch einen individuellen Vergleich der Testleistungen auf Seiten mit normalen und verzauberten Pferden kann ein Kennwert für Ablenkbarkeit bestimmt werden.

Ein ganz ähnlicher Aufgabentyp, bei dem zwar auch nach einem bestimmten Zeichen zu suchen ist, aber dieses Zeichen von Aufgabe zu Aufgabe wechselt, wird beim Differentiellen Konzentrationstest für Kinder (DKT-K) von Funsch und Arias Martin (2017) verwendet. Dort ist ein Ausgangsbild (z. B. eine bunte Zeichnung eines Autos) zu sehen und die Kinder müssen prüfen, ob sich dieses Bild unter 3 Vergleichsbildern befindet oder nicht (► Abb. 3.7). Welches Bild zu suchen ist, ändert sich mit jedem Item. Die vorzunehmenden Vergleiche sind unterschiedlich schwierig und komplex, was sich in den 4 Subtests widerspiegelt.

Figurale Reize vergleichen statt suchen

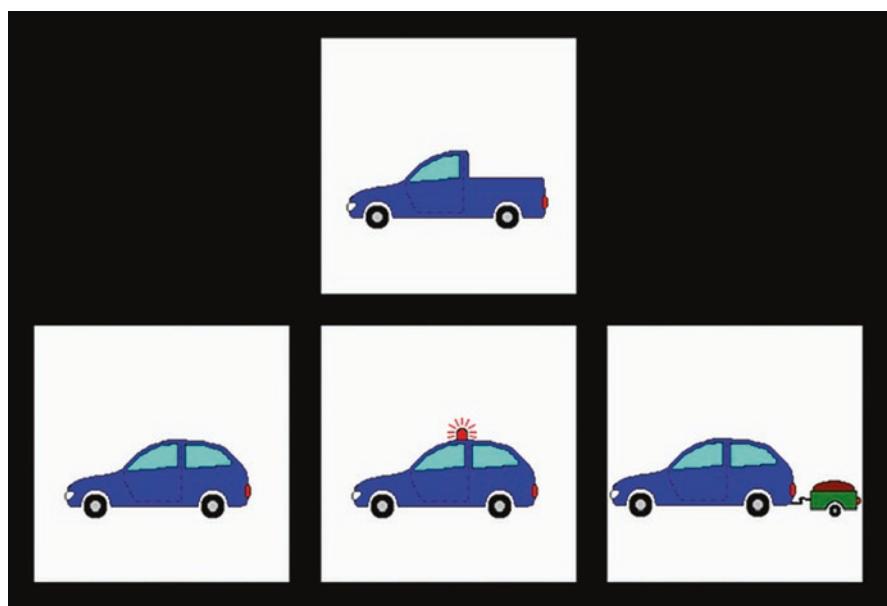


Abb. 3.7 Item aus dem DKT-K. (Aus Funsch und Arias Martin 2017, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

■ Konzentrationstests mit Rechenaufgaben

Einige Konzentrationstests verlangen das Prüfen oder Lösen einfacher Rechenaufgaben. Solche „Rechentests“ setzen stillschweigend voraus, dass alle Testpersonen einen etwa gleich hohen Automatisierungsgrad hinsichtlich der erforderlichen Rechenfertigkeiten erreicht haben und sich nur in ihrer Konzentrationsfähigkeit unterscheiden – eine Annahme, die angesichts sehr unterschiedlicher schulischer Biografien als problematisch angesehen werden muss. Die Resultate des KLT-R (s. u.) korrelieren dementsprechend auch mit der Mathematiknote. Scheinbar unterdurchschnittliche Konzentrationsleistungen können deshalb auch auf geringen numerischen Fertigkeiten beruhen. In der faktorenanalytischen Studie von Schmidt-Atzert et al. (2006; □ Tab. 3.6) zeigten zwei Rechenkonzentrationstests in der Tat hohe Ladungen auf einem eigenen numerischen Faktor.

Testleistung hängt von Rechenfertigkeit ab

Konzentrations-Leistungs-Test (KLT) Der KLT wurde von Düker und Lienert (1965) entwickelt und sollte Koordination verlangen. Darunter verstehen die beiden Autoren „das zu einer Gesamttätigkeit geordnete Zusammenwirken der Einzeltätigkeiten, die zur Erreichung eines bestimmten Zweckes erforderlich sind“ (Düker und Lienert 1965, S. 3). Um diese Koordination bewerkstelligen zu können, bedarf es laut der Autoren der Konzentration. Im Falle des KLT wird diese mit Aufgaben gemessen, die ihrerseits interne Koordinationsprozesse verlangen. Als zu koordinierende Einzeltätigkeiten können beim KLT „Auffassen“, „Rechnen“, „Merken“, „Regelabruf“ und „Entscheiden“ spezifiziert werden.

Konzentration als Koordinationsleistung

Lukesch und Mayrhofer haben den Test leicht modifiziert und neu normiert (Düker et al. 2001). Der KLT-R liegt jeweils in den Parallelformen A und B und in 2 unterschiedlichen Schwierigkeitsstufen vor. Jede Version enthält 180 Aufgaben, die sich auf 9 Blöcke mit je 20 Aufgaben verteilen. Die leichtere Version KLT-R 4–6 ist für die 4., 5. und 6. Schulklasse bestimmt, die Version KLT-R 6–13 für die 6.–13. Schulklasse. Jede der Formen des KLT enthält Aufgaben des folgenden Typs:

Der KLT wurde modifiziert und neu normiert

Beispielaufgaben aus dem KLT

Beispiel A: $8 + 9 - 2$

$5 - 4 + 3$

Beispiel B: $3 + 6 - 8$

$9 + 1 + 7$

Die Testpersonen müssen zunächst pro Zeile die Ergebnisse ausrechnen und jeweils im Kopf behalten (Beispiel A: 15 bzw. 4; Beispiel B: 1 bzw. 17). Im Anschluss daran ist mit den Teilergebnissen nach unterschiedlichen Vorschriften zu verfahren:

Rechnen, Merken, Vergleichen

- KLT-R 4–6: Das kleinere Zwischenergebnis ist vom größeren zu subtrahieren. Im Beispiel A lautet die Lösung also 11 ($15 - 4 = 11$) und im Beispiel B 16 ($17 - 1 = 16$).
- KLT-R 6–13: Falls das 1. Zwischenergebnis größer ist als das 2. (wie im Beispiel A), ist die Differenz zu bilden (Beispiel A: $15 - 4 = 11$). Falls das 1. Zwischenergebnis kleiner ist als das 2. (wie im Beispiel B), sind beide zu addieren (Beispiel B: $17 + 1 = 18$). Diese Aufgabenstellung wurde vom „alten“ KLT übernommen. Nur das Endergebnis ist in das Kästchen neben den Aufgaben einzutragen.

Konzentrations- oder Rechentest?

Weitere Rechenkonzentrationstests mit einfachen Aufgaben

Schritte bei der Lösung eines Testitems

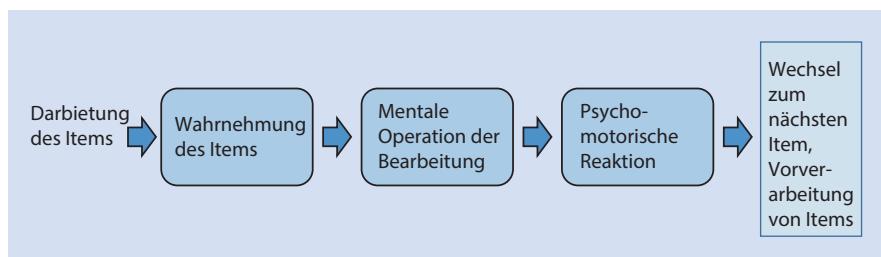
Die Validitätsbelege (Kennwert hier jeweils Richtige) deuten darauf hin, dass der Test eher die Rechenfertigkeit und die Intelligenz als die Konzentration erfasst. So fand sich zwischen dem KLT-R 6–13 eine Korrelation von nur $r = .27$ mit dem „alten“ d2, während die Korrelationen mit einem standardisierten Rechentest für die 4., 5. und 6. Klasse .57, .33 und .53 betragen.

Ein Rechenkonzentrationstest muss nicht zwangsläufig so schwierig und komplex sein wie der KLT-R. Der Revisions-Test (Marschner 1980), der Pauli-Test (Arnold 1975; Christiansen 1983) und die numerischen Aufgaben zur Bearbeitungsgeschwindigkeit im Berliner Intelligenzstrukturtest (Jäger et al. 1997) belegen, dass man die Konzentrationsfähigkeit auch mit sehr einfachen Rechenaufgaben messen kann. Diese Tests sind jedoch älteren Datums, sodass sie heute nur in der Forschung eingesetzt werden, wo (aktuelle) Normen zumeist irrelevant sind.

Exkurs: Ein Prozessmodell der Itembearbeitung in Konzentrationstests

Ein oft propagierter, aber selten angewandter Ansatz ist die Analyse der psychologischen Prozesse, die zu einer Antwort in einem Konzentrationstest führt (► Abschn. 2.6.3.2). Blotenberg und Schmidt-Atzert (2019a, c) verfolgen einen solchen Ansatz und nehmen an, dass mehrere Schritte bei der Lösung eines einzelnen Items nacheinander ablaufen und jeweils Zeit konsumieren. Hinzu kommen 2 Prozesse, die aber auf das Durcharbeiten von simultan dargebotenen Items beschränkt sind (Itemwechsel und Vorverarbeitung; □ Abb. 3.8).

Ein Item wird zunächst wahrgenommen. Nachdem es nun sozusagen „im Kopf der Testperson angelangt ist“, wird es gelöst. Die dazu erforderliche mentale Operation ist testspezifisch. Beispielsweise wird bei Rechenkonzentrationstests eine Rechenoperation durchgeführt und bei einem Test mit Selektionsaufgabe geprüft, ob ein Item festgelegte Merkmale aufweist (z. B. ein d mit 2 Strichen vorliegt). Liegt die Lösung vor, muss sie irgendwie angezeigt werden. Je nach Test muss eine Zahl aufgeschrieben, ein Zeichen durchgestrichen oder am Bildschirm angeklickt werden. Es handelt sich um eine psychomotorische Reaktion, die angebahnt (geplant) und ausgeführt werden. Bei Tests, die simultan mehrere Items präsentieren, die nacheinander (self-paced) zu bearbeiten sind, können zwischen der Bearbeitung von 2 Items 2 Prozesse ablaufen: Die Testperson wendet sich vom bearbeiteten Item ab und dem nächsten zu. Das kann mehr oder weniger viel Zeit kosten, die man interpretationsfrei als Itemwechselkosten verbuchen oder inhaltlich als Erholungspause interpretieren kann. Ähnlich wie beim Lesen von Texten ist es vielleicht möglich, nachfolgende Items unbeabsichtigt vorzuverarbeiten, was die Bearbeitungszeit verkürzen könnte.



□ Abb. 3.8 Vermuteter Prozess bei der Bearbeitung von Aufmerksamkeits- und Konzentrationstestitems. (Nach Blotenberg und Schmidt-Atzert 2019a, c)

So weit die Theorie. Die einzelnen Verarbeitungsschritte können nachgewiesen werden, indem die jeweils dafür benötigte Zeit gemessen oder geschätzt und erfolgreich zur Prädiktion der Leistung in 3 verschiedenen Konzentrationstests (numerisch, figural und verbal) verwendet wird (s. u.):

- Die individuelle *Wahrnehmungszeit* wurde mit dem Inspection-Time-Paradigma exakt gemessen: Auf dem Bildschirm erscheint eine dem griechischen Zeichen Pi ähnliche Figur mit 2 unterschiedlich langen Schenkeln. Die Testperson gibt ohne Zeitdruck an, welcher Schenkel länger war. Die Darbietungszeit wird in einem adaptiven Prozess mit vielen Items so variiert, dass die minimale Darbietungszeit für das sichere Erkennen ermittelt werden kann. Im Durchschnitt lag die Wahrnehmungszeit bei 63 ms.
- Die *psychomotorische Reaktionszeit* wurde als Reaktionszeit auf einen einfachen Reiz ermittelt ($M = 243$ ms).
- Zur Schätzung der für die *mentale Operation benötigen Zeit* wurden Items aus dem d2-R auf dem Bildschirm dargeboten und die Testpersonen gaben an, ob ein Zielobjekt (d mit 2 Strichen) dabei ist oder nicht. Die für die Wahrnehmung und die Antwort benötigte Zeit wurde rechnerisch herausgenommen (die „reine“ Bearbeitungszeit pro Item, betrug durchschnittlich 1016 ms).
- Zur Schätzung der für den *Itemwechsel* benötigten Zeit wurden die Items zur mentalen Operation mit unterschiedlich langen Zwischenpausen dargeboten. Wie zu erwarten, war die Reaktionszeit auf ein Item bei einer Darbietung ohne Pause am längsten und nahm mit der Pausenlänge zunächst stark und dann immer weniger ab. Das individuelle Optimum für die Pausenlänge lag im Durchschnitt bei 614 ms.

Beschreibung der Verarbeitungselemente

Wichtig ist hier die Feststellung, dass alle genannten Zeiten Mittelwerte sind und die individuellen Werte eine Streuung aufwiesen.

Ziel der beiden Studien bei Blotenberg und Schmidt-Atzert (2019c) war es, die Leistung in Konzentrationstests durch die einzelnen Verarbeitungsschritte zu erklären. Die Testpersonen bearbeiten ein figuralen (d2-R), einen numerischem (Revisions-Test; Additionsaufgaben mit einstelligen Zahlen wie $5 + 2 = 8$ sind als richtig oder falsch zu bewerten) und einen verbalen Konzentrationstest (3 Subtests aus dem BIS-Test zur verbalen Bearbeitungsgeschwindigkeit; z. B. sind beim „Klassifizieren von Wörtern“ in 30 s in einer Liste von 100 Wörtern möglichst viele Pflanzennamen durchzustreichen). Aus den 3 Tests wurde ein Gesamtwert für die Konzentrationsfähigkeit errechnet und in einem Strukturgleichungsmodell durch die oben beschriebenen Verarbeitungselemente vorhergesagt. Die Varianz der Konzentrationsfähigkeit konnte in Studie 1 zu 74 % aufgeklärt werden. Die Zeit für die Wahrnehmung und für die mentale Operation wiesen mit je $r = .58$ die höchsten Korrelationen mit der Konzentrationsfähigkeit auf; die Beiträge der Reaktionsschnelligkeit ($r = .16$) und der Itemwechselzeit ($r = -.06$) waren klein und nicht signifikant.

Für Studie 2 wurden die Aufgaben zur mentalen Operation etwas verändert. Die Testpersonen mussten immer nur auf 1 Item reagieren (d mit 2 Strichen oder nicht). Diese erschienen ohne oder mit Pausen (je 500 ms) auf dem Monitor. Aus der Differenz wurde die Zeit für den Itemwechsel geschätzt. Trotz dieser Änderungen des methodischen Vorgehens konnten die Ergebnisse

Bestimmung und Vorhersage der Konzentrationsfähigkeit anhand der Verarbeitungselemente

der 1. Studie im Wesentlichen repliziert werden: Die Varianz der Konzentrationstestleistung konnte zu 68 % aufgeklärt werden. Die Wahrnehmungsgeschwindigkeit korrelierte um $r=.46$ und die Zeit für mentale Operation um $r=.65$ mit der Konzentrationsfähigkeit (für die Reaktionsschnelligkeit ergab sich $r=.20$ und für die Itemwechselzeit $r=.06$, beide waren damit nicht signifikant).

Damit lässt sich festhalten, dass sich das Erklärungsmodell (Abb. 3.8) in Bezug auf 2 Punkte sehr gut bewährt hat: Die Wahrnehmungsgeschwindigkeit und die für die mentale Operation benötigte Zeit tragen zur Erklärung von Konzentrationstestleistungen wesentlich bei. Die reine Reaktionszeit spielt allenfalls eine kleine Rolle. Die Annahme, dass die benötigte „Pausenzeit“ zwischen der Bearbeitung von 2 Items zur Varianzaufklärung beträgt, konnte trotz unterschiedlicher Operationalisierungen in keiner der beiden Studien bestätigt werden.

Das Modell wurde bei Blotenberg und Schmidt-Atzert (2019a) um die Vorverarbeitung von Items erweitert. Verbessert sich die Testleistung, wenn auch schon die nächsten zu bearbeitenden Items sichtbar sind? Konkret war nun auch eine Zeile mit 10 Items zu sehen. Das zu bearbeitende Item war mit einem Pfeil gekennzeichnet. Dieser rückte nach jeder Antwort eine Position vor. Eine gute Kontrollbedingung ist die Darbietung von 3 Items, von denen nur das mittlere, ebenfalls mit einem Pfeil markierte Item zu bearbeiten war. Die durchschnittliche Antwortzeit verkürzte sich von 664 ms (Einzeldarbietung) auf 504 ms (Darbietung in einer Reihe; der Unterschied ist signifikant). Die Testpersonen profitierten aber unterschiedlich stark von der reihenweisen Darbietung. Die individuelle Verkürzung der Antwortzeit korrelierte zu $r=.50$ mit der Leistung im d2-R (ebenfalls computerbasiert mit zeilenweiser Bearbeitung der Items), aber auch signifikant mit der Leistung im Revisions- und im BIS-Test ($r=.35$ bzw. $.26$).

Bei vielen Konzentrationstests werden Items simultan gezeigt und nacheinander bearbeitet. Sie erlauben also eine Vorverarbeitung. Die Leistung in solchen Tests (damit ist nicht die gemessene Fähigkeit gemeint, sondern das Testergebnis!) kann also zum Teil durch die individuelle Effizienz der Vorverarbeitung erklärt werden. Eine Vorverarbeitung von Items ist jedoch nicht bei allen Tests möglich: Bei einem Test wie dem Cognitrone erscheint auf dem Bildschirm immer nur ein einzelnes Item. Deshalb müssen wir die vorausschauende Verarbeitung als ein testspezifisches Merkmal ansehen, das bei den meisten, aber nicht allen Konzentrationstests eine Rolle spielt. Deshalb sollte die vorausschauende Verarbeitung von Items nicht als eine Komponente der Konzentrationsfähigkeit angesehen werden, sondern als eine testspezifische „technische“ Variable.

Die hier vorgenommene Zerlegung der Konzentrationstestleistung hilft bei der Erklärung von Validitätsbefunden und bei der Erklärung eines Phänomens, nämlich den Übungsgewinnen bei Testwiederholung. Blotenberg und Schmidt-Atzert (2019b) untersuchten, wie groß die Übungsgewinne bei den einzelnen Komponenten der Konzentrationstestleistung (nach 30 min) sind. Den Ergebnissen zufolge ist der Übungsgewinn für die mentale Operation (bei d2-Testaufgaben) mit $d=1,71$ (durchschnittlich 97 ms schneller) mit Abstand am höchsten, gefolgt von der einfachen Reaktionszeit mit $d=0,50$ (12 ms schneller) und der Wahrnehmungszeit mit $d=0,32$ (6 ms schneller). Die vorausschauende Verarbeitung und die Itemwechselzeiten verbesserten sich durch Testwiederholung dagegen nicht signifikant.

Vorverarbeitung von Items relevant

Vorverarbeitung nur ein „technisches“ Testmerkmal

Erklärung von Übungseffekten

Die hier berichteten Ergebnisse sind relevant für die Beurteilung der Validität von Konzentrationstests. Versteht man Konzentrationsfähigkeit im Wesentlichen als die Fähigkeit, eine Information (einen Reiz) schnell wahrnehmen und eine geforderte einfache mentale Operation schnell durchführen zu können, so sollten Konzentrationstests, die keine weiteren Anforderungen an Testpersonen stellen, hoch miteinander korrelieren. Viele Konzentrationstests unterscheiden sich aber in „technischer“ Hinsicht. Ein wichtiger Unterschied ist die Art der mentalen Operation zur Lösung der Items. Damit kommen andere Fähigkeiten bzw. Fertigkeiten wie die Rechenfertigkeit, der Wortschatz oder die Fähigkeit, figurale Reize unterscheiden zu können, ins Spiel. Ferner hängt die Leistung bei Tests, in denen simultan gezeigte Items nacheinander zu bearbeiten sind, von einer vorausschauenden Verarbeitung der Items ab. Eine fortwährende Bearbeitung von Items ist auch ohne diese Anforderung realisierbar. Es genügt, in einem Computertest Items mit Pausen von z. B. 500 ms einzeln nacheinander darzubieten. Damit wäre keine vorausschauende Verarbeitung möglich. Auch in der motorischen Reaktion können deutliche Unterschiede bestehen. Es gibt Konzentrationstest, bei denen jedes der Items zu beantworten ist, während bei anderen nur die „richtigen“ zu markieren sind. In manchen Tests kommen kaum Fehler vor, in anderen dagegen viele. Die Fehler werden unterschiedlich verrechnet, indem etwa nur die richtigen Antworten zählen oder Minuspunkte für Fehler vergeben werden. Mühelos lassen sich weitere Unterscheidungsmerkmale finden. Mit anderen Worten: Jeder Test misst die Konzentrationsfähigkeit etwas anders.

Würde man einen Konzentrationstest an einem beliebig anderen validieren, vielleicht weil dieser sehr bekannt ist, wäre das Ergebnis mehr oder wenig beliebig. Es gibt keinen Test, der als „Goldstandard“ fungieren könnte. Deshalb bietet sich eine andere Strategie an: der Einsatz mehrerer Konzentrationstests, die konzeptuell wichtige Anforderungen erfüllen und sich in „technischer“ Hinsicht voneinander unterscheiden. Idealerweise ist jeder Test mit einem anderen „technischen“ Merkmal konfundiert. Das Gemeinsame dieser Testbatterie kann als Konzentrationsfaktor und damit als Operationalisierung der Konzentrationsfähigkeit verstanden werden. Die Korrelation eines Tests mit diesem Konzentrationsfaktor wäre die bestmögliche Schätzung der Validität eines Konzentrationstests.

Gemeinsamkeit: schnelle
Verarbeitung von Informationen

Goldstandard

3.2.3 Intelligenztests

Intelligenztests sind vermutlich die erfolgreichsten Verfahren in der psychologischen Diagnostik. Sie sind erfolgreich, weil sie in wichtigen Lebensbereichen erstaunlich gute Vorhersagen erlauben und zudem sehr zeitstabile Kennwerte liefern. (► Abschn. 2.3.3).

Sehr erfolgreiche Verfahren

Die herausragende Bedeutung der Intelligenz wird durch die monumentale Terman-Studie eindrucksvoll belegt: In den Jahren 1921 und 1922 wurden rund 1400 Kinder nach ihren Ergebnissen im Stanford-Binet-Staffeltest ausgewählt. Ihr IQ musste mindestens 135 betragen; damit gehörten sie

Intelligenz ist wichtig für Erfolg im Leben

Intelligenztests werden häufig eingesetzt

zu dem oberen Prozent in der Intelligenzverteilung. Bei der genauen Verfolgung ihres Lebenswegs über viele Jahre hinweg zeigte sich, dass sie in fast jeder Hinsicht erfolgreicher und zufriedener waren als die Durchschnittsbevölkerung. Beispielsweise gehörten im Jahre 1960 von den männlichen Teilnehmern 47 % in die obere von 5 Berufsgruppen; dazu zählten Rechtsanwälte und Richter (10 % der 738 hochbegabten Männer mit Beschäftigung), Ingenieure (8 %), Universitätsmitglieder (7 %), Naturwissenschaftler (6 %) und Ärzte (5 %; Oden 1968). Erstaunlich ist nicht nur, dass die Intelligenz ein derart starker Prädiktor für Erfolg ist, sondern auch, dass sie langfristige Prognosen ermöglicht.

Viele in der Praxis tätige Psychologinnen und Psychologen setzen Intelligenztests ein, wie Umfragen zeigen (vgl. ▶ Abschn. 3.2.1; □ Tab. 3.3). Die Nennungshäufigkeiten sind jedoch nicht gleichzusetzen mit der Anzahl der Anwendungen. Sie besagen lediglich, wie viele der Befragten einen bestimmten Test überhaupt verwenden. Zumindest für einen Anwendungsbereich liegen auch Erkenntnisse über die Einsatzhäufigkeit vor: Bölte et al. (2000) befragten Psychologinnen und Psychologen in ambulanten und stationären kinder- und jugendpsychiatrischen Einrichtungen. Erstens zeigte sich, dass Intelligenztests die Liste der dort eingesetzten Testverfahren anführen. Zweitens wurde deutlich, dass sie diese Tests auch sehr oft einsetzen: 74 % der Befragten gaben an, Intelligenztests „immer“ einzusetzen, die restlichen 26 % antworteten „oft“.

3.2.3.1 Systematik der Intelligenztests

Es gibt viele Auffassungen darüber, was man unter Intelligenz versteht. Dies zeigt sich u. a. in der Vielzahl von Intelligenzmodellen, die sich zum Teil erheblich voneinander unterscheiden. Dementsprechend unterschiedlich sind die Intelligenztests, die zur Verfügung stehen. Eine Übersicht über verschiedene Intelligenzmodelle und zugehörige Test geben Süß und Beauducel (2011; siehe auch ▶ Abschn. 3.2.4.1 zur Darstellung des CHC-Modells). Darüber hinaus unterscheiden sich Intelligenztests in vielen weiteren Aspekten voneinander. Zusammenfassend sind die Kriterien in □ Tab. 3.7 aufgeführt. Sie werden im Folgenden kurz erläutert.

Angaben zur konzeptuellen Einordnung erforderlich

Messintention Grundsätzlich sollten Testanwenderinnen und -anwender darüber informiert werden, welche Art von Intelligenz das Verfahren misst. Da es „die“ Intelligenz nicht gibt, soll das Manual präzise Angaben zur konzeptuellen Einordnung des Tests enthalten – d. h., Nutzerinnen und Nutzer sollen über die Messintention informiert werden. Nicht allen Tests liegt explizit ein bestimmtes Intelligenzmodell zugrunde.

□ Tab. 3.7 Wichtige Merkmale zur Einordnung von Intelligenztests

Betreff	Testmerkmal
Messintention	<ul style="list-style-type: none"> – Allgemeine Intelligenz (g) oder eine bestimmte Intelligenzkomponente – Ein Globalmaß oder (auch) Intelligenzstruktur bzw. mehrere Komponenten – Intelligenz sprachfrei/kulturfair oder bildungsabhängig messen
Durchführungsbedingungen	<ul style="list-style-type: none"> – Einzel- oder Gruppentestung – Speed- oder Powertest – Papier-und-Bleistift-Test oder Computertest – Dauer der Testdurchführung
Zielgruppe	<ul style="list-style-type: none"> – Bestimmter Altersbereich – Bestimmter Intelligenzbereich – Gesamtbevölkerung oder spezielle Personengruppe

Für viele Fragestellungen ist es nützlich, ein Maß für die Allgemeine Intelligenz zu erheben. Einige Tests sind dazu auch geeignet. Grundsätzlich sind hier folgende Ansätze zu erkennen: Einerseits wird versucht, den „Kernbereich“ der Intelligenz, das schlussfolgernde Denken (Reasoning), zu erfassen. Diese Konzeption liegt beispielsweise den Standard Progressive Matrices (SPM) zugrunde. Andererseits strebt man eine „breite“ Messung mit Aufgabengruppen zu verschiedenen Bereichen (Komponenten) der Intelligenz an. Die „Breite“ kann dabei unterschiedlich gefüllt werden, d. h., die Auswahl der Intelligenzkomponenten variiert von Test zu Test.

Die „breiten“ Tests liefern zusätzlich Informationen über mehrere Intelligenzkomponenten (beispielsweise sprachliches, rechnerisches und räumliches Denken). Sie firmieren als Strukturtests, wenn mehrere Intelligenzkomponenten erfasst und Unterschiede zwischen den Untertests interpretiert werden können; häufig findet in diesen Fällen eine Darstellung der Untertestleistungen in Form eines Profils statt.

Einige Tests sollen nur eine bestimmte Komponente der Intelligenz messen. Die sog. „Culture Fair Tests“, die in verschiedenen Varianten vorliegen (der CFT 20 wird unten beschrieben), sollen die fluide Intelligenz erfassen, also die von Bildungseinflüssen relativ freie geistige Leistungsfähigkeit. Dazu finden oft sprachfreie Aufgaben zum schlussfolgernden Denken Verwendung. Manchmal ist es nicht sinnvoll, einen Test einzusetzen, dessen Ergebnis von der (Schul-)Bildung oder der Beherrschung der (deutschen) Testsprache abhängt. Wenn Testpersonen die Testsprache nicht hinreichend beherrschen oder aus einer anderen Kultur kommen, wäre es unfair, ihre Intelligenz mit einem Test zu messen, in dem z. B. nach dem Namen des Bundespräsidenten oder nach der Bedeutung des Wortes „Katakombe“ gefragt wird. In vielen Fällen ist es diagnostisch aufschlussreich, gesonderte Informationen über die fluide und die kristallisierte (gleichbedeutend mit „kristalline“) Intelligenz zu haben. Andere Tests sind so konzipiert, dass sie etwa nur den Wortschatz (als Indikator für erworbenes Wissen oder kristallisierte Intelligenz) prüfen. So soll der Mehrfachwahl-Wortschatz-Intelligenztest (MWT) den Wortschatz mit Items wie „Oher – Ohr – Ehr – Ereh – Hor“ (das einzige richtige Wort ist zu markieren) erfassen. (Vom MWT existieren verschiedene Formen; die am häufigsten verwendete ist der MWT-B; Lehrl 1977, die 5., unveränderte Auflage stammt aus dem Jahr 2005). Tests zur fluiden Intelligenz können auch als zusätzliches Modul einen Test zur kristallisierten Intelligenz (Wissen) enthalten.

Durchführungsbedingungen Aus ökonomischen Gründen ist oft eine Gruppenetestung zu bevorzugen. Die dafür geeigneten Tests lassen sich selbstverständlich auch als Einzeltest verwenden. Die Wechsler-Tests (► Abschn. 3.2.3.2) wurden bewusst für Einzeluntersuchungen konzipiert. Die Durchführung verläuft als weitgehend standardisierter Dialog: Der Testleiter bzw. die Testleiterin fragt etwas, und die Testperson gibt eine Antwort darauf. Dieses Vorgehen kann aus motivationalen Gründen nötig sein, insbesondere bei Kindern und bei Erwachsenen mit einer psychischen Störung oder einer Intelligenzminderung. Es hat den weiteren Vorteil, dass man Einblick in das Arbeitsverhalten bekommt und das Testergebnis vor dem Hintergrund der beobachteten Anstrengung und der eingesetzten Lösungsstrategien interpretieren kann. Ein Ergebnis von 80 IQ-Punkten, das mit höchster Anstrengung „erkämpft“ wurde, ist anders zu werten als das gleiche Ergebnis, das mit geringerer Motivation oder einem unkonzentrierten Arbeitsstil „entstanden“ ist.

Grundlegende Ansätze zur Erfassung der Allgemeinen Intelligenz

Intelligenzkomponenten

Fluide und kristallisierte Intelligenz

Einzel- oder Gruppenuntersuchung

Bei den meisten Intelligenztests ist die Bearbeitungszeit knapp bemessen; es kommt also bei der Bearbeitung auch auf Schnelligkeit an. Für manche Testpersonen stellt Zeitdruck eine ungerechtfertigte Benachteiligung dar. Beispielsweise können sie aufgrund von Seh- oder Sprachschwierigkeiten nur verlangsamt lesen; andere sind motorisch beeinträchtigt, sodass sie für das Ankreuzen von Items oder die Betätigung von Tasten mehr Zeit benötigen als andere Menschen. Auch eine Verlangsamung von Denkprozessen durch bestimmte Erkrankungen (insbesondere Depression) oder bestimmte Medikamente ist möglich. Schließlich kann in manchen Fällen Zeitdruck in Kombination mit einer starken Testangst zu einer Leistungsbeeinträchtigung führen. In diesen Fällen ist der Einsatz von Tests ohne (starke) Zeitbegrenzung (Powertests) sinnvoll. Dabei steigt die Schwierigkeit von Item zu Item derart an, dass die letzten Aufgaben selbst von sehr fähigen Testpersonen kaum noch zu lösen sind. Viele Tests sind als eine Kombination von Speed- und Powertests konzipiert: Die Items werden zunehmend schwerer, und die Bearbeitungszeit ist für alle Testpersonen begrenzt. Die Zeit ist so bemessen, dass man zügig, aber nicht besonders schnell arbeiten muss. Diese Variante hat den Vorteil, dass die Testpersonen in einer Gruppenuntersuchung fast gleichzeitig fertig werden.

Computerbasierte Tests

Viele ursprünglich als Papier-und-Bleistift-Verfahren entwickelte Tests sind auch als Computerversion erhältlich. Mittlerweile werden zahlreiche Tests von Anfang an als computerbasierte Verfahren entwickelt. Viele Vorteile computergestützter Diagnostik liegen auf der Hand:

- Durchführung hoch standardisiert
- Entlastung für den Testleiter/die Testleiterin (diese können während der Testdurchführung andere Aufgaben erledigen)
- Auswertung völlig standardisiert und nicht fehleranfällig
- Auswertung sehr ökonomisch (keine Arbeitszeit erforderlich)
- Ergebnisse sofort verfügbar
- Bei Bedarf exakte Erfassung von Einzelreaktionen inklusive der zugehörigen Zeit möglich
- Bei Bedarf Darbietung von sich bewegenden Reizen oder von Videosequenzen
- Adaptives Testen möglich
- Verwendung von komplexen Problemlöseszenarien möglich

Dem stehen kaum Nachteile gegenüber. Unter bestimmten Umständen kann die computerunterstützte Diagnostik zu Mehrkosten gegenüber der Papier-und-Bleistift-Version führen. Allerdings sollte man nicht nur die Anschaffungskosten für den Test und ggf. Kosten für die Anschaffung der notwendigen Hardware, sondern auch Personalkosten für die Auswertung berücksichtigen.

Für computerbasierte Tests werden oft die Normen einer vorausgegangenen Papier-und-Bleistift-Version ohne Äquivalenzprüfung übernommen. Ob die Normen dann angemessen sind, lässt sich im Einzelfall zumeist schwer beurteilen.

Diagnostische Verfahren

Klinck (2002) hat sich schon in einer Zeit, als computerbasierte Diagnostik nicht verbreitet war, mit der Äquivalenz und Akzeptanz von computerbasierter Intelligenzdiagnostik befasst. In einer großen und sorgfältig geplanten Studie im psychologischen Dienst der Arbeitsämter konnte sie zeigen, dass die beiden verwendeten Intelligenztestversionen zu den gleichen Ergebnissen führten, die computerbasierte Testung kein Akzeptanzproblem zur Folge hatte und keine Benachteiligung bestimmter Personengruppen zu befürchten ist.

Für einige Tests, die typischerweise oder ausschließlich in Einzelsitzungen durchgeführt werden, ist auch eine „computerassistierte“ Testung möglich. So wird etwa für den WISC-V ein „Q-interactive“ genanntes System angeboten, bei dem der Testleiter/die Testleiterin computerassistiert mit der Testperson interagiert. Auf dem Monitor für die Testleiterin bzw. den Testleiter stehen z. B. Instruktionen, die vorzulesen sind. Dort können die Leistungen der Testperson auch direkt eingetragen werden. Auf dem Tablet-PC der Testperson werden bestimmte Aufgaben direkt bearbeitet, und die Ergebnisse werden drahtlos auf den Tablet-PC der Testleitung übertragen. Für andere Aufgaben kommt weiterhin das konventionelle Testmaterial zum Einsatz. Daniel und Wahlstrom (2018) haben mit gut vergleichbaren Probandengruppen untersucht, ob sich die Testergebnisse beider Versionen unterscheiden. Sie stellten bei einzelnen Untertests unterschiedlich große Effekte fest (d zwischen $- .20$ und $+ .20$), was bis zu 3 IQ-Punkten entspricht. Über den kompletten Test gliedern sich die Unterschiede weitgehend aus.

Die Durchführungszeit stellt in der Praxis ein wichtiges Kriterium für die Testauswahl dar. Gerade wenn für eine umfangreiche diagnostische Untersuchung verschiedene Verfahren notwendig sind, kann das Zeitargument in den Vordergrund treten. In der Regel müssen eine höhere Reliabilität und eine größere „Breite“ des Tests bei der Messung der Intelligenz mit mehr Items und Subtests und damit mit mehr Zeit „bezahlt“ werden. Deshalb ist zu bedenken, zu welchem Zweck ein Intelligenztest eingesetzt werden soll. Wird nur ein Screeningverfahren gesucht oder kommt dem Testergebnis eine große Bedeutung zu? Insbesondere adaptive Testverfahren können helfen, die Durchführungszeit zu verkürzen. Wegen des großen Entwicklungsaufwands sind adaptive Verfahren aber rar.

Zielgruppe Intelligenztests unterscheiden sich auch bezüglich der Zielgruppe, für die sie konzipiert wurden. Es liegt auf der Hand, dass sich Intelligenztests für Kinder von denen für Erwachsene in der Aufgabenart unterscheiden. Ebenso einleuchtend ist, dass Intelligenztests zur Hochbegabtendiagnostik zumindest teilweise andere (schwierigere) Aufgaben enthalten als Intelligenztests für eine breitere Zielgruppe.

Je nach Zielgruppe verfügen Intelligenztests über unterschiedliche Normen. Die zur Normierung herangezogene Stichprobe muss hinreichend groß und möglichst repräsentativ für die Zielgruppe sein. Bei einigen Tests liegen für bestimmte Altersgruppen – meist sind es die unteren und oberen Ränder der Altersverteilung – nur sehr kleine Eichstichproben vor. Wenn Intelligenztests eine breite Zielgruppe ansprechen, müssen für Subgruppen ggf. getrennte Normen erstellt werden. Besonders im Schulbereich sind getrennte Normen für einzelne Schultypen hilfreich, sodass man beispielsweise feststellen kann, wie begabt eine Testperson im Vergleich zu altersgleichen Gymnasiasten/Gymnasiastinnen ist. Insgesamt ist zu beachten, dass die Normierung

Äquivalenz mit Papier-und-Bleistift-Tests

Computerassistiertes Testen

Geht es um eine genaue Messung oder ein Screening?

Sind die Normen an die Altersgruppe, den Bildungsstand und den Intelligenzbereich angepasst?

Akzeptanz wichtig

die Verwendung des Tests für bestimmte Zielgruppen einschränkt: Viele Tests sind für den unteren oder oberen Intelligenzbereich nicht oder wenig geeignet, da nicht genügend Personen mit niedriger und/oder hoher Intelligenz in die Normierungsstichprobe eingegangen sind.

Für Forschungszwecke oder in der Personalauswahl spielt die Akzeptanz oft eine erhebliche Rolle. Eine für die Zielgruppe angemessene Aufgabenschwierigkeit sowie Iteminhalte, die möglichst aus dem Lebensbereich der Testpersonen stammen, sind dafür entscheidend.

3.2.3.2 Ausgewählte Intelligenztests

Bei der Auswahl der nachfolgend ausführlicher dargestellten Tests spielt die theoretische und praktische Bedeutsamkeit (► Tab. 3.3) eine Rolle, aber auch die Unterschiedlichkeit der Tests. Das Ziel besteht darin, die Verschiedenheit deutlich zu machen. Im Anschluss an die Beschreibung eines Tests werden auch Alternativen zu dem vorgestellten Verfahren kurz vorgestellt.

Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V)

Zum besseren Verständnis des hier ausgewählten Tests ist es hilfreich, seine Vorgesichte und seine Einbettung in eine ganze Testfamilie zu kennen. Deshalb wird zunächst der Hintergrund beschrieben, um dann auf den ausgewählten Test näher einzugehen.

■ Geschichte der Wechsler-Intelligenztests

Die Wechsler-Intelligenztests werden seit vielen Jahren gerne in der Praxis eingesetzt (► Tab. 3.3). Die Wechsler-Tests stellen eine ganze Familie von Tests dar, die für Erwachsene, Kinder und Vorschulkinder entwickelt und inzwischen mehrfach überarbeitet worden sind. Die heutigen Tests gehen auf die Wechsler-Bellevue Intelligence Scales von 1939 zurück. David Wechsler (► Abb. 3.9; für eine Kurzbiografie s. ► https://doi.org/10.1007/978-1-4419-1698-3_258) hatte den Test am Bellevue Hospital in New York erstellt. Er wollte keinen völlig neuen Test entwickeln, sondern suchte nur ganz pragmatisch in den vorhandenen Tests nach brauchbaren Aufgaben. Als Vorbilder dienten insbesondere der Test von Binet und die Army-Alpha- und

WISC-V und die Wechsler-Tests



► Abb. 3.9 David Wechsler. (Courtesy of the National Library of Medicine)

Diagnostische Verfahren

-Beta-Tests (► Abschn. 1.6). Aus den beiden Armeetests hat er Dutzende von Items „übernommen“. Viele dieser Items finden sich noch in den späteren Versionen der Wechsler-Tests (Gregory 2004).

Der Name „Wechsler Adult Intelligence Scale“ (WAIS) tauchte 1955 erstmals auf. Der Test ist in vielen Sprachen erschienen und seit dem Erscheinen der WAIS 1955 folgten mehr oder weniger zeitnah auch deutschsprachige Versionen. Bereits 1949 kam eine sehr ähnlich aufgebaute Variante für Schul Kinder, die Wechsler Intelligence Scale for Children (WISC), hinzu und 1967 folgte eine für Vorschulkinder, die Wechsler Preschool Primary Scale of Intelligence (WPPSI). Diese wurden ebenfalls im Laufe der Zeit Revisionen unterworfen – und es erschienen stets auch deutschsprachige Ausgaben.

Bei den älteren deutschen Versionen stehen die ersten 4 Buchstaben des Testnamens für **Hamburg** (den Ort, an dem die erste Eindeutschung erfolgte), **Wechsler** (den Autor) und **Intelligenztest**; der letzte Buchstabe bezeichnet die Erwachsenen- (HAWIE) bzw. die **Kinderversion** (HAWIK). Zusätze wie -R oder die römische Zahl -III kennzeichnen die Version des Tests. Erst mit dem Wechsler-Intelligenztest für Erwachsene (WIE) ist der Bezug auf Hamburg weggefallen; die Tests werden nun offenbar enger an das amerikanische Original angelehnt. Die deutschen Versionen der Wechsler-Tests werden jetzt von dem deutschen Tochterunternehmen eines großen amerikanischen Verlages vermarktet, der über die Rechte an den Wechsler-Tests verfügt. Die Wechsler-Tests liefern ein Maß für die Allgemeine Intelligenz sowie weitere Angaben zu einzelnen Fähigkeiten oder Bündeln von Fähigkeiten. Sie werden mit der Testperson in einer Einzelsitzung in Form eines weitgehend standardisierten Dialogs durchgeführt.

Das Geheimnis des großen Erfolgs der Wechsler-Tests lautete zumindest für lange Zeit: Konstanz. Das Grundkonzept blieb lange und über verschiedene Zielgruppen (Kleinkinder, Kinder, Erwachsene) weitgehend unverändert. Obwohl die Entwicklung über eine lange Zeitspanne lief und Tests für unterschiedliche Altersstufen vorgelegt wurden, wiesen die Verfahren eine große Ähnlichkeit untereinander auf. Sie bestanden lange Zeit aus 10–12 Subtests, die sich etwa zu gleichen Teilen auf den Verbal- und den Handlungsteil (s. u.) verteilten und zu einem Gesamt-IQ, also einem Maß für die Allgemeine Intelligenz, verrechnet wurden. Es mussten jedoch mit der Zeit einzelne Items ausgetauscht werden, weil sie nicht mehr zeitgemäß waren (► Abb. 3.10). Auch der Name Wechsler wird weiter bei der Autorenangabe geführt, obwohl der „Erfinder“ dieser Tests bereits 1981 verstorben ist und seitdem auch erhebliche Änderungen an den Wechsler-Tests vorgenommen wurden (s. u.). Wer mit einem der Tests gearbeitet hat, kann sich schnell in eine neue Version oder einen Test für eine andere Altersgruppe einarbeiten. Selbst Psychologinnen und Psychologen, die schon seit 30 Jahren im Beruf stehen, können prinzipiell noch von dem Wissen profitieren, das sie einmal im Studium erworben haben. Die große Verbreitung und Akzeptanz lassen sich zum Teil auch mit einer großen Präsenz der Wechsler-Intelligenztests in der Forschung erklären. Zu den Kinder- und Erwachsenentests WAIS bzw. WISC lassen sich heute (Stand: Mai 2020) rund 10.000 bzw. über 5000 Publikationen nachweisen.

Erst mit der Einführung neuer Strukturmodelle wurde mit einer alten Tradition der Wechsler-Tests gebrochen, unterhalb der Allgemeinen Intelligenz nur zwischen einem Handlungs- und einem Verbal-IQ zu unterscheiden. Das alte Strukturmodell war pragmatischer Art. Ein Teil der Subtests verlangte eine manuelle Tätigkeit, etwa das Nachlegen von Mustern; diese wurden zum Handlungsteil gezählt. Bei den übrigen Subtests musste die Testperson Fragen beantworten – das war der Verbalteil. Was sich konzeptuell dahinter

Erwachsenen-, Kinder- und Vorschulkinder-Versionen

Deutsche Versionen

Erfolg durch Konstanz

Altes Modell mit Verbal- und Handlungs-IQ



Abb. 3.10 Item aus dem HAWIK-R (Subtest „Bilder ergänzen“). Auf dem Bild fehlt das Kabel am Mikrofon. Seit sich schnurlose Mikrofone durchgesetzt haben, ist dieses Item nicht mehr zeitgemäß, und es fehlt in der Nachfolgeversion HAWIK-III. (© 1983 NCS Pearson, Inc. Reproduced with permission. All rights reserved.)

Aktuell Gesamt-IQ und 4–5 Indizes

verbirgt, wurde nie angemessen geklärt. Dadurch bestand die Gefahr von Fehlinterpretationen, etwa dass jemand einen vergleichsweise hohen Wert im Handlungsteil als Hinweis auf eine eher praktische Begabung deutete.

Faktorenanalytische Studien wiesen zumindest auf einen 3. Faktor hin, der u. a. „freedom from distractibility“ genannt wurde (Waller und Waldman 1990). Einen Durchbruch brachte eine Studie von Gignac (2006), in der die Passung verschiedener Modelle mithilfe von konfirmatorischen Faktorenanalysen geprüft wurde. Dazu konnten die Korrelationsmatrizen aus den Daten der Normstichprobe der WAIS-III (2450 Erwachsene) verwendet werden. Das damals naheliegende Modell mit Verbal- und Handlungsteil und darüber die Allgemeine Intelligenz passte nicht zu den Daten. Am besten passte ein Modell, bei dem ebenfalls die Allgemeine Intelligenz (aus den 14 Subtests) als ein Faktor extrahiert wurde. Zusätzlich wurden die Subtests noch einmal zu 4 Gruppen („sprachliches Verständnis“, „Arbeitsgeschwindigkeit“, „Arbeitsgedächtnis“ und „Wahrnehmungsorganisation“) zusammengefasst. In der Folge wurden neue Untertests zunächst probeweise mit eingesetzt und die Angemessenheit verschiedener Strukturmodelle geprüft. Für die Versionen WISC-V (Wechsler 2017) und WPPSI-IV (Wechsler 2018) gilt aktuell ein Modell mit Gesamt-IQ und den 5 Testgruppen (Indizes) „Arbeitsgedächtnis“, „Sprachverständnis“, „Verarbeitungsgeschwindigkeit“, „visuell-räumliches Denken“ und „fluides Schlussfolgern“. Die Erwachsenenversion WAIS-IV (Wechsler 2012) kommt neben dem Gesamt-IQ (noch) mit nur 4 Indizes aus, und zwar „Sprachverständnis“, „wahrnehmungsgebundenes logisches Denken“, „Arbeitsgedächtnis“ und „Verarbeitungsgeschwindigkeit“.

Der Aufbau und die Auswertung der Wechsler-Tests sollen am Beispiel der WISC-V erläutert werden.

Aufbau und Auswertung der Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V)

Steckbrief WISC-V: Wechsler Intelligence Scale for Children – Fifth Edition. Deutsche Bearbeitung von F. Petermann (Wechsler 2017)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Erfassung der Allgemeinen Intelligenz und spezifischer kognitiver Fähigkeiten
Anwendungsbe-reich	Breiter Anwendungsbereich für Kinder und Jugendliche zwischen 6 und 16 Jahren
Theoretischer Hin-tergrund	Konzeptuelle Einordnung in die 5 Bereiche „Sprachverständnis“, „visuell-räumliche Verarbeitung“, „fluides Schlussfolgern“, „Arbeitsgedächtnis“ und „Verarbeitungsgeschwindigkeit“ auf Grundlage unterschiedlicher theoretischer Ansätze zur Intelligenzstruktur
Testentwicklung	Adaptation durch Hinzunahme, Modifikation und Ausschluss von Items und Untertests gegenüber der 4. Auflage
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche und mündliche Instruktionen mit ausführlichen Beschreibungen
Auswertung	Genaue Angaben zum Ausfüllen des Protokollbogens, Auswertungsschablonen und -beispiele; computergestützte Auswertung
Interpretation	Vorgaben zur Benennung der erfassten Merkmale und zur Verbalisierung ihrer Ausprägung; konkrete Hinweise auf nicht instruktionsgemäße Testbearbeitung; computergenerierter Report auf Basis dieser Vorgaben
Reliabilität	
Konsistenz	Indizes: .89 bis .93; Gesamt-IQ: .96; Subtests: .80 bis .93 (Split-Half-Reliabilitätsschätzung)
Retest	Indizes: .72 bis .82; Gesamt-IQ: .90; Subtests: .73 bis .92 (Retestung nach durchschnittlich 26 Tagen, N=94)
Validität	
Konstruktvalidität	Die angenommene Intelligenzstruktur konnte durch konfirmatorische Faktorenanalysen und Interkorrelationen der Subtests weitgehend belegt werden. Im Manual finden sich außerdem einige Studien zur Korrelation der WISC-V mit anderen Intelligenztests (u. a. KABC-II, WPPSI-III und WAIS-V): $r=.78$ bis $.89$
Kriteriumsvalidität	Hochbegabung und Intelligenzminderung wurden als Kriterien ausgewählt. In Bezug auf hochbegabte Kinder konnten signifikant höhere Indexwerte als für die gematchte Kontrollgruppe ermittelt werden. Ebenso liefern signifikant niedrigere Indexwerte für Kinder mit einer Intelligenzminderung im Vergleich zur Kontrollgruppe Hinweise auf die Kriteriumsvalidität.
Normen	
Zusammensetzung	Normen ($N=1087$) für 33 Altersgruppen in 4-Monats-Intervallen zwischen 6;0 und 16;11 Jahren; damit recht geringer Stichprobenumfang pro Altersgruppe; Repräsentativität hinsichtlich Geschlecht, elterlichem Bildungshintergrund, Schulform, regionaler Herkunft und Migrationshintergrund angestrebt
Erhebungszeitraum	Oktober 2015 bis September 2017
Sonstiges	
Formen	Es gibt keine unterschiedlichen Testformen
Testrezension	
Quelle	Renner und Schroeder (2018)

Deutsche Version der WISC-V mit breitem IQ-Bereich

Bei der WISC-V handelt es sich um die deutsche Version der amerikanischen Fassung aus dem Jahr 2014. Der Test dient der Messung der Allgemeinen Intelligenz und von 5 „primären“ plus 5 „sekundären“ kognitiven Fähigkeiten, die als Teilespekte der Intelligenz aufzufassen sind und im Manual „Indizes“ genannt werden. Der Test deckt mit seinen Normen einen breiten Intelligenzbereich (IQ von 40 bis 160) ab und ist damit potenziell zur Diagnostik von geistiger Behinderung bis hin zur Hochbegabung geeignet.

Testaufbau und Gliederung Die deutschsprachige Form der WISC-V umfasst 15 Untertests. Davon wurden 12 aus der WISC-IV in überarbeiteter Form übernommen, 3 („visuelle Puzzles“, „Formenwaage“ und „Bilderfolgen“) sind neu und ersetzen die Untertests „Bildkonzepte“, „Bilder ergänzen“ sowie „Begriffe erkennen“ aus der Vorgängerversion WISC-IV. Die Untertests werden in primäre und sekundäre Untertests unterteilt. Von den 10 primären Untertests finden 7 zur Bestimmung des Gesamt-IQ Verwendung. Bei Bedarf können dabei einzelne Untertests durch einen anderen ersetzt werden. Mit anderen Worten: Obwohl der Test 15 Untertests hat, wird der Gesamt-IQ nur anhand von 7 Untertests bestimmt.

Die Untertests werden ferner zu sog. „Indexwerten“ verrechnet. Für jeden Indexwert findet immer nur eine Auswahl von Untertests Verwendung. Die 15 Untertests werden zu nicht weniger als 11 Kennwerten der kognitiven Leistungsfähigkeit verrechnet. Einzelne Untertests werden dabei mehrfach eingesetzt; beispielsweise geht der Untertest „Formenwaage“ in nicht weniger als 4 Indizes ein. In □ Tab. 3.8 sind die Subtests und deren Zuordnung zu den Indexwerten aufgeführt. Zusätzlich können zu 5 der Untertests noch einmal insgesamt 12 sog. „Prozesswerte“ berechnet werden.

Angesichts dieses außerordentlich komplexen Aufbaus stellt sich die Frage, welches Intelligenzmodell dem Test in der Testentwicklungsphase zugrunde gelegt wurde. Die Antwort lautet: kein bestimmtes. Stattdessen wurden Anleihen aus der Forschung zur Intelligenzforschung und zur kognitiven Entwicklung sowie aus den kognitiven Neurowissenschaften gemacht. In der amerikanischen Originalversion wird Dombrowski et al. (2018) zu folge versucht, die Intelligenzmodelle von a) Spearman (1904), Carroll, Cattell und Horn, b) Horn, Horn und Blankson sowie, c) Horn und Cattell zu berücksichtigen (zu den Modellen von Carroll, Cattell und Horn s. ▶ Abschn. 3.2.4.1).

Da kein bestimmtes Intelligenzmodell als Ausgangsbasis diente, wurde mit den Daten der Eichstichprobe versucht, ein passendes Strukturmodell zu finden. Im Manual der deutschen Version ist die Forschung dazu nicht dokumentiert. Canivez et al. (2018) kritisieren das methodische Vorgehen sowie die schlechte Dokumentation der Ergebnisse, die im Manual der amerikanischen Version das Modell mit 5 Faktoren (s. o., theoretischer Hintergrund im Steckbrief) begründen sollen. Sie beziehen sich in ihrer Kritik auf eine Reihe von einschlägigen Publikationen.

Canivez et al. (2018) überprüften die Faktorstruktur anhand der Testdaten von 2200 Testpersonen aus der amerikanischen Eichstichprobe. Weil der Testverlag die Herausgabe der Rohdaten verweigerte, verwendeten sie die im Testmanual publizierten Korrelationstabellen als Ausgangsbasis für ihre Strukturanalysen. In die Analysen gingen alle 15 Subtests ein. Die Autoren kommen zu dem Schluss, dass das Modell mit 5 Indexwerten (Faktoren) in allen Altersgruppen nicht gut zu den Daten passt. Angemessen ist ein

IQ, Index- und Prozesswerte

Schwer nachzuvollziehendes Strukturmodell

Tab. 3.8 Aufbau der WISC-V. (Wechsler 2017, © Pearson)

Abkürzung	Subtest	Aufgabenbeschreibung bzw. Itembeispiel	Index
MT	Mosaik-Test	Ein zweifarbiges Muster mit bis zu 9 Klötzen soll nach einer Vorlage nachgelegt werden. Die Flächen sind rot, weiß und rot/weiß.	<i>VRV</i> , AFI, NVI
GF	Gemeinsamkeiten finden	„Was haben Hemd und Schuhe gemeinsam?“ – „Kleidungsstücke.“	SV, AFI
MZ	Matrizen-Test	Ein unvollständiges Muster wird vorgelegt; die Testperson wählt unter 5 Vorlagen den fehlenden Teil des Musters aus.	FS, AFI, NVI
ZN	Zahlen nachsprechen	Die Zahlenfolge „3 – 4 – 1 – 7“ ist nachzusprechen.	AGD, KLI, AAGD
ZST	Zahlen-Symbol-Test	Eine Umwandlungstabelle für Zahlen in Symbole liegt vor; Symbole sind in begrenzter Zeit in leere Felder unter den Zahlen einzutragen.	VG, KLI, NVI
WT	Wortschatz-Test	„Was ist eine Uhr?“ – „Hat Zeiger und macht ticktack.“	SV, AFI
FW	Formenwaage	Auf einer Waage fehlen auf der einen Seite jeweils die Gewichte; die Testperson wählt aus einer Liste die Formen aus, die die Waage ins Gleichgewicht bringen würden.	FS, AFI, NVI, QS
VP	Visuelle Puzzles	Die Testperson wählt jeweils aus einer Liste 3 Puzzleteile aus, die zusammengesetzt ein gezeigtes Puzzlebild ergeben würden.	NVI, (<i>VRV</i>)
BF	Bilderfolgen	Die Testperson sieht eine Vorlage mit einem oder mehreren Bildern und wählt anschließend auf dem Antwortbogen die gezeigten Bilder möglichst in der richtigen Reihenfolge aus.	KLI, NVI, (<i>AGD</i>)
SYS	Symbolsuche	In begrenzter Zeit sollen Symbole in einer Zeile daraufhin überprüft werden, ob ein oder mehrere Zielsymbole darin enthalten sind.	KLI, (<i>VG</i>)
AW	Allgemeines Wissen	„Welcher Tag kommt nach Donnerstag?“ – „Freitag.“	(<i>SV</i>)
BZF	Buchstaben-Zahlen-Folgen	Eine Abfolge von Zahlen und Buchstaben wird vorgelesen. Die Testperson soll die Zahlen in aufsteigender Folge und die Buchstaben in alphabetischer Reihenfolge wiederholen.	AAGD
DT	Durchstreich-Test	In einer Menge an Bildern sind Zielobjekte durchzustreichen.	(<i>VG</i>)
AV	Allgemeines Verständnis	Es sind Fragen zu beantworten, die sich auf die Lösung von alltäglichen Problemen oder das Verständnis sozialer Regeln und Konzepte beziehen.	(<i>SV</i>)
RD	Rechnerisches Denken	Mündlich vorgetragene oder bildhaft dargestellte Rechenaufgaben sind im Kopf zu lösen.	QS, (<i>FS</i>)

Subtests aufgeführt in der Reihenfolge ihrer Darbietung. Abkürzung der primären und sekundären Indizes (*primäre* kursiv gedruckt): AAGD = Auditives Arbeitsgedächtnis, AFI = Allgemeiner Fähigkeitsindex, AGD = Arbeitsgedächtnis, FS = Fluides Schlussfolgern, KLI = Kognitiver Leistungsindex, NVI = Nonverbaler Index, QS = Quantitatives Schlussfolgern, SV = Sprachverständnis, VG = Verarbeitungsgeschwindigkeit, VRV = Visuell-Räumliche Verarbeitung. Der Gesamt-IQ wird über die Leistung in 7 Subtests zu den primären Indizes bestimmt (MT, GF, MZ, ZN, ZST, WT, FW). Subtests, deren Zugehörigkeit zu einem primären Index in Klammern gesetzt sind, können einen dieser Untertest zur Berechnung des IQ ersetzen.

Vier-Faktor-Modell, das dem Modell der Vorgängerversion WISC-IV weitgehend entspricht. Neben „Sprachverständnis“, „Arbeitsgedächtnis“ und „Verarbeitungsgeschwindigkeit“, die in der WISC-IV und der WISC-V vorkommen, sprechen ihre Analysen für einen Faktor, der in der WISC-IV „wahrnehmungsgebundenes logisches Denken“ genannt wurde. Dafür entfallen die Faktoren „visuell-räumliche Verarbeitung“ und „fluides Schlussfolgern“. Die 3 Subtests zur Verarbeitungsgeschwindigkeit weisen relativ wenig gemeinsame Varianz mit der Allgemeinen Intelligenz auf.

Starke Zweifel am Modell mit 5 Faktoren und sekundären Indizes

Man könnte nun einwenden, dass in der Studie alle 15 Subtests einbezogen wurden, obwohl für die 5 Indexwerte in der WISC-V nur 10 Subtests verrechnet werden. Die Studie von Dombrowski et al. (2018) unterscheidet sich von der vorigen darin, dass nur die 10 primären Subtests verrechnet wurden. Die nach dem Manual zu erwartenden 5 Faktoren konnten nur in der Altersgruppe von 15 bis 16 Jahren repliziert werden. In den Altersgruppen von 6 bis 8, von 9 bis 11 und von 12 bis 14 Jahren entsprachen die Ergebnisse denen, die Canivez et al. (2018) bei der Analyse aller 15 Subtests gefunden hatten (s. o.). In beiden Studien fanden sich keine Belege für die Existenz der propagierten sekundären Indexwerte.

Standardisierter Dialog

Durchführung und Auswertung Die einzelnen Subtests (► Tab. 3.8) werden in Einzelsitzungen mit der Testleiterin bzw. dem Testleiter und in fester Reihenfolge in Form eines standardisierten Dialogs vorgegeben und zum Teil durch Demonstrations-, Übungs- und Lernaufgaben verdeutlicht. In Abhängigkeit vom Alter der Testperson werden unterschiedliche Startpunkte eines Subtests gewählt, sodass teilweise mit einer etwas schwierigeren Aufgabe innerhalb eines Untertests begonnen wird. Die Punkte für die davorliegenden Items werden einer Testperson dann gutgeschrieben, wenn sie anschließend die schwierigeren Aufgaben gelöst hat. Andernfalls gelten bei den meisten Untertest „Umkehrregeln“, die dann festlegen, dass auch die vor der altersspezifischen Startaufgabe liegenden Aufgaben durchgeführt werden müssen.

Der Testleiter/die Testleiterin muss bei einigen Untertests (z. B. „allgemeines Wissen“) die Antworten unmittelbar bewerten, da der Untertest nach einer bestimmten Anzahl von falschen oder fehlenden Antworten vorzeitig beendet wird (Abbruchregel). Bei 3 Subtests sehen die Bewertungsrichtlinien für die gültigen Antworten je nach ihrer Qualität 1 oder 2 Punkte vor. Bei einigen Untertests misst die Testleiterin bzw. der Testleiter mit einer Stoppuhr die Zeit, um eine Zeitbegrenzung einzuhalten; bei einem dieser Tests (Mosaik-Test) wird zudem die Bearbeitungszeit gemessen, um Bonuspunkte für eine schnelle Bearbeitung zu vergeben. Die WISC-V stellt also erhebliche Anforderungen an die Testleitung. Vor dem ersten „richtigen“ Einsatz sind eine gründliche Einarbeitung und Übung erforderlich. Da sich der Test von der Vorgängerversion WISC-IV zum Teil deutlich unterscheidet, ist auch bei großer Testerfahrung mit der WISC-IV eine erneute Einarbeitung nötig.

Besonders zu erwähnen ist, dass der Test auch mit Computerunterstützung durchgeführt werden kann. Die Software Q-interactive erlaubt es, mit einem Tablet-PC für Testleiter/-innen und einem für die Testperson, die drahtlos miteinander verbunden sind, einen Teil der Untertests auf dem Tablet-PC durchzuführen. Der standardisierte Dialog wird jedoch beibehalten; die Instruktionen werden vom Tablet-PC anstatt aus dem Manual abgelesen. Die Vorteile liegen in einer Entlastung der Testleitung und einer erleichterten Auswertung. Die Eingaben der am Tablet-PC durchgeföhrten Subtests werden direkt erfasst. Bei den übrigen werden die Bewertungen direkt in den Computer eingegeben. Die Auswertung erfolgt dann mittels Software.

Hohe Anforderungen an Testleiter und Testleiterinnen

Computerunterstützte Durchführung möglich

Nachdem jede Aufgabe mit 0, 1 oder bei einigen Subtests auch mit 2 Punkten bewertet worden ist, sind die Rohpunkte zu addieren. Bei der manuellen Auswertung werden die Rohpunktsummen subtestweise anhand von altersspezifischen Umrechnungstabellen in Wertpunkte umgewandelt. Zur Bestimmung von Indexwerten sind die Wertpunkte der zu einem Index gehörenden Subtests zu addieren. Der Summenwert kann mithilfe weiterer Tabellen in einen IQ-Wert transformiert werden. Die Tabellen enthalten auch Angaben zum Konfidenzintervall des ermittelten Wertes, was als sehr anwendungsfreundlich zu bewerten ist. Für weitere Details sei auf das Manual verwiesen.

Damit sind noch nicht alle Auswertungsmöglichkeiten beschrieben. Ein kostenpflichtiges Auswertungsprogramm steht zur Verfügung, das alle Transformationen vornimmt und anhand hinterlegter Normtabellen den IQ und die Indexwerte bestimmt.

Interpretation Zur Interpretation der Testergebnisse gibt das Manual eine umfangreiche Anleitung. Folgende Möglichkeiten der Interpretation bestehen:

- Gesamt-IQ
- Indexwerte (☞ Tab. 3.8)
- Diskrepanzen zwischen den Indexwerten
- Ermittlung von Stärken und Schwächen durch Analyse des Ergebnisprofils für die Untertests
- Spezifische Vergleiche von Untertests
- Wertemuster innerhalb der Untertests (z. B. Konsistenz oder unregelmäßig richtige und falsche Antworten)
- Prozessanalyse (z. B. Betrachtung des Lösungswegs)

Objektivität Die Durchführungsobjektivität kann nicht perfekt sein, da die Durchführungsrichtlinien komplex sind und die Testleiter/-innen sich den Kindern bzw. Jugendlichen gegenüber von Fall zu Fall nicht immer gleich verhalten können und auch nicht sollen.

Die hohen Anforderungen machen die Wechsler-Tests für Kinder anfällig für Testleitungseffekte. McDermott et al. (2014) untersuchten, wie stark die Ergebnisse von 448 Psychologinnen und Psychologen variierten, die in 2 großen Schulbezirken insgesamt 2.783 Kinder mit der WISC-IV getestet hatten. Die für das Testergebnis vermutlich relevanten biografischen Merkmale der Kinder wurden als Kovariate berücksichtigt. Insbesondere der Gesamt-IQ und der Indexwert für Sprachverständnis erwiesen sich als anfällig für Testleitungseffekte. Bei diesen Kennwerten waren über 10 % der Varianz zwischen den Testergebnissen auf die Testleitungen und Testleiter zurückzuführen.

Die manuelle Auswertung der Wechsler-Tests für Kinder ist fehleranfällig, wie Styck und Walsh (2016) in einer Metaanalyse über Studien mit verschiedenen Vorgängerversionen der WISC-V feststellten. Im Durchschnitt ergab eine Kontrolle, dass in 99,7 % aller Auswertungen mindestens ein Fehler vor kam. Psychologinnen und Psychologen unterliefen mehr Fehler als Studierenden. Am fehleranfälligsten waren die Bestimmung des Gesamt-IQ sowie die des Sprachverständnisses. Insgesamt hielt sich der „Schaden“ aber in Grenzen. Die Auswertungsfehler führten beim Gesamtwert „nur“ zu einer durchschnittlichen Abweichung von einem IQ-Punkt.

Zur Interpretationsobjektivität liegen (noch) keine Studien vor.

Verrechnung der Wertpunkte

Auswertungsprogramm verfügbar

Viele Möglichkeiten bei der Interpretation

Starke Testleitereffekte

Manuelle Auswertung häufig mit kleinen Fehlern

Eine Interpretation nach Manual schränkt die Validität der Interpretationen ein

Manual oder Forschungsergebnissen folgen?

Dombrowski et al. (2018) weisen darauf hin, dass die Testanwenderinnen und -anwender dem einschlägigen Standard zufolge für eine angemessene Testinterpretation verantwortlich sind. Angesichts der bekannten Mängel dürfen sie sich nicht allein auf das Testmanual verlassen. Weil die Empfehlungen im Manual zum Teil nicht mit den Erkenntnissen aus wissenschaftlichen Studien vereinbar sind (s. die vorherigen Ausführungen zum Strukturmodell), sollte die Interpretation nicht allein anhand des Manuals erfolgen. Vor allem ist das Fünf-Faktoren-Modell, das den Indexwerten zugrunde liegt, nicht angemessen, und es fehlen Belege für die sekundären Indexwerte. Es darf aber bezweifelt werden, dass sich die Testanwenderinnen und -anwender in einheitlicher Weise über die einschlägige Forschung informieren und die gleichen Schlüsse daraus ziehen. Deshalb besteht hier ein Dilemma: Die Orientierung am Manual ist gut für die Interpretationsobjektivität, aber schlecht für die Validität der Interpretationen. Eine Orientierung an den Forschungsergebnissen ist gut für die Validität, aber vermutlich schlecht für die Interpretationsobjektivität.

Vielzahl an Kennwerten schlecht für die Validität der Interpretationen

Ein anderes gravierendes Problem resultiert aus der Vielzahl von Kennwerten. Da in der Regel bei der Bestimmung von Konfidenzintervallen oder kritischen Differenzen keine Alpha-Adjustierung vorgenommen wird, besteht die große Gefahr, dass Zufallsbefunde als „wahr“ interpretiert werden.

- » Mit jedem neuen Kennwert steigt die Wahrscheinlichkeit, dass sich irgendwo im Leistungsprofil eines Kindes Schwächen oder statistisch signifikante Profildifferenzen finden, auch wenn keine nennenswerten Beeinträchtigungen vorliegen (Renner und Schroeder 2018, S. 18).

Reliabilität des Gesamt-IQ sehr hoch

Reliabilität Die Reliabilität des Gesamt-IQ liegt in einem für Intelligenztests sehr hohen Bereich, die der Indexwerte in einem hohen. Bei manchen Subtests fallen die Werte relativ niedrig aus, wenn man die Schätzungen aus den Altersgruppen zugrunde legt (was angemessen ist). Das führt teilweise zu großen Konfidenzintervallen.

Zu wenige Belege für die Validität

Validität Die Angaben zur Faktorenstruktur im Manual spiegeln nicht den Stand der Forschung wider (s. zuvor erwähntes Strukturmodell). Die Korrelationen des Gesamt-IQ mit den Gesamtwerten in anderen Intelligenztests sind hoch. Für viele Anwendungen wären Korrelationen mit Schulleistungen sehr informativ gewesen. Über den Gesamt-IQ hinaus liegen nur spärliche Informationen vor.

Fein gestufte Altersnormen

Normen Der Test wurde vom Oktober 2015 bis zum September 2016 an insgesamt 1087 Kindern und Jugendlichen im Alter von 6;0–16;11 Jahren normiert. Die Eichstichprobe ist in Bezug auf die Bildung der Eltern, die besuchte Schulform, das Geschlecht, die regionale Verteilung und den Migrationshintergrund weitgehend repräsentativ. Ein Test für Kinder und Jugendliche verlangt nach fein abgestuften Altersnormen. Die Unterteilung in die Altersgruppen ist in der Tat relativ fein; die Gruppen unterscheiden sich bezüglich des Alters jeweils um 4 Monate. Pro Altersgruppe (in Jahren!) umfasst die Stichprobe allerdings nur jeweils 33 Testpersonen. Normen liegen für zahlreiche Kennwerte vor. Die Angaben werden durch Konfidenzintervalle und kritische Differenzen ergänzt.

Fazit Viele Jahrzehnte lang zeichneten sich die Wechsler-Intelligenztests durch eine große Konstanz aus. Dieser Weg wurde zum Teil verlassen. Den Verbal- und den Handlungs-IQ abzulösen, war überfällig. Leider ist es mit der WISC-V noch nicht gelungen, ein Strukturmodell zu finden, das durch gute empirische Belege überzeugt. Scheinbar bietet der Test eine enorm große Informationsausbeute. Allerdings muss man konstatieren, dass alles, was unterhalb des Gesamt-IQ gemessen wird, mehr oder weniger unklar ist. Anwenderinnen und Anwender sollten extrem vorsichtig sein, wenn sie mehr als die 4 von der WISC-IV bekannten Indexwerte interpretieren wollen. Vor allem ist davor zu warnen, ohne gute Gründe, also gut begründete konkrete Fragestellungen oder Hypothesen, alles auszuwerten und zu interpretieren, was der Test hergibt. Die Gefahr, Zufallsbefunde aufzusitzen, ist groß. Ein revidiertes Manual sollte auf der Grundlage der inzwischen schon ansehnlichen und ständig weiterwachsenden internationalen Forschung aufzeigen, was welche Belege für die Interpretation der vielen Kennwerten existieren.

Viele Unklarheiten jenseits des Gesamt-IQ

■ Alternativen zur WISC-V

Wenn die Testperson in eine andere Altersgruppe fällt, bieten sich die im Altersbereich passenden Wechsler-Tests an. Für Erwachsene sind das die Wechsler Adult Intelligence Scale (WAIS-IV; Wechsler 2012) und für den Vorschulbereich die WPPSI-IV (Wechsler 2018).

Im Folgenden werden 2 Alternativen für die Wechsler-Tests kurz vorgestellt, die ebenfalls als Einzeltests in einem strukturierten Dialog mit der Testperson durchzuführen sind. Die Tests wurden für Kinder und Jugendliche konzipiert und stellen damit eine Alternative zur WISC-V dar. Jeder dieser Tests steht auf eigene Weise mit den Wechsler-Tests in einem „Verwandtschaftsverhältnis“. Die Kaufman Assessment Battery for Children (K-ABC) wurde von dem Psychologen-Ehepaar Nadine und Alan Kaufman, 2 ehemaligen Mitarbeitern von David Wechsler, in den Jahren 1978 und 1979 entwickelt und erschien 1983 in den USA (deutsche Fassung: Kaufman und Kaufman 2015). Das Ehepaar Kaufman kannte also die Vor- und Nachteile der Wechsler-Tests gut und war zugleich völlig frei bei der Entwicklung eines eigenen Tests. Das Adaptive Intelligenz Diagnostikum (AID) stammt aus dem deutschen Sprachraum und erschien erstmals 1995 (Kubinger und Wurst 1995).

K-ABC und AID mit Bezug zu den Wechsler-Tests

Wie bei der Entwicklung von Intelligenztests nicht unüblich, orientierten sich die Autorinnen und Autoren an bereits vorliegenden Tests und räumen dies auch ein. Beim AID ist die Ähnlichkeit mit dem damaligen HAWIE bzw. HAWIE-R und dessen Nachfolgern unverkennbar. So gibt es in den Wechsler-Tests für Kinder einen Untertest „allgemeines Wissen“ mit Items wie „In welcher Himmelsrichtung geht die Sonne unter?“ Im AID-2 findet man mit dem „Alltagswissen“ einen ähnlichen Untertest (Itembeispiel: „Wie viele Minuten hat eine Stunde?“). Auch bei weiteren Untertests ist die Ähnlichkeit mit entsprechenden Untertests der Wechsler-Tests offensichtlich. Allerdings kann auch Wechsler nicht als Erfinder vieler dieser Tests gelten, da er bei anderen Vorbildern Anleihen gemacht hatte (s. o.). Im Folgenden stellen wir die aktuell (Stand: Mai 2020) neuesten Versionen der beiden Tests vor.

Adaptives Intelligenz Diagnostikum 3 (AID 3)

Wir stellen den Test zunächst im Steckbrief kurz vor und kontrastieren ihn dann mit der WISC-V.

Steckbrief AID 3: Adaptives Intelligenz Diagnostikum 3 (Kubinger und Holocher-Ertl 2014)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Komplexe und basale Kognitionen („Intelligenz“) in den Bereichen Wahrnehmen, Merken und Verarbeiten
Anwendungsbereich	Breiter Anwendungsbereich mit Fokus auf förderungsorientierter Diagnostik
Theoretischer Hintergrund	Konzeptuelle Orientierung an Wechslers Intelligenzmodell auf Basis des Rasch-Modells und weiteren pragmatischen Modifikationen
Testentwicklung	Erweiterung um Untertests, Zusatztests und Items sowie Optimierung der Verzweigungsschemata im Sinne des adaptiven Testens (branched testing)
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche und mündliche Instruktionsvorgaben und Angaben zur Gestaltung der Testsituation
Auswertung	Verrechnungsvorschriften für 13 der 17 Untertests; Auswertungsprogramme zur automatischen Auswertung
Interpretation	Vorgaben zur Benennung und Interpretation der erfassten Merkmale
Reliabilität	
Konsistenz	Es werden Standardschätzfehler der Fähigkeitsparameter als Schätzungen der Reliabilität angegeben. Subtests: minimal .55 bis .89, maximal 1,10 bis 2,48; Kurzform: minimal .66 bis .76, maximal 1,15 bis 1,57; Parallelform: minimal .59 bis .89, maximal 1,19 bis 1,59
Retest	Angaben nur zu Vorgängerversionen
Validität	
Konstruktvalidität	Angaben beziehen sich überwiegend auf die Vorgängerversionen. Mittels Faktorenanalyse wurde ein 4-faktorielles Modell der Vorgängerversion repliziert, das aber spezifisch für das AID ist (vgl. Ziegler und Reichert 2017).
Kriteriumsvalidität	Keine Angaben
Normen	
Zusammensetzung	Normen ($N=2165$) für unterschiedliche Altersgruppen zwischen 6;0 und 15;11 Jahren aus Deutschland und Österreich; Repräsentativität hinsichtlich Alter und Stadt-Land-Verteilung gegeben
Erhebungszeitraum	2010 bis 2011
Sonstiges	
Formen	Kurz- und Parallelform; das AID-G (Kubinger und Hagenmüller 2019) stellt eine Überarbeitung des AID-3 für Gruppentests dar
Testrezension	
Quelle	Ziegler und Reichert (2017), Renner und Renner (2015)

Profilinterpretation und
Mindestwerte anstatt Gesamt-IQ

Das AID 3 von Kubinger und Holocher-Ertl (2014) deckt mit Ausnahme der 16-Jährigen den gleichen Altersbereich wie die WISC-V ab. Obwohl es mit seinen zahlreichen Untertests (kurze Beschreibungen bei Renner und Renner 2015) an die Wechsler-Tests angelehnt ist, unterscheidet es sich in Bezug auf das Intelligenzmodell grundlegend von dem in der WISC-V verwendeten. Intelligenz wird als Bündel aller kognitiven Voraussetzungen verstanden, die notwendig sind, um Wissen zu erwerben und Handlungskompetenzen zu

entwickeln. Eine Ähnlichkeit mit Cattells Investmenttheorie der Intelligenz ist erkennbar. Allerdings wird beim AID 3 Wert auf die einzelnen Glieder der „kognitiven Voraussetzungen“ gelegt und eine Profilinterpretation vorgeschlagen, die Stärken und Schwächen der Testperson erkennen lässt. Kubinger und Holocher-Ertl (2014) sind dem Gesamt-IQ als Durchschnitt aller im Test gemessenen Leistungen gegenüber sehr kritisch eingestellt. Sie bevorzugen die niedrigste oder zweitniedrigste Testleistung als Maß für die „Intelligenzquantität“, also die kognitive Mindestfähigkeit. Da jedoch bei vielen Testanwendern ein Interesse an dem Gesamt-IQ-Wert besteht, zeigen sie auch einen Weg für dessen Bestimmung auf.

- !** Wenn sich ein Test nicht an ein bekanntes Intelligenzstrukturmodell oder einschlägige empirische Forschung zu Intelligenzkomponenten anlehnt, sollte genau dargelegt werden, was mit den Untertests gemessen werden soll und wie gut das gelingt. Diesbezüglich besteht beim AID 3 noch Entwicklungspotenzial. Das Gleiche gilt auch für das AID-spezifische Konzept der Intelligenzquantität.

Das AID 3 hebt sich von den Wechsler-Tests und auch von vielen anderen Intelligenztests vor allem durch den Ansatz des *adaptiven Testens* ab. Das heißt, die Auswahl der Aufgaben(-gruppen), die einer Testperson vorzugeben sind, richtet sich nach den Leistungen dieser Person in vorangegangenen Aufgaben. Weil auf viele (individuell) zu leichte bzw. zu schwere Items verzichtet wird, ergibt sich zudem eine besondere Ökonomie, die je nach Ziel der Testvorgabe in eine verkürzte Testzeit oder eine besondere Messgenauigkeit umgesetzt werden kann. Das adaptive Vorgehen erfordert die sofortige Bewertung der gelieferten Antworten als „richtig“ oder „falsch“. Das adaptive Testen wird optional durch ein Computerprogramm für Testleiter/-innen „AID_3_tailored – Testleiterprogramm zum AID 3“ unterstützt, das gegen Aufpreis erhältlich ist. Das AID 3 bietet jedoch auch die Möglichkeit, 10 Subtests und einen Zusatzttest konventionell durchzuführen.

Adaptives Testen

Bei den meisten Untertests erfolgte die Itemselektion nach dem Rasch-Modell. Die Items wurden nach ihrer Verträglichkeit mit dem Rasch-Modell ausgewählt; die Subtests sind somit eindimensional, die zu einem Subtest gehörigen Items messen also die gleiche Fähigkeit.

Rasch-Modell

Kaufman Assessment Battery for Children – Second Edition (KABC-II)

Die K-ABC (deutsche Version von Kaufman et al. 2001) stellt ein eigenständiges, nicht an die Wechsler-Tests angelehntes Verfahren dar. Bei der Auflage aus dem Jahr 2015 handelt es sich um eine revidierte Fassung (KABC-II; Kaufman und Kaufman 2015). Für eine Kurzbeschreibung erfolgt zunächst eine Darstellung im Format unseres Steckbriefs. Die Gemeinsamkeiten mit der WISC-V sowie die wesentlichen Unterschiede dazu werden im Folgenden dargelegt: Die augenfälligste Gemeinsamkeit mit der WISC-V besteht darin, dass beide Verfahren sehr viele Subtests umfassen, die zudem bei der Überarbeitung zu einem erheblichen Teil neu hinzugekommen sind. Beide Tests wurden für ähnliche Anwendungsbereiche konzipiert und sind für einen weiten Altersbereich normiert. Beide liefern einen Gesamt-IQ und darüber hinaus noch weitere Kennwerte. Die Normen sind ebenfalls aktuell und bieten eine feine Altersabstufung.

Viele Kennwerte, ähnliche Anwendungsbereiche, aktuelle Normen

Steckbrief KABC-II: Kaufman Assessment Battery for Children-II (Kaufman und Kaufman 2015)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Informationsverarbeitende und kognitive Fähigkeiten bei Kindern und Jugendlichen
Anwendungsbereich	Breiter Anwendungsbereich in psychologischen, klinischen, psychisch-pädagogischen und neuropsychologischen Kontexten
Theoretischer Hintergrund	Duale theoretische Fundierung anhand des Carroll-Horn-Cattell-Modells (CHC-Modells) als hierarchisches Strukturmodell kognitiver Fähigkeiten mit der Gesamtintelligenz als Generalfaktor (g-Faktor) und anhand des Luria-Modells über neuropsychologische Verarbeitungsprozesse
Testentwicklung	Testverlängerung durch Hinzunahme von Items und Untertests gegenüber Vorgängerversion
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Besonders klare, genaue und differenzierte Angaben zur Durchführung
Auswertung	Feste Regeln zur Auswertung; automatische Auswertung am Computer möglich
Interpretation	Differenzierte Hinweise zur Interpretation sowie Fallbeispiele
Reliabilität	
Split-Half	Skalen: .84 bis .96; Gesamtindizes: .94 bis .98; sprachfreie Skala: .90 bis .96
Validität	
Konstruktvalidität	Im Manual finden sich zahlreiche Studien zur Korrelation der KABC-II mit anderen Intelligenztests (u. a. KABC-II Gesamtskalen mit Gesamtskalen der WISC-IV, Kaufman-Test zur Intelligenzmessung für Jugendliche und Erwachsene [K-TIM] und Intelligence and Development Scales [IDS]): $r=.66$ bis $.92$. Die konfirmatorischen Faktorenanalysen sprechen überwiegend für die strukturellen Annahmen einer 4- bzw. 5-faktoriellen Lösung. Die Subtests einer Skala hängen untereinander und mit der Skala mehrheitlich stärker zusammen als mit anderen Skalen.
Kriteriumsvalidität	Keine Angaben
Normen	
Zusammensetzung	Normen ($N=1745$) für 16 Altersgruppen in unterschiedlichen Intervallen zwischen 3;6 und 18;6 Jahren; Repräsentativität hinsichtlich Alter und Geschlecht gegeben
Erhebungszeitraum	2013 bis 2014
Sonstiges	
Formen	Keine weiteren Formen
Testrezension	
Quelle	Kuschel et al. (2017)

Anderes Intelligenzmodell

Die wichtigsten Unterschiede zur WISC-V sind:

- Der KABC-II liegt ein anderes Intelligenzmodell zugrunde. Der Test kann sowohl nach dem Carroll-Horn-Cattell-Modell (CHC-Modell) der Intelligenz (McGrew 2005; ► Abschn. 3.2.4.1) als auch nach Lurias Theorie zu neuropsychologischen Verarbeitungsprozessen (Luria 1970) interpretiert werden. Für die Testauswahl dürfte dies ein wichtiger Aspekt sein. Anwenderinnen und Anwender haben in der Regel ein bestimmtes Erkenntnisinteresse, das über den Gesamt-IQ hinausgeht.
- Der Einsatzbereich der KABC-II ist mit einer Altersspanne von 3;6 bis 18;6 Jahren etwas größer. Damit ist der Test auch für die Entwicklungsdiagnostik in der frühen Kindheit sowie für die Berufseignungsdiagnostik einsetzbar.

Wir verlassen nun die WISC-V und ähnliche Tests für Kinder und Jugendliche und wenden uns einem Intelligenztest zu, der aus einer ganz anderen Tradition stammt.

Intelligenz-Struktur-Test 2000 – Revision (I-S-T 2000 R)

Steckbrief I-S-T 2000 R: Intelligenz-Struktur-Test 2000 – Revision (2. erweiterte und überarbeitete Auflage; Liepmann et al. 2007)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Intelligenzkomponenten: genauer Schlussfolgerndes Denken, Allgemeinwissen und Merkfähigkeit; Unterscheidung nach verbalen, numerischen und figuralen Fähigkeiten
Anwendungsbereich	Berufseignungsdiagnostik, auch klinischer Bereich
Theoretischer Hintergrund	Starke Überarbeitung des Vorgängers IST 70, der an das Thurstone-Modell angelehnt war, führte zum IST 2000, der noch einmal überarbeitet wurde (I-S-T 2000 R); neues eklektizistisches Strukturmodell
Testentwicklung	Hinzunahme von Items und Untertests gegenüber Vorgängerversion, Itemselektion nach Trennschärfe und Itemschwierigkeit
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Klare Angaben zur Durchführung; schriftliche Instruktion zum Mitlesen
Auswertung	Multiple-Choice-Items, Antwortbogen wird mit Schablone ausgewertet, Punkt für jede richtige Antwort, Anleitung zur Schätzung der fluiden und kristallisierten Intelligenz unter Verwendung von Faktorwerten; kostenpflichtiges Auswerteprogramm verfügbar
Interpretation	Normtabellen, keine Vorgaben zur Verbalisierung der Merkmale und ihrer Ausprägung
Reliabilität	
Konsistenz	Angaben nur für die ganze Eichstichprobe, $\alpha=.87$ bis $.96$ für die Kennwerte, Untertest (sollen nicht interpretiert werden): $\alpha=.71$ bis $.94$
Retest	Keine Angaben
Validität	
Konstruktvalidität	Konfirmatorische und explorative Faktorenanalysen zur Struktur, Korrelationen des I-S-T 2000 u. a. mit Matrizaufgaben des CFT 20 ($N=180$): $r=.58$ (fluide) bzw. $.28$ (kristallisierte Intelligenz) und $.63$ (Reasoning), mit Schulnoten ($n=151$ bis 202): $r=-.14$ (Deutsch) bis $-.45$ Mathematik; diskriminant: $r=.22/.21$ ($n=172/484$) zwischen Schlussfolgerndem Denken und Test d2 (Konzentration)
Kriteriumsvalidität	Keine Angaben
Normen	
Zusammensetzung	Altersbereich von 15 bis >50 Jahren, $N=3484$ (Grundmodul Form A und B) bzw. 2363 (Grundmodul Form C, n pro Gruppe mindestens 69); Wissen: $n=1107$; separate Normen für Schüler/-innen an Gymnasien vs. andere Schulformen; Gesamtgruppe nicht repräsentativ bezüglich Bildung
Erhebungszeitraum	Keine Angabe
Sonstiges	
Formen	Pseudoparalleltests Form A und B, zusätzlich Parallelform C sowie Computerversion; auch Kurztest mit 3 Subtests für schlussfolgerndes Denken verfügbar (Liepmann et al. 2012)
Testrezension	
Quelle	Schmidt-Atzert und Rauch (2008)

Separat einsetzbare Module

Baut auf früheren Versionen auf

Der I-S-T 2000 R (erw.) wurde, anders als die bisher vorgestellten Intelligenztests, explizit für Erwachsene entwickelt, kann aber auch bei Jugendlichen ab 15 Jahren eingesetzt werden. Er ist so konstruiert, dass er als Gruppentest verwendbar ist. Es handelt sich um einen breit angelegten Intelligenztest, der mehrere Kennwerte zur Intelligenz liefert. Er besteht aus 2 separaten einsetzbaren Teilen: Das „Grundmodul“ dient der Messung des schlussfolgernden Denkens sowie bei Bedarf, mit 2 separaten Subtests, auch der Merkfähigkeit. Das „Erweiterungsmodul“ erfasst allgemeines Wissen. Werden das Grund- und das Erweiterungsmodul durchgeführt, können zusätzlich Kennwerte für fluide und kristallisierte Intelligenz ermittelt werden.

Der I-S-T 2cauer (1953) entwickelten Test, der in der Vergangenheit eine herausragende Bedeutung hatte. Der I-S-T 2000 R stellt eine Weiterentwicklung des in Deutschland früher mit Abstand am häufigsten angewendeten Intelligenztests dar, dem I-S-T 70. Die 1. Ausgabe des I-S-T erschien 1953, eine geringfügig überarbeitete Fassung wurde 1970 veröffentlicht (I-S-T 70). In diesen beiden Versionen zählte der I-S-T zu den am häufigsten eingesetzten Leistungstests (► Tab. 3.3). Bereits im Manual des I-S-T 70 wurde von bis dahin nicht weniger als 1,5 Mio. Anwendungen berichtet. Die Normen waren seit der Auflage aus dem Jahr 1970 nicht mehr aktualisiert worden. Die vom Autor vehement propagierte Profilauswertung erwies sich empirisch als ungeeignet zur Prognose von Ausbildungserfolg (Schmidt-Atzert und Deter 1993). Bei einigen Subtests wurden psychometrische Mängel aufgedeckt, und der Test galt als revisionsbedürftig (Schmidt-Atzert et al. 1995). Obwohl sich Amthauer, der Autor des I-S-T und I-S-T 70, bei der Testentwicklung offenbar an Thurstones Intelligenzmodell angelehnt hatte, war die verbale Fähigkeit mit insgesamt 4 von 9 Untertests stark überrepräsentiert. In dem 1999 erschienenen I-S-T 2000 wurden 6 der 9 „alten“ Untertests nach mehr oder weniger starken Modifikationen (bei 2 Subtests wurde nur die Itemabfolge verändert) übernommen, 2 weitere Untertests (Rechenaufgaben und verbale Merkfähigkeit) wurden mit neuen Items ausgestattet, und ein alter Untertest entfiel ganz. Dafür ergänzten die Autoren das „Grundmodul“ (s. u.) um 2 neue Aufgabengruppen und erweiterten die bislang nur verbalen Aufgaben des Untertests „Merkfähigkeit“ um figurale Aufgaben (► Beispieldaufgaben zu ausgewählten Untertests). Völlig neu war auch ein Erweiterungsmodul, das Wissen prüft. Der I-S-T 2000 R unterscheidet sich im Wesentlichen vom I-S-T 2000 nur hinsichtlich der nun sehr viel größeren Normierungsstichprobe sowie einiger Verbesserungen eher technischer Art. Der Wissens- test im I-S-T 2000 R wurde gegenüber der Vorgängerversion stark überarbeitet. Für die 2007 erschienene, erweiterte und überarbeitete Auflage wurde eine Parallelform für das Grundmodul entwickelt.

Beispieldaufgaben zu ausgewählten Untertests

Analogien

Bei 3 Wörtern besteht zwischen den ersten beiden eine Beziehung. Aus 5 Wörtern ist dasjenige Wort zu finden, das zu dem 3. Wort in ähnlicher Beziehung steht wie das 2. zum 1.

Beispiel: Wald: Bäume = Wiese: ?

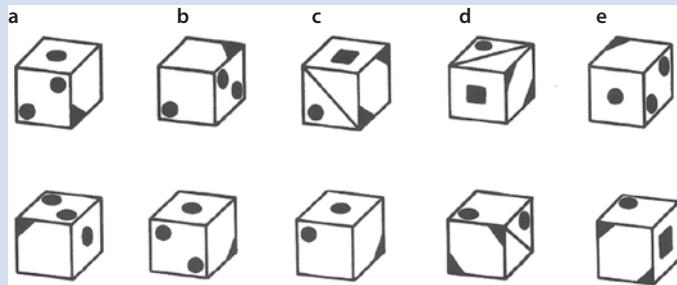
- | | | | | |
|------------|---------|------------|----------|-----------|
| (a) Gräser | (b) Heu | (c) Futter | (d) Grün | (e) Weide |
|------------|---------|------------|----------|-----------|

Rechenzeichen

Die Aufgaben bestehen aus Gleichungen im Bereich der rationalen Zahlen, bei denen die Verknüpfungen weggelassen sind. Das Lösen erfordert das

Einsetzen von Rechenzeichen der 4 Grundrechenarten.

Beispiel: $6 ? 2 ? 3 = 5$.



Würfelaufgaben

Die Testperson soll erkennen, welchem von 5 Auswahlwürfeln ein vorgegebener Würfel gleicht. Der Würfel kann gekippt, gedreht oder gekippt und gedreht sein.

(Aus Liepmann et al. 2007, © Hogrefe).

Merkfähigkeit (figural)

Während der Lernphase werden Figurenpaare eingeprägt. Die Prüfung erfolgt jeweils durch Vorgabe eines der Elemente; das fehlende Element ist unter 5 vorgegebenen auszuwählen.

Theoretischer Hintergrund und Gliederung Mit dem im Jahr 2000 publizierten I-S-T 2000 und anschließend mit dem I-S-T 2000 R wollen die Autoren nicht nur die Schwachstelle überalterter Normen beheben, sondern vor allem das Testkonzept erweitern und den im Zuge der modernen Intelligenzforschung aufgetretenen Konvergenzen inhaltlicher und struktureller Art Rechnung tragen. Die Autoren haben dazu ein Modell formuliert, das sie „Hierarchisches Rahmen- bzw. Protomodell der Intelligenzstrukturforschung (HPI)“ nennen. Warum sie nicht auf eines der vielen etablierten Intelligenzstrukturmodelle zurückgegriffen haben, sondern diese zu einem neuen Modell kombinieren, bleibt unklar.

Inhaltlich lässt sich der Test am besten verstehen, wenn man auf Intelligenzmodelle zurückgreift, die für das HPI-Modell Pate gestanden haben. Mit dem Test sollen 5 der 7 Primärfaktoren von Thurstone erfasst werden, nämlich verbale, numerische und figurale Intelligenz, dazu Merkfähigkeit und – mit etwas höherer Generalität und als Summenscore der 3 erstgenannten Faktoren – Reasoning (schlussfolgerndes Denken). Die jeweils 3 Subtests zur Erfassung der verbalen, numerischen und figuralen Intelligenz werden zu 3 entsprechenden Skalen zusammengefügt. Zur Messung der Merkfähigkeit stehen 2 Aufgabengruppen mit verbalem bzw. figuralem Material zur Verfügung. Diese Batterie bildet das *Grundmodul*. Der Kennwert für das schlussfolgernde Denken errechnet sich aus den Subtests des Grundmoduls, allerdings ohne die beiden zur Merkfähigkeit (Abb. 3.11).

Das *Erweiterungsmodul* enthält Wissensfragen verbaler, numerischer und figuraler Art aus den Gebieten Alltag, Geografie/Geschichte, Kunst/Kultur, Wirtschaft, Naturwissenschaften und Mathematik. Neben einem Gesamtwert für das Wissen können verbales, numerisches und figurales Wissen durch eigene Kennwerte abgebildet werden.

Die Autoren berücksichtigen das Modell von Horn und Cattell (1966), das mit der fluiden und kristallisierten Intelligenz 2 Generalfaktoren der Intelligenz unterscheidet. Die fluide Intelligenz wird als Fähigkeit verstanden, neuen Problemen oder Situationen gerecht zu werden, ohne dass es dazu im wesentlichen Ausmaß früherer Lernerfahrungen bedarf. Die kristallisierte Intelligenz vereinigt dagegen kognitive Fertigkeiten, in denen sich die kumulierten Effekte vorangegangenen Lernens verfestigt haben. Es liegt nahe, dass mit dem Grundmodul (ohne Merkfähigkeit) eher die fluide und mit dem Erweiterungsmodul eher die kristallisierte Intelligenz erfasst wird. Allerdings wird die Bearbeitung der Aufgaben des Grundmoduls durch Wissen und die des Erweiterungsmoduls durch Kombinationsfähigkeit oder schlussfolgerndes

HPI-Modell

Grundmodul mit 5 Primärfaktoren nach Thurstone

Erweiterungsmodul zu Wissen

Fluide und kristallisierte Intelligenz

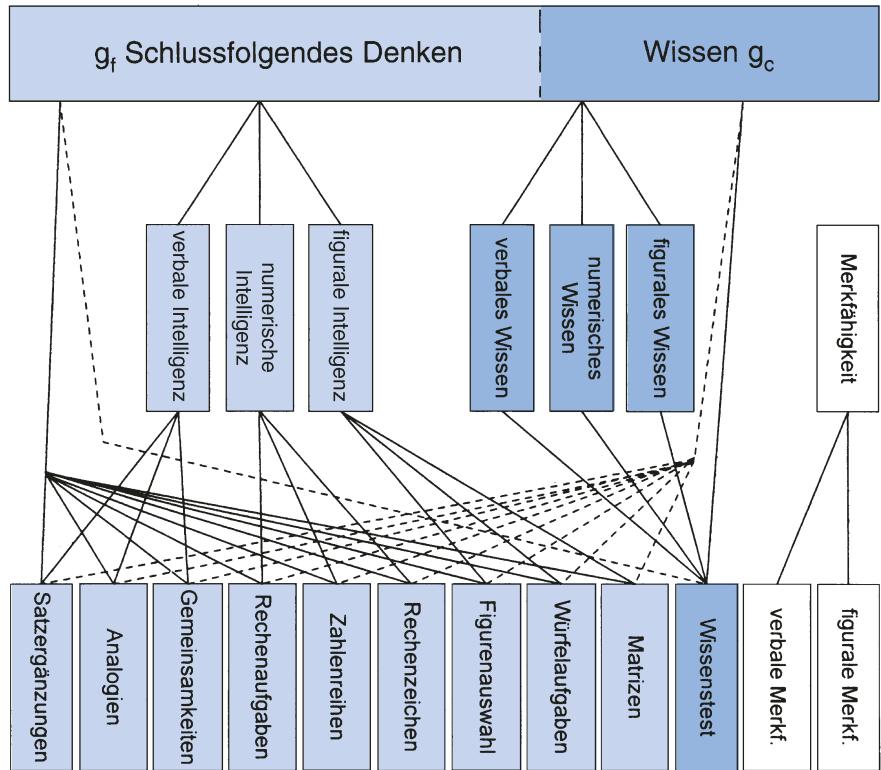


Abb. 3.11 Die mit dem I-S-T 2000 erfasste Fähigkeitsstruktur. (Nach Amthauer et al. 2001, S. 13, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

Denken erleichtert. Die Autoren vertreten die Auffassung, dass die Auspartialisierung der Wissensvarianz aus dem Maß für schlussfolgerndes Denken zu einem optimalen Indikator für fluide Intelligenz und die Auspartialisierung von schlussfolgerndem Denken aus dem Wissen zu einem optimalen Indikator für kristallisierte Intelligenz führt. Die jeweils „bereinigten“ Komponenten erhalten die Notationen „ g_f “ bzw. „ g_c “. Abb. 3.11 gibt in schematischer Form die skizzierte Gesamtstruktur wieder. Die durchgezogenen Linien stehen jeweils für einen positiven Zusammenhang, die gestrichelten Linien für die statistische Auspartialisierung von Fähigkeitskomponenten.

Auswertung manuell, mit PC-Auswerteprogramm oder unter Einsatz der Computerversion

Durchführung und Auswertung Bei einer Gruppenuntersuchung empfiehlt sich die Verwendung von Paralleltests, um Abschreiben zu verhindern. Die Testpersonen kreuzen die Antworten auf einem separaten Antwortbogen an, was die Auswertung mit Schablonen erleichtert. Zur Bestimmung der Faktorwerte für fluide und kristallisierte Intelligenz sind Subteststandardwerte unter Verwendung von im Manual angegebenen Beta-Gewichten zu transformieren. Ein Auswertungsprogramm, bei dem die Antworten in einen PC einzugeben sind, erleichtert die Auswertung. Komfortabler ist der Einsatz der Computertestversion, bei der die Eingaben der Testpersonen automatisch verrechnet und vollständig ausgewertet werden. Die Testzeit beträgt für das Grundmodul ca. 2 h (darin enthalten: 10 min Pause und 12 min für die Bearbeitung der Merkaufgaben). Für das Erweiterungsmodul mit den Wissenstests sind zusätzlich knapp 40 min erforderlich.

Hohe interne Konsistenz

Reliabilität Die innere Konsistenz (Cronbachs α) des Grundmodulgesamtwertes (Schlussfolgendes Denken) beträgt .96 (Form C = .94) und ist damit

sehr hoch. Für die verbale Intelligenz wird α mit .88 angegeben, für die numerische Intelligenz und die Merkfähigkeit mit .95 und für die figurale Intelligenz mit .87. Der Wissenstest weist eine Reliabilität von .93 auf. Die Koeffizienten für die jeweiligen Aufgabengruppen innerhalb der Skalen verbal, numerisch und figural liegen zum Teil deutlich darunter; die Autoren raten daher davon ab, einzelne Untertestergebnisse zu interpretieren. Für die Faktorwerte von fluider und kristallisierter Intelligenz beträgt die interne Konsistenz .96 bzw. .91. Zur Retest-Reliabilität liegen keine Daten vor.

Validität Konfirmatorische Faktorenanalysen mit den 9 Aufgabengruppen des Grundmoduls bestätigten die postulierte Struktur: verbale, numerische und figurale Fähigkeit sowie Merkfähigkeit. Zur Einordnung der Subtests des Grund- und Erweiterungsmoduls in das Modell der fluiden und kristallisierten Intelligenz führten die Autoren eine explorative Faktorenanalyse durch. Die Aufgabengruppen des Grundmoduls und die des Erweiterungsmoduls zeigten erwartungsgemäß Ladungen auf getrennten Faktoren. Warum die Aufgabengruppen zur Merkfähigkeit nicht einbezogen wurden, bleibt ungeklärt. Die beiden extrahierten Faktoren (fluide und kristallisierte Intelligenz) korrelierten zu .54, dennoch sehen die Autoren davon ab, einen übergeordneten Generalfaktor der Intelligenz zu bestimmen.

Darüber hinaus werden Korrelationen mit anderen Tests berichtet. Demzufolge korreliert beispielsweise Reasoning (Gesamtwert) zu .63 mit dem Matrizentest CFT 20 von Weiß (1997). Die Faktorwerte für fluide und kristallisierte Intelligenz korrelieren mit den CFT-20-Matrizen zu .58 bzw. .24, mit einem Wortschatztest (MWT-B) zu .16 bzw. .54 – ein erwartungskonformes Muster. Es finden sich auch Angaben zu Korrelationen mit Schulnoten. So korreliert das Schlussfolgernde Denken mit der Deutschnote zu -.14 und der Mathematiknote zu -.45 (dies stellt den höchsten Zusammenhang mit Noten dar). Als Beleg für die diskriminante Validität kann eine niedrige Korrelation ($r = .22$) zwischen Schlussfolgerndem Denken und dem Konzentrations- test d2 angesehen werden.

Bislang liegt für den IST 2000 R unseres Wissens lediglich eine Untersuchung zu berufsbezogenen Außenkriterien vor. In einer Studie (Steinmayr und Amelang 2006) bearbeiteten berufstätige Frauen und Männer ($N=219$) im Alter von durchschnittlich 34 Jahren das Grund- und Erweiterungsmodul des I-S-T 2000 R (Liepmann et al. 2007). Gleichzeitig wurden Maße für das Ausbildungsniveau und den sozialen Status (Aggregat aus Einkommen und Prestige des ausgeübten Berufs) erhoben. Die höchste Korrelation bestand mit $r = .59$ zwischen Wissen (Erweiterungsmodul) und Ausbildungsniveau. Mit dem sozialen Status der ausgeübten Berufstätigkeit korrelierte das Schlussfolgernde Denken höher ($r = .47$) als das Wissen ($r = .35$).

Normierung Die Normen stützen sich auf die Vorgabe des Grundmoduls an eine Stichprobe von insgesamt $N=3484$ Probanden für die Form A bzw. B und von 2363 Probanden für die Form C. Die Erhebungen fanden in 7 Bundesländern statt. Normentabellen liegen für Schüler/-innen an Gymnasien und anderen Schulformen (außer Gymnasien) und für die Gesamtgruppe vor. Die Aufteilung nach dem Alter fällt unterschiedlich aus. Am differenziertesten ist sie mit 8 Altersgruppen (15–16, 17–18, 19–20, 21–25, 26–30, 31–40, 41–50 und > 50 Jahre) für das Grundmodul bei den Gymnasiasten/Gymnasiastinnen. Die Normen zur Merkfähigkeit unterscheiden nur 3 breite Altersgruppen. Die Wissenstests wurden mit insgesamt 1107 Personen durchgeführt. Daraus wurden 5 nicht nach Bildung differenzierte Altersgruppen sowie 2 Bildungsgruppen (Gymnasium und andere Schulform) ohne Altersdifferenzierung gebildet. Die Normgruppen sind unterschiedlich groß; der

Strukturanalysen

Korrelation mit Intelligenztests und Schulnoten

Korrelation mit beruflichem Status

Normen unterschiedlich differenziert

Stichprobenumfang liegt beim Grundmodul zwischen $N=69$ (Form C, Gymnasium, 51+ Jahre) und $N=578$ (Form A/B, andere Schulform, 31–40 Jahre).

3

Sorgfältig konstruiert

Bewertung Beim I-S-T 2000 R handelt es sich um ein sehr sorgfältig konstruiertes Instrument, das sich zur reliablen Erfassung von 5 Primärfaktoren der Intelligenz sowie der beiden Sekundärfaktoren fluide und kristallisierte Intelligenz eignet. In einer Rezension nach dem Testbeurteilungssystem des Testkuratoriums (Schmidt-Atzert und Rauch 2008) erfährt der Test insgesamt eine gute Bewertung. Die Anforderungen an die Reliabilität und die Validität wurden jedoch nicht als „voll“, sondern nur als „weitgehend erfüllt“ eingestuft. Ergänzend wird aus der Abschlussbewertung zitiert:

- » Die Konsistenzmaße für die globalen Kennwerte liegen über .90 und damit in einem sehr hohen Bereich. Leider werden keine Retest-Reliabilitäten berichtet, auch keine Reliabilitäten für die einzelnen Normgruppen (zur Berechnung von Konfidenzintervallen). [...] Für die Wissenstests fehlen auf der Ebene der verbalen, numerischen und figuralen Facetten noch Validitätsbelege in Form von Korrelationen mit anderen Tests oder Kriterien. [...] Für viele Altersgruppen sind die Normierungsstichproben hinreichend groß, für einige aber mit $N < 100$ zu klein ... Wünschenswert wären genaue Angaben zu Anwendungszweck und Zielgruppen (Schmidt-Atzert und Rauch 2008, S. 304).

■ Alternativen zum I-S-T 2000 R

Als Alternativen kommen andere „breite“ Intelligenztests infrage, die wie der I-S-T 2000 R auch als Gruppentest einsetzbar sind:

Am Carroll-Modell orientiert (auch g-Maß)

Leistungsprüfsystem 2 (LPS-2) Der Vorgänger, das LPS von Horn (1983), orientierte sich noch am Thurstone-Modell der Intelligenz, das LPS-2 (Kreuzpointner et al. 2013) bezieht sich nun auf Carrolls Strukturmodell der Intelligenz. Es liefert einen Gesamtwert, der als Maß der Allgemeinen Intelligenz interpretiert wird. Darin unterscheidet es sich vom I-S-T 2000 R, der auf der höchsten Ebene nur Reasoning, fluide und kristallisierte Intelligenz kennt. Auf der Ebene unterhalb der Allgemeinen Intelligenz wird mit den Komponenten „kristalline Intelligenz“, „fluide Intelligenz“, „visuelle Wahrnehmungsfähigkeit“ und „kognitive Schnelligkeit“ ein Teil der von Carroll gefundenen Intelligenzkomponenten abgebildet. Die Bearbeitung der 11 Subtests dauert ähnlich lange wie die des Grundmoduls im I-S-T 2000 R. Das LPS-2 wurde an 2583 Personen normiert, und zwar überwiegend an Schülerinnen und Schülern zwischen 16 und 19 Jahren ($n=2134$). Es wird von den Testautoren zur Eignungsdiagnostik im schulischen und beruflichen Kontext sowie zur Leistungsdiagnostik im schulischen, beruflichen und im klinischen Kontext empfohlen. Damit deckt sich der Anwendungsbereich zum Teil mit dem des I-S-T 2000 R. Allerdings fehlen Angaben zur Kriteriumsvalidität (Vorhersage von Schulleistungen oder Ausbildungserfolg). Die Verwendung als Gruppentest wird durch die Verfügbarkeit von 2 Testformen (A und B) erleichtert. Eine Computerversion ist (noch) nicht verfügbar.

LPS 50+ für einen Altersbereich zwischen 50 und 90 Jahren

Leistungsprüfsystem 50+(LPS 50+) Mit dem LPS 50+ von Sturm et al. (2015) liegt eine Variante des LPS für den Altersbereich zwischen 50 und 90 Jahren vor. Das Testmaterial ist so gestaltet, dass es auch für ältere Personen gut lesbar ist. Allerdings leitet sich das LPS 50+ nicht aus dem LPS-2, sondern aus

der Vorgängerversion von 1993 her. Es steht damit nach wie vor in der Tradition des Thurstone-Modells der Intelligenz. Die Normen stammen aus dem Jahr 2009 und gliedern sich in 4 Altersgruppen ($n=53$ bis 90). 7 Untertests können als Kurzform eingesetzt werden. Die Gesamttestzeit beträgt dann nur etwa 38 anstatt 80 min für alle 15 Subtests. Das LPS 50+ ist eines der wenigen Intelligenztestverfahren, die auch im höheren Alter eingesetzt werden können, und liefert neben einem Gesamt-IQ noch Informationen zur Intelligenzstruktur.

Wilde-Intelligenz-Test 2 (WIT-2) Der WIT-2 von Kersting et al. (2008) weist inhaltlich eine deutliche Überlappung mit dem I-S-T 2000 R auf. Die berichteten Korrelationen von 4 korrespondierenden Kennwerten (sprachliches, rechnerisches, räumliches und schlussfolgerndes Denken) mit vergleichbaren I-S-T 2000 R Kennwerten liegen dementsprechend bei $r=.79, .81, .60$ und $.81; n=78$. Der WIT-2 repräsentiert 5 der 7 Primärfähigkeiten von Thurstone (verbal comprehension, number, space, reasoning, memory). Das schlussfolgernde Denken wird im WIT-2 (anders als bei Thurstone) als eine dem verbalen, rechnerischen und räumlichen Denken übergeordnete Skala konzipiert. Darüber hinaus werden die Dimensionen „Arbeitseffizienz“ und „Wissen“ (Wirtschaft sowie Informationstechnologie) erfasst. Die Autoren sprechen von einem „modifizierten Modell der primary mental abilities“ nach Thurstone, das dem Test zugrunde liege. Der Test wurde vorrangig für die berufsbezogene Diagnostik entwickelt. Dazu passt, dass die Testaufgaben teilweise in eine Semantik aus dem Berufs- und Arbeitsleben eingekleidet sind und der überwiegende Teil der Normdaten ($n=\text{mindestens } 2234$) im Kontext des Ernstfalls von beruflichen Bewerbungssituationen erhoben wurde. Der Test ist als Baukastensystem konzipiert; je nach Bedarf können unterschiedliche Module eingesetzt werden. Aussagekräftige Studien für den vorgesehnen Einsatzbereich Personalauswahl fehlen im Manual. Der Erstautor verweist auf seiner Webseite ([► https://kersting-internet.de/testentwicklungen/intelligenztest-wit-2/aktuelles/](https://kersting-internet.de/testentwicklungen/intelligenztest-wit-2/aktuelles/)) auf eine Metaanalyse zur Kriteriumsvalidität des WIT hin, der dem Nachfolger WIT-2 immerhin sehr ähnlich ist. Lang et al. (2010, Tab. 2, S. 619) zufolge korrelieren die Kennwerte des „alten“ WIT zwischen .21 (Wahrnehmungsgeschwindigkeit) und .39 (sprachliches sowie schlussfolgerndes Denken) mit Berufserfolg.

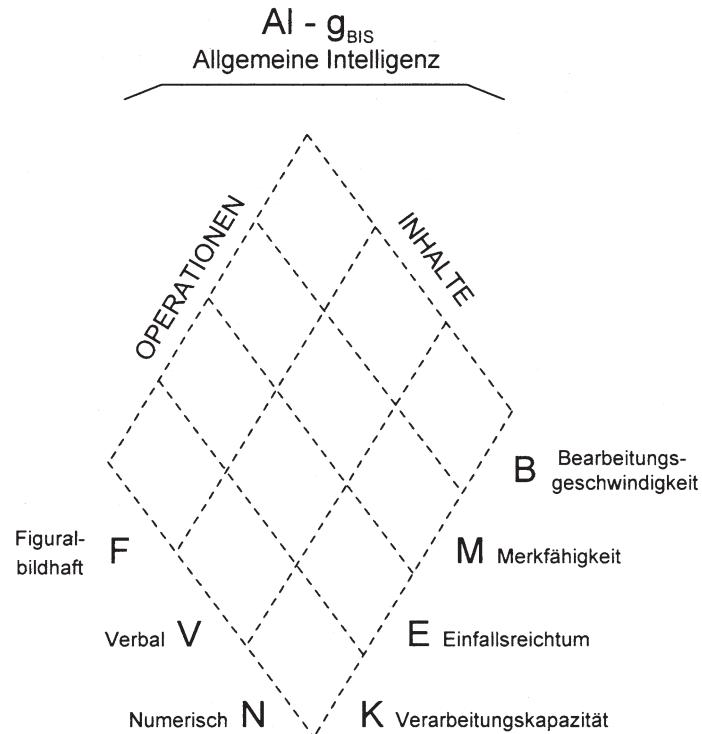
Für berufsbezogene Diagnostik entwickelt

Berliner Intelligenzstruktur-Test, Form 4 (BIS-4) Der BIS-4 von Jäger et al. (1997) ist mit seinen sehr alten Normen für weite Bereiche der Praxis wenig brauchbar. Er wird dennoch hier kurz dargestellt, weil er z. B. für Forschungszwecke (wo üblicherweise Rohwerte ausreichen) sehr nützlich sein kann. Ihm liegt das „Berliner Intelligenzstrukturmodell“ (► Abb. 3.12) zugrunde. Mithilfe dieses bekannten und anerkannten Modells können andere Intelligenztests oder Aufgabengruppen konzeptuell leicht eingeordnet werden. Ebenso lässt sich deren Konstruktvalidität an entsprechend ausgewählten Subtests des BIS-4 (beispielsweise für Konzentrationstests Subtests zur Bearbeitungsgeschwindigkeit) überprüfen.

Berliner Intelligenzstrukturmodell als Referenzsystem

Die Autoren unterscheiden 4 Arten von „Operationen“, die jeweils mit 3 unterschiedlichen „Inhalten“ kombinierbar sind. So kann etwa die Merkfähigkeit als mentale Operation mit numerischen, verbalen und figuralen Aufgaben gemessen werden. Die Anordnung der Operationen im Modell soll übrigens keine Hierarchie ausdrücken. Die rautenförmige Anordnung symbolisiert, dass die Operationen und Inhalte nicht orthogonal zueinander stehen,

Allgemeine Intelligenz plus 7 weitere Kennwerte



■ Abb. 3.12 Berliner Intelligenzstrukturmodell. (Nach Jäger et al. 1997, S. 5, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Hogrefe Testzentrale, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

sondern Gemeinsamkeiten aufweisen. Beispielsweise ergibt sich die Bearbeitungsgeschwindigkeit einer Person als Mittelwert ihrer Testleistungen in allen figuralen, verbalen und numerischen Aufgaben zur Bearbeitungsgeschwindigkeit. Die numerischen Fähigkeiten etwa können als Aggregat aller Testleistungen mit numerischen Aufgaben aus den Bereichen Bearbeitungsgeschwindigkeit, Merkfähigkeit, Einfallsreichtum und Verarbeitungskapazität bestimmt werden. Folglich liefert der Test für jede Testperson neben einem Maß der Allgemeinen Intelligenz 7 Kennwerte: figurale, verbale, numerische Fähigkeiten, Verarbeitungskapazität (konzeptuell dem schlussfolgenden Denken ähnlich), Einfallsreichtum, Merkfähigkeit und Bearbeitungsgeschwindigkeit (Konzentrationsfähigkeit; ► Abschn. 3.2.2.2). Jeder der insgesamt 45 Subtests kann in einer der 12 Zellen verortet werden.

Version für Hochbegabte

Zur Hochbegabungsdiagnostik steht mit dem Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB; Jäger et al. 2006) eine Testversion zur Verfügung, die nicht nur an durchschnittlich begabten, sondern auch an über 500 hochbegabten Schülerinnen und Schülern normiert wurde. Der Test ist für den Altersbereich von 12 bis 16 Jahren geeignet.

Grundintelligenztest Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest (CFT 20-R)

Steckbrief CFT 20-R mit WS/ZF-R: Grundintelligenztest Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest (2., überarbeitete Auflage mit aktualisierten und erweiterten Normen; Weiß 2019)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Grundintelligenz bzw. fluide Intelligenz, ergänzt um kristalline Intelligenz
Anwendungsbereich	Insbesondere Bildungsbereich
Theoretischer Hintergrund	Cattells Intelligenztheorie mit der Unterscheidung von fludier Intelligenz im Sinne eines intellektuellen Potenzials und der daraus durch Bildungseinflüsse hervorgegangenen kristallinen Intelligenz
Testentwicklung	Items des Tests überwiegend aus Vorgängerversion (Weiß 2006) übernommen, um neue erprobte ergänzt; Testmaterial für die aktualisierte Auflage von 2019 nicht verändert
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Klare Angaben zur Durchführung (Instruktion)
Auswertung	Multiple-Choice-Items, Auswertung des Antwortbogens mit Schablone
Interpretation	Normtabellen, keine Vorgaben zur Verbalisierung der Merkmale und ihrer Ausprägung
Reliabilität	
Konsistenz	Reliabilität für Testbatterien = .95 für CFT 20-R (Teil 1 + 2); Split-Half-Reliabilität = .87 (Wortschatztest) und .92 (Zahlenfolgetest)
Retest	Weiß (2006): r_{tt} für Gesamttest nach 2 Monaten = .91 ($N=38$ Schüler), für Zusatztest WS $r_{tt}=.83$ und ZF $r_{tt}=.85$ (4 Stichproben, Intervall: 2–4 Monate); Weiß (2019): $r_{tt}=.80$ bis .82 (Intervall: 3 Monate)
Validität	
Konstruktvalidität	Weiß (2006): Korrelationen mit PSB-R 4–6: $r=.50$ ($N=490$), mit Mathematiknote durchschnittlich $r=.49$ und damit deutlich höher als mit Deutschnote $r=.35$ ($n=855$, unterschiedliche Klassen); Weiß (2019): Korrelationen mit Schulnote in Mathematik $r=.50$ bis .57
Kriteriumsvalidität	Gute prognostische Validitätswerte in Schulbeobachtungen (6–10 Jahre; Weiß 2019)
Normen	
Zusammensetzung	Weiß (2006): 4350 Schülerinnen und Schüler aller Schularten aus 6 Bundesländern, 8;5 bis 19 Jahre; bei Normen Schulart nach Angaben des statistischen Bundesamtes gewichtet; für den Altersbereich von 20–60 Jahren wurden die Normen anhand vorherigen Testversion ermittelt; Zusatztests WS und ZF: $N=2724$;
Erhebungszeitraum	2003 und 2004; Weiß (2019): Gültigkeit der Normen wurde in 2015 für den Altersbereich von 8;5 bis 19 Jahren weitgehend bestätigt ($N = 870$); neue Normen liegen für den Altersbereich von 20 bis 64 Jahren vor (Teil 1 des Tests: empirische Normen von 5858 Personen; Teil 2 des Tests: rechnerisch ermittelte Normen).
Sonstiges	
Formen	Computerversion verfügbar; Teil 1 als Kurzversion verwendbar
Testrezension	
Quelle	Gruber und Tausch (2016)
<i>Anmerkung.</i> Das Manual liegt in 2 Auflagen vor, von Weiß (2006) sowie von Weiß (2019). Die Angaben zur 2. Auflage von 2019 basieren auf Informationen des Testverlages.	

„Sprachfreie“ Messung der fluiden Intelligenz

3

Beim CFT 20-R (das Kürzel steht für „Culture Fair Test“) handelt es sich um ein Mitglied einer ganzen „Testfamilie“ zur sprachfreien Messung der fluiden Intelligenz. Cattell stellte den ersten Test dieser Art bereits 1940 vor. Der ursprüngliche Anspruch, die Intelligenz „kulturfrei“, also unabhängig von Einflüssen des soziokulturellen, schulischen und erziehungsspezifischen Erfahrungshintergrunds zu messen, erwies sich als überhöht und die ursprüngliche Testbezeichnung „culture free“ (Cattell 1940) wurde später zu „culture fair“ abgeschwächt. Um dem Anspruch der kulturfairen Messung wenigstens konzeptuell gerecht zu werden, sind die Items sprachfrei. „Sprachfrei“ bedeutet nicht, dass der Test ohne verbale Instruktionen auskommt; es bezieht sich lediglich auf die figuralen Items. Bis zur 1998 erschienenen, 4. Auflage wurden die Items immer wieder unverändert übernommen; die Überarbeitung betraf lediglich das Testmanual. Der Testautor hatte aber bereits in der Vergangenheit weitere, überwiegend schwerere Items erprobt, indem er sie in Forschungsversionen des Tests unter die alten Items gemischt hatte. Für die aktualisierte Auflage von Weiß (2019) wurden im Manual auch neuere Forschungsergebnisse berücksichtigt sowie die Normen deutlich überarbeitet. Für geistig behinderte Personen existiert zudem eine PC-Version, für die auch Normen vorliegen. Die folgenden Ausführungen beziehen sich auf die 2006 Auflage von CFT-R, da uns die überarbeitete Auflage von 2019 beim Abschluss des Buchmanuskripts noch nicht vorlag.

Mit 4 Subtests breite Messung der fluiden Intelligenz

Gliederung Der CFT 20-R besteht aus 4 Subtests, bei denen Figurenreihen fortgesetzt, Figuren klassifiziert, Figurenmatrizen vervollständigt und topologische Schlussfolgerungen gezogen werden sollen (s. Beispiele in Abb. 3.13).

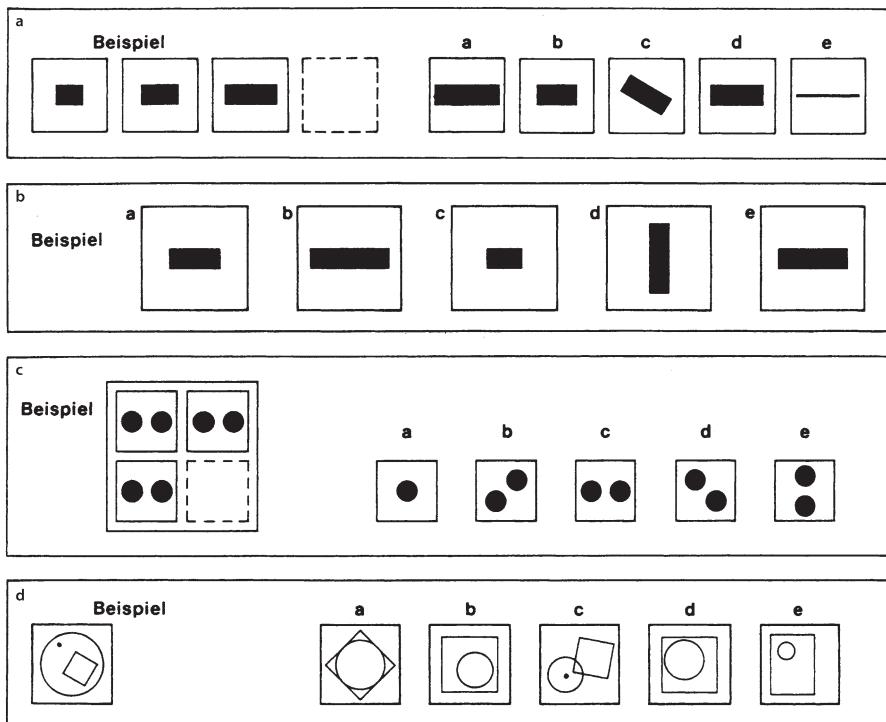


Abb. 3.13 a-d Itembeispiele aus dem CFT 20-R (Weiß 2019, © Hogrefe). a Es ist die Figur zu suchen, die die Reihe richtig fortsetzt (a). b Es ist die Figur zu finden, die nicht in die Reihe passt (d). c Gesucht ist die das Muster richtig ergänzende Figur (c). d Zu finden ist die Figur, in der der Punkt ähnlich wie im Beispiel (im Kreis, aber außerhalb des Quadrats) gesetzt werden kann (c).

Die Items sind innerhalb der Subtests nach Schwierigkeit angeordnet. Der Test gliedert sich in 2 nach dem gleichen Prinzip gestaltete Teile mit 56 bzw. 45 Items. Teil 1 mit etwas leichteren Items kann als Kurzform verwendet werden, die Langform setzt sich aus Teil 1 und 2 zusammen. Sowohl bei der Kurz- als auch der Langform kann die Testzeit verlängert werden; entsprechende Normen liegen vor. Bei Testpersonen, die Schwierigkeiten mit der Instruktion haben, kann Teil 1 als Übungsphase betrachtet werden; ausgewertet wird dann nur Teil 2. Mit 4 sprachfreien (figuralen) Subtests unterscheidet sich der Test von einem anderen bekannten sprachfreien Intelligenztest, den Raven's Progressive Matrices (s. u.), in dem nur Matrizenaufgaben verwendet werden.

Der sprachfreie Teil wird um 2 fakultative Tests zur kristallisierten Intelligenz ergänzt, für die ein separates Manual vorliegt. Beim Wortschatztest ist bei jeder der 30 Aufgaben zu einem vorgegebenen Wort unter mehreren Auswahlwörtern das ähnlichste herauszufinden. Der Zahlenfolgentest besteht aus 21 Zahlenreihen, die fortzusetzen sind. Entsprechende Items könnten sein: Auto: (a) Straße, (b) Fahrrad, (c) Fahrer, (d) Kraftfahrzeug, (e) Eisenbahn bzw. 5, 7, 9, 11, 13, 15 ?.

2 zusätzliche Tests zur kristallisierten Intelligenz

Durchführung und Auswertung Der Testbearbeitung gehen (wie üblich) Übungsaufgaben voraus, die in die Eigenart jedes Subtests sowie in die Technik der Übertragung der Antworten in ein Antwortblatt einführen. Die Bearbeitung der beiden Testhälften dauert einschließlich Instruktion und Übungsaufgaben etwa 50 min (bei Gruppenuntersuchungen etwa 60 min). Bei Verwendung der Kurzform (Teil 1) verkürzt sich die Zeit auf 35–40 min. Für die fakultativen Wortschatz- und Zahlenfolgentests werden im Manual jeweils 15 min Durchführungszeit bei Gruppenuntersuchungen veranschlagt (inkl. Instruktionen).

Einfache Auswertung dank Antwortbogen

Reliabilität Die Reliabilitätsschätzung schätzung für den Gesamttest liegt mit .95 in einem sehr hohen Bereich. In einer Studie mit 1886 Schülerinnen und Schülern aus 88 Schulklassen haben Kuhn et al. (2008) für Testteil 1 „nur“ eine interne Konsistenz von $\alpha = .82$ ermittelt. Der Schlussfolgerung, dass die „Reliabilität der Kurzform [...] für einen Intelligenztest nicht hoch“ ist (Kuhn et al. 2008, S. 190), kann nur zugestimmt werden. Eine Erklärung der Autoren ist, dass sich die Subtests nach dem Rasch-Modell als nicht homogen erwiesen, was zu einer Unterschätzung der Reliabilität durch Konsistenzkoeffizienten führe. Die Angaben zur Retest-Reliabilität liegen in einem für Intelligenztests angemessenen Bereich.

Hohe interne Konsistenz – aber nicht für die Kurzform

Validität Die Korrelationen des CFT 20-R fielen mit der Mathematiknote höher aus als mit der Deutschnote, was der Erwartung entspricht (s. Steckbrief CFT 20-R mit WS/ZF-R). Eine Faktorenanalyse mit den beiden Testteilen des CFT 20-R, dem Wortschatz- und dem Zahlenfolgentest als Variablen, ergab ein für den Zahlenfolgentest problematisches Ergebnis (Weiß 2008): Dieser Test wies mit .65 eine hohe Ladung auf dem CFT-Faktor auf – ein Hinweis, dass die fluide Intelligenz stark in die Testleistung einfließt und die Leistung in diesem Zusatztest offenbar weniger bildungsabhängig ist als gewünscht.

Korrelation mit Mathematik- und Deutschnote

Misst intellektuelles Potenzial

Bewertung Der eigentliche Wert des Verfahrens liegt darin, die grundlegende intellektuelle Leistungsfähigkeit im Sinne der fluiden Intelligenz relativ unabhängig von kultur- bzw. schichtspezifischen Einflüssen und Schulkenntnissen prüfen zu können. Dadurch ist es zur Untersuchung der Intelligenz von Kindern mit verminderten Kenntnissen der Testsprache geeignet. Allerdings erfolgt die Instruktion notwendigerweise in der Sprache, die die Testleiterin bzw. der Testleiter beherrscht. Deshalb ist es sehr zu begrüßen, dass für die Auflage von 2019 eine Computerversion u. a. auch in arabischer Sprache vorliegt. Die Ergebnisse im CFT 20-R können im Einzelfall helfen, das intellektuelle Potenzial von Testpersonen mit niedrigen Testwerten in bildungsabhängigen Intelligenztests und/oder schlechten Schulleistungen einzuschätzen. Ob die beiden Zusatztests (Wortschatz- und Zahlenfolgentest) eine sinnvolle Ergänzung zur Erfassung der kristallisierten Intelligenz darstellen, kann bezweifelt werden. Die Stärke des Tests liegt darin, dass 4 unterschiedliche Aufgabentypen eingesetzt werden, um eine Fähigkeit zu messen. Den Zusatztests liegt dieses Prinzip nicht zugrunde. Die verbale und numerische Intelligenz werden mit jeweils nur einem einzigen Test abgedeckt. Für viele Fragestellungen wird es vorteilhaft sein, sprachliche Fähigkeiten und die Beherrschung der Grundrechenarten mit entsprechenden Schulleistungstests (► Abschn. 7.4.1.1) zu messen.

In einer Testrezension von Gruber und Tausch (2016) erfährt der CFT 20-R (Weiß 2008) insgesamt eine gute Bewertung. Die Autorinnen loben die einfache Handhabbarkeit, die genauen Instruktionshinweise, die Ökonomie und die Verfügbarkeit verschiedener Versionen. Als problematisch werten sie, dass keine aktuellen Angaben zur prognostischen Validität vorliegen.

■ Alternativen zum CFT 20-R

CFT 1-R für jüngere Kinder

Soll ein Kind im Altersbereich von 5;3 bis 9 bzw. 6;6 bis 12 Jahren untersucht werden, bietet sich aus der gleichen Testfamilie der Grundintelligenztest Skala 1 – Revision (CFT 1-R) von Weiß und Osterland (2012) an. Dieser Test wurde 2010 an 4641 Kindern normiert. Den Anspruch, die Intelligenz weitgehend sprachfrei und bildungsunabhängig messen zu können, erheben aber auch andere Testverfahren. Ist bei einem Kind eine Einzeltestung vorgesehen, kommen z. B. auch sprachfreie Untertests der K-ABC (s. o.) infrage. Darüber hinaus stehen im deutschen Sprachraum 2 weitere Verfahren – nicht nur zum Einsatz bei Kindern – zur Verfügung, die nun kurz mit dem CFT 20-R kontrastiert werden.

Nur ein Itemtyp verwendet

Raven's Progressive Matrices Unter dem Überbegriff „Raven“ lassen sich verschiedene Tests finden, die anders als die CFT-Tests nur aus Matrizenaufgaben bestehen. Obwohl bei den Raven-Matrizen und den CFT-Tests figurale, sprachfreie Aufgaben eingesetzt werden, die schlussfolgerndes Denken erfordern, bezieht sich Raven auf Spearmans g-Faktor und Cattells CFT auf die fluide Intelligenz. Dieser Unterschied erklärt sich aus den unterschiedlichen Forschungstraditionen der Testautoren und ist für die diagnostische Praxis allenfalls etwas irritierend.

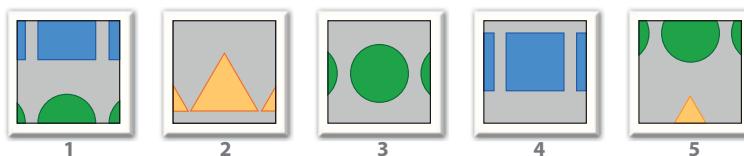
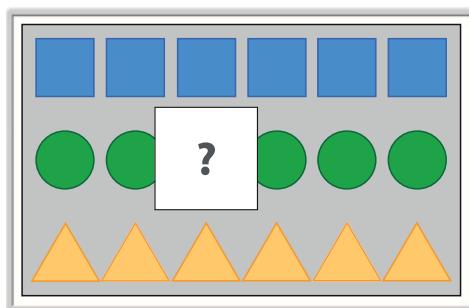
Die Raven-Matrizen-Tests unterscheiden sich in ihrer Schwierigkeit und ihrem Einsatzbereich. Wie der CFT haben die Verfahren eine lange Tradition; die Erstveröffentlichung stammt aus dem Jahr 1938. Eine weitere Gemeinsamkeit ist das Vorliegen von internationalen Forschungsarbeiten zu den Tests. Wegen ihrer sehr alten Normen kommen die deutschen Tests hauptsächlich bei Forschungszwecken zum Einsatz. Die Coloured Progressive Matrices (CPM) von Raven et al. (2002) wurden für das Alter von 3;9

Verschiedene Varianten der Matrizen-Tests

bis 11;8 Jahren entwickelt; die Normen stammen aus 2 Erhebungen aus den Jahren 1998 und 1999 in Deutschland und Frankreich. Die Advanced Progressive Matrices (APM) von Raven et al. (1998) dienen der Diagnostik von Jugendlichen ab 12 Jahren und von Erwachsenen; die Normen stammen aus dem Jahr 1997. Die Standard Progressive Matrices (SPM) von Raven (2009) sind für Kinder im Schulalter, Jugendliche und Erwachsene konzipiert; die Normen stammen aus den Jahren 1998 und 1999. Bei dem Testanbieter Schuhfried (► <https://www.schuhfried.com/>) finden sich Computerversionen verschiedener Versionen von Raven's Progressiven Matrices im Programm; über das Jahr der Normierung ist auf der Webseite nichts zu erfahren. Der Pearson-Verlag hat 2019 eine deutsche Version der Raven's 2 Progressive Matrices, Clinical Edition (Raven's 2) für den Altersbereich von 4;0 bis 69;11 Jahren publiziert (Pearson Clinical Assessment Deutschland 2019). Der Test stellt eine Neuentwicklung auf Basis der APM, CPM und SPM dar. Die Items sind farbig (s. □ Abb. 3.14 für ein Itembeispiel). Der Test wurde 2018 bis 2019 an einer Stichprobe von $N=1200$ Personen aus verschiedenen europäischen Ländern normiert. Er ist auch als Computerversion verfügbar.

Bochumer Matrizentest (BOMAT) Eine weitere Familie von Matrizentests wurde in Bochum entwickelt; daher die Namensgebung. Die Tests verwenden den gleichen Typ von Matrizenaufgaben wie der CFT 20-R. Der BOMAT – Standard (Hossiep und Hasella 2010) stellt die neueste Version dar. Er wurde an 3439 Schülerinnen und Schülern zwischen 14 und 20 Jahren an Hauptschulen, Realschulen und Gymnasien normiert. Als Messgegenstand wird das „kognitive Leistungspotenzial“ genannt – eine andere Formulierung für fluide Intelligenz. □ Abb. 3.15 zeigt ein Itembeispiel. Der Test wird für die Auswahl und Beratung von Auszubildenden sowie für die Schullaufbahnberatung empfohlen. Diese Testversion zeichnet sich durch einen umfangreichen Übungsteil aus; für die Instruktion und die Übung sind 15 min zu veranschlagen, während der eigentliche Test mit seinen 30 Aufgaben 30 min dauert.

Für Schullaufbahnberatung
sowie Auswahl und Beratung von
Auszubildenden



□ Abb. 3.14 Itembeispiel aus den Raven's Progressive Matrices 2, Clinical Edition (Ravens's 2). (Copyright © 2018 NCS Pearson, Inc. Deutsche Fassung Copyright © 2019 NCS Pearson, Inc. Alle Rechte vorbehalten. Übersetzung, Adaptation und Produktion durch Pearson Deutschland GmbH, Frankfurt am Main, mit freundlicher Genehmigung und Lizenz der NCS Pearson, Inc.)

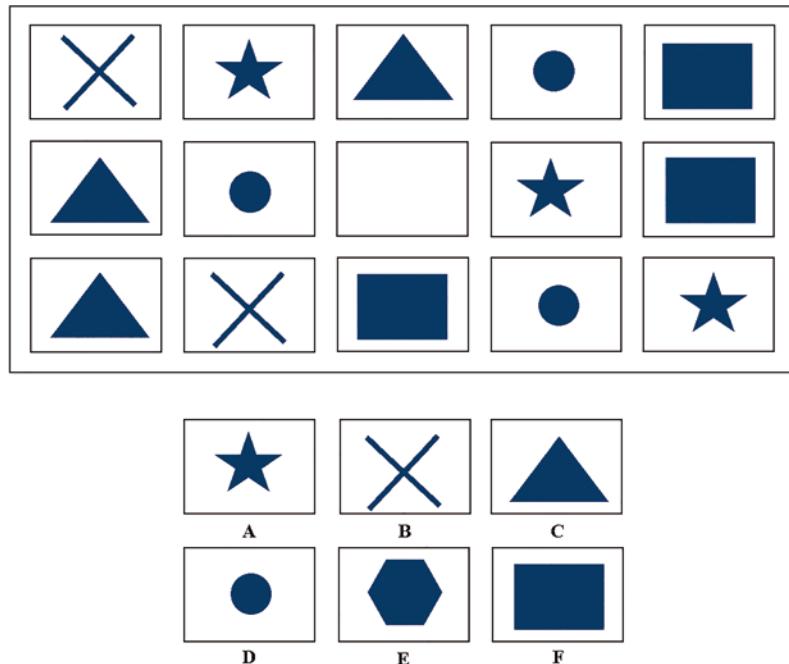


Abb. 3.15 Item aus BOMAT – Standard. (Aus Hossiep und Hasella 2010, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4,37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

BOMAT – advanced und dessen Kurzform für den oberen Intelligenzbereich

Dem BOMAT – Standard ähnlich ist der BOMAT– advanced (Hossiep et al. 1999). Der Test wurde für den oberen Intelligenzbereich konstruiert und an 303 Studierenden und Absolventinnen bzw. Absolventen von Universitäten und Fachhochschulen normiert. Aus dem Itempool des BOMAT wählten der Autor und die Autorinnen Items für die Kurzform, den BOMAT – advanced – short version, aus (Hossiep et al. 2001). Es existieren 2 Parallelformen mit je 29 Items (eines davon dient nur dem Einstieg und wird nicht gewertet). Die reine Bearbeitungszeit beträgt 45 gegenüber 80 min beim BOMAT – advanced. Normiert wurde der Test an 668 Hochschülern/-schülerinnen und (Fach-)Hochschulabsolventen/-absolventinnen. In einer Testrezension resümiert Fay (2003), dass der BOMAT – advanced – short version in seiner Schwierigkeit besser als der BOMAT auf die Zielgruppe abgestimmt, ansprechend aufgemacht und sorgfältig konstruiert ist.

3.2.4 Spezielle Fähigkeitstests

Neben den zuvor vorgestellten allgemeinen Leistungstests und den Intelligenztests liegt noch eine Vielzahl anderer Leistungstests vor. Da sie ein breites Spektrum von Fähigkeiten abdecken, stellen wir nachfolgend ein Intelligenzstrukturmodell vor, das bei der Einordnung dieser Tests helfen kann. Auch die bereits vorgestellten allgemeinen Leistungs- und Intelligenztests lassen sich in breite Intelligenzstrukturmodelle wie beispielsweise das Carroll-Modell (Carroll 1993, 1996) oder das daraus weiterentwickelte Carroll-Horn-Cattell-Modell (CHC-Modell; McGrew 2005) einordnen.

3.2.4.1 Intelligenzstrukturmodelle zur Einordnung von kognitiven Leistungstests

Zur Einordnung von kognitiven Leistungstests bietet sich unter anderem das Modell von Carroll (1993) an. Es handelt sich dabei um ein hierarchisches Modell mit der Allgemeinen Intelligenz g an der Spitze. Auf der darunterliegenden Ebene unterscheidet Carroll zwischen 8 Fähigkeitsbereichen, die er nach ihrer Nähe zur Allgemeinen Intelligenz ordnet. Den höchsten Zusammenhang mit g weist die fluide Intelligenz auf, den niedrigsten die Verarbeitungsgeschwindigkeit. Jeder Fähigkeitsbereich wird durch mehrere Kategorien von Tests spezifiziert. Beispielsweise ordnet Carroll (1993) Tests zum schlussfolgernden Denken der fluiden Intelligenz zu. McGrew (2005) hat in einem viel beachteten Beitrag versucht, das Strukturmodell von Carroll mit Modellen von Cattell und Horn (s. McGrew 2005) zu vereinen. Ergebnis ist das CHC-Modell. In □ Tab. 3.9 sind das Carroll- und das CHC-Modell dargestellt.

Carroll-Modell als Ordnungsschema

Leistungstests lassen sich in vielen Fällen zumindest vorläufig in die Taxonomie des Carroll- und/oder des CHC-Modells einordnen. Allerdings liegt ein Problem darin, dass sich Testautorinnen und Testautoren selten auf das

□ Tab. 3.9 Fähigkeitsbereiche (Ebene II) nach Carroll (1993) und im CHC-Modell zur Einordnung von kognitiven Leistungstests

Carroll-Modell	CHC-Modell	Erläuterung zum CHC-Modell
2 F Fluide Intelligenz	Gf Fluide Intelligenz	Absichtsvolle kontrollierte mentale Prozesse zum Lösen von Problemen einschließlich induktivem und deduktivem Schlussfolgern, Hypothesen generieren und prüfen ...
2 C Kristalline Intelligenz	Gc Verstehen-Wissen	Erworbenes deklaratives und prozedurales Wissen und dessen Anwendung
2Y Allgemeines Gedächtnis und Lernen	Gsm Kurzzeitgedächtnis	Fähigkeit, Informationen aufzunehmen und kurzzeitig zu speichern
2 V Breite visuelle Wahrnehmung	Gv Fähigkeit zur Verarbeitung visueller Informationen	Fähigkeit, visuelle Eindrücke und Vorstellungen zu generieren, zu speichern, abzurufen und zu transformieren
2U Breite auditive Wahrnehmung	Ga Fähigkeit zur Verarbeitung auditiver Informationen	Fähigkeit, akustische Informationen kontrolliert aufzunehmen, zu verarbeiten (u. a. Muster zu erkennen) und Tonelemente zu erzeugen
2R Breite Wiedergabefähigkeit	Glr Langzeitgedächtnis und Abruf	Fähigkeit, neue Informationen einzuspeichern, im Langzeitgedächtnis zu behalten und flüssig abzurufen
2 S Breite kognitive Schnelligkeit	Gs Verarbeitungsgeschwindigkeit	Einfache kognitive Aufgaben schnell lösen
2 T Verarbeitungsgeschwindigkeit (Entscheidungsgeschwindigkeit)	Gt Allgemeine kognitive Schnelligkeit	Auf einfache Reize schnell (und richtig) reagieren
–	Gq Quantitatives Wissen	Erworbenes quantitatives und mathematisches Wissen
–	Grw Lesen und Schreiben	Erworbenes Wissen zum Lesen und Schreiben
<i>Vorläufige, nicht hinreichend gesicherte Faktoren</i>		
–	Gkn Allgemeines (bereichsspezifisches) Wissen	
–	Gh Taktile Fähigkeiten	
–	Gk Kinästhetische Fähigkeiten	
–	Go Olfaktorische Fähigkeiten	
–	Gp Psychomotorische Fähigkeiten	
–	Gps Psychomotorische Schnelligkeit	

Andere Tests in Carroll- oder BIS-Modell einordnen

Modell von Carroll beziehen und von den Bezeichnungen in dem Modell von Carroll oder von anderen Modellen abweichende Testnamen wählen. Testnamen, oft sogar die Ausführungen zur konzeptuellen Einordnung eines Tests, tragen daher manchmal mehr zur Verwirrung als zur Klärung der Frage bei, welche Fähigkeit mit dem Test erfasst werden soll.

Die bisher vorgestellten Intelligenztests fallen z. B. in den Bereich der fluiden Intelligenz (z. B. CFT 20-R) oder stellen mit ihren Subtests einen Mix aus Tests zu den Bereichen fluide Intelligenz (z. B. mehrere Subtests des I-S-T 2000 R), kristallisierte Intelligenz (z. B. die Wissenstests im I-S-T 2000 R) oder Gedächtnis und Lernen (z. B. Zahlen nachsprechen in der WISC-V) dar.

Weiterführende Literatur

Weitere, in diesem Buch nicht vorgestellte Tests lassen sich konzeptuell in dieses Modell und/oder in das Berliner Intelligenzstrukturmodell (BIS-Modell, □ Abb. 3.15) einordnen. Süß und Beau-ducel (2011) haben für ausgewählte Intelligenztests und deren Untertests den Versuch unternommen, sie in die 3 genannten Modelle (Carroll-, CHC- und BIS-Modell) einzuordnen. In diesem Beitrag werden zudem die zuvor genannten und weitere Intelligenzmodelle ausführlicher erläutert. Bei Mickley und Renner (2019) findet sich eine Einordnung von Intelligenztests für Kinder und Jugendliche in das CHC-Modell.

3.2.4.2 Spezielle Intelligenztests

Manchmal ist nicht sofort ersichtlich, wohin ein Test gehört. Wie bereits erwähnt, verrät der Testname oftmals nicht, was eigentlich gemessen wird.

Test zur Praktischen Alltagsintelligenz (PAI 30) Nehmen wir als Beispiel den PAI 30 von Mariacher und Neubauer (2005). Die Autorin und der Autor verweisen auf die allgemeine Beobachtung, dass sich kluge, intelligente Menschen in praktischen Angelegenheiten manchmal ungeschickt anstellen, und dass es umgekehrt wenig intelligente Menschen gibt, die sich sehr effektiv mit ihrer materiellen Umwelt auseinandersetzen. Sie vermuten, dass die „praktische Alltagsintelligenz“ in der Regel nicht durch eine bestimmte Ausbildung, sondern durch die Auseinandersetzung mit der eigenen Lebensumwelt erworben wird. Aber auch viele Berufsausbildungen würden die praktische Alltagsintelligenz schulen. Damit liegt der Verdacht nahe, dass der Test einen bestimmten Bereich der kristallisierten Intelligenz erfasst, die nicht als reines Wissen verstanden werden sollte, sondern als die Fähigkeit, Probleme auf der Basis von Wissen zu lösen.

Die Aufgaben des PAI 30 sehen wie folgt aus: In jeder Aufgabe wird ein Problem geschildert, das zumeist durch ein Foto oder eine Skizze veranschaulicht wird. Die Testperson soll eine Lösung finden und die Antwort im Antwortheft entweder durch Ankreuzen einer Antwortalternative oder durch freie, stichwortartige Beschreibung der Lösung eingetragen (s. Beispielaufgabe).

Itembeispiel (Übungsaufgabe) aus dem PAI

Sie versuchen, mit einem Löffel Speiseeis aus einer großen Eispackung in Röllchenform abzurollen. Das Eis bleibt jedoch am Löffel kleben und bricht, wenn Sie stärker andrücken, sodass Sie keine Röllchen formen können. Was unternehmen Sie, damit Sie das Eis doch in Röllchenform abheben können?

Zur Verfügung steht gewöhnliches Kücheninventar.

Lösung: „Den Löffel in Wasser tauchen.“

Anmerkung: Wird zwischen Löffel und Eis ein Wasserfilm aufgebracht, verringert sich die Haftung zwischen Eis und Löffel, sodass das Eis nicht am Löffel kleben bleibt.

(© Hogrefe. Ein weiteres Beispiel findet sich in ▶ Abschn. 2.6.5.5.)

Die Validitätsbefunde zeigen eine enge Korrelation mit einem Test zum technischen Verständnis. Auch technisches Verständnis kann konzeptuell der kristallisierten Intelligenz zugeordnet werden, da es kultur- oder bildungsabhängig ist.

Die Testaufgaben verlangen überwiegend die freie Beantwortung von Aufgaben bei Verarbeitung von figuralem oder auch von sprachlichem Material wie im Beispiel. Deshalb könnte der Test auch im Berliner Intelligenzstrukturmodell (► Abb. 3.12) der Operation Einfallsreichtum und hier den Facetten „figural-bildhaft“ und „verbal“ zugeordnet werden.

Das Beispiel PAI 30 zeigt, dass nicht unbedingt eine neue Intelligenz postuliert werden muss. Vielmehr können etablierte Intelligenzstrukturmodelle einen Platz dafür bieten.

Was misst der Test zur Praktischen Alltagsintelligenz PAI 30?

3.2.4.3 Weitere Intelligenzkonstrukte

Das Wort „Intelligenz“ wird auch gelegentlich für Fähigkeiten oder Fertigkeiten verwendet, deren Bezug zum etablierten Intelligenzkonstrukt schwer nachvollziehbar ist. Gardner (2002) hat mit seinen „multiplen Intelligenzen“ ein Tor aufgestoßen, indem der scheinbar neue Intelligenzen „entdeckt“ hat. Ein Beispiel für eine seiner bisher 8 speziellen Intelligenzen ist die „musikalisch-rhythmische Intelligenz“, die etwa bei Musikerinnen/Musikern und Komponistinnen/Komponisten oft in hoher Ausprägung anzutreffen sei. Fähigkeiten oder auch (erlernte) Fertigkeiten, die in irgendeinem Lebensbereich zu Erfolg führen, werden so als Intelligenz geadelt.

„Multiple Intelligenzen“ nötig?

Zu diesen Spezialintelligenzen gehört auch die „emotionale Intelligenz“, die jedoch nicht auf Gardner (2002) zurückgeht, sondern durch ein populärwissenschaftliches Buch von Goleman (1995) in den USA und später auch weltweit bekannt wurde. In der Folge war eine rege Forschungstätigkeit zu verzeichnen. Es wurden diverse Tests entwickelt, die zum Teil Leistungstestcharakter haben und zum Teil eher den Persönlichkeitsfragebögen zuzuordnen sind. Das international wohl bekannteste Verfahren, der Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz (MSCEIT), ein Leistungstest, wurde auch in einer deutschsprachigen Version publiziert (Steinmayr et al. 2011). Es soll die emotionale Intelligenz erfassen, die als Fähigkeit verstanden wird, Emotionen zielführend in soziale und analytische Problemlöseprozesse einzubeziehen.

Emotionale Intelligenz

3.2.4.4 Motoriktests

Über den engeren kognitiven Leistungsbereich hinaus lassen sich etwa im sonderpädagogischen, sportpsychologischen oder Berufseignungsbereich Merkmale finden, die nicht direkt der Intelligenz zugeordnet werden. Hier ist besonders die Motorik zu nennen. Für bestimmte diagnostische Fragestellungen ist eine Beurteilung von motorischen Fähigkeiten und Fertigkeiten wichtig. Beispielsweise soll in der Neuropsychologie manchmal die motorische Beeinträchtigung quantifiziert werden, die ein Patient etwa durch eine Schädel-Hirn-Verletzung erlitten hat. Mithilfe computerbasierter Alertness-Tests, die Bestandteil von Testbatterien zur Aufmerksamkeitsprüfung (► Abschn. 3.2.2) sein können, lässt sich die psychomotorische Reaktionsschnelligkeit messen.

Psychomotorische Reaktionsschnelligkeit

In der beruflichen Eignungsdiagnostik, Reha-Diagnostik und in der Sportpsychologie können Anforderungsanalysen auf die Notwendigkeit hinweisen, bestimmte fein- oder grobmotorische Fähigkeiten zu prüfen. Für solche Zwecke stehen zahlreiche Testverfahren zur Verfügung, die so unterschiedliche Aspekte der Motorik wie Tremor, Zweihand- oder Körperkoordination messen. Mit der Bezeichnung „Psychomotorik“ wird darauf hingewiesen, dass die Motorik vom Gehirn gesteuert wird. Aber auch organische

Tests zu unterschiedlichen Aspekten der Motorik

Faktoren wie eine Lähmung, eine Verletzung oder eine körperliche Behinderung können sich auf die motorische Leistungsfähigkeit auswirken. Im sonderpädagogischen Bereich kann eine Fragestellung lauten, wie stark ein Kind motorisch eingeschränkt ist. Ein standardisiertes Testverfahren wie der Körperkoordinationstest für Kinder (KTK) von Kiphard und Schilling (2017) erlaubt es, verschiedene Aspekte – hier der Körperkoordination – zu quantifizieren.

Weiterführende Literatur

Aus Platzgründen ist es nicht möglich, eine detaillierte Übersicht über diese Verfahren zu geben oder einzelne Tests vorzustellen. Stattdessen wird auf das von Bös (2017) herausgegebene, umfangreiche Handbuch verwiesen, in dem über 300 Verfahren zur Beurteilung der Motorik beschrieben werden.

3.2.5 Entwicklungstests

Entwicklungsretardierungen erkennen

Mit Entwicklungstests soll festgestellt werden, ob sich ein Kind allgemein oder in einem speziellen Bereich altersgemäß entwickelt hat. Solche Tests sollten deshalb Items aufweisen, die vor allem mit dem Lebensalter hoch korrelieren, also beispielsweise zwischen benachbarten Altersstufen differenzieren, wie dieses bei den Binet-Tests der Fall ist (Stemmler et al. 2016). Durch Vergleich der individuellen Ergebnisse mit den Leistungen Gleichaltriger lassen sich Hinweise auf eventuell behandlungsbedürftige Entwicklungsverzögerungen finden.

Diese Anwendung setzt voraus, dass die eingesetzten Verfahren nicht nur für ein bestimmtes Lebensalter normiert sind, sondern auch Normwerte für längere Entwicklungsperioden bereitstellen. Mit allgemeinen Entwicklungstests soll der kindliche Entwicklungsstand dabei breit erfasst werden, spezielle Entwicklungstests fokussieren dagegen nur einen Ausschnitt. Intelligenztests für Kinder, die bereits in ► Abschn. 3.2.3 behandelt wurden, können grundsätzlich auch zur Entwicklungsdiagnostik eingesetzt werden, wenn die allgemeine kognitive Entwicklung oder die Entwicklung von intellektuellen Teilsfähigkeiten zu beurteilen ist.

Wenn keine anderen Verfahren zur Verfügung stehen, wird in der Praxis gelegentlich der Entwicklungsstand von Erwachsenen mit einer Intelligenzminderung sowie älteren Kindern und Jugendlichen mit Tests eingeschätzt werden, die nur für jüngere Kinder entwickelt und normiert sind.

Entwicklungsalter einer Testperson schätzen

Unter dem Entwicklungsalter versteht man das Alter, in dem eine durchschnittliche Testperson die gleiche Testleistung wie die tatsächliche Testperson erzielt. Man schaut in den Normtabellen nach, in welcher Altersgruppe der tatsächlich ermittelte Testrohwert zu einem durchschnittlichen Normwert (bei Standard- und IQ-Werten also 100) führen würde. Das Alter dieser Normgruppe entspricht dem aktuellen Entwicklungsalter der Testperson. Falls sich für den ermittelten Testrohwert keine Altersgruppe findet, in der dieser Wert zu einem exakt durchschnittlichen Normwert führt, kann man das Entwicklungsalter durch Interpolieren finden. Liegt der Testrohwert beispielsweise genau zwischen den durchschnittlichen Rohwerten der 5- und der 6-Jährigen, kann das Entwicklungsalter auf 5 Jahre und 6 Monate geschätzt werden.

► Beispiel

Ein 17-Jähriger erreicht in einem Intelligenztest für Kinder 65 Punkte. In den Normtabellen sucht man die Altersgruppe, in der ein IQ von 100 bei 65 Punkten zuerkannt wird. Die Diagnostikerin bzw. der Diagnostiker stellt fest, dass dies in der Altersgruppe von 9;0 bis 10;0 Jahren der Fall ist. Die Schlussfolgerung lautet, dass die Testperson in diesem Test den Leistungsstand einer durchschnittlichen 9-jährigen Testperson gezeigt hat. Ihr Entwicklungsalter oder hier konkreter ihr Intelligenzalter beträgt somit 9 Jahre.

3.2.5.1 Allgemeine Entwicklungstests

Einige Entwicklungstests dienen als Breitbanddiagnostikum und erfassen mit ihren Subtests mehrere Entwicklungsbereiche. Gerade bei jüngeren Kindern ist die aktive Bearbeitung von Aufgaben schwer realisierbar. Deshalb wird bei Entwicklungstests auch mit einer (meist systematischen) Verhaltensbeobachtung durch die Testleiterin oder den Testleiter gearbeitet.

Griffiths-Entwicklungsskalen (GES) zur Beurteilung der Entwicklung in den ersten beiden Lebensjahren

Obwohl die GES schon sehr alt sind und die Normen dringend einer Überprüfung bedürfen, stellen wir das Verfahren vor. Erstens handelt es sich um einen Klassiker unter den Entwicklungstest für den frühkindlichen Bereich, zweitens ist der Ansatz deutlich von dem klassischen Leistungstests verschieden und drittens wird das Verfahren zumindest in Großbritannien weiter gepflegt (s. u.).

Die GES in der deutschen Version von Brandt und Sticker (2001) gehen auf die Griffiths Mental Development Scale zurück, die erstmalig von Griffiths (1954) unter dem Titel *The Abilities of Babies* veröffentlicht wurde. Diese Version kam in Deutschland zwischen 1967 und 1979 im Rahmen einer Längsschnittstudie zur Entwicklung von Früh- und Reifgeborenen zum Einsatz. Neben den 257 Items der Originalversion wurden 102 Zusatzaufgaben erprobt. Die Ergebnisse dieser Studie lagen der 1983 erschienenen 1. deutschen Version zugrunde. Dabei fanden auch die Kürzungen Berücksichtigung, die Griffiths in einer 1970 erschienenen Überarbeitung des Tests vorgenommen hatte. Die Autorinnen der deutschen Fassung bemühten sich um eine möglichst enge Anlehnung an die englische Originalfassung und nahmen nur unbedingt erforderliche Änderungen vor.

Für die ersten beiden Lebensjahre

Mit den Griffiths Scales of Child Development, Third Edition (Griffith III) von Green et al. (2015) steht ein englischsprachiger Nachfolgetest zur Verfügung. Diese Entwicklungsskalen decken sogar die ersten 6 Lebensjahre ab und wurden in Großbritannien und Irland normiert. Ein kurzer Film, der über die Griffith III informiert, ist beim Hogrefe Verlag abrufbar unter: ► <https://www.hogrefe.co.uk/shop/griffiths-scales-of-child-development-third-edition.html>.

Griffiths III nur als englische Version verfügbar

Gliederung Die insgesamt 208 Aufgaben der GES erlauben die Untersuchung folgender Bereiche: Motorik, sozialer Kontakt, Hören und Sprechen, Auge-Hand-Koordination und kognitive Entwicklung. Jede der 5 Skalen misst einen eigenen Entwicklungsbereich und kann auch separat durchgeführt werden. In ▶ Tab. 3.10 werden Aufgabenbeispiele aufgeführt. Die Aufgabennummer informiert über die Position des Items in der Skala. Aufgaben mit 2 Nummern (z. B. A 31/32) werden mit 2 Punkten bewertet. Als „Normalbereich“ wird hier der Altersbereich bezeichnet, in dem 90 % der Kinder das instruierte Verhalten zeigen; bei jeweils 5 % der Kinder ist das instruierte Verhalten folglich schon früher bzw. erst später zu beobachten (5. bis 95. Perzentil).

5 Subtests

Tab. 3.10 Skalen und Itembeispiele der GES. (Brandt und Sticker 2001, © Hogrefe)

Skala	Itembeispiele (Nr.)	Normalbereich (Monate)	Median (Monate)
Motorik	Sitzt frei, mindestens eine Minute (A 14)	6–10	7,3
	Kann rückwärts gehen (A 31/32)	14–19	15,4
Persönlich-Sozial	Lächelt (B 3) Unterscheidet Fremde von Bekannten (B 14)	1–3 5–9	2,0 7,0
Hören und Sprechen	Reagiert, wenn es gerufen wird (C 12)	4–8	6,0
	Sagt Mama oder Papa klar bzw. ein anderes Wort (C 17/18)	7–15	9,0
Auge und Hand	Nimmt den Ring, den man reicht (D 7)	2–5	3,5
	Vollständiger Pinzettengriff (D 19)	9–12	10,1
Leistungen	Hält den runden Holzstab für einige Sekunden (E 6)	1–5	3,2
	Findet das versteckte Spielzeug unter der Tasse (E 21)	7–13	9,7

Standardisiertes Testmaterial in Form von Spielsachen

Durchführung Das Verfahren ist ein Individualtest für die ersten beiden Lebensjahre und soll in Gegenwart einer vertrauten Person durchgeführt werden, die notfalls die Testleiterin bzw. den Testleiter unterstützen kann. Für die Durchführung wird standardisiertes Testmaterial (z. B. ein kleiner, rot lackierter runder Holzstab, eine Schachtel mit 12 Spielsachen) benötigt. Die Untersuchung beginnt normalerweise mit Aufgaben, die etwa 2 Monate unter dem Lebensalter des Kindes liegen, und wird beendet, wenn mehr als 2 aufeinanderfolgende Aufgaben in einem Untertest nicht mehr gelöst werden. Wenn aus der Beobachtung des Kindes bekannt ist, dass es eine bestimmte Aufgabe lösen kann, braucht diese nicht durchgeführt werden. Bei den Aufgaben mit Testmaterial finden sich genaue Anweisungen zum Vorgehen. Die reine Durchführungszeit ist bei den meisten Kindern im 1. Lebensjahr mit 20–30 min zu veranschlagen, bei älteren Kindern mit etwa 45 min.

Entwicklungsalter

Auswertung Eine Aufgabe gilt als gelöst, wenn die Bewertungskriterien erfüllt sind. Für einige Aufgaben werden 2 Punkte vergeben. Die Skalen sind so aufgebaut, dass für jeden Lebensmonat 2 Aufgaben bzw. eine mit 2 Punkten bewertete Aufgabe vorliegen. Das Entwicklungsalter eines Kindes lässt sich daher relativ einfach feststellen, indem die erreichte Punktzahl durch 2 dividiert wird. Für den Gesamttest (5 Bereiche) ist die Summe der gelösten Aufgaben durch 10 zu dividieren. Erreicht ein Kind beispielsweise insgesamt 125 Punkte, hat es ein Entwicklungsalter von 12,5 Monaten. Das Entwicklungsalter ist in Relation zum Lebensalter (z. B. 18 Monate) zu sehen. Ein Entwicklungsquotient kann berechnet werden, indem das Entwicklungsalter durch das Lebensalter dividiert und das Ergebnis mit 100 multipliziert wird. Im Beispiel ergibt sich ein Entwicklungsquotient von 69: $(12,5/18) \times 100 = 69,44$.

Retest-Reliabilität altersabhängig

Reliabilität Die Retest-Reliabilität des Gesamtentwicklungsquotienten variiert bei einer Testwiederholung nach 3 Monaten zwischen .49 (Alter bei der 1. Messung: 3 Monate) und .81 (Alter: 15 Monate). Im Durchschnitt liegen die Koeffizienten im 2. Lebensjahr mit .80 höher als im 1. Lebensjahr (.62).

Validität Im Manual wird lediglich auf die Entwicklungsprofile verschiedener behinderter Kinder verwiesen, die den Erwartungen entsprechen.

Wenige Validitätsbelege

Normierung Die deutsche Normierung fand im Rahmen einer Längsschnittstudie zwischen 1967 und 1979 an 102 Kindern statt, die zunächst in Monatssintervallen und später in größeren Abständen wiederholt untersucht wurden. Normen in dem Sinne, dass Testwerte in Standardwerte transformiert werden, existieren nicht. Die Erhebung diente dazu, für jede Aufgabe das Alter zu ermitteln, in dem 50% der Kinder sie lösten. Mit der oben beschriebenen Auswertungsprozedur wird die individuelle Testleistung jedoch mit den Leistungen anderer Kinder verglichen.

Altersgruppe suchen

Bewertung Bei den GES handelt es sich um ein sehr sorgfältig konstruiertes Verfahren. Dennoch erscheint eine umfangreichere deutsche Nacheichung sowie die Ermittlung eigener Werte zur Abklärung der Validität der deutschen Version geboten.

Sorgfältig konstruiert

Alternativen zu den GES

Für die ersten 2–3 Lebensjahre sind nur wenige andere Verfahren neueren Datums verfügbar.

Bayley Scales of Infant and Toddler Development – Third Edition (Bayley-III) Die Bayley-III sind ein Verfahren mit langer Tradition. Die amerikanische Psychologin Nancy Bayley hatte seit 1928 zur frühkindlichen Entwicklung geforscht. Drei von ihr entwickelte Tests fasste sie zu einem Verfahren zusammen und veröffentlichte dieses 1969, kurz nachdem sie in Ruhestand gegangen war. Die 3. Auflage der Skalen stammt von 2006 und ist Grundlage für die deutsche Version (Bayley 2014). Die Skalen wurden für Kinder im Alter von 1 bis 42 Monaten entwickelt und auch normiert. Die Eichstichprobe setzt sich aus $N=878$ deutschen Kindern zusammen, die im unteren Altersbereich um niederländische Säuglinge ($N=131$) ergänzt wurde.

Traditionsreiches Verfahren in deutscher Version

Die Bayley-III umfassen 3 Skalen mit insgesamt 5 Untertests zu Kognition, Sprache (rezeptiv, expressiv) und Motorik (Feinmotorik, Grobmotorik) mit 324 Items. Das Besondere an dem Verfahren ist, dass die 5 Untertests für alle Altersbereiche aus den gleichen Aufgaben bestehen. Diese sind nach ihrer Schwierigkeit geordnet. Einstiegsregeln legen fest, mit welchen Aufgaben bei einem bestimmten Alter des Kindes zu beginnen ist. Geregelt ist auch, wann bei Bedarf auf leichtere Aufgaben zurückgegangen wird. Ein Untertest wird beendet, wenn das Kind 5 aufeinanderfolgende Aufgaben nicht gelöst hat. Somit muss ein Kind nie alle Items bearbeiten, sondern nur einen mehr oder weniger kleinen Teil. Die Bearbeitungsdauer liegt bei etwa 50 min im 1. Lebensjahr und etwa 90 min bei älteren Kindern. Daneben kann auch der Bayley-III-Screening-Test als Kurzform mit weniger Items und entsprechend kürzeren Bearbeitungszeiten (15–20 bzw. ca. 30 min) durchgeführt werden. Die Gesamtpunktzahl für einen Untertest ergibt sich aus den gelösten Aufgaben plus den einfacheren Aufgaben, die dem Kind aus Altersgründen erst gar nicht vorgegeben wurden. Für die Ermittlung von Normwerten werden die Rohpunktsummen in Altersnormen transformiert. Die Ergebnisse liegen auch grafisch als Profil über die 3 Skalen und die 5 Untertests vor.

Items für alle Kinder – nur ein Teil wird bearbeitet

Positive Bewertung in Rezension

In einer Rezension loben Macha und Petermann (2015) die solide Konstruktion und die sehr umfangreichen Validierungsstudien zur englischsprachigen Version, die im deutschen Testmanual zusammenfassend dargestellt werden.

3

Testaufgaben und Elternfragebogen

Frühkindliches Entwicklungsdiagnostikum für Kinder von 0–3 Jahren (FREDI 0–3) Im FREDI 0–3 von Mähler et al. (2016) werden wie in den Griffiths-Entwicklungsmaßnahmen standardisierte Testaufgaben mit kindgerechtem Testmaterial und Verhaltensbeobachtung als Methoden der Leistungsbeurteilung gemeinsam herangezogen. Hinzu kommt ein Elternfragebogen. Testitems und Items aus dem Elternfragebogen werden jeweils zu einem Gesamtwert verrechnet (z. B. gibt es für den Altersbereich 2–3 Monate 14 Testitems und 8 Fragebogenitems zur Motorik; zur Kognition und zur Sprache liegen je 3 und 3 und zur sozial-emotionalen Entwicklung 1 und 7 Items vor). Das Verfahren wurde auf der Grundlage der einschlägigen Forschung zur frühkindlichen Entwicklung aufgebaut und stellt eine deutsche Neuentwicklung dar.

Für die Testaufgaben dienten andere Entwicklungstests als Vorbild. Das FREDI 0–3 liefert Informationen über 4 Entwicklungsbereiche, nämlich Motorik, Kognition, Sprache und die sozial-emotionale Entwicklung. Je nach Alter des Kindes kommen zum Teil andere Aufgaben zum Einsatz. Der Elternfragebogen liegt in verschiedenen Versionen für einzelne Altersgruppen vor. Es gibt 14 Altersgruppen, 6 für das 1. Jahr (ab 1. Lebenstag, jeweils 2 Monate umfassend) und insgesamt 8 für das 2. und 3. Lebensjahr (bis Ende 35. Monat, jeweils 3 Monate umfassend). Für jede Altersgruppe existieren eigene Altersnormen ($n=46$ bis 60; $N=717$). Die Ergebnisse liegen in Form eines Entwicklungsprofils mit den 4 Entwicklungsbereichen als Merkmale vor; ein Gesamtwert wird nicht berechnet. Da die Aufgaben von Altersgruppe zu Altersgruppe variieren, kann auch kein Entwicklungsalter bestimmt werden. Es liegt eine Testrezession (Macha und Petermann 2017) und eine Replik zu dieser vor (Mähler 2017).

3.2.5.2 Verfahren für den Altersbereich bis 5 Jahre

Vom Altersbereich schließen sich der Wiener Entwicklungstest und 2 weitere Verfahren an die oben vorgestellten Entwicklungstests an, die im Folgenden kurz vorgestellt werden.

Wiener Entwicklungstest (WET)

Der WET von Kastner-Koller und Deimann (2012) liegt in der 3., überarbeiteten und erweiterten Auflage vor. Für die 3. Auflage wurde der Test nur leicht modifiziert (Subtest zur Erfassung der mathematischen Entwicklung neu konstruiert, Subtest „Muster legen“ um Aufgaben für 5- bis 6-Jährige erweitert). Der Test soll bei Vorschulkindern im Alter von 3;0 bis 5;11 Jahren den Entwicklungsstand in 6 Funktionsbereichen erfassen (► Tab. 3.11).

Die verwendeten Aufgabentypen hatten sich bereits in vorliegenden Entwicklungstests bewährt; zum Teil handelt es sich auch um Neuentwicklungen. Die Entwicklung der Skalen erfolgte auf Grundlage der Probabilistischen Testtheorie.

Altersbereich: 3–5 Jahre

Standardisierte Testmaterialien

Durchführung Der WET wird in einer Einzelsitzung durchgeführt. Die Durchführungszeit beträgt bei Kindern bis 3;6 Jahren ca. 90 min, bei älteren Kindern ca. 75 min. Dabei kommen standardisierte Testmaterialien

Tab. 3.11 Subtests des Wiener Entwicklungstests. (WET, Kastner-Koller und Deimann 2012, © Hogrefe)

Funktionsbereich und Subtest	Messgegenstand	Itembeispiel
Motorik		
Turnen	Grobmotorische Fähigkeiten	Einbeiniges, freihändiges Stehen mit geschlossenen Augen für mindestens 3 s
Lernbär	Feinmotorische Fähigkeiten	Am Teddybär mit einer Kordel (als Halsband) einen Knoten binden
Visuelle Wahrnehmung/Visuomotorik		
Nachzeichnen	Visuomotorische Koordination	Ein Kreuz von einer Vorlage abzeichnen
Bilderlotto	Differenzierte Raum-Lage-Wahrnehmung	Einzelne Kärtchen zum Thema „Meer“ auf einer Bildtafel mit 6 Feldern ordnen
Lernen und Gedächtnis		
Schatzkästchen	Visuell-räumliche Speicherkapazität	Unmittelbar, nach maximal 10 Lerndurchgängen sowie 20 min später 6 verschiedene, in Schubladen versteckte Spielgegenstände wiederfinden
Zahlen merken	Phonologische Speicherkapazität	Vorgesprochene Zahlenfolgen (2 bis maximal 6 Zahlen) sollen unmittelbar nachgesprochen werden
Kognitive Entwicklung		
Muster legen	Räumliches Denken (2-D)	Nach Vorlagen Muster mit Mosaiksteinen nachlegen
Bunte Formen	Induktives Denken (Kreuzklassifikationen)	Matrizenaufgaben: Aus jeweils 5 vorgegebenen Lösungsmöglichkeiten soll das Element bestimmt werden, das eine 3×3 -Matrix sinnvoll ergänzt
Gegensätze	Analoges Denken	Der Satz „Der Papa ist ein Mann, die Mama ist ...“ ist zu ergänzen
Quiz	Orientierung in der Lebenswelt	„Warum sollte man nicht so viele Süßigkeiten essen, wie man gerne möchte?“
Rechnen	Mathematische Entwicklung	Auf der gezeigten Karte sind 3 „Käferfamilien“ mit 2, 3 bzw. 3 Käfern zu sehen. Frage: „Welche Familien haben gleich viele Käfer?“
Sprache		
Wörter erklären	Sprachliche Begriffsbildung	Das Wort „Bilderbuch“ ist zu erklären
Puppenspiel	Verständnis grammatischer Strukturformen	„Die Mutter erlaubt, dass das Mädchen sich hinlegt“ mit Spielmaterial darstellen
Sozial-emotionale Entwicklung		
Fotoalbum	Verständnis mimischer Gefühlsausdrücke	Foto einer Person; Gefühl („Freude“) benennen
Elternfragebogen	Selbstständigkeitsentwicklung des Kindes	„Mein Kind zieht sich ohne Hilfe aus.“

wie etwa ein Lernbär oder ein Schatzkästchen zur Anwendung. Bei der Entwicklung und Auswahl der Subtests legten die Autorinnen besonderen Wert auf eine Verankerung der Aufgabeninhalte im konkreten Lebensraum 3- bis 6-jähriger Kinder und eine spielerische Gestaltung der Testsituation.

Auswertung Jedes richtig gelöste Item wird mit 1 Punkt bewertet; lediglich beim Subtest „Wörter erklären“ sind auch 2 Punkte für eine Antwort möglich. Die anhand der Normtabellen in C-Werte (Mittelwert = 5; Standardabweichung = 2) umgewandelten Rohwerte werden in ein Profilblatt eingetragen und ergeben das sog. „Entwicklungsprofil“. Fakultativ kann ein „Gesamtentwicklungsscore“ berechnet werden, indem der Mittelwert aller C-Werte ohne den Elternfragebogen bestimmt und anhand einer Tabelle in

Entwicklungsprofil und Gesamtwert

Reliabilität der Subtests variiert

einen Standardwert transformiert wird. Die Variabilität des Profils kann als Differenz zwischen dem besten und schlechtesten Subtestergebnis ermittelt werden; für den „Range“ liegen Prozentrangnormen vor.

Verschiedene Validitätsbelege

Objektivität und Reliabilität Bei den meisten Subtests ist die richtige Lösung anhand des Manuals eindeutig feststellbar. Die verbalen Subtests und das „Nachzeichnen“ lassen einen gewissen Spielraum bei der Bewertung zu, was die Auswertungsobjektivität leicht einschränken wird. Zur Reliabilität liegen Ergebnisse von Konsistenzanalysen sowie zum Subtest „Zahlen merken“ Re-test-Ergebnisse vor ($r_{tt} = .67$). Cronbachs α variiert zwischen .66 („Lernbär“) und .90 („Elternfragebogen“). Diese Angaben beziehen sich auf die gesamte Normierungsstichprobe.

Repräsentative Normstichprobe

Validität Die Autorinnen werten die Zunahme der Subtestleistungen mit dem Alter als Validitätsbeleg. Faktorenanalysen sprechen dafür, dass der Test verschiedene Aspekte der Entwicklung erfasst. Extrahiert wurden 6 Faktoren; die Ladungen der Subtests passen relativ gut zu den 6 a priori angenommenen Funktionsbereichen. Korrelationen mit Skalen der K-ABC (► Abschn. 3.2.3.2) sowie Korrelationen mit weiteren Tests mit ähnlichem Messanspruch wie die WET-Subtests werden ebenso als Validitätsbelege aufgeführt wie unterschiedliche Testleistungen von Kindern mit und ohne eine klinische Beeinträchtigung (z. B. Autismus).

Gute Orientierung über die aktuellen Stärken und Schwächen eines Kindes

Normierung Für die 2. Auflage wurde die Normierungsstichprobe aus Österreich ($N = 274$) um eine aus Deutschland auf insgesamt $N = 1245$ Kinder im Alter von 3;0 bis 5;11 Jahren) erweitert. Die Stichprobe wurde so zusammengestellt, dass sie bezüglich des Bildungsstands des Vaters repräsentativ ist; in Deutschland stammen die Normierungsdaten aus 19 Städten, die sich über das Bundesgebiet verteilen. Für die 3. Auflage wurde die Angemessenheit der Normen 2009 an einer repräsentativen Stichprobe deutscher und österreichischer Kinder ($N=385$) überprüft. Die in der 3. Auflage neuen Subtests „Rechnen“ und „Musterlegen neu“ wurden an 261 österreichischen und 124 deutschen Kindern eigens normiert.

Bewertung Der WET ist ein Breitbandverfahren, das im Vorschulalter Hinweise auf Entwicklungsrückstände in mehreren Bereichen liefern kann. Das Testmaterial und die Aufgaben selbst sind sehr kindgerecht. Die Eichstichprobe wurde sorgfältig zusammengestellt. In einer sehr fundierten Testrezension zur 2. Auflage des Tests hatte Renziehausen (2003) darauf hingewiesen, dass der Test mit bis zu 90 min relativ lange dauert und sich die Testdurchführung gerade bei jüngeren und bei schwächer begabten Kindern daher schwierig gestalten kann. Sie kritisierte weiter, dass die angegebenen C-Werte häufig schlecht differenzieren. Ihr Fazit fiel dennoch positiv aus:

- » Der WET ist vor allem für förderdiagnostische Fragestellungen konzipiert und liefert eine Statusdiagnose der kindlichen Entwicklung in relevanten Bereichen. Das Verfahren bietet eine gute Orientierung über die aktuellen Stärken und Schwächen eines Kindes (Renziehausen 2003, S. 145).

Da der Test in einigen Bereichen nicht hinreichend gut differenziert und die Skalen teilweise nicht sehr reliabel sind, ist anzuraten, den Test als Screening-Instrument einzusetzen und bei auffällig niedrigen Skalenwerten diesen Bereich mit anderen Verfahren näher zu untersuchen. Eine Rezension zu der nur wenig veränderten 3. Auflage liegt derzeit (Stand: Mai 2020) nicht vor.

Entwicklungstest 6 Monate bis 6 Jahre – Revision (ET 6–6-R)

Der ET 6-6-R von Petermann und Macha (2015) überlappt im unteren Altersbereich mit den Griffiths-Entwicklungsskalen und den Bayley-III (► Abschn. 3.2.5.1) und reicht bis zum Beginn des Grundschulalters. Das Verfahren erfasst 5 Entwicklungsbereiche: Körpermotorik und Handmotorik, kognitive Entwicklung, Sprachentwicklung, sowie sozial-emotionale Entwicklung (über die Elternauskunft). Ab 4 Jahren steht zusätzlich ein Untertest „Nachzeichnen“ zur Verfügung, dessen Messanspruch aber relativ vage formuliert ist. Die interne Konsistenz der Skalen ist mit $\alpha=.66$ bis .77 relativ niedrig. Die Normen der 2. Auflage von 2015 stammen den Verlagsangaben zufolge aus dem Jahr 2013 ($N=1053$). Die Anforderungen an die Validität sind nur teilweise erfüllt, wie in einer Testrezension festgestellt wird; die Belege bestehen im Wesentlichen aus „Mittelwertunterschieden (ohne Effektstärken) zwischen gesunden und entwicklungsaußfälligen Kindern“ (Hasselhorn und Margraf-Stiksrud 2015, S. 164).

ET 6–6-R soll 6 Entwicklungsbereiche erfassen

Dortmunder Entwicklungsscreening für den Kindergarten – Revision (DESK 3–6 R)

Das DESK 3-6 R von Tröster et al. (2016) verwendet altersspezifische Beobachtungsskalen (für 3, 4, 5 und 6 Jahre) zur Überprüfung der motorischen, sprachlichen, sozial-emotionalen und kognitiven Kompetenzen. Bei den 5- und 6-jährigen Kindern liegt der Fokus auf der Überprüfung schulischer Lernvoraussetzungen (z. B. schriftsprachliche Basiskompetenzen). Es kann von den Erzieherinnen und Erziehern durchgeführt werden. Die Beobachtung der Kinder erfolgt sowohl im Kindergartenalltag als auch in standardisierten Situationen.

DESK 3–6 R: ein Beobachtungsverfahren für den Kindergarten

Beispielaufgabe aus dem DESK 3–6 R

Im Zirkusbogen für 3-jährige Kinder kommt etwa die Aufgabe *Wie die Pferde über ein Hindernis springen* vor: „Die Pferde im Zirkus können gut über Hindernisse springen. Wir wollen jetzt alle wie die Zirkuspferde über ein Hindernis springen.“

Material: ein ca. 5 cm hohes Hindernis, z. B. ein Spielzeug.

Durchführung in der Gruppe: Die Kinder werden nacheinander aufgefordert, wie ein Zirkuspferd über das Hindernis zu springen. Etwa 2 m Anlauf sollten genommen werden.

Bewertung: Das Kind springt im Wechselschritt über 5 cm Höhe, ohne hinzufallen.

(Quelle: Tröster et al. 2005, S. 143).

Intelligence and Development Scales – 2 (IDS-2)

Die IDS-2 überlappen sich vom Altersbereich zum Teil mit den zuvor vorgestellten Verfahren, reichen aber nach oben bis zum Alter von 21 Jahren.

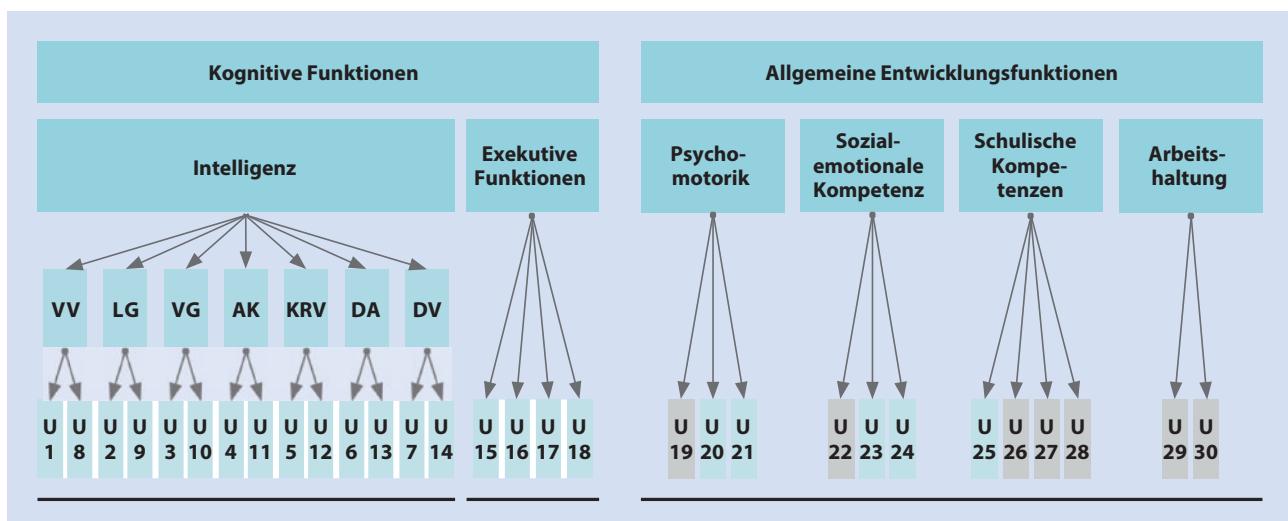
Steckbrief IDS-2: Intelligence and Development Scales – 2. Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche (Grob und Hagmann-von Arx 2018)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Intelligenz sowie Kompetenzen in 5 entwicklungsrelevanten Funktionsbereichen bei Kindern und Jugendlichen
Anwendungsbereich	Breiter Anwendungsbereich im gesamten Spektrum der Entwicklungs- und Leistungsdiagnostik, der Schuleingangsdagnostik und im klinischen Bereich
Theoretischer Hintergrund	Theoretische Fundierung der Intelligenz anhand des CHC-Modells (► Abschn. 3.2.4.1), Auswahl der Kompetenzbereiche eher pragmatisch
Testentwicklung	Weiterentwicklung der IDS mit Erweiterung des Altersbereichs; neue Funktionsbereiche „exeektive Funktionen“ und „Arbeitshaltung“, erweiterte Normen
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche und mündliche Instruktion; mit Hinweisen für den Umgang mit Fragen
Auswertung	Detaillierte Bewertungsregeln und die anschließende Auswertung der Rohwerte erfolgt ausschließlich durch ein Auswertungsprogramm
Interpretation	Genaue Vorgaben zur Benennung der erfassten Merkmale und zur Verbalisierung ihrer Ausprägung und Fallbeispiele
Reliabilität	
Konsistenz	Hauptkennwert IQ resp. IQ-Profil: $\alpha=.97$ resp. .98; IQ-Screening: $\alpha = .95$ Intelligenzfaktoren: $\alpha=.91$ bis .98 (jeweils für die Gesamtstichprobe); exeektive Funktionen: $\alpha=.88$ (für die Gesamtstichprobe); allgemeine Entwicklungsfunktionen: $\alpha=.82$ bis .97 (für die einzelnen Altersgruppen)
Retest	IQ resp. IQ-Profil: $r=.89$ resp. .86; IQ-Screening: $r = .85$; Intelligenzfaktoren: $r=.65$ bis .89 (durchschnittlich: 24 Tage, Intervall: 9–63 Tage, $N=69$)
Validität	
Konstruktvalidität	Überprüfung der Faktorenstruktur der Intelligenzwerte IQ und IQ-Profil durch konfirmatorische Faktorenanalysen ($N=1672$); einige Studien zur Korrelation der IDS-2 mit anderen Intelligenztests (u. a. IDS-2 IQ/IQ-Profil mit Kennwerten der Gesamtintelligenz in Referenztests WISC-IV, Reynolds Intellectual Assessment Scales and Screening [RIAS], Non-verbal Intelligenztest [SON-R 6–40]): $r=.51$ bis .69
Kriteriumsvalidität	Für den Funktionsbereich „schulische Kompetenzen“ Analyse der Schulleistungen durch Elterneinschätzungen und Schulnoten
Normen	
Zusammensetzung	Normen ($N=1672$) für Kinder und Jugendliche zwischen 5;0 und 20;11 Jahren
Erhebungszeitraum	Mitte 2015 bis Mitte 2017
Sonstiges	
Formen	Es gibt keine unterschiedlichen Testformen
Testrezension	
Quelle	Renner (2019)

Anders als der Testname „Intelligence and Development Scales“ vermuten lässt, handelt es sich bei den IDS-2 nicht um ein englischsprachiges Verfahren, das ins Deutsche übertragen wurde. Vielmehr wurde der Test an der Universität Basel entwickelt. Die 1. Version der IDS erschien 2009 und orientierte sich an den in Lehrbüchern und Nachschlagewerken zur Entwicklungspsychologie behandelten Funktionsbereichen mit Schwerpunkt auf der kognitiven Entwicklung (Grob et al. 2009). In einer Rezension sprach Renner (2011) zwar anerkennend von einer „ambitionierten Testentwicklung“, verwies aber gleichzeitig auf eine Reihe von Verbesserungsmöglichkeiten. Die Rezension von Koch et al. (2011) fiel deutlich kritischer aus. Den Vorwurf, die IDS stelle eine Revision und Neubearbeitung des in der Schweiz beliebten Kramer-Intelligenztests von 1972 dar und der Rückgriff auf Alfred Binet wirke befreindlich, weisen Grob und Hagmann-von Arx (2011) in einer Replik zurück. Sie stellen klar, dass die ursprünglich geplante Überarbeitung und Aktualisierung des Kramer-Tests aufgegeben wurde und mit den IDS eine „vollständige Testneukonzipierung in Angriff genommen wurde“, wie auch im Manual nachzulesen ist. Mit der Würdigung der Pionierarbeit von Alfred Binet wurden auch nicht „100 Jahre entwicklungspsychologisches Wissen unbeachtet“ gelassen, wie Koch et al. (2011, S. 112) kritisiert hatten. Schließlich wehren sich die Autorin und der Autor gegen den Vorwurf, die Testaufgaben in der IDS seien „aus aktuellen Verfahren übernommen“ (Koch et al. 2011, S. 108) worden. Vielmehr wurden sie „in Anlehnung an weitere Testverfahren konstruiert und neu entwickelt“ (Grob und Hagmann-von Arx 2011, S. 112).

Teils überzogene Kritik an der 1. Auflage

Gliederung Die IDS-2 sind modular aufgebaut. Je nach Fragestellung können damit verschiedene Entwicklungsbereiche untersucht werden. Im kognitiven Bereich werden die Intelligenz sowie exekutive Funktionen erfasst. Für die Intelligenzdiagnostik orientiert sich das Verfahren an dem CHC-Modell (► Abschn. 3.2.4.1). Mit jeweils 2 Subtests sollen die Faktoren „Verarbeitung Visuell“, „Langzeitgedächtnis“, „Verarbeitungsgeschwindigkeit“, „Kurzzeitgedächtnis Auditiv“, „Kurzzeitgedächtnis Räumlich-Visuell“, „Denken Abstrakt“ und „Denken Verbal“ erfasst werden. □ Abb. 3.16 zeigt den Aufbau des kompletten Tests.

Modularer Aufbau, CHC-Modell zur Intelligenz



□ Abb. 3.16 Aufbau der IDS-2. Abkürzungen für die Faktoren: VV = Verarbeitung Visuell, LG = Langzeitgedächtnis, VG = Verarbeitungsgeschwindigkeit, KA = Kurzzeitgedächtnis Auditiv, KRV = Kurzzeitgedächtnis Räumlich-Visuell, DA = Denken Abstrakt, DV = Denken Verbal; (U = Untertest). (Aus Grob und Hagmann-von Arx 2018, S. 14, mit freundlicher Genehmigung des Hogrefe Verlages)

IQ-Screening, IQ und Intelligenzprofil

4 exekutive Funktionen

Psychomotorik, sozial-emotionale Kompetenz, schulische Kompetenzen und Arbeitshaltung

► Beispiel

Der Faktor „Denken Abstrakt“ wird mit den Untertests „Matrizen ergänzen“ und „Unpassende Bilder erkennen“ abgedeckt. Bei letzterem besteht ein Item aus 6 abgedruckten farbigen Bildkärtchen. Kinder bzw. Jugendliche sollen das Bild finden, das sich von den anderen unterscheidet. Bei einer sehr leichten Aufgabe sind beispielsweise 5 rote und 1 gelber Punkt zu sehen. ◀

Ist man an der Intelligenz interessiert, eröffnen sich mit den IDS verschiedene Möglichkeiten: Für ein Screening reicht es aus, 2 bestimmte Untertests durchzuführen (Dauer: ca. 10 min). Soll der IQ zuverlässig bestimmt werden, setzt man 7 bestimmte Untertests ein, die alle 7 Intelligenzfaktoren abdecken (Dauer: ca. 50 min). Für ein Intelligenzprofil dagegen sind alle 14 Untertests durchzuführen (Dauer: ca. 90 min). Anhand des Intelligenzprofils ist gut zu erkennen, in welchem der 7 Bereiche ein Kind Stärken und Schwächen aufweist.

Im Bereich der exekutiven Funktionen werden die Inhibition, das Arbeitsgedächtnis, die kognitive Flexibilität und das Planen erfasst. Die Testdurchführung nimmt etwa 30 min in Anspruch.

► Beispiel

Zur Inhibition etwa wird eine Aufgabe verwendet, die an den Farb-Wort-Interferenz-Test angelehnt ist und den Stroop-Effekt nutzt. Allerdings sind keine Farbwörter zu benennen, sondern die jeweilige Farbe von Tieren auf einer Karte. Dessen Farbe interferiert in der Interferenzbedingung mit deren „richtiger“ Farbe in der Natur. Ein Frosch kann rot, eine Marienkäfer blau abgebildet sein etc. ◀

Unter dem Überbegriff „allgemeine Entwicklungsfunktionen“ finden sich die Bereiche „Psychomotorik“, „sozial-emotionale Kompetenz“, „schulische Kompetenzen“ und „Arbeitshaltung“. Sollen alle Bereiche erfasst werden, nimmt die Durchführung – je nach Alter – etwa 72 bis 102 min in Anspruch:

- Die *Psychomotorik* umfasst Aufgaben zur Grobmotorik (u. a. auf einem auf dem Boden liegenden Band einen Fuß vor den anderen setzend laufen), Feinmotorik (z. B. Perlen auffädeln) und Visuomotorik (Nachzeichnen von Vorlagen).
- Für den Bereich *sozial-emotionale Kompetenz* stehen Aufgaben zum Erkennen von Emotionen anhand von Gesichtsbildern, zur Regulation von auf Fotos gezeigten Emotionen und zum sozial kompetenten Handeln bereit. Zur Messung der letztgenannten Kompetenz werden Bilder von Situationen gezeigt, in denen eine Person helfen oder einschreiten könnte. Beispielsweise setzt ein Mädchen dazu an, am Strand eine Sandburg zu zerstören. Die Frage ist, was ein Junge im Hintergrund hier tun könnte.
- *Schulische Kompetenzen* untergliedern sich in logisch-mathematisches Denken, sprachliche Fähigkeiten, Lesen und Rechtschreiben. Die Lesekompetenz etwa wird mit Aufgaben zum Wörterlesen, zum Lesen von Pseudowörtern (z. B. bune, sonir, ...) und zum sinnverstehenden Textlesen überprüft.
- Der Bereich *Arbeitshaltung* umfasst Gewissenhaftigkeit (Beispiel: „Mein Zimmer ist ordentlich und sauber“) und Leistungsmotivation (Beispiel: „Es ist mir wichtig, viel zu wissen und zu können“). Die Aussagen werden vorgelesen und das Kind zeigt auf einer grafischen Ratingskala die Antwort an.

Durchführung, Auswertung und Interpretation Einige Aufgabengruppen der IDS-2 sind nur für einen bestimmten Altersbereich vorgesehen. Das Testmaterial ist altersgerecht und die Instruktionen liegen schriftlich vor. Die Durchführung ist insbesondere bei der Auswahl vieler Testmodule anspruchsvoll. Der Testverlag bietet für eine „schnelle und sichere Einarbeitung zur Anwendung der IDS-2“ ein eintägiges Seminar an. Die Ergebnisse in den einzelnen Untertests werden auf einem Protokollbogen dokumentiert. Zusätzlich beurteilt die Testleiterin/der Testleiter die Mitarbeit während der gesamten Untersuchung. Für die Auswertung steht ein Auswertungsprogramm zur Verfügung, das auch einen kompletten Ergebnisbericht generiert. Renner (2019) kritisiert, dass die Eingaben in das Auswertungsprogramm nicht auf Plausibilität geprüft werden. Die Interpretation wird durch Vorgaben zur Verbalisierung der erfassten Merkmale sowie zu deren Ausprägung sehr gut standardisiert. Zudem erleichtern Fallbeispiele die Interpretation.

Elektronische Testauswertung und gute Interpretationshilfen

Testgütekriterien Die interne Konsistenz der Intelligenzwerte aus dem Screening, zum Gesamt-IQ und zum Profil sind sehr hoch. Selbst das kurze Screening erreicht mit $\alpha = .95$ einen sehr hohen Wert. Auch die Retest-Reliabilität nach 24 Tagen ($n = 69$) liegt mit $r_{tt} = .85$ bis $.89$ in einem sehr hohen Bereich. Das gilt jedoch nicht für alle Untertests; hier variiert sie zwischen $r_{tt} = .39$ bis $.89$.

Hohe Reliabilität der Intelligenzkennwerte

Für die zahlreichen zusätzlich erfassten Merkmale variieren die Reliabilitätsabschätzungen. Sie reichen von $\alpha = .82$ bis $.97$. Eine Ausnahme findet sich bei den exekutiven Funktionen, bei denen die interne Konsistenz von $.64$ bis $.91$ reicht. Für die exekutiven Funktionen wurden moderate ($r_{tt} = .72$ bis $.75$) und für sozial-emotionale Kompetenz moderate bis hohe ($r_{tt} = .71$ bis $.85$) Retest-Koeffizienten ermittelt. Bei der Auswahl von Testmodulen für bestimmte Fragestellungen sind die Reliabilitätskennwerte zu beachten.

Überwiegend hohe Reliabilität der übrigen Kennwerte

Zur Konstruktvalidität liegen umfangreiche Befunde vor, die sich sowohl auf die Überprüfung der postulierten Intelligenzfaktoren als auch auf die Korrelation mit konvergenten und diskriminanten Verfahren beziehen. Für den Funktionsbereich „schulische Kompetenzen“ liegen auch Korrelationen mit simultan erhobenen Elterneinschätzungen und Schulnoten vor, die wir ebenfalls der Konstruktvalidität zuordnen möchten.

Validität gut untersucht

Die IDS-2 wurden von 2015 bis 2017 an 1672 Kindern und Jugendlichen im Alter von 5;0 bis 20;11 Jahren aus der deutschsprachigen Schweiz ($n = 973$), aus Deutschland ($n = 614$) und aus Österreich ($n = 85$) normiert. Der dabei verwendete Continuous-Norming-Ansatz (► Abschn. 2.6.4.5) erlaubt eine sehr feine Altersabstufung. Das Prinzip besteht darin, eine mathematische Funktion zu finden, die den Zusammenhang zwischen den Testwerten und dem Alter über die gesamte Altersspanne beschreibt. Bei den IDS-2 werden Normwerte für 1-Monats-Intervalle geschätzt.

Fein gestufte Altersnormen für 5- bis 20-Jährige

Fazit Bei den IDS-2 handelt es sich um ein sehr sorgfältig konstruiertes und theoretisch fundiertes Verfahren mit guten Testkennwerten. Auch wenn nicht alle Aufgaben in allen Altersgruppen anwendbar sind, erlaubt es doch eine differenzierte Intelligenzdiagnostik sowie die Erfassung eines breiten Spektrums an weiteren kognitiven und nichtkognitiven Funktionen. Damit erschließt sich das Verfahren vielfältige Anwendungsmöglichkeiten. Exemplarisch seien die klinische Kinder- und Jugendpsychologie, die Sonder- und Heilpädagogik sowie die Schulpsychologie genannt.

Vielfältige Anwendungsmöglichkeiten

Ausblick Nach Erscheinen der IDS-2 haben Grieder und Grob (2019) die Struktur des Intelligenzbereichs der IDS-2 nun auch mit explorativen Faktorenanalysen untersucht. Sie verwendeten dazu Daten zur Intelligenz und

Überprüfung der Intelligenzstruktur mit explorativen Faktorenanalysen

zusätzlich auch zu den intelligenznahen schulischen Kompetenzen aus der Eichstichprobe. Explorative Faktorenanalysen (► Abschn. 2.5.4) sind „unvorgenommen“ gegenüber den Erwartungen der Anwenderinnen und Anwender und suchen nur die Struktur, die in den Daten steckt. Konfirmatorische Faktorenanalysen prüfen dagegen, ob ein postulierte Modell (notfalls) zu den Daten passt. Die Ergebnisse bestätigen weitgehend das postulierte Modell, legen aber auch einige Änderungen nahe. Grieder und Grob (2019) fanden einen sehr starken g-Faktor, der die Berechnung eines Gesamt-IQ als angemessen bestätigt.

Anstelle der 7 Gruppenfaktoren nach dem CHC-Modell sprechen die Analysen für ein Strukturmodell mit 5 Faktoren. Dabei fallen die „alten“ Faktoren „Verarbeitung Visuell“ und „Denken Abstrakt“ zu einem Faktor zusammen, der als „abstraktes visuelles Denken“ bezeichnet wird. Ebenso konvergieren die „alten“ Faktoren „Denken Verbal“ und „Langzeitgedächtnis“ zu einem Faktor, der „semantisches Langzeitgedächtnis“ genannt wird.

Wurden die Aufgaben zu den intelligenznahen schulischen Kompetenzen hinzugenommen, änderte sich das Bild leicht. Der Untertest „logisch-mathematisches Denken“ wanderte zum neuen Faktor „abstraktes visuelles Denken“, während die Untests zum Lesen und Schreiben einen eigenen Faktor bildeten. Die Aufgaben zum logisch-mathematischen Denken haben also eine stärkere Beziehung zur Intelligenz (und hier speziell zum abstrakt-visuellen Denken) als zu sprachlichen schulischen Kompetenzen.

Die IDS-2 wird auch für andere Länder adaptiert und normiert. Eine holländische Fassung wurde 2018 publiziert, eine polnische 2019. Für die italienische und die englische Version ist eine Publikation Ende 2019 geplant. 2020 bis 2022 sollen die IDS-2 in portugiesischer (Brasilien), schwedischer, norwegischer, finnischer, dänischer, französischer, spanischer und tschechischer Sprache mit eigenen Normierungen erscheinen (persönliche Mitteilung Prof. Grob, März 2019). Wir erwähnen dies, weil es ein starker Hinweis auf die (anstehende) Verbreitung des Tests ist, die wiederum internationale Forschungsaktivitäten erwarten lässt, was der Pflege des Verfahrens zugutekommen wird.

3.2.5.3 Breitbandverfahren für das Vorschulalter – spezielle Entwicklungstests

Bei der Erfassung eines mehr oder weniger eng umschriebenen Entwicklungsbereichs kommen sog. „spezielle Entwicklungstests“ zum Einsatz. Unter diesen nehmen Tests zum kognitiven Entwicklungsstand quantitativ eine dominante Rolle ein (im ► Abschn. 3.2.3 wurden bereits einige Tests vorgestellt, die für das Vorschul- und Schulalter geeignet sind). Mehrere Verfahren dienen der Erfassung der Sprachentwicklung. Die folgenden beiden Verfahrenstypen sind dabei zu unterscheiden: Fragebögen zur Beurteilung des Entwicklungsstands, die meist von den Eltern auszufüllen sind, und Leistungstests. Daneben gibt es Tests zur Beurteilung des motorischen Entwicklungsstands.

Fragebogen zur frühkindlichen Sprachentwicklung FRANKIS

Fragebogen zur frühkindlichen Sprachentwicklung (FRANKIS) Der FRANKIS von Szagun et al. (2009) ist ein Beispiel für die erste Kategorie. Allerdings wird in einer Testrezension nach den Standards des Testkuratoriums (Deimann et al. 2010) zu Recht darauf hingewiesen, dass dieses Verfahren für den empfohlenen Einsatz nicht ausreichend validiert wurde und dass die Normen nicht repräsentativ sind.

Sprachstandserhebungstest für Fünf- bis Zehnjährige (SET 5–10)

Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren (SET 5–10) Als Beispiel für einen Sprachtest, in dem die Kinder unter standardisierten Bedingungen Aufgaben bewältigen müssen, sei der SET 5–10 von Petermann et al. (2010) genannt. Mit diesem Test werden mehrere Aspekte

der Sprache geprüft: Wortschatz, semantische Relationen, Sprachverständnis, Sprachproduktion, Morphologie, Verarbeitungsgeschwindigkeit und auditive Merkfähigkeit. Neben der Sprache findet die Motorik eine besondere Beachtung.

Lincoln-Oseretzky-Skala Kurzform (LOS KF 18) Ein Beispiel für einen Test zur Untersuchung des motorischen Entwicklungsstands ist die LOS KF 18 von Eggert (1974). Dieser Test enthält Aufgaben wie mit geschlossenen Augen die Nasenspitze berühren, mit offenen Augen 10 s auf einem Bein stehen oder Streichhölzer mit einer Hand sortieren.

Untersuchung des motorischen Entwicklungsstands

Weiterführende Literatur

Über weitere Motoriktests informiert das Handbuch von Bös (2017).

3.2.6 Schultests

Die für die Schule relevanten Tests werden in ▶ Kap. 7 (Diagnostik in der Pädagogischen Psychologie) vorgestellt, weil sie sich dort thematisch gut einfügen. Hier soll nur eine kurze Übersicht gegeben werden.

- *Schuleingangstests* (▶ Abschn. 7.4.1.1) sollen mögliche Entwicklungsdefizite aufzeigen, die ggf. vor der Einschulung durch geeignete Trainingsmaßnahmen kompensiert werden können. Sie sind Entwicklungstests ähnlich, fokussieren aber auf die Bereiche, die als wichtige Voraussetzungen für Lesen, Schreiben und Rechnen angesehen werden. Sie kommen im Vorschulbereich zum Einsatz.
- *Schulleistungstests* (▶ Abschn. 7.4.1.2) dienen dazu, den Lernstand eines Kindes in einem bestimmten Bereich (insbesondere Rechtschreibung oder Mathematik) durch einen Vergleich mit Kindern der gleichen Klassenstufe festzustellen. Die Testinhalte orientieren sich an den Lehrplänen der Klassen und sind bundesweit normiert.

Darüber hinaus werden in ▶ Abschn. 7.4.2 auch Tests zur Evaluierung des Bildungssystems (u. a. PISA) behandelt, die aber nicht der Individualdiagnostik dienen.

Weiterführende Literatur

In der Einleitung (▶ Abschn. 3.1) wurden bereits Informationsquellen zu standardisierten diagnostischen Verfahren und damit auch Leistungstests vorgestellt und bewertet. Testkompendien und Bücher, in denen Tests rezensiert werden, sind schnell nicht mehr aktuell, wenn sie nicht immer wieder aktualisiert werden. Deshalb können ältere Werke nicht mehr empfohlen werden. Das Buch *Diagnostische Erhebungsverfahren* von Petermann und Daseking (2015) gibt einen guten Überblick über die Diagnostik in verschiedenen Anwendungsbereichen und beschreibt ausgewählte Verfahren.

Zu speziellen Themen finden sich in der Buchreihe „Kompendien Psychologische Diagnostik“ nützliche Informationen. Ein Teil der Bücher thematisiert Leistungstests: *Aufmerksamkeitsdiagnostik* (Heubrock und Petermann 2001), *Intelligenzdiagnostik* (Holling et al. 2005), *Diagnostik von Rechenstörungen* (Jacobs und Petermann 2012), *Entwicklungsdiagnostik* (Esser und Petermann 2010), *Sprachdiagnostik im Kindesalter* (Petermann et al. 2016) und *Demenzdiagnostik* (Stemmler und Kornhuber 2018).

Über Leistungstests, die in der Personalauswahl Verwendung finden, informieren Krumm und Schmidt-Atzert (2009). Mehrere Werke befassen sich mit Tests, die (auch) in der Neuropsychologie eingesetzt werden, wobei insbesondere die von Schellig et al. (2018a, b) sowie Schellig et al. (2009) herausgegebenen Bände zu erwähnen sind.

Informationen und Besprechungen zu Tests, die im pädagogischen Bereich eingesetzt werden können, bietet die seit 2000 von Hasselhorn, Schneider und Trautwein herausgegebene Buchreihe „Tests und Trends in der pädagogisch-psychologischen Diagnostik“. Bisher (Stand: Mai 2020) sind 18 Themenbände erschienen, die von Expertinnen und Experten auf dem jeweiligen Gebiet herausgegeben wurden.

?

Übungsfragen

— Abschn. 3.1:

- Welche beiden diagnostische Leistungstests werden einer Befragung von Roth und Herzberg (2008) zufolge in der Praxis am häufigsten eingesetzt?
- Nennen Sie Informationsquellen zu standardisierten diagnostischen Verfahren!

— Abschn. 3.2:

- Für was sind die Leistungen in Leistungstests ein Indikator?
- Wie wirken sich Übung und Vorbereitung auf die Testleistung aus?
- Was haben Aufmerksamkeit und Konzentrationsfähigkeit gemeinsam, und wie lassen sie sich voneinander abgrenzen?
- Welche Aufmerksamkeitsfunktionen werden unterschieden, und was versteht man jeweils darunter?
- Welche Aufgabentypen kommen bei Konzentrationstests vor? Nennen Sie jeweils ein Testbeispiel!
- Wie viele Konzentrationsfaktoren wurden bei einer Faktorisierung verschiedener Tests gefunden?
- Welches Merkmal misst der d2-R?
- Beschreiben Sie die Aufgabe der Testpersonen bei der Bearbeitung des d2-R! Was sind die Durchführungsbedingungen?
- Welche Kennwerte werden beim d2-R bestimmt (dazu liegen auch Normen vor), und wie sind sie definiert?
- Welche beiden Leistungen muss man bei der Bearbeitung des KLT-R (Konzentrations-Leistungs-Test – revidierte Fassung) erbringen?
- Mit welchen 2 Merkmalen (oder Tests) korreliert der KLT-R relativ hoch?
- Nach welchen 3 Merkmalen kann man Intelligenztests einteilen?
- Welche Merkmale (Formen der Intelligenz) erfassen verschiedene Intelligenztests?
- Welche Kennwerte liefert die deutsche Version der WISC-V (Wechsler Intelligence Scale for Children) auf den beiden höchsten Ebenen?
- Für welchen Alters- und welchen Intelligenzbereich wurde die WISC-V entwickelt?
- Welche Kritik am Strukturmodell der WISC-V haben Canivez et al. (2018) vorgebracht?
- Welche 2 gut vergleichbare Alternativen zur Intelligenzdiagnostik bei Kindern gibt es zur WISC-V?
- Aus welchen 2 Modulen besteht der Intelligenz-Struktur-Test 2000-R, und welche Intelligenzkomponenten werden damit gemessen?
- Nennen Sie 3 weitere „breite“ Intelligenztests als Alternativen zum I-S-T 2000-R!
- Wie ist der CFT 20-R aufgebaut, und was soll der Test messen?
- Nach welchen 2 Modellen der kognitiven Fähigkeiten lassen sich sowohl konventionelle Intelligenz- als auch spezielle Fähigkeitstests einordnen?
- Welche Art von Aufgaben verwendet der Test zur Praktischen Alltagsintelligenz (PAI 30), und wie werden die Antworten dort ausgewertet?
- Beschreiben Sie den Aufbau und die Aufgaben der Griffiths-EntwicklungsSkalen (GES) zur Beurteilung der Entwicklung in den ersten beiden Lebensjahren!
- Für welchen Altersbereich ist der Wiener Entwicklungstest (WET) vorgesehen, und welche Merkmale erfasst er?
- Welche Merkmale sollen die Intelligence and Development Scales – 2 (IDS-2) erfassen?

3.3 Persönlichkeitsfragebögen

In diesem Abschnitt befassen wir uns zunächst mit dem Konzept bzw. „Persönlichkeitsmerkmal“. Dabei wird hier noch einmal die Frage aufgeworfen, welche Rolle situative Faktoren neben den (klassischen) Persönlichkeitsmerkmalen bei der Erklärung und Vorhersage von Verhalten spielen. Anschließend versuchen wir, die Persönlichkeit inhaltlich näher zu bestimmen und gehen der Frage nach, welche Persönlichkeitsmerkmale in der Psychologischen Diagnostik eine besondere Rolle spielen (vgl. ▶ Abschn. 1.4). Bei deren Messung gibt es die Möglichkeit, Verhalten und/oder Erleben zu erfassen. Dabei können entweder die betroffene Person selbst oder externe Beurteilerinnen und Beurteiler bzw. Beobachterinnen und Beobachter Auskunft geben. Neben Persönlichkeitsfragebögen existieren weitere Messmethoden zur Erfassung von Persönlichkeitsmerkmalen, die in ▶ Abschn. 3.4 und 3.6 behandelt werden.

3.3.1 Persönlichkeitsmerkmale und ihre Messung

Bevor man etwas zu messen versucht, sollte man versuchen, den Messgegenstand zu definieren oder zumindest zu beschreiben. Wie bei vielen anderen Konstrukten ist dies auch für Persönlichkeit(-smerkmale) nicht einfach. Wir stellen hier eine relativ verständlich formulierte Definition eines einschlägigen Persönlichkeitspsychogen vor und kommentieren sie im Anschluss kurz.

— Definition —

„**Persönlichkeit** [engl. *personality*; lat. *persona* Maske, Rolle, Person, *personare* hindurch tönen] ist die Gesamtheit aller überdauernden individuellen Besonderheiten im Erleben und Verhalten eines Menschen (der Persönlichkeitseigenschaften, syn. Persönlichkeitsmerkmale [engl. *traits*]). Beispiele für Persönlichkeitseigenschaften sind intelligent (Intelligenz), aggressiv (Aggressivität), gesellig (Geselligkeit), leistungsmotiviert (Leistungsmotivation), konservativ. „Überdauernd“ bezieht sich in dieser Definition auf Zeiträume von wenigen Wochen oder Monaten. Persönlichkeit setzt also eine kurzfristige Stabilität dieser Besonderheiten voraus. Damit können viele Persönlichkeitseigenschaften als Dispositionen aufgefasst werden, d. h. Tendenzen, bestimmte Situationen in bestimmter Weise zu erleben und sich dort in bestimmter Weise zu verhalten. Das schließt langfristige Veränderungen der Persönlichkeit nicht aus (Persönlichkeitsentwicklung). Mit ‚individueller Besonderheit‘ ist gemeint, dass es sich um Merkmale handelt, die zwischen den Mitgliedern einer Bezugsgruppe variieren (meist wird Persönlichkeit bezogen auf Unterschiede innerhalb derselben Altersgruppe und Kultur). Beschrieben wird die Persönlichkeit eines Individuums durch ein Persönlichkeitsprofil, d. h. die Ausprägung der Person in vielen Persönlichkeitseigenschaften.“ (Asendorpf, 2020; vom Autor verwendete Abkürzungen werden hier ausgeschrieben, mit freundlicher Genehmigung des Hogrefe Verlages).

Zu dieser Definition ist Folgendes anzumerken:

- Wir bevorzugen es, in diesem Kapitel von „Persönlichkeitsmerkmalen“ zu sprechen, weil sich der Begriff „Persönlichkeit“ auf die Gesamtheit (s. o.) bezieht und viele Verfahren nicht den Anspruch haben, die „ganze“ Persönlichkeit zu erfassen, sondern in der Regel nur einen Ausschnitt.

Persönlichkeit als Gesamtheit aller Persönlichkeitsmerkmale

Unter den Beispielen für Persönlichkeitseigenschaften taucht auch Intelligenz auf. Der Begriff „Persönlichkeit“ wird unterschiedlich eng gefasst und kann Intelligenz einschließen. In der Psychologischen Diagnostik wird sie zumeist separat betrachtet und eine Unterscheidung zwischen Leistungs- und Persönlichkeitstests getroffen. Dieser Unterteilung schließen wir uns an.

- Die obige Definition betont, dass Persönlichkeitsmerkmale „überdauernd“ in dem Sinne sind, dass sie in Zeiträumen „von wenigen Wochen oder Monaten“ stabil sind, sich aber längerfristig sehr wohl verändern können. In der Tat verändert sich die Ausprägung von manchen Persönlichkeitsmerkmalen mit dem Alter (Roberts et al. 2006). Auch Lebensereignisse können Persönlichkeitsmerkmale verändern (Bleidorn et al. 2018). Für die Evaluation von diagnostischen Instrumenten bedeutet dies, dass eventuell Altersnormen erforderlich und zur Schätzung der Retest-Reliabilität eher kurze Zeitintervalle informativ sind.

3.3.1.1 Bedeutung der Situation

Auch die Situation erklärt Verhalten und Erleben

Persönlichkeitsmerkmale sind nur einschränkt tauglich, Verhalten zu erklären oder vorherzusagen (Mischel 2004). Die von Mischel einmal genannte Grenze von .30 als maximale Korrelation zwischen Persönlichkeitsmerkmalen und relevantem Verhalten (z. B. Berufserfolg) gilt im Wesentlichen noch immer. Wie sich Menschen verhalten, lässt sich oftmals durch die Situation erklären, in der sie sich befinden. Wenn viele Menschen in einem Flugzeug sitzen, dessen Triebwerke Feuer gefangen haben und das in einen steilen Sinkflug übergeht, werden (fast) alle sehr starke Angst erleben. Mit dem Konzept „Situationsstärke“ (► Abschn. 1.4) lässt sich gut beschreiben, dass der Einfluss der Situation mehr oder weniger stark sein kann. In ► Abschn. 1.4 wurde auch eine umfangreiche Metaanalyse erwähnt, der zufolge sozialpsychologische Effekte insgesamt etwas besser durch situative Faktoren als durch Persönlichkeitsmerkmale erklärt werden.

Interaktion von Person und Situation bei manchen Items berücksichtigt

Persönlichkeit und Situation kann man als eigenständige Faktoren betrachten. Aber auch die Interaktion, also das Zusammenwirken von Person und Situation kann helfen, Verhalten und Erleben zu erklären (vgl. ► Abschn. 1.4). Diese Einsicht steckt bereits in vielen Items von Persönlichkeitfragebögen. Es ist selbstverständlich, dass viele Verhaltensweisen, die wir als Indikatoren von Persönlichkeitsmerkmalen ansehen, nur in bestimmten Situationsklassen relevant sind. Items wie „Es fällt mir schwer, mich zu beherrschen“, „Ich fühle mich nicht ganz behaglich“ oder „Ich mache viele Fehler“ würden die Testpersonen wahrscheinlich irritieren, „weil man das so allgemein nicht sagen kann“. Selbst wenn sie beantwortet würden, fielen diese Items einer Itemanalyse (► Abschn. 2.5) wegen schlechter Kennwerte vermutlich zum Opfer. Die Items stammen aus dem deutschen HEXACO-Fragebogen (s. u.) und lesen sich im vollen Wortlaut so: „Es fällt mir schwer, mich zu beherrschen, wenn Leute mich beleidigen“, „Ich fühle mich nicht ganz behaglich, wenn ich vor einer Gruppe von Leuten spreche“, „Ich mache viele Fehler, weil ich nicht nachdenke, bevor ich handele“. Bei diesen Items wird also gefragt, wie sich jemand in bestimmten Situationen verhält. Dagegen sind Items wie „Ich halte mich selbst für eine etwas exzentrische Person“ oder „Im Allgemeinen bin ich mit mir ziemlich zufrieden“ so formuliert, dass sie über viele, ganz unterschiedliche Situationen hinweg gelten.

Testkonstrukteurinnen und Testkonstrukteure haben noch einen anderen Weg gefunden, das Zusammenwirken von Persönlichkeitseigenschaften und Situationen zu berücksichtigen. Sie streben bewusst keine Generalisierung über alle denkbaren Situation an, sondern fokussieren bei der Messung der Persönlichkeit auf ausgewählte Situationen. Sie wollen damit auch nur das Verhalten und Erleben in einer Situation oder in einer Situationsklasse beschreiben, erklären oder vorhersagen. Beispielsweise soll mit dem Stressverarbeitungsfragebogen (SVF; Erdmann und Janke 2008) erfasst werden, wie Menschen in Stresssituationen reagieren. In der Testanweisung werden Stresssituationen sehr allgemein und abstrakt durch folgende Formulierung umschrieben: „Wenn ich durch irgendetwas oder irgendjemanden beeinträchtigt, innerlich erregt oder aus dem Gleichgewicht gebracht worden bin...“. Die Testpersonen gibt über die Beantwortung von Items an, zu welchen Bewältigungsstrategien sie in solchen Situationen neigt. Beispiele für die Skalen sind Resignation (Itembeispiel: „...erscheint mir alles so hoffnungslos“), soziale Abkapselung („...schließe ich mich von meiner Umgebung ab“) und Bagatellisierung („...sage ich mir, es geht schon alles wieder in Ordnung“). Mit anderen Worten: Hier wird auf eine spezifische Klasse von Situationen, nämlich Stresssituationen, fokussiert. Innerhalb dieser Situationsklasse wird jedoch eine Generalisierung über viele entfernt ähnliche (Stress-)Situationen angestrebt. Solche Situationen können eine Kündigung, ein Familienstreit, die Androhung von Gewalt, eine bevorstehende Operation oder eine beliebige andere Situation, die sich eine Testperson unter „Beeinträchtigungen“ vorstellt, sein.

Noch etwas spezifischer ist der Messanspruch des Fragebogens Arbeitsbezogenes Verhaltens- und Erlebensmuster (AVEM; Schaaerschmidt und Fischer 2008). Mit dem AVEM soll (nur) die Bewältigung von Arbeits- und Berufsanforderungen erfasst werden. Im Vergleich zum SVF werden auch die möglichen Reaktionen enger gefasst. Es gibt 11 Skalen, die sich mit gesundheitsförderlichen bzw. -gefährdenden Verhaltens- und Erlebensweisen befassen. Als Beispiele genannt seien „beruflicher Ehrgeiz“ (Itembeispiel: „Ich möchte beruflich weiterkommen, als es die meisten meiner Bekannten geschafft haben“) und „Resignationstendenz bei Misserfolgen“ (Itembeispiel: „Wenn ich keinen Erfolg habe, resigniere ich schnell“).

Eine andere Variante von Interaktionismus wird bei den sog. „Situational-Judgment-Tests“ (► Abschn. 6.2.1.2) praktiziert. Diese Tests finden in der Berufseignungsdiagnostik Verwendung. Zunächst wird jeweils eine Situation geschildert oder in Bildern bzw. einem Video gezeigt. Die Testpersonen sollen dann durch Auswahl einer Antwortalternative angeben, wie sie sich in dieser Situation verhalten würden oder (in einer anderen Variante) welches Verhalten hier angemessen wäre.

Fazit Als Fazit lässt sich festhalten, dass Persönlichkeitsmerkmale als situationsübergreifendes Erklärungskonzept für Verhalten verstanden werden und dass es zudem unstrittig ist, dass man auch mit der jeweiligen Situation relativ gut erklären kann, warum sich Menschen so und nicht anders verhalten. In manchen Fällen wirken beide Faktoren zusammen, wie die interaktionistische Sichtweise betont.

Verhalten und Erleben (nur) in Stresssituationen

Verhalten und Erleben (nur) in beruflichen Stresssituationen

Verhalten (nur) in spezifischen berufsrelevanten Situationen

Persönlichkeit, Situation und deren Interaktion

3.3.1.2 Struktur der Persönlichkeit

Wenn wir uns selbst oder andere Menschen beschreiben, verwenden wir dazu Eigenschaftswörter wie „beharrlich“, „risikoscheu“ oder „unberechenbar“ – das sind übrigens Begriffe, die laut einem Artikel im *Spiegel* vom 29.11.2010 amerikanische Diplomaten einmal zur Charakterisierung deutscher Spitzopolitiker verwendet haben (Fischer 2010). Die Forschung zur Ähnlichkeit bzw. Unähnlichkeit der Beschreibung von Personen anhand solcher Begriffe wird als psycholexikalischer Ansatz bezeichnet.

Psycholexikalischer Ansatz

Psycholexikalischer Ansatz

Unter dem psycholexikalischen Ansatz versteht man die Analyse der Sprache zur Erforschung der Struktur der Persönlichkeit. Ein erster Schritt besteht darin, möglichst alle Wörter zu finden, die sich auf Unterschiede zwischen Menschen beziehen. Besonders einflussreich wurde eine Arbeit von Allport und Odbert (1936). Die Autoren fanden in einem Wörterbuch (Webster's New International Dictionary) 17.953 Wörter, die ihnen geeignet erschienen, das Verhalten eines Menschen von dem eines anderen zu unterscheiden. Das waren immerhin etwa 4,5 % aller englischen Wörter. Allport und Odbert (1936) identifizierten darunter 4504 Wörter, die ihrer Meinung nach eindeutig reale Persönlichkeitsmerkmale „symbolisieren“ (z. B. aggressive, introverted) und führten sie im Anhang ihrer Publikation auf. Dies war weder die erste noch die letzte Studie dieser Art. Während Allport und Odbert (1936) überwiegend Adjektive gesammelt hatten, haben andere Forscherinnen und Forscher Aussagen wie „act as a leader“ oder „don't show my feelings“ zusammengestellt. Die Beispiele sind dem „International Personality Item Pool“ ([► https://ipip.org/](https://ipip.org/)) entnommen, der 3320 solcher Items enthält (s. Goldberg et al. 2006).

Lässt man Personen sich selbst oder auch andere Menschen anhand von Adjektiven oder Aussagen beurteilen (z. B. „trifft zu“ – „trifft nicht zu“), so kann die Korrelation zweier Items als Maß für die Ähnlichkeit der Begriffe oder Aussagen angesehen werden. Mittels Faktorenanalyse werden die der Persönlichkeit zugrunde liegenden Dimensionen gesucht – immer unter der Annahme, dass die Sprache ein Abbild der realen Welt liefert.

5 globale Dimensionen

Fünf-Faktoren-Modell Nach heutigem Stand kann man die Begriffe zur Beschreibung von Persönlichkeit folgenden 5 globalen Dimensionen zuordnen: Neurotismus, Extraversion, Verträglichkeit, Offenheit für Erfahrungen und Gewissenhaftigkeit. Das Modell ist auch unter dem Akronym *OCEAN* bekannt. Dabei werden die Anfangsbuchstaben der englischsprachigen Dimensionsnamen (*Openness to Experience, Conscientiousness, Extraversion, Agreeableness und Neuroticism*) verwendet. Diese *Big Five* können jeweils noch einmal in verschiedene Inhaltsbereiche, Facetten genannt, zerlegt werden. Beispielsweise umfasst Extraversion in einem verbreiteten Persönlichkeitssinventar die Eigenschaften Herzlichkeit, Geselligkeit, Durchsetzungsfähigkeit, Aktivität, Erlebnishunger und Frohsinn ([► Abschn. 3.3.3.5](#)).

Persönlichkeitseigenschaften nur bei Menschen?

Wie bereits erwähnt, entstammen die Big Five dem auf der Alltagssprache basierenden psycholexikalischen Ansatz. Man kann sich jedoch fragen, ob der psycholexikalische Ansatz zwingend zur Beschreibung der Realität führt oder „nur“ unsere Sprache oder Denkweise reflektiert. Es ist erstaunlich, dass dieses Strukturmodell zur Beschreibung der menschlichen Persönlichkeit auch sehr gut geeignet ist, die Persönlichkeit von Hündinnen und Hunden zu beschreiben, wie u. a. in einer viel zitierten Untersuchung (Gosling et al. 2003) belegt wurde. Beurteilerinnen und Beurteiler waren natürlich Menschen. Denkbar ist, dass Hunde und Menschen aufgrund ihrer (entfernten) genetischen Verwandtschaft tatsächlich Ähnlichkeiten in ihrem Verhalten

zeigen. Aber ebenso plausibel ist der Schluss, dass die auffällige Übereinstimmung etwas über das menschliche Denken verrät, das in der Sprache seinen Niederschlag findet. Immerhin schreiben wir auch Objekten und Phänomenen „menschliche“ Eigenschaften zu: Die Säure ist „aggressiv“, ein Felsbrocken „mächtig“ und das Wetter ist „launisch“.

HEXACO-Modell Es existieren aber auch Persönlichkeitsmodelle, in denen mehr als 5 Dimensionen berücksichtigt werden. Wir beschränken uns hier auf ein Modell, zu dem auch ein deutschsprachiger Fragebogen vorliegt. Im HEXACO-Modell (Ashton und Lee 2007) existiert eine 6. Dimension in Form der „Ehrlichkeit-Bescheidenheit“ (englisch Honesty-Humility). Zur inhaltlichen Erläuterung mögen die folgenden beiden Itembeispiele genügen: „Ich würde niemals Bestechungsgeld annehmen, auch wenn es sehr viel wäre“ und „Viel Geld zu haben ist nicht besonders wichtig für mich“ (aus der deutschsprachigen Version von Moshagen et al. 2014). Der Modellname ist ebenfalls ein Akronym und steht für Honesty-Humility, EXtraversion, Agreeableness, Conscientiousness und Openness to Experience. Zum HEXACO-Modell wurde ein Fragebogen entwickelt, der später überarbeitet wurde. Das HEXACO Personality Inventory-Revised (HEXACO-PI-R; Ashton und Lee 2004) liegt in einer Selbst- und einer Fremdbeurteilungsform vor; zu beiden Formen existieren Versionen mit 60 und mit 100 Items. Diese wurden in viele Sprachen übersetzt, darunter auch ins Deutsche (s. ► <https://hexaco.org/>).

Mehr als 5 Dimensionen

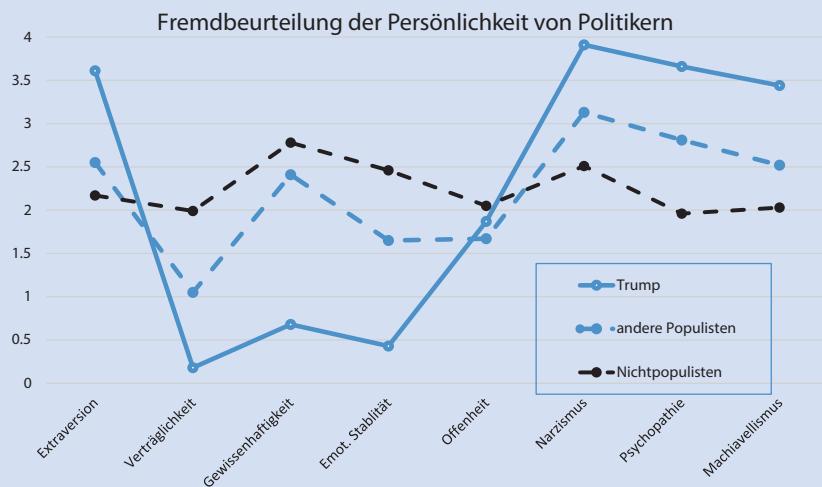
Dunkle Triade Ein völlig anderes Persönlichkeitsmodell ist die sog. „Dunkle Triade“ (engl. Dark Triad; Paulhus und Williams 2002). Dieses Modell erhebt jedoch nicht den Anspruch, alle Bereiche der Persönlichkeit abzudecken. Die 3 Merkmale „Machiavellismus“, „subklinischer Narzissmus“ und „subklinische Psychopathie“ sind sehr viel spezifischer als die „breiten“ Big Five, stellen aber für verschiedene Fragestellungen eine Alternative oder auch eine Ergänzung (s. u.) dar. Machiavellistinnen und Machiavellisten sind skrupellose Machtmenschen. Der italienische Philosoph Niccolò di Bernardo dei Machiavelli hatte in der Renaissance die Verhaltensweisen und die Strategien von erfolgreichen Fürsten analysiert und beschrieben. In seinem Werk *Der Fürst* stellte er fest, dass ein Herrscher zugunsten der Staatsraison auch die Gesetze der traditionellen Moral verletzen dürfe (von Lenz 2013). Narzisstinnen und Narzissen sind selbstverliebte Menschen, die von anderen bewundert werden wollen. Psycho-pathinnen und Psychopathen schließlich sind rücksichtslose und wenig empathische Menschen. Die 3 Merkmale weisen untereinander eine gewisse konzeptuelle Ähnlichkeit auf, sind jedoch hinreichend eigenständig. In einer Metaanalyse (O'Boyle et al. 2012) wurden niedrige bis moderate Korrelationen gefunden ($r = .23$ zwischen Machiavellismus und Narzissmus bis $r = .46$ zwischen Machiavellismus und Psychopathie; korrigiert: $r = .30$ bis $.59$). Die Dunkle Triade wird vor allem für den beruflichen Kontext als bedeutsam angesehen. Auch wenn man vermuten könnte, dass Menschen mit diesen Eigenschaften beruflich erfolgreich sind, so lehrt die Forschung, dass das Gegenteil zutrifft. Die 3 Persönlichkeitsmerkmale korrelieren entweder nicht (Narzissmus) oder schwach negativ mit Berufserfolg. Die 3 Merkmale korrelieren zudem positiv und in geringer (Psychopathie) bzw. moderater Höhe mit kontraproduktivem Verhalten im Beruf (O'Boyle et al. 2012). Zur Dunklen Triade existieren verschiedene Fragebögen. Für den deutschen Sprachraum haben Schwarzinger und Schuler (2016) mit dem Dark Triad of Personality at Work (TOP) einen normierten und validierten Fragebogen vorgelegt. Im Gegensatz zu der zuvor erwähnten Metaanalyse prüften die Autoren auch nichtlineare Zusammenhänge zur beruflichen Leistung. Sie fanden zumindest für die Skala „narzisstische Arbeitshaltung“ einen deutlichen, umgekehrt u-förmigen Zusammenhang dergestalt, dass Menschen mit einem mittleren Wert in narzisstischer Arbeitshaltung höhere Leistungen zeigten als Menschen mit geringen oder hohen Werten.

Die Dunkle Triade mit
Machiavellismus, Narzissmus und
Psychopathie

Reichen die Big Five zur Charakterisierung von Politikerinnen und Politikern?

Persönlichkeitsfragebögen können in der Fremdbeurteilungsversion zur Beschreibung der Persönlichkeit von Politikerinnen bzw. Politikern genutzt werden. Visser et al. (2017) konnten 10 HEXACO-Expertinnen bzw. -Experten dafür gewinnen, die Persönlichkeit von Hillary Clinton und Donald Trump im US-Wahlkampf 2016 zu beurteilen. Sie verwendeten dafür den zuvor bereits erwähnten HEXACO-Fragebogen mit 100 Items. Sie fanden vor allem auf den Dimensionen Verträglichkeit, Gewissenhaftigkeit und Offenheit für Erfahrungen sehr große Unterschiede (Trump erhielt im Vergleich zu Clinton sehr niedrige Werte). Auch die Dimension „Ehrlichkeit-Bescheidenheit“ lieferte interessante Erkenntnisse: Trump wies hier nicht nur niedrigere Werte auf als Clinton, sondern fiel auch mit einem extrem niedrigen Wert, der einem Prozentsatzwert von 0,2 entspricht, auf.

Es ist sehr interessant, Trump mit vielen weiteren Politikerinnen und Politikern im Wahlkampf zu vergleichen. Nai, Martínez i Coma und Maier (2019) konnten umfangreiche Datensätze auswerten. Sie haben die anderen Politikerinnen und Politiker in 2 Gruppen unterteilt, nämlich Populistinnen und Populisten ($N=21$, darunter Marine Le Pen aus Frankreich und Alexander Gauland aus Deutschland) sowie Nichtpopulistinnen und Nichtpopulisten ($N=82$, darunter auch Angela Merkel). Als Beurteilungsinstrument fanden Kurzversionen mit je 2 Items pro Skala zu den Big Five und zu der Dunklen Triade Verwendung. In folgender Grafik ist zu sehen, dass sich Trump nicht nur deutlich von Nichtpopulistinnen und Nichtpopulisten, sondern auch von anderen Populistinnen und Populisten unterscheidet.



Beurteilung von Donald Trump im Vergleich zu anderen populistischen und nichtpopulistischen Politikerinnen und Politikern. Skala von 0 (sehr niedrig) bis 4 (sehr hoch). (Nach Nai et al. 2019, Mittelwerte aus Tab. 2 und 3).

Die Dunkle Triade hilft sehr, über die großen 5 Persönlichkeitsdimensionen hinaus die Unterschiede deutlich zu machen. Bei Narzissmus etwa beträgt der Unterschied zu anderen Populistinnen und Populisten $d=2,37$.

Dunkle Triade und Ehrlichkeit-Bescheidenheit zur Charakterisierung von Politikern hilfreich

Weitere Strukturmodelle Gegenwärtig sind andere, theoriegeleitete Strukturmodelle der Persönlichkeit in der psychologischen Diagnostik in den Hintergrund geraten. Ein Grund dafür ist sicherlich, dass sie in das Big-Five-System „eingepasst“ wurden. Dazu beigetragen hat ein einflussreicher Beitrag von Digman (1990), in dem die Ähnlichkeit bekannter Persönlichkeitsmodelle zum Fünf-Faktoren-Modell aufgezeigt (und vielleicht überbetont) wurde. Ein gutes Beispiel ist die Personality Research Form (PRF), dessen deutsche Version Stumpf et al. (1985) abgefasst haben. Zu diesem Verfahren liegen sehr viele Forschungsergebnisse vor; es hat also eine große internationale Beachtung erfahren. Der Fragebogen umfasst in der amerikanischen Originalversion 22 Skalen und wurde für die deutsche Version auf 14 Inhaltsskalen reduziert. Theoretische Grundlage war die Persönlichkeitstheorie von Murray (1938). Dieser ging davon aus, dass Motive (Bedürfnisse) unser Verhalten mit steuern. Skalen wie „Leistungsstreben“, „Geselligkeit“ oder „Aggressivität“ lassen sich mühelos in das Big-Five-System einordnen – aber wie fügen sich etwa „Ausdauer“, „Bedürfnis nach Beachtung“ oder „spielerische Grundhaltung“ oder (in der amerikanischen Version) „Sinnhaftigkeit“ und „Sexualität“ in das System ein? In einer Studie von Ashton et al. (1998) wurden die Personality Research Form und das Jackson Personality Inventory zusammen mit einem Big-Five-Verfahren einer gemeinsamen Faktorenanalyse (vgl. ▶ Abschn. 2.5.4) unterzogen. Über die Big Five hinaus bildeten sich dabei mehrere Faktoren, die sich nicht den Big-Five zuordnen ließen. Auch Paunonen und Jackson (2000) haben einen Beitrag betitelt mit „What is beyond the big five?“ und gleich die Antwort folgen lassen: „Plenty!“ Durch eine Reanalyse der Daten zur Clusterung von persönlichkeitsbeschreibenden Adjektiven, die zur Begründung der Big Five verwendet wurde, konnten sie eine Reihe weiterer Cluster finden.

Trotz der zuvor genannten alternativen Persönlichkeitsmodelle und ergänzender Faktoren jenseits der „Big Five“ hat sich das klassische Big-Five-Modell aus den folgenden beiden Gründen für die Psychologische Diagnostik als nützlich erwiesen: Erstens liegen sehr viele Forschungsergebnisse zu diesen 5 Persönlichkeitsdimensionen vor. Zweitens steht damit ein weitverbreitetes Einordnungsschema zur Verfügung. Das heißt, man kann damit Verfahren, die nicht explizit diesem Modell verpflichtet sind, konzeptuell einordnen. Manchmal trägt eine Skala nur einen anderen Namen, lässt sich aber inhaltlich einer der 5 Persönlichkeitsfaktoren oder einer Facette davon zuordnen. Beispielsweise entspricht im FPI-R (▶ Abschn. 3.3.3.3) die Skala „emotionale Stabilität“ der Dimension „Neurotizismus“. In anderen Fällen kann eine Skala als Konglomerat aus 2 oder mehr Dimensionen oder auch Facetten des Fünf-Faktoren-Modells der Persönlichkeit identifiziert werden.

Fazit Fazit ist, dass wir mit dem Big-Five-Modell ein sehr nützliches Referenzsystem haben, das zur Verortung anderer Verfahren, d. h. zur Beurteilung ihrer Konstruktvalidität, verwendet werden kann. Die Korrelationen der Skalen eines neuen Verfahrens mit einem Big-Five-Fragebogen geben Aufschluss darüber, ob die Skalen etwas Neues messen oder weitgehend das Gleiche wie die Big Five. Jenseits der Big Five gibt es aber noch andere Persönlichkeitsmerkmale.

Andere Persönlichkeitsmodelle werden heute weniger beachtet

Fünf-Faktoren-Modell als Referenzsystem nützlich

3.3.1.3 Selbst- und Fremdeinschätzung

Selbsteinsicht Mit Ausnahme der zuvor erwähnten Hunde haben wir zuvor vorausgesetzt, dass Befragte in der Lage sind, ihre Persönlichkeitsmerkmale selbst zu beschreiben. Es wird also davon ausgegangen, dass der Gegenstand der Psychologie, das „Erleben und Verhalten“, selbst reflektiert und

Erleben und Verhalten können durch Selbstauskünfte erfasst werden

beschrieben werden kann. Die Messung der Persönlichkeit zielt genau darauf ab. Das Erleben kann durch Aussagen wie „Ich leide unter Selbstzweifel“ oder „Ich mache mir viele Sorgen“ erfasst werden. Das Verhalten kann mit Aussagen wie „Ich schreie manchmal andere Leute an“ oder „Wenn mich jemand provoziert, werde ich schnell handgreiflich“ beschrieben werden. Solche Items lassen sich Persönlichkeitsmerkmalen wie Neurotizismus bzw. Aggressivität zuordnen.

Fremdbeurteilung der Persönlichkeit

Fremdbeurteilungen In vielen Fällen werden Erleben und Verhalten aber auch mittels Fremdbeurteilung erfasst. Eine Fremdbeurteilung liegt vor, wenn eine andere Person etwa in einem Fragebogen ankreuzt, „Er/sie schreit manchmal andere Leute an“ oder „Wenn er/sie provoziert wird, wird er/sie schnell handgreiflich“. Einige Persönlichkeitsfragebögen liegen in einer Selbst- und einer Fremdbeurteilungsversion vor. Die Items werden in der Regel nur in die entsprechende sprachliche Form transformiert („er/sie“ statt „ich“). Das Verhalten ist zumeist auch durch andere Menschen beobachtbar. Deshalb ist es leicht nachvollziehbar, dass Fremdbeurteilungen möglich sind, sofern die beurteilende Person viel Zeit gemeinsam mit der anderen Person verbringt. Das kann etwa bei Arbeitskolleginnen und -kollegen, Paaren, Eltern und Kindern der Fall sein. Fragebogenskalen in der Selbst- und Fremdbeurteilungsform korrelieren in der Größenordnung von $r = .50$ miteinander. Die Höhe der Korrelationen hängt vom untersuchten Persönlichkeitsmerkmal und vom Grad der Bekanntheit ab. Stammen die Fremdbeurteilungen von engen Angehörigen, liegen die beobachteten Korrelationen bei .44 (Neurotizismus), .54 (Extraversion), .46 (Offenheit), .35 (Verträglichkeit) und .37 (Gewissenhaftigkeit). Nach Korrektur für die Reliabilität der Urteile (Cronbachs α , bei den Fremdbeurteilungen wurden Übereinstimmungen der Beurteilenden herangezogen) stiegen die Korrelationen auf .69, .74, .72, .51 bzw. .54 (Connolly et al. 2007).

Wahl zwischen Selbst- und Fremdbeurteilung

Damit wird deutlich, dass beides 2 Seiten einer Münze sind. Selbst- und Fremdbeurteilung stellen in der Psychologischen Diagnostik sich ergänzende Perspektiven dar. In der Partnerschaftsdiagnostik sowie im beruflichen Kontext können sich beide Ansätze beispielsweise gut ergänzen. Zumeist wird jedoch die Selbstbeurteilungsvariante von Fragebögen eingesetzt. Dafür sprechen ökonomische Gründe und die nur begrenzte Verfügbarkeit von anderen Personen, die eine Fremdbeurteilung abgeben könnten. In bestimmten Situationen werden nur Fremdbeurteilungsinstrumente eingesetzt. Das ist bei der Persönlichkeitsdiagnostik von Kindern und von Erwachsenen, die keine zutreffende Selbstauskunft geben können oder wollen, der Fall. Ersteres könnte beispielsweise der Fall sein, wenn Personen mit der Beantwortung kognitiv überfordert sind oder unzureichende einschlägige Sprachkenntnisse haben. Letzteres liegt vor, wenn Personen an einem bestimmten Ergebnis der Befragung interessiert sind (z. B. einer positiven Selbstdarstellung).

Erleben und Verhalten nicht identisch

Doch wie sind Aussagen wie die folgenden über andere Personen zu beurteilen: „Er/sie/es leidet unter Selbstzweifeln“, „.... macht sich viele Sorgen“ oder „.... ist ein fröhlicher Mensch“? Wie bei der Selbstbeurteilung können solche Aussagen bei der Bearbeitung eines Fragebogens, in einem diagnostischen Interview oder in einer freien schriftlichen Beschreibung fallen. Es fällt es uns nicht schwer, ein Urteil darüber abzugeben, ob eine andere Person Angst hat, wütend ist oder sich etwa schämt. In Wahrheit sind das Schlussfolgerungen über das Erleben eines anderen Menschen, wobei wir Mimik, Gestik, Stimme, den situativen Kontext etc. implizit als Informationsquelle nutzen. Die beurteilte Person verwendet zumindest teilweise andere Informationen, wenn sie ein Urteil über ihr Erleben abgibt. Wenn sie etwa Angst

empfindet, nimmt sie wahrscheinlich körperliche Symptome wie Herzklopfen, „Kloß im Hals“ oder feuchte Hände wahr. Sie merkt vielleicht, dass ihre Gedanken immer wieder um die vermeintliche Bedrohung und die eigene Hilflosigkeit in der aktuellen Situation kreisen. All dies ist für Außenstehende unsichtbar. Übrigens ist der Zusammenhang zwischen dem emotionalen Erleben einer Person und sogar objektiv erfassten „Ausdruckserscheinungen“ von Emotionen (z. B. bestimmte mimische Reaktionen) sehr schwach (s. z. B. Schmidt-Atzert et al. 2014). Deshalb ist es angemessen, zu sagen, dass das Erleben nicht der Fremdbeurteilung zugänglich ist. Beurteilt wird vielmehr das Ausdrucksverhalten, das aber nicht als Spiegel des Erlebens anzusehen ist.

► Beispiel

Der Erstautor dieses Buchkapitels erinnert sich gut an folgende Situation: Eine Studentin hatte ein Referat zu halten. In der nicht verpflichtenden Vorbesprechung hatte sie sich sehr besorgt gezeigt und davon berichtet, dass sie Angst vor dem Referat habe. Das Referat war von hoher Qualität. Die Referentin wirkte souverän, ruhig und entspannt. Wie üblich bot er nach dem Referat eine Rückmeldung an, für die sie sich interessierte. Die Feststellung, sie habe gar nicht nervös, unsicher oder gar ängstlich gewirkt, überraschte sie. Sinngemäß antwortete sie, dass sie das nicht verstehe; „in ihr“ habe es ganz anders ausgesehen. Sie habe sich angespannt, unsicher und ängstlich gefühlt. ◀

Weder eine Selbst- noch eine Fremdeinschätzung liefert ein Bild der „wahren“ Persönlichkeit eines Menschen. Besonders bei diskrepanten Ergebnissen stellt sich die Frage nach dem Warum. Diskrepanzen können diagnostisch sehr bedeutsam sein, wenn sich schlüssige Erklärung dafür finden lassen. Im Folgenden werden einige Interpretationsmöglichkeiten genannt:

- Unterschiedliche Urteilsbasis: Das Erleben und das Verhalten sind nicht deckungsgleich (s. o.).
- Unrealistisches Selbstkonzept: Die Person nimmt sich selbst auf eine Weise wahr, die mit dem Fremdurteil anderer Personen in Widerspruch steht. Die Selbstwahrnehmung kann in die positive, aber auch in die negative Richtung verzerrt sein. Ein sozial erwünschtes Selbstkonzept kann motivational als Selbstdäuschung (► Abschn. 3.3.2) erklärt werden. Ein sehr negatives Selbstkonzept ist eventuell Ausdruck einer psychischen Störung.
- Strategische Selbstdarstellung: Die Angaben im Fragebogen werden an den erwarteten Nutzen angepasst; die Person beschreibt sich nicht so, wie sie sich selbst sieht, sondern wie sie gerade gerne gesehen werden möchte (Impression Management; ► Abschn. 3.3.2).
- Urteilsfehler bei den Beobachtenden: Es können verschiedene Beurteilungsfehler auftreten. Beispielsweise tendieren manche Personen eher zu positiven Urteilen (Milde-Effekt; ► Abschn. 3.6.4).
- Messfehler: Weder Selbst- noch Fremdbeurteilungen sind hoch reliabel. Diskrepanzen können daher alleine aufgrund der begrenzten Messgenauigkeit der Verfahren auftreten. Durch Berechnung von kritischen Differenzen (► Abschn. 2.6.2.2) lässt sich feststellen, ob eine beobachtete Diskrepanz wahrscheinlich zufällig zustande gekommen ist oder nicht. Vorsicht ist geboten, wenn bei mehrdimensionalen Persönlichkeitsfragebögen Profile verglichen werden. Je mehr Skalen verglichen werden, desto größer ist die Gefahr, eine Diskrepanz zu „entdecken“, die nur messfehlerbedingt zustande gekommen ist.

Warum Selbst- und Fremdbeschreibungen voneinander abweichen

Fremdbeurteilungen des Erlebens

Hinweis auf weitere diagnostische Verfahren

Grundsätzlich sollten Erklärungen für eine Diskrepanz zwischen Selbst- und Fremdbericht als Hypothese behandelt werden, die es zu prüfen gilt. So kann von einer anderen Person eine Fremdbeurteilung eingeholt werden, wenn man einen Urteilsfehler vermutet. Stehen Messfehler im Verdacht, kann die Selbst- und/oder Fremdbeurteilung mithilfe einer Parallelform des Fragebogens repliziert werden.

Auch wenn die Fremdbeurteilung des Erlebens kein Ersatz für die Selbstbeurteilung sein kann, können sich beide Sichtweisen sehr gut ergänzen.

3.3.1.4 Andere Möglichkeiten zur Erfassung von Persönlichkeitsmerkmalen

Zuvor wurde die Diagnostik von Persönlichkeitsmerkmalen mittels subjektiver Selbst- oder auch Fremdbeurteilungen vorgestellt. Die Persönlichkeit kann aber auch mit anderen Methoden erfasst werden. Objektive Persönlichkeitstests (► Abschn. 3.4) und projektive Verfahren (► Abschn. 3.5) gehören zu diesen Testverfahren. In *objektiven Persönlichkeitstests* soll die Testperson bestimmte Aufgaben bearbeiten, die so gewählt sind, dass das Lösungsverhalten beispielsweise etwas über die Leistungsmotivation der Testperson aussagt. Bei den *projektiven Verfahren* werden standardisierte Reize, z. B. Bilder von Tintenklecksen oder Zeichnungen von Menschen in einer Situation, anstelle von Textitems verwendet. Die Testperson soll zumeist eine freie Antwort geben. Dazu ist ein genau definierter Auftrag („Eine Geschichte zu jedem Bild erzählen“) oder eine Frage (bei den Tintenklecksen: „Was könnte das sein“) nötig. Das Antwortverhalten wird inhaltsanalytisch ausgewertet. Aus den Reaktionen der Testperson kann auf bestimmte Persönlichkeitseigenschaften geschlossen. *Verhaltensbeobachtung und -beurteilung* (► Abschn. 3.6) werden oft nicht in so hoch standardisierter Form durchgeführt wie Tests. Sie dienen der Messung von Verhalten. Die Ergebnisse können dazu verwendet werden, Schlussfolgerungen über Persönlichkeitseigenschaften abzuleiten. Schließlich kann man auch mithilfe von Gesprächen Erkenntnisse über die Persönlichkeit gewinnen; das *diagnostische Interview* wird in ► Abschn. 3.7 behandelt. Alle genannten Verfahren haben Vor- und Nachteile, die in □ Tab. 3.12 in kurzer Form aufgeführt sind. Für eine ausführliche Darstellung sei auf die entsprechenden folgenden Abschnitte verwiesen. Wir bleiben zunächst weiter bei den Persönlichkeitsfragebögen und betrachten ausführlich deren Vor- und Nachteile. Anschließend werden verschiedene Fragebögen vorgestellt.

3.3.2 Allgemeine Vor- und Nachteile von Persönlichkeitsfragebögen

Skalen zu vielen Persönlichkeitsmerkmalen

Fragebögen wurden zur Messung von vielen verschiedenen Persönlichkeitsmerkmalen konstruiert. Mehrdimensionale Verfahren erfassen gleichzeitig viele Merkmale; das NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO PI-R; ► Abschn. 3.3.3.5) etwa liefert Informationen über 30 Einzelaspekte (Facetten) der Persönlichkeit sowie über 6 übergeordnete Dimensionen. Bei Instrumenten, die nur ein Merkmal messen sollen, ist eine wahre Inflation zu verzeichnen: Ständig werden neue Konstrukte entwickelt und Fragebögen dazu konstruiert. Viele dieser Verfahren sind allerdings meist nur für Forschungszwecke geeignet, solange sie nicht normiert sind.

Tab. 3.12 Übersicht über diagnostische Verfahren zur Messung der Persönlichkeit

Verfahren	Vorteile	Nachteile/Einschränkungen
Persönlichkeitsfragebogen	<ul style="list-style-type: none"> – Für viele Persönlichkeitsmerkmale verfügbar – Erfassung von (für andere Menschen) nicht beobachtbarem Verhalten/Erleben – Mittels Normen Vergleich mit anderen Menschen möglich – Ökonomisch 	<ul style="list-style-type: none"> – Selbsteinsicht nötig – Anfällig für Selbsttäuschung – Verfälschbar
Diagnostisches Interview	<ul style="list-style-type: none"> – Im Prinzip auf alle Persönlichkeitsmerkmale anwendbar – Erfassung von (für andere Menschen) nicht beobachtbarem Verhalten/Erleben – Eventuelle mangelnde Selbsteinsicht erkennbar 	<ul style="list-style-type: none"> – Selbsteinsicht nötig – Anfällig für Selbsttäuschung – Verfälschbar – Vergleich mit anderen Menschen nur bedingt möglich
Verhaltensbeobachtung und -beurteilung	<ul style="list-style-type: none"> – Im Prinzip auf alle Persönlichkeitsmerkmale anwendbar – Unabhängig von Selbsteinsicht 	<ul style="list-style-type: none"> – Nur für sehr wenige Merkmale standardisierte Verfahren vorhanden – Verfälschbar – Beschränkung auf beobachtbares Verhalten – Vergleich mit anderen Menschen nur bedingt möglich – Konstruktionsaufwand hoch
Objektive Persönlichkeitstests	<ul style="list-style-type: none"> – Unabhängig von Selbsteinsicht – Mittels Normen Vergleich mit anderen Menschen möglich 	<ul style="list-style-type: none"> – Nur wenige Tests verfügbar – Validität meist noch nicht hinreichend geklärt
Projektive Verfahren	<ul style="list-style-type: none"> – Unabhängig von Selbsteinsicht – Teils mittels Normen Vergleich mit anderen Menschen möglich 	<ul style="list-style-type: none"> – Nur für wenige Merkmale Verfahren vorhanden – Validitätsprobleme

3.3.2.1 Vorteile

Ein Vorteil der Fragebogenmethode liegt in der *standardisierten Erhebung*. So zeichnet sich die Fragebogenmethode dadurch aus, dass vorformulierte Fragen oder Feststellungen in schriftlicher Form vorgelegt werden. In einer Anleitung wird die gewünschte Art der Bearbeitung zuvor erläutert. Darin kann festgelegt sein, dass die Fragen in der vorgegebenen Reihenfolge, ehrlich und ohne langes Überlegen zu beantworten sind. Meistens wird betont, dass die Testpersonen alle Items beantworten sollen, und wenn die Entscheidung einmal schwerfällt, soll die am ehesten zutreffende Antwortalternative gewählt werden.

Das *Antwortformat* ist ebenfalls festgelegt. Prinzipiell können freie Antworten vorgesehen sein oder das Ankreuzen einer von mehreren Alternativantworten. Die freie Beantwortung spielt jedoch praktisch keine Rolle, weil sie eine aufwendige Auswertung nach sich zieht. Bei den gebundenen Antworten sind dichotome Antwortformate wie „Ja“ – „Nein“ oder „Trifft zu“ – „Trifft nicht zu“, Ratingskalen (z. B. von 0 = „trifft überhaupt nicht zu“, bis 6 = „trifft völlig zu“) und Forced-Choice-Antworten gebräuchlich. Letztere zeichnen sich dadurch aus, dass mehrere ausformulierte Antworten zur Auswahl stehen, die etwa gleich stark sozial erwünscht sind. Eine der Antworten indiziert das zu messende Merkmal (zu Antwortformaten vgl. ► Abschn. 2.4.2.6). Die Testpersonen fragen manchmal, wie eine Frage oder Aussage genau zu verstehen ist. Deshalb enthält das Testmanual zumeist Hinweise, wie mit Nachfragen umzugehen ist. Diese Maßnahmen führen dazu, dass die Bearbeitung in hohem Maße standardisiert und somit die Durchführungsobjektivität gewährleistet ist.

Durchführung und Auswertung leicht zu standardisieren

In der Regel wird jedes Merkmal, das mit dem Fragebogen erfasst werden soll, durch mehrere Items repräsentiert. Daher werden nicht einzelne Antworten interpretiert, sondern es werden die Antworten gezählt, die für eine hohe Merkmalsausprägung sprechen. Bei einem dichotomen Antwortformat werden die Antworten meist mit 0 und 1 (Ankreuzen spricht für hohe Merkmalsausprägung) kodiert. Bei mehrstufigen Antworten werden den Antwortkategorien in der Regel Zahlen (etwa 0 – 1 – 2 – 3 – 4 bzw. 4 – 3 – 2 – 1 – 0 bei invertierten Items) zugewiesen, die zu addieren sind. Die *Auswertung* ist in der Regel standardisiert; zumeist werden mithilfe von Schablonen oder einem Auswertungsprogramm merkmalsspezifische Antworten ausgezählt. Es wird geregelt, wie mit unbeantworteten Items oder Mehrfachankreuzungen umzugehen ist. Damit wird die Auswertungsobjektivität sichergestellt. Bei computerbasierten Tests übernimmt die Software die Summenbildung und auch gleich die Transformation in Normwerte. Nähere Informationen zur Konstruktion von Items finden sich in ▶ Abschn. 2.4.2.

Zugang zu Informationen, die der Beobachtung nicht zugänglich sind

Ein weiterer Vorteil besteht darin, dass Informationen erhoben werden können, die einer Beobachtung durch Außenstehende unzugänglich sind. So können alle Ereignisse, die in der Vergangenheit liegen, heute nicht mehr beobachtet werden. In einem Persönlichkeitsfragebogen wird die Testperson normalerweise Fragen zu ihrem früheren Verhalten und zu zurückliegenden Ereignissen beantworten können. Verhaltensweisen wie Drogenkonsum, die Ausübung von Sexualpraktiken oder das Begehen von Straftaten entziehen sich meist der Beobachtung: Die zu untersuchende Person wird einer Beobachtung nicht zustimmen. Die Bereitschaft, in einem Fragebogen Angaben zu diesen Themen zu machen, wird dagegen größer sein. Was Menschen empfinden (Schmerz, Hunger, Durst etc.), welche Gefühle sie haben (Angst, Traurigkeit, Ärger etc.), was sie denken (wie sie andere Menschen oder Situationen beurteilen, Vorlieben und Aversionen, Einstellungen etc.) und von welchen Motiven sie sich leiten lassen, ist per se nicht beobachtbar. Fragebögen stellen in all diesen Fällen eine gute Zugangsmöglichkeit dar.

Vergleich mit anderen Menschen durch Normen

Für viele diagnostische Fragestellungen will man wissen, wie ausgeprägt ein Persönlichkeitsmerkmal ist. Normierte Fragebögen liefern diese Information. Oftmals erlauben unterschiedliche Normtabellen Vergleiche mit unterschiedlichen Bezugsgruppen, etwa mit Personen gleichen Geschlechts und/oder gleichen Alters.

Ökonomisch

Persönlichkeitsfragebögen sind eine sehr *ökonomische Methode*. Der Testleiter bzw. die Testleiterin braucht in der Regel nur ein paar einführende Worte zu sagen, vielleicht ist noch die Instruktion vorzulesen. Die Bearbeitung kann die Testperson in der Regel ohne weitere Hilfe alleine vornehmen. Eine Durchführung in Gruppen ist möglich. Unter Umständen kann ein Persönlichkeitsfragebogen zu Hause aufgefüllt werden – eventuell sogar über das Internet. Die Bearbeitung ist zudem meist nicht zeitintensiv. Selbst für die Beantwortung der über 500 Items des Minnesota Multiphasic Personality Inventory-2 (MMPI-2; ▶ Abschn. 3.3.3.1) benötigen gesunde Personen nur etwa 1 h, Patientinnen und Patienten allerdings etwas länger. Die Auswertung erfolgt automatisch, wenn der Fragebogen am Computer bearbeitet wurde.

Robuste Methode, die Fehler verzeiht

Bei der *Formulierung von Fragen* können viele Fehler passieren, und es gibt viele Ratschläge, was man alles beachten soll. Auch wenn die Ratschläge beachtet werden, sind die Items nicht zwangsläufig perfekt. Begriffe wie „gewöhnlich“, „häufig“, „selten“ und selbst „nie“ bedeuten nicht für alle

Menschen das Gleiche. Die einfache Aussage „Ich bin morgens oft müde“ kann an 3 Stellen unterschiedlich interpretiert werden: Von wann bis wann ist „morgens“? Wie häufig ist „oft“? Was bedeutet „müde“? Aus diesen Überlegungen könnte man folgern, dass Fragebögen für alle möglichen Fehler sehr anfällig sind. Die Empirie deckt sich aber nicht mit dieser Vermutung. Tatsächlich finden sich selbst in dem publizierten NEO-Fünf-Faktoren-Inventar (NEO-FFI; ▶ Abschn. 3.3.3.5) „schlechte“ Items und der Fragebogen „funktioniert“ gut. Pargent et al. (2018) haben dieses Inventar in einer experimentellen Studie einmal im Original, einmal in einer sprachlich verbesserten Version und einmal in einer verschlechterten Version bearbeiten lassen. In der verschlechterten Version wurden absichtlich viele Regeln der Itemkonstruktion verletzt. Erstaunlicherweise fanden sich kaum Unterschiede bei den psychometrischen Kennwerten.

3.3.2.2 Einschränkungen

Gültige Antworten in einem Fragebogen kann nur eine Person geben, die sich selbst beobachtet hat und über Wissen verfügt, das auf diesen Beobachtungen aufbaut. Menschen, bei denen eine geistige Behinderung vorliegt oder die an einer schweren psychiatrischen Störung leiden, erfüllen diese Voraussetzungen nicht unbedingt. Einige Testautorinnen und -autoren raten explizit vom Einsatz ihres Persönlichkeitsfragebogens ab, wenn die Testperson nicht über ein näher spezifiziertes Mindestmaß an Intelligenz verfügt. Damit wird sichergestellt, dass die Testpersonen den Sinn der Fragen verstehen. Aber auch bei Personen mit einer wenigstens durchschnittlichen Intelligenz sind manchmal Zweifel erlaubt, ob sie über die nötige Selbsteinsicht verfügen. Die Schwierigkeiten, Selbstbeobachtungen vorzunehmen, deren Resultate korrekt abzuspeichern und zu erinnern sowie Urteile darüber abzugeben, sollten nicht unterschätzt werden.

Selbsteinsicht nötig

Beispiel für trügerische Erinnerungen

Die Ergebnisse einer schwedischen Studie zum Rauchen in der Schwangerschaft belegen eindrucksvoll, dass selbst eindeutige und persönlich bedeutsame Verhaltensweisen manchmal falsch erinnert werden. Post et al. (2008) erhoben mit einem Fragebogen, ob und ggf. wie stark Mütter während der Schwangerschaft geraucht hatten. Die Befragung wurde durchgeführt, als deren Kinder bereits 11 Jahre alt waren. Die retrospektiven Angaben der Mütter konnten mit Angaben verglichen werden, die sie während der Schwangerschaft im Rahmen von medizinischen Routinebefragungen gemacht hatten. Wenn die Angaben in „Rauchen“ oder „nicht Rauchen“ dichotomisiert wurden, fanden sich in 9,4 % der Fälle Diskrepanzen zwischen beiden Erhebungen.

Erinnern fehleranfällig

Untersuchungen zum autobiografischen Gedächtnis haben gezeigt, dass es sich beim Erinnern nicht lediglich um die Aktivierung von Gespeichertem handelt, sondern um die Rekonstruktion vergangener Ereignisse mit heuristischen Strategien (Schwarz und Sudman 1994). Dabei kommt es schon während des Einspeicherns (Encoding), später auch beim Abruf (Retrieval) zu Ungenauigkeiten und systematischen Verzerrungen der Gedächtnisinformation. In einem (englischsprachigen) Wikipedia-Beitrag findet sich eine eindrucksvoll lange Liste mit „memory biases“ (▶ https://en.wikipedia.org/wiki/List_of_memory_biases). Des gesamte Prozess der Erinnerung, also die

Retrospektive Aussagen sind mentale Repräsentationen von subjektivem Erleben und Verhalten

3

Begriffe wie „häufig“ und „nie“ variieren in ihrer Bedeutung

Komplexe Urteilsprozesse

Verzerrung durch Selbstdäuschung

Verfälschung verändert Ergebnisse

Enkodierung, die Speicherung und der Abruf von Informationen, kann zudem durch Emotionen beeinflusst werden, wie verschiedene Experimente zeigen (s. Schmidt-Atzert et al. 2014, S. 256 ff.). Aufgrund all dieser und zahlreicher weiterer Forschungsbefunde können die retrospektiven Aussagen über Verhalten nicht mit dem Verhalten selbst gleichgesetzt werden. Vielmehr handelt es sich nur um mentale Repräsentationen von subjektivem Erleben und Verhalten.

Urteile, die in Fragebögen verlangt werden, sind oft zu quantifizieren. Beispielsweise ist auf einer Häufigkeitsskala ein Kreuz zu setzen, oder in einem Item ist eine Häufigkeitsaussage enthalten (z. B. „Ich bin oft erschöpft“). Gezielten Untersuchungen zufolge verstehen die Menschen Unterschiedliches unter Begriffen wie „gewöhnlich“, „häufig“ oder „selten“, und selbst „nie“ bedeutet keineswegs durchgängig die Auftretenswahrscheinlichkeit 0.

Manchmal sind Fragebogenitems mehrdeutig und daher komplex. So dürfte es schwierig sein, auf Items wie „Übernehmen Sie bei gemeinsamen Aktionen gern die Führung?“ eine angemessene Antwort zu geben. Handelt es sich um eine „gemeinsame“ Aktion, wenn eine weitere Person, etwa die Partnerin, mit von der Partie ist, oder ist dazu eine größere Gruppe nötig? Verlangt eine „Aktion“, dass ein Verhalten aus eigenen Stücken erfolgt, oder darf es auch durch externe Zwänge bestimmt sein? Den höchsten Komplexitätsgrad erreichen schließlich Beurteilungen, die unmittelbar eine Einstufung auf der entsprechenden Eigenschaftsdimension erfordern (z. B. „Im Großen und Ganzen bin ich ein ehrlicher Mensch“). Hier müssten eigentlich aus dem Gedächtnisspeicher ganze Serien von situativen und temporären Verhaltensstichproben abgerufen werden. Oder die befragte Person erinnert nur einige prototypische Ereignisse, die etwas über die eigene Ehrlichkeit aussagen. Für das Gesamтурteil sind die Häufigkeit und die Schwere von aufrichtigen Verhaltensweisen abzuschätzen. Zusätzlich sind auch noch Annahmen über die durchschnittliche Ehrlichkeit anderer nötig. Die Stärke eigener Merkmalsausprägungen erfährt nämlich in Ermangelung von absoluten Anhaltspunkten eine Relativierung durch die bei den Mitmenschen wahrgenommene (oder nur vermutete) Eigenschaftsausprägung.

Angesichts solch komplexer Urteilsprozesse ist die Annahme, dass die Befragten eine realistische Beschreibung ihres Verhaltens bzw. ihrer Verhaltensgewohnheiten liefern, wenig plausibel. Die Antworten geben vielmehr *mentale Repräsentationen* wieder, z. B. ein Bild, das sich Menschen über sich gemacht haben. Verständlicherweise gestehen sich viele Menschen ihre Schwächen nicht gerne ein. Daher findet oft eine Verzerrung in Richtung eines sozial erwünschten Bildes statt. Paulhus (1984) hat dieses „Vor-sich-selbst-gut-dastehenden-Wollen“ Selbstdäuschung (engl. self-deception) genannt und von dem Wunsch unterschieden, vor anderen einen guten Eindruck machen zu wollen (Impression Management).

In der Tat muss man konstatieren, dass die Items der meisten Persönlichkeitsfragebogen durchschaubar sind: Ein durchschnittlich intelligenter Mensch kann erkennen, ob eine zustimmende oder ablehnende Antwort für ihn vorteilhaft ist. Damit besteht bei vielen Untersuchungsanlässen die Gefahr, dass die Testperson absichtlich versucht, einen guten oder einen schlechten Eindruck zu erwecken. Ein negativer Gesamteindruck kann im Interesse eines Klienten oder einer Klientin liegen, wenn ein Therapiewunsch besteht. Eine typische Situation, die zu einer positiven Selbstdarstellung verführt, ist eine eignungsdiagnostische Untersuchung zur Personalauswahl. Mit anderen Worten: Manchmal besteht ein Interesse an einer falschen Selbstdarstellung.

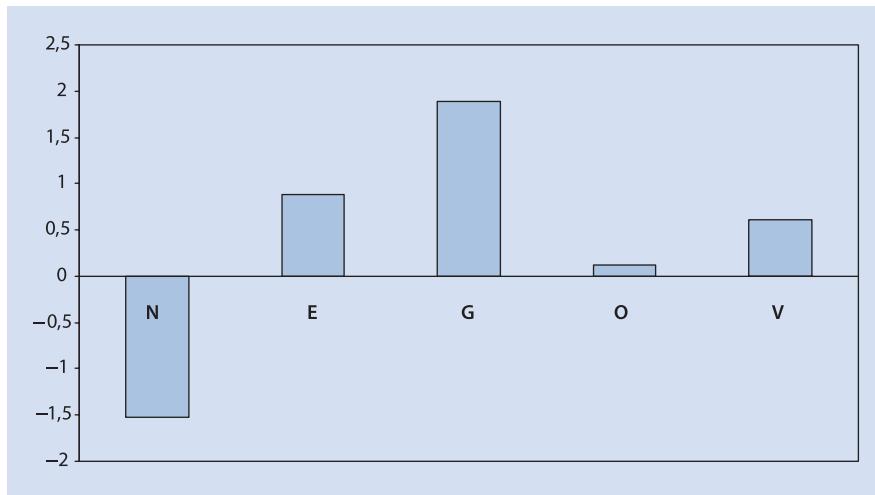
Um die Anfälligkeit diagnostischer Instrumente für Verfälschung zu prüfen, wird oft mit „Faking-Instruktionen“ gearbeitet; das Vorgehen wird anhand einer Untersuchung von Ziegler und Bühner (2009) erläutert: Studentische Versuchspersonen wurden in eine Experimental- und eine Kontrollgruppe aufgeteilt. In der Kontrollgruppe wurden sie gebeten, ehrlich zu antworten, in der Experimentalgruppe sollten sie sich vorstellen, dass sie sich um einen Studienplatz bewerben. Sie sollten versuchen, einen guten Eindruck zu machen, aber dabei so geschickt vorgehen, dass es eine Expertin oder ein Experte nicht merkt. Die Ergebnisse in □ Abb. 3.17 machen deutlich, dass die Angaben im Persönlichkeitsfragebogen situationsgerecht verfälscht wurden: Die „Studienplatzbewerber“ beschrieben sich im Vergleich zur Kontrollgruppe vor allem als gewissenhafter und weniger neurotisch. Offenheit für Erfahrung war von der Verfälschung nicht betroffen; die „Bewerberinnen und Bewerber“ versprachen sich vermutlich keinen Nutzen davon, als besonders offen für Erfahrungen zu gelten. Das hier beschriebene Verhalten würde in der Terminologie von Paulhus (1984) als *Impression Management* eingeordnet werden.

Das Bemühen, einen guten Eindruck zu hinterlassen, führt nicht nur zu einer Verschiebung von Skalenmittelwerten, sondern verändert auch die Konstruktvalidität des Fragebogens. Skalen, die üblicherweise unabhängig voneinander sind, korrelieren nun moderat bis hoch. In der Studie von Ziegler und Bühner (2009) korrelierten Neurotizismus und Gewissenhaftigkeit in der Experimentalgruppe zu $-.72$ miteinander.

Es wäre interessant, zu erfahren, wie sich Bewerberinnen und Bewerber in einer realen Auswahlsituation verhalten. Aus ethischen und rechtlichen Gründen ist es nicht möglich, sie in einer realen Auswahlsituation zu faking good aufzufordern. Aber es ist möglich, den gleichen Persönlichkeitsfragebogen nach Zulassung zum Studium noch einmal bearbeiten zu lassen und dabei zu faking good aufzufordern.

Anfälligkeit von Fragebögen für Verfälschung prüfbar

Konstruktvalidität verändert sich bei faking



□ Abb. 3.17 Verfälschung eines Persönlichkeitsinventars in einer imaginären Auswahlsituation. Angegeben sind die Abweichungen von einer neutralen Situation in Standardabweichungen der Kontrollbedingung. N = Neurotizismus, E = Extraversion, G = Gewissenhaftigkeit, O = Offenheit für Erfahrung, V = Verträglichkeit. (Nach Ziegler und Bühner 2009, Tab. 1)

Wiederholung von Tests unter verschiedenen Bedingungen

Genau das haben Krammer et al. (2017) getan. In einer Kontrollbedingung sollten die Lehramtsstudierenden den Fragebogen möglichst so wie beim ersten Mal ausfüllen (Reproduktion), in 2 weiteren Bedingungen wurden sie explizit zu einer ehrlichen Beantwortung oder zu faking good aufgefordert. Bei ehrlicher Bearbeitung des Fragebogens im Auswahlverfahren sollten die Skalenwerte am besten mit der Testwiederholung unter der „Ehrlichkeitsbedingung“ übereinstimmen und bei faking good mit der gleichnamigen Bedingung. Das Auswahlinstrument umfasste 3 Skalen, von denen wir hier nur 2 betrachten.

Bei der Skala „Gesundheitsvorsorge bei Warnsignalen“ bestand die höchste Korrelation mit der Reproduktionsbedingung ($r=.70$), die beiden anderen Korrelationen fielen niedriger aus ($r=.39$). Die Mittelwerte der Reproduktions- und der Ehrlichkeitsbedingung waren identisch; bei faking good fiel der Mittelwert im Vergleich dazu etwas höher aus. Dieses Ergebnismuster spricht eher für eine in Richtung faking good gehende Testbearbeitung im Ernstfall.

Die Skala „Selbstvertrauen bei Prüfungsanforderung“ korrelierte unter der Faking-good-Bedingung praktisch nicht mit der Ersttestung ($r=.12$), während bei den beiden anderen Bedingungen moderate Korrelationen bestanden ($r=.53$ und $.44$; beide unterschieden sich nicht signifikant). In der Faking-good-Bedingung war der Skalenmittelwert etwas höher als bei den beiden anderen Bedingungen. Das spricht dafür, dass alle Testpersonen bei faking good etwas mehr Selbstvertrauen angaben.

Insgesamt wirken die Ergebnisse eher uneinheitlich. Man kann sagen, dass die Studie Hinweise darauf liefert, dass in der Auswahlsituation wohl kein ganz ehrliches Antwortverhalten, aber auch kein erhebliches faking good vorliegt. Die Effekte treten nicht pauschal auf, sondern sind in Abhängigkeit von den verwendeten Skalen oder präziser deren (vermuteter) Relevanz für den Studienplatz zu sehen. Die Autoren schlussfolgern, dass man die schädlichen Effekte von faking good nicht überbewerten sollte. Mit anderen Worten: Es gibt solche Effekte in einem Auswahlverfahren, sie sind aber nicht dramatisch.

Faking good in einer realen Auswahlsituation

Steht diese Einschätzung nicht im Widerspruch zu Erkenntnissen aus Faking-Studien im Labor (s. o.)? Auf den ersten Blick mag das so sein. Man muss aber bedenken, dass das Verhalten im Labor nicht einfach auf das Verhalten im Feld generalisiert werden kann. Auch Ones et al. (2007) argumentieren, dass die Rolle von Faking und sozial erwünschtem Verhalten in realen Auswahlsituationen überschätzt wird.

3.3.2.3 Verhinderung, Kontrolle und Ignorieren einer sozial erwünschten Selbstdarstellung

Wenn im Einzelfall dennoch zu befürchten ist, dass sich getestete Personen sozial erwünscht darstellen, stellt sich die Frage, wie man damit umgehen soll. Dazu sind 3 Lösungswege vorgeschlagen worden: verhindern, kontrollieren oder ignorieren.

„Ehrlich antworten“ oder Forced-Choice-Format verwenden

Verhindern Zur Verhinderung einer sozial erwünschten Selbstdarstellung stehen 2 praktikable Maßnahmen zur Verfügung. Erstens kann man in der Instruktion standardmäßig darauf hinweisen, dass eine ehrliche Beantwortung wichtig ist. Es gibt keine richtigen und falschen Antworten, sondern jeder soll sich so beschreiben, wie er wirklich ist. Zumindest in einer Beratungssituation kann man so eine ehrliche Beantwortung fördern. Zweitens kommt

ein Forced-Choice-Antwortformat infrage (vgl. ▶ Abschn. 2.4.2.6). Die Testpersonen wählen nicht zwischen „Ja“ und „Nein“ oder geben den Grad ihrer Zustimmung an, sondern entscheiden sich zwischen Antwortalternativen, die ähnlich sozial erwünscht oder unerwünscht sind. In einer experimentellen Untersuchung ließen Martin et al. (2002) einen berufsbezogenen Persönlichkeitsfragebogen in der üblichen Ratingskalen- oder in einer Forced-Choice-Version bearbeiten. Die Versuchspersonen sollten ehrlich antworten oder einen guten Eindruck machen, um zu einem Job zu passen. Bei Verwendung der Ratingskalenversion unterschieden sich die Ergebnisse in der Faking-good-Bedingung erwartungsgemäß von der ehrlichen Bearbeitung. Die Forced-Choice-Version erwies sich dagegen als resistent gegenüber faking good. Wie unten noch ausgeführt wird, sorgt eine ehrliche Selbstdarstellung in Fragebögen allerdings nicht unbedingt für eine bessere Kriteriumsvalidität. Dazu passen metaanalytische Befunde (Salgado und Táuriz 2014), denen zufolge sich im beruflichen Bereich die Kriteriumsvalidität von Persönlichkeitsfragebögen durch ein Forced-Choice-Format nicht verbessern lässt.

Wenig Erfolg verspricht eine Begrenzung der Antwortzeit. In einer Faking-Studie, bei der die zur Verfügung stehenden Antwortzeiten gekürzt wurden, ließ sich Verfälschung nicht verhindern (Holden et al. 2001). Komar et al. (2010) versuchten erneut, die Verfälschung durch Begrenzung der Bearbeitungszeit zu reduzieren. Zur Bearbeitung von 64 Items eines Big-Five-Fragebogens plus einer Skala zur Entdeckung von Impression Management standen den Testpersonen entweder nur 10 min zur Verfügung oder es gab keine Zeitbegrenzung. Es trat jedoch nur ein Haupteffekt auf: Alle Skalen wurden in der Faking-Bedingung verfälscht, und die Werte für Impression Management stiegen an – aber unabhängig davon, ob Zeitdruck ausgeübt wurde oder nicht. Schon die implizite Annahme, dass Verfälschung längeres Nachdenken erfordert als ehrliches Antworten, ist problematisch. In mehreren Untersuchungen wurden die Zeiten zur Beantwortung der Items unter Verfälschungs- und Standardbedingungen verglichen. Die Befunde sind widersprüchlich; sowohl kürzere als auch längere Antwortzeiten wurden beim Verfälschen beobachtet (s. Holden et al. 2001).

Begrenzung der Antwortzeit nicht effektiv

Kontrolle Wenn man eine Verfälschung nicht effektiv verhindern kann, so kann man zumindest versuchen, sie zu entdecken. Zu diesem Zweck stehen eine Reihe von *Kontrollskalen* zur Verfügung, die in unterschiedlichem Maße Selbstäuschung und Impression Management erfassen (Paulhus 1991). Gibt man solche Skalen mit der Anweisung vor, beim Ausfüllen einen guten Eindruck zu machen („faking good“), fallen die Testwerte deutlich höher aus als unter einer Standardbedingung. Pauls und Crost (2004) fanden für eine bekannte Impression-Management-Skala einen Anstieg der Testwerte, der 26 Standardwertpunkten entspricht. Dies ist ein deutlicher Validitätsbeleg für diese Skala.

Kontrollskalen sollen Lügen aufdecken

Eine sehr bekannte Kontrollskala ist die Marlowe-Crowne-Skala zur sozialen Erwünschtheit, die auch als deutsche Version verfügbar ist (Lück und Timaeus 1969). Die 23 Items (Beispiel: „Ich bin immer höflich, auch zu unangenehmen Leuten“) sind durch Ankreuzen mit „richtig“ oder „falsch“ zu beantworten. Einige Fragebögen (z. B. MMPI-2 und FPI-R; ▶ Abschn. 3.3.3) enthalten Kontrollskalen, die der Marlowe-Crowne-Skala ähnlich sind. Ein erhöhter Wert auf einer Kontrollskala kann als Warnhinweis verstanden werden; wer hier hohe Werte aufweist, hat möglicherweise den ganzen Fragebogen nicht ehrlich ausgefüllt. Dabei ist zu beachten, dass es auch andere

Marlowe-Crowne-Skala

Gründe für erhöhte Werte auf einer Erwünschtheitsskala geben kann. Menschen, die sich stark an moralischen Standards orientieren, verhalten sich vielleicht wirklich so, wie sie es im Fragebogen angeben: Sie nutzen tatsächlich die Gelegenheit nicht aus, umsonst mit der Straßenbahn zu fahren; sie halten sich streng an Verabredungen, fluchen nicht etc. Bei ihnen versagt das Messprinzip der Erwünschtheitsskalen, und sie werden zu Unrecht als Lügner oder Uneinsichtige verdächtigt. Deshalb sollten erhöhte Werte auf einer solchen Skala als Warnhinweis und nicht als Beweis verstanden werden.

Ignorieren Der Vorschlag, das Problem der Verfälschbarkeit von Persönlichkeitsfragebogen zu *ignorieren*, basiert auf empirischen Befunden zur Kriteriumsvalidität von Skalen zur sozialen Erwünschtheit, die in der Tat verblüffend sind.

Sozial erwünschtes Antworten auspartialisieren?

Ones et al. (1996) haben in einer Metaanalyse folgende Fakten zusammengetragen: Soziale Erwünschtheit korreliert mit emotionaler Stabilität und mit Gewissenhaftigkeit (minderungskorrigiert zu .37 und .20). Allerdings korreliert soziale Erwünschtheit auch mit Ausbildungserfolg ($r_{corr} = .22$). Das heißt, je sozial erwünschter sich jemand im Fragebogen darstellt, desto erfolgreicher wird er oder sie die Ausbildung abschließen. Deshalb ist es wenig Erfolg versprechend, die Ergebnisse in Persönlichkeitsfragebögen für soziale Erwünschtheit zu korrigieren. Dementsprechend zeigt sich, dass ein Auspartialisieren der sozialen Erwünschtheit aus den großen 5 Persönlichkeitsmerkmalen die Vorhersage von Berufserfolg (erfasst durch eine Vorgesetztenbeurteilung) nicht verbessert. Der beste Prädiktor ist die Gewissenhaftigkeit ($r_{corr} = .23$); nach Auspartialisierung der sozialen Erwünschtheit bleibt der Zusammenhang exakt gleich ($r_{corr} = .23$).

Wie lassen sich diese Befunde interpretieren? Eine Antwort lautet, dass Skalen zur sozialen Erwünschtheit Aspekte der Persönlichkeit miterfassen, die für den beruflichen Erfolg nützlich sind, nämlich emotionale Stabilität und Gewissenhaftigkeit. Eine andere Erklärung ist, dass Menschen, die sich sozial erwünscht darstellen, damit eine nützliche Fähigkeit zeigen. Sie erkennen, worauf es bei der Stelle ankommt und passen sich bei der Beantwortung des Fragebogens an die Anforderungen der Stelle an. Beim Kriterium, also der Beurteilung des Berufserfolgs durch Vorgesetzte, wirkt sich diese Fähigkeit ebenfalls günstig aus. Wer sich am Arbeitsplatz sozial erwünscht verhält, erfährt eher eine gute Beurteilung.

Jedenfalls scheint es unvorteilhaft zu sein, die soziale Erwünschtheit durch Auspartialisierung aus der Vorhersage herauszurechnen, um die „wahren“ Ausprägungen der Persönlichkeitsmerkmale als Prädiktor zu verwenden. Auf diese Weise entfernt man zugleich nützliche Varianzanteile. Dennoch bleibt ein tiefes Unbehagen, wenn Bewerberinnen und Bewerber einen Persönlichkeitsfragebogen bearbeiten und die Diagnostikerin oder der Diagnostiker im Einzelfall nicht weiß, ob beispielsweise der hohe Gewissenhaftigkeitswert Ausdruck einer hohen Gewissenhaftigkeit oder einer geschickten Selbstdarstellung ist. Hat hier vielleicht eine unzuverlässige, unordentliche Person erkannt, dass es auf Gewissenhaftigkeit ankommt, und sich entsprechend dargestellt? Für eine vertiefende Diskussion dieses Themas sei auf Marcus (2003) und Kanning (2003) verwiesen.

Soziale Erwünschtheit als positive Eigenschaft?

Es sei an dieser Stelle nochmals betont: Die Tatsache, dass Persönlichkeitsfragebögen verfälschbar sind, bedeutet nicht, dass sie auch tatsächlich verfälscht werden. Verfälschung ist stark situationsabhängig; wenn die Testpersonen keinen Nutzen darin sehen, einen Fragebogen zu verfälschen, werden sie in der Regel ehrlich antworten. Aus der Umfrageforschung liegen mehrere Untersuchungen vor, in denen die gleichen Informationen entweder durch ein Interview oder durch einen Fragebogen erhoben wurden. Bei Fragen nach dem Gebrauch illegaler Drogen fielen die Angaben durchschnittlich um 30 % höher aus als in Interviews (Tourangeau und Yan 2007). Auch andere Formen sozial unerwünschten Verhaltens (Abtreibung, Rauchen bei Teenagern) werden in Fragebögen eher eingeräumt als im Interview. Das Gleiche gilt für Symptome psychischer Störungen.

Für die Offenheit der befragten Personen spielt es jedoch keine Rolle, ob Fragen computergestützt oder in konventionellen Papier-und-Bleistift-Fragebögen dargeboten werden. Speziell bei Persönlichkeitsfragebögen fand sich in mehreren Studien praktisch kein Unterschied zwischen beiden Darbietungsformen ($d = -0.02$; Tourangeau und Yan 2007).

3.3.2.4 Gütekriterien

Persönlichkeitsfragebögen haben in der Regel eine hohe *Objektivität*, da Durchführung, Auswertung und Interpretation leicht zu standardisieren sind. Die *Reliabilität* eines Persönlichkeitsfragebogens kann auf unterschiedliche Weise geschätzt werden. Fast immer finden sich Angaben zur *internen Konsistenz* (meist Cronbachs Alpha). Oftmals ist es jedoch nicht angebracht, eine hohe interne Konsistenz als Qualitätsmerkmal zu bewerten: Testautorinnen und Testautoren können die interne Konsistenz maximieren, indem sie sukzessiv Items mit niedrigen Trennschärfen eliminieren. Die verbleibenden Items korrelieren hoch miteinander und bilden einen homogenen Fragebogen. Diese Strategie ist nur angemessen, wenn das zu messende Merkmal ebenfalls sehr homogen ist. Ansonsten werden wichtige Aspekte des Persönlichkeitsmerkmals ausgeblendet, was einem Verlust an Inhaltsvalidität gleichkommt. Wenn ein 2-Item-Fragebogen zu Neurotizismus (ein Merkmal mit vielen Facetten!) eine interne Konsistenz von .90 aufweist, so dürfte dieser hohe Wert mit einer starken Einengung auf einen Ausschnitt von Neurotizismus „bezahlt“ worden sein. Die *Retest-Reliabilität* hängt von der Stabilität des Merkmals und der Länge des Zeitintervalls zwischen den Testungen ab. Zu lange Intervalle sind nicht angemessen, weil dann die „natürliche“ Merkmalsfluktuation, die nicht dem Fragebogen anzulasten ist, die Reliabilität mindert (► Abschn. 3.3.1).

Zur Beurteilung der *Validität* von mehrdimensionalen Persönlichkeitsfragebögen ist oftmals die Analyse der Struktur relevant (► Abschn. 2.6.3.3). Soweit einzelne Skalen betrachtet werden, ist die Übereinstimmung mit Skalen gleichen Messanspruchs und zugleich die Abgrenzung von Skalen mit anderem, aber nicht völlig verschiedenem Messanspruch (!) aufschlussreich. Für die Konstruktvalidität von Persönlichkeitsfragebögen sind auch Korrelationen mit Fremdeinschätzungen von Bekannten, Verwandten oder Freunden relevant. Metaanalytische Befunde (► Tab. 2.20) belegen, dass die Korrelationen in der Größenordnung von .50 bis .60 liegen, wobei Unterschiede zwischen den Persönlichkeitsmerkmalen zu verzeichnen sind.

Verfälschbare Fragebögen werden nicht unbedingt verfälscht

Die Offenheit in Computer- bzw. Papier-und-Bleistift-Tests ist gleich

Interne Konsistenz soll nicht immer hoch sein

Konstrukt- und Kriteriumsvalidität

Bei Fragebögen, die für bestimmte Anwendungsfelder konstruiert wurden, ist die Prüfung der Kriteriumsvalidität unverzichtbar. Besonders im Bereich der Berufseignungsdiagnostik liegen viele einschlägige Studien vor, die meta-analytisch zusammengefasst wurden (► Abschn. 6.3.1). Sie liefern ein gutes Referenzsystem zur Beurteilung von berufsbezogenen Persönlichkeitsfragebögen.

3

Weiterführende Literatur

Eine ebenso umfassende wie kritische Erörterung der „Konstruktion und methodenbewussten Anwendung von Persönlichkeitsfragebögen“ geben Fahrenberg et al. (2010, S. 152 ff.) in Kap. 8 des Handbuchs *FPI-R: Freiburger Persönlichkeitseinventar*; der Beitrag geht u. a. auf Mess- und Skalierungsprobleme, den Stellenwert von Item- und Faktorenanalysen sowie die angemessene Interpretation erhaltener Daten ein und wäre damit geeignet, jedes einschlägige Lehrbuch zu bereichern. In grundlegender Art befasst sich das Buch von Mummendey und Grau (2014) mit dem Fragebogen als Messinstrument.

3.3.3 Persönlichkeitstestsysteme

Fragebögen zur Messung von Persönlichkeitsmerkmalen können sich auf ein einzelnes Merkmal beziehen oder gleichzeitig auf mehrere; letztere werden „mehrdimensionale Persönlichkeitsfragebögen“ oder „Persönlichkeitstestsysteme“ genannt. In der Praxis sind die verfügbaren Persönlichkeitsfragebögen unterschiedlich bedeutsam, wie Umfragen unter Praktikern und Praktikerinnen zeigen (► Tab. 3.13). Allerdings reichen diese Umfragen zeitlich relativ weit zurück und ignorieren zwangsläufig neue Trends. Im Bereich der Berufseignungsdiagnostik hat das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP; Hossiep und Paschen 2003) bzw. eine Kurzform dieses Fragebogens in Deutschland eine hohe Bekanntheit und Verbreitung gefunden (Hossiep et al. 2015). Das ist erfreulich, weil sich im Personalbereich immer noch etliche Verfahren großer Beliebtheit erfreuen, die die Anforderungen an psychometrische Kriterien nicht erfüllen.

► Tab. 3.13 Die in Deutschland am häufigsten verwendeten Persönlichkeitsfragebögen

Test	Rang	Verwendungshäufigkeit (Nennungen in Prozent)		
		Roth und Herzberg (2008)	Steck (1997)	Schorr (1995)
FPI-R ^a , FPI	1	25	44	34
BDI	2	21	11	–
SCL-90	3	19	–	–
Angstfragebogen für Schüler	4	8	8	5
MMPI ^a	5	7	18	7
PFK 9–14	6	6	6	5
Gießen-Test ^a	7	5	23	14

Prozentualer Anteil der Befragten, die angeben, den Test zu verwenden, Mehrfachnennungen möglich; bei Schorr: Nennen Sie 5 Tests, die Sie am häufigsten verwenden. $N=398$ (Roth und Herzberg 2008), 169 (Steck 1997) und 661 (Schorr 1995). Tests geordnet nach Nennungshäufigkeit in der neuesten Studie. FPI = Freiburger Persönlichkeitseinventar, BDI = Beck Depressions-Inventar, SCL-90 = Symptom-Checkliste, MMPI = Minnesota Multiphasic Personality Inventory, PFK 9–14 = Persönlichkeitfragebogen für Kinder zwischen 9 und 14 Jahren. Auf Quellenangaben wird verzichtet, da unterschiedliche Versionen und Auflagen der Tests in Gebrauch waren.

^aDiese Tests werden im Folgenden ausführlich behandelt.

Aus der großen Zahl verfügbarer Persönlichkeitstestsysteme werden hier exemplarisch nur bestimmte Inventare herausgegriffen. Bei dem MMPI-2 (amerikanisches Original: Butcher et al. 1989; deutsche Ausgabe: Hathaway et al. 2000) handelt es sich um das weltweit gebräuchlichste überhaupt. So erscheinen pro Jahr allein ca. 1000 Untersuchungen, die sich mit Einsatzmöglichkeiten und Erfahrungen beschäftigen, und zwar insbesondere an klinisch auffälligen Gruppen. Das MMPI-2 (sowie die Vorgängerversion MMPI) ist zudem das einzige praktisch bedeutsame Persönlichkeitsinventar, das nach externalen Prinzipien (► Abschn. 2.4.2.4) konzipiert wurde. In den USA wurde mit dem Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF; Ben-Porath und Tellegen 2008) eine deutlich revidierte Form vorgelegt, die unter gleichem Namen auch in einer deutschen Ausgabe erschienen ist (Engel 2019). Wir stellen beide Tests vor, weil das MMPI-2 vermutlich auch noch weiter Verwendung finden wird und sehr viele Publikationen dazu vorliegen. Vor allem lässt sich die neue Version leichter darstellen und auch begreifen, wenn man das MMPI-2 kennt. Ein weiteres Verfahren, das FPI-R (Fahrenberg et al. 2020), stellt das im deutschen Sprachraum am häufigsten verwendete dar. Das oben bereits erwähnte BIP bzw. dessen verkürzte Weiterentwicklung BIP-6 F wurde ausgewählt, weil es im Bereich der Personalpsychologie eine herausragende Stellung einnimmt. Vertiefend vorgestellt wird außerdem das NEO-PI-R (deutsche Version: Ostendorf und Angleitner 2004), weil es besonders in der Forschung sehr bekannt ist. Es basiert auf internationalen Forschungsaktivitäten und dem Fünf-Faktoren-Modell der Persönlichkeit, das gegenwärtig noch immer als „Goldstandard“ angesehen wird. Es wird daher auch gerne zur Validierung anderer Verfahren herangezogen.

Auswahlkriterien für die Persönlichkeitsfragebögen

3.3.3.1 Minnesota Multiphasic Personality Inventory-2 (MMPI-2)

Beim MMPI-2 von Hathaway et al. (2000) handelt es sich um einen Persönlichkeitsfragebogen für den Bereich der klinisch-psychologischen Diagnostik. Es dient hauptsächlich dazu, Probandinnen und Probanden hinsichtlich diverser psychiatrischer Kategorien zu beurteilen oder ggf. auch festzustellen, dass sie zur unauffälligen „Normalpopulation“ gehören.

Klinischer Persönlichkeitsfragebogen

Das MMPI bzw. das MMPI-2 ist ein sehr bekanntes und intensiv beforstetes Verfahren mit einer langen Tradition. Allein zum MMPI-2 lassen sich (Stand: Juni 2020) mit der Suchmaschine PsycINFO über 2700 Litteraturnachweise (mit „MMPI-2“ im Titel oder Abstract) finden. Das MMPI hebt sich von den meisten Persönlichkeitsfragebögen durch das angewendete Konstruktionsverfahren ab, das als exterale Konstruktion bezeichnet wird (► Abschn. 2.4.2.4). Das Vorgehen wird später noch beschrieben.

Weiterführende Internetressourcen

Die University of Minnesota Press Test Division hat Videoaufnahmen veröffentlicht, in denen die beiden Testautoren über die Entwicklung des MMPI-2-R unterhalten, siehe ► <https://conservancy.umn.edu/handle/11299/187605> (Ben-Porath und Tellegen 2016). Weitere Gesprächssequenzen befassen sich mit anderen Themen rund um das MMPI.

Steckbrief MMPI-2: Minnesota Multiphasic Personality Inventory-2 (Hathaway et al. 2000)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Breitbandverfahren zur Beschreibung wichtiger Persönlichkeitseigenschaften und psychischer Störungen
Anwendungsbereich	Klinische Psychologie und Psychiatrie
Theoretischer Hintergrund	Klassifikationssystem psychischer Störungen – bei der Testentwicklung nach dem Stand in den USA Ende der 1930er-Jahre
Testentwicklung	Externale Konstruktion der Skalen
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche Instruktion; Hinweise zur Gestaltung der Testsituation
Auswertung	Antwortbogen kann mit Schablonen oder MMPI-2-Faxdienst ausgewertet werden; Computertest mit automatischer Auswertung
Interpretation	Gültigkeit der Ergebnisse werden anhand von Validitätsskalen beurteilt; einzelne Skalen oder das gesamte Profil werden anhand von Angaben im Manual beurteilt
Reliabilität	
Konsistenz	Cronbachs α der 13 Basis-Skalen von .45 (Skala „Männlich / weibliche Interessen“) bis .88 (Skalen Schizophrenie und Psychopathie); im Durchschnitt nach Hank und Schwenkmezger (2003) bei .75 (inkl. der Zusatzskalen; deutsche Normierungsstichprobe)
Retest	$r_{tt} = .66$ bis .90 bei Männern ($N=49$) und .71 bis .92 bei Frauen ($N=56$); Retest-Intervall: 10 Tage
Validität	
Konstruktvalidität	Faktorenanalysen über die Skalenwerte
Kriteriumsvalidität	keine Angabe
Normen	
Zusammensetzung	$N=958$, Alter von 18 bis 70 Jahren; repräsentativ für Deutschland bezüglich Geschlecht, Alter und Region
Erhebungszeitraum	1996
Sonstiges	
Formen	Papier-und-Bleistift- und Computerversion
Testrezension	
Quelle	Hank und Schwenkmezger (2003); Replik: Engel (2003)

Vergleich klinisch auffälliger Gruppen mit Kontrollpersonen

Konstruktionsansatz Die Testentwicklung begann 1937, das MMPI wurde erstmals 1942 publiziert (Hathaway und McKinley 1942); ein Jahr später erschien eine überarbeitete Version. Am Anfang aller Entwicklungsarbeiten stand das Anlegen einer Liste von 1000 Items, die sich auf allgemeine Gesundheit, familiäre und eheliche Beziehungen, sexuelle und religiöse Einstellungen sowie emotionale Zustände bezogen und letztlich psychopathologische Symptome erfassen sollten. Später kamen Items zu Geschlechtsrollencharakteristika und abwehrender Selbstdarstellung hinzu. Gruppen von klinisch auffälligen Personen, bei denen damals von Psychiaterinnen und Psychiatern eine Schizophrenie, Hysterie, Hypochondrie usw. diagnostiziert worden war, bearbeiteten die Items ebenso wie „unauffällig-normale“ Kontrollpersonen (Einwohnerinnen und Einwohner von Minnesota, Bewerberinnen und Bewerber um einen Studienplatz, Besucherinnen und Besucher des Krankenhauses). Jene 566 Items, in denen sich die Patientinnen und Patienten von den Kontrollpersonen signifikant unterschieden, wurden schließlich in Skalen zusammengestellt.

Weil zahlreiche Fragen zugleich mehrere der Patientinnen- und Patientengruppen von Nichtpatientinnen und -patienten unterschieden, sind diese

dementsprechend auch Bestandteil mehrerer Skalen. Das heißt, die einmaliige Antwort zu einem Item wie „Ich schlafte unruhig und werde oft wach“ (Ja/Nein) wird mehrfach verrechnet (in diesem Fall unter den Skalen Hypochondrie, Depression und Hysterie; zu den Skalen s. u.). Ein solcher „Item-Overlap“ treibt die Interkorrelationen zwischen den Skalen in die Höhe.

Bei der Revision zum MMPI-2 erfolgte eine Überarbeitung der Items. Einige „alte“ Items (z. B. zu sexuellen Gewohnheiten, religiösen Einstellungen) waren unangemessen, andere nicht mehr zeitgemäß (manche Freizeitbeschäftigungen waren nicht mehr aktuell). Einige Items mussten sprachlich revidiert werden (z. B. waren Redewendungen nicht mehr gebräuchlich). Die daraus resultierende Forschungsversion enthielt zusätzlich 154 neue Items, die neue Inhaltsbereiche wie Ess- und Arbeitsstörungen und den familiären Bereich abdecken. Bei der Revision sollte sowohl die Kontinuität gewahrt bleiben als auch eine Modernisierung erreicht werden. Die 567 Items der revidierten Form setzen sich aus 459 „alten“ und 108 neuen Items zusammen. Bei der deutschen Fassung des MMPI-2 handelt es sich um eine Übersetzung des revidierten amerikanischen Originals; bei den „alten“ Items hielten sich die Autoren an die alte deutsche Testversion des MMPI-Saarbücken (Spreen et al. 1963).

Bei Revision Kompromiss zwischen Kontinuität und Neuerung

Struktur und Items Sowohl das MMPI wie auch die revidierte Form MMPI-2 umfasst 3 Arten von Skalen: 4 Validitätsskalen (Weiß nicht-, Lügen-, Seltenheits-, Korrekturskala), 13 Basisskalen sowie zahlreiche Zusatzskalen (z. B. soziale Verantwortlichkeit, Posttraumatische Belastungsstörung, Suchtgefährdung). Die Testpersonen geben durch Ankreuzen von „richtig“ oder „falsch“ an, ob eine Aussage auf sie zutrifft oder nicht.

Validitäts- und klinische Skalen

Übersicht

Skalen des MMPI-2 (Hathaway et al. 2000) mit Beispielitems (in Klammern wird angegeben, welche Antwort für das Merkmal spricht)

- ? – Weiß nicht-Skala (Anzahl nicht oder ungültig beantworteter Items)
- L – Lügenskala (15 Items), „Manchmal möchte ich am liebsten fluchen“ (richtig)
- F – Seltenheitsskala (60 Items), „Ich leide unter Anfällen von Übelkeit und Erbrechen“ (richtig)
- K – Korrekturskala (30 Items), „Zuweilen möchte ich am liebsten etwas kaputtschlagen“ (richtig)
- Hd – Hypochondrie (32 Items), „Ich leide unter Anfällen von Übelkeit und Erbrechen“ (richtig)
- D – Depression (57 Items), „Ich habe einen guten Appetit“ (falsch)
- Hy – Hysterie, Konversionsstörung (60 Items), „Ich habe häufig das Gefühl, als ob ich einen Kloß im Halse hätte“ (richtig)
- Pp – Psychopathie, Soziopathie, antisoziale Persönlichkeitsstörung (50 Items), „Manchmal habe ich mir sehr gewünscht, von zu Hause fortzugehen“ (richtig)
- Mf – Maskulinität/Femininität (56 Items), „Ich lese gern Liebesgeschichten“ (richtig = feminine Interessen)
- Pa – Paranoia (40 Items), „Niemand scheint mich zu verstehen“ (richtig)
- Pt – Psychasthenie (48 Items), „Ich habe sicherlich zu wenig Selbstvertrauen“ (richtig)
- Sc – Schizophrenie (78 Items), „Ich habe Angst, den Verstand zu verlieren“ (richtig)
- Ma – Hypomanie (46 Items), „Manchmal habe ich Lach- oder Weinanfälle, die ich nicht beherrschen kann“ (richtig)
- Si – Soziale Introversion, „Ich gehe gern zu Parties und anderen Gelegenheiten, bei denen es laut und lustig zugeht“ (falsch)

(© Hogrefe)

3

Items sind nicht immer inhaltlich nachvollziehbar

Viele Interpretationshinweise zu den Validitätsskalen

Separates Antwortblatt

Rohwerte werden direkt in Profilblatt eingetragen

Profilauswertung

Die hier ausgewählten Items sind größtenteils typisch für die zu messenden Merkmale. Bei einigen anderen Items ist dagegen kaum nachzuvollziehen, warum sie das Merkmal indizieren. Beispielsweise spricht die Verneinung des Items „Ich lese gern Zeitungsartikel über Gerichts- und Kriminalfälle“ für Hysterie, und wer bei „Mein Sexuelleben ist zufriedenstellend“ „falsch“ ankreuzt, bekommt dafür einen Punkt auf der Psychopathieskala.

Im Manual zum MMPI-2 (Hathaway et al. 2000) finden sich zu den Validitätsskalen zahlreiche Interpretationshinweise (► Tab. 3.14 zur L-Skala).

Durchführung Das MMPI-2 kann in Einzel- und Gruppensitzungen durchgeführt werden. Die Bearbeitung dauert etwa 1 h, bei Patientinnen und Patienten etwas länger. Testpersonen kreuzen auf einem separaten Antwortblatt für jedes Item „richtig“ oder „falsch“ an.

Auswertung Die Rohwerte werden skalenweise mit Schablonen ermittelt und direkt in ein Profilblatt für Frauen oder Männer eingetragen. Darin sind die Rohwerte bei jeder Skala grafisch so angeordnet, dass praktisch eine Transformation in T-Werte erfolgt. Normtabellen sind also überflüssig. Bei einigen Skalen sind zuvor die Rohwerte um eine bestimmte Punktzahl zu erhöhen. Bei dieser sog. „K-Korrektur“ wird der Rohwert einer Basisskala für mangelnde Offenheit der Testperson nach oben korrigiert. Dazu wird der Punktwert der K-Skala mit dem angegebenen Faktor (z. B. × 4) multipliziert. Der resultierende Korrekturwert wird zum Rohwert der Basisskala addiert. Die mühsame manuelle Auswertung mit Schablonen lässt sich durch Verwendung der Computerversion des MMPI-2 umgehen, bei der die Auswertung per Mausklick erfolgt.

Interpretation Zuerst wird anhand der Validitätsskalen geprüft, ob das Protokoll gültig ist. Die Basisskalen können einzeln interpretiert werden. Dazu stehen im Manual Interpretationshinweise nach dem in ► Tab. 3.14 gezeigten

► Tab. 3.14 L-(Lügen-)Skala: Interpretation der Skalenwerte

Niveau (T-Wert)	Gültigkeit des Profils	Mögliche Ursachen	Mögliche Interpretation
Sehr hoch (über 79)	Wahrscheinlich ungültig	Dissimulation ^a	Widerstand gegen den Test oder Naivität
Hoch (70–79)	Gültigkeit fraglich	Zufällige Beantwortung, Leugnen von Fehlern	Verwirrtheit, mangelnde Einsicht, Verdrängung
Erhöht (60–69)	Wahrscheinlich gültig	Abwehrende Untersuchungshaltung	Konventionell und konformistisch, absolut tugendhaft
Mittel (50–59)	Gültig	Typische, normale Einstellung gegenüber dem Test	Keine Probleme mit dem eigenen Selbstbild
Niedrig (unter 50)	Möglicherweise Simulation ^a	Zustimmungstendenz, Aufmerksamkeitserheischung	Überbetonung von Krankheitssymptomen, selbstbewusst und unabhängig, zynisch, sarkastisch

Quelle: Hathaway et al. (2000, S. 24–26, © Hogrefe)

^aSimulation bedeutet, dass eine Störung vorgetäuscht, Dissimulation, dass sie verborgen wird

Schema zur Verfügung. Bei Bedarf können bestimmte Zusatzskalen ausgewertet und interpretiert werden. Schließlich kann eine Profilauswertung vorgenommen werden. Dazu werden die Nummern der 3 Skalen mit den höchsten T-Werten notiert. In entsprechenden Handbüchern finden sich Erläuterungen und Fallbeispiele für die jeweiligen Punktodes (Skalenummern mit Symbolen für T-Wert-Bereich).

Reliabilität Die wegen der externalen Skalenkonstruktion besonders wichtige Retest-Reliabilität der Skalen wird im Manual mit .66 (Ma) bis .90 (D) bei Männern und mit .71 (Hy) bis .92 (Sc) bei Frauen angegeben (bei jeweils sehr kleinen Stichprobenumfängen).

Validität Im Testmanual werden lediglich Faktorenanalysen auf Basis der Skalenwerte berichtet; eine Faktorenanalyse über die einzelnen Items wird nicht berichtet. Es fanden sich ähnliche Strukturen für Männer und Frauen sowie gute Übereinstimmung mit US-amerikanischen Ergebnissen. Die 4 Faktoren sind:

- F1: Psychotische Gedankeninhalte (Sc, Pp, Pa, F)
- F2: Neurotische Verhaltensweisen (Hy, L, K)
- F3: Introversion (Si, D)
- F4: Geschlechtsrollenidentifikation (Mf)

4 Faktoren

Wegen der Itemüberlappung sind die Ergebnisse der Faktorenanalysen auf Basis der Skalenwerte mit Vorsicht zu bewerten.

Normierung Die Eichstichprobe ($N=958$) zur Normierung des deutschen MMPI-2 ist bezüglich Alter, Geschlecht und geografischer Herkunft repräsentativ für die 18- bis 70-jährige deutsche Bevölkerung. Bei der Erhebung entstand ein gewisser Schwund dadurch, dass immerhin 192 Personen die Bearbeitung des MMPI ablehnten und dass Protokolle wegen extrem hoher F-Werte oder zu vielen unbeantworteten Items eliminiert werden mussten.

Repräsentative Normstichprobe

Bewertung Der in *Report Psychologie* erschienenen Testrezension (Hank und Schwenkmezger 2003) und der Replik (Engel 2003) wurde eine auffällig groß gedruckte Überschrift vorangestellt: „Vertretbar nach 40 Jahren Anwendung? Meinungen über MMPI-2 gehen weit auseinander“ Das MMPI wurde schon früher kontrovers diskutiert. Die Bewertungen reichten von „psychometrischer Albtraum“ bis „nützliches Verfahren“. Wir bewerten die extrem große Informationsausbeute durch die vielen klinischen Skalen, Validitäts- und Zusatzskalen positiv. Das Verfahren hat ein sehr großes Entwicklungspotenzial, weil der Itempool zur Entwicklung neuer Skalen genutzt werden kann. Am aktuellen Manual zum MMPI-2 ist zu bemängeln, dass darin zu wenige Angaben zur Validität enthalten sind. Ein grundsätzliches Problem des MMPI ist, dass das Verfahren auf veralteten diagnostischen Kriterien basiert und eine klinische Diagnostik nach ICD-11 oder DSM-5 nicht unterstützt. Allerdings ist dieser Mangel behebbar. Zum Nachfolgeverfahren MMPI-2-RF existiert bereits eine umfangreiche Forschung, u. a. zur Messung von Persönlichkeitssstörungen.

Kein Bezug zu ICD-11 und DSM-5

3.3.3.2 Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF) in der deutschsprachigen Adaptation

3

Steckbrief MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2 Restructured Form (Deutschsprachige Adaptation; Engel 2019)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Breitbandverfahren zur Beschreibung wichtiger Persönlichkeitseigenschaften und psychischer Störungen
Anwendungsbereich	Klinische Psychologie und Psychiatrie
Theoretischer Hintergrund	Überarbeitung der Skalen des MMPI-2 mit dem Ziel, relativ homogene Skalen zu bilden und übergeordnete Komponenten der „alten“ Skalen in eigene Skalen auszulagern
Testentwicklung	Überwiegend faktorenanalytische Konstruktion der Skalen auf Basis der Skalen des MMPI-2
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche Instruktion; Hinweise zur Gestaltung der Testsituation
Auswertung	Computertest mit automatischer Auswertung oder Eingabe der Antworten auf dem Antwortbogen in ein Auswerteprogramm
Interpretation	Gültigkeit der Ergebnisse werden anhand von Validitätsskalen beurteilt; zu einzelnen Skalen detaillierte Angaben im „klinischen Report mit Testinterpretation“
Reliabilität	
Konsistenz	Cronbachs α der klinischen Skalen und der 3 Skalen höherer Ordnung im Durchschnitt (Median) .78 (deutsche Normierungsstichprobe)
Retest	$r_{tt} = .83$ (Median der Skalen wie oben; $N = 105$); Retest-Intervall: 10 Tage
Validität	
Konstruktvalidität	Sehr viele Belege; nicht auf wenige Zahlen reduzierbar
Kriteriumsvalidität	Mittelwerte (T-Werte) für zahlreiche Patientinnen- und Patientengruppen mit verschiedenen ICD-Diagnosen
Normen	
Zusammensetzung	$N = 916$ (gleich viele Frauen und Männer aus der Normierungsstichprobe des MMPI-2; für Details s. Beschreibung des MMPI-2; ▶ Abschn. 3.3.3.1)
Erhebungszeitraum	1996
Sonstiges	
Formen	Papier-und-Bleistift- und Computerversion
Testrezension	
Quelle	–

51 Skalen aus 338 MMPI-2-Items neu konstruiert

Konstruktionsansatz und Struktur des MMPI-2-RF Beim MMPI-2-RF handelt es sich um eine grundlegende Überarbeitung des MMPI-2. Übernommen wurden lediglich 338 der 567 Items des MMPI-2. Für das amerikanische MMPI-2-RF wurden alle 51 Skalen neu konstruiert. Für die deutsche Version wurde die Zuordnung der Items zu den Skalen (außer bei einer Validitätsskala) unverändert vom amerikanischen MMPI-2-RF übernommen. Der Zusatz im Testnamen „Restructured Form“ klingt unspektakulär. Tatsächlich verbirgt sich dahinter eine große Umwälzung. Der für das MMPI charakteristische externe

Ansatz der Skalenkonstruktion (► Abschn. 2.4.2.4) wurde endgültig verlassen. Die Gründe dafür werden im Folgenden genannt und das neue Vorgehen bei der Testkonstruktion beschrieben. Grundlage sind die Ausführungen im Testmanual (Engel 2019, insbesondere S. 105 ff.).

Die *klinischen Skalen* des „alten“ MMPI-2 (► Abschn. 3.3.3.1) weisen 2 Probleme auf. Erstens korrelieren sie höher miteinander, als es aufgrund der Ähnlichkeit der Merkmale zu erwarten wäre. Dafür lassen sich mindestens 2 Ursachen benennen:

- Die Skalen verwenden zum Teil die gleichen Items, d. h. ein Item, das beispielsweise zur Messung von Hysterie verwendet wird, taucht auch in der Skala Paranoia auf.
- Zudem zeigte die Forschung zum MMPI, dass die Skalen etwas messen, was vielen psychiatrischen Patienten gemeinsam ist und sie zugleich von gesunden Kontrollpersonen unterscheidet. Das ist nicht überraschend, weil bei einer extevnalen Skalenkonstruktion diejenigen Items identifiziert werden, durch die sich bestimmte klinische Gruppen von den Kontrollpersonen unterscheiden. Dabei spielt es keine Rolle, ob die Items spezifisch für die Störung sind oder ob sie bei allen Störungen relevant sind. Das wichtigste störungsunspezifische Merkmal war Entmutigung (engl. demoralization). Die Patientinnen und Patienten beschreiben sich als entmutigt, hilflos, verzweifelt, mit geringer Selbstachtung und von Versagensängsten geplagt.

Das 2. Problem war die große inhaltliche Heterogenität der meisten Skalen – ebenfalls eine Folge der externalen Skalenkonstruktion. Die Skalen sind nicht nur mit Aussagen angereichert, die auf Entmutigung hinweisen, sondern auch mit weiteren Komponenten. Dabei müssen die Items nicht falsch ausgewählt worden sein. Wenn ein Merkmal, z. B. Hysterie, schon bei der Testkonstruktion als komplex angesehen wurde, spiegelt sich das folglich auch in den Skalen wider.

Die genannten Probleme (zu hohe Skaleninterkorrelation und zu große Heterogenität der Skalen) versuchte man bei der „Rekonstruktion“ konsequent zu lösen. Jede Skala sollte nur eine Dimension messen. Die klinischen Skalen des neuen MMPI-2-RF (für eine Übersicht bzw. Auszüge der Skalen s. □ Tab. 3.15) wurden faktorenanalytisch in 3 Schritten gebildet:

1. Eine Skala „Entmutigung“ wurde „aus einem Itemset konstruiert, das mit den klinischen Skalen 2 (D) und 7 (Pt) die entsprechende Varianz teilte“ (Engel 2019, S. 108; die Beschreibung ist leider nicht präziser):
 - Mit sämtlichen Items jeder einzelnen klinischen Skala wurde eine Hauptkomponentenanalyse durchgeführt. Der 1. rotierte Faktor enthielt die Entmutigungsitems. Die verbleibenden Items bildeten im besten Fall nur einen weiteren Faktor, der inhaltlich sinnvoll interpretierbar war. Es fanden sich aber auch bis zu 3 weitere Faktoren (Skala „Maskulinität/Femininität“).
 - Für jeden der 12 so gefundenen Faktoren wurde eine sog. „Keimskala“ (eine Auswahl passender Items) gebildet, wobei bei der Auswahl der zugehörigen Items auf die Differenzierbarkeit der Skalen Wert gelegt wurde.
2. Für 9 Keimskalen („Entmutigung“ und 8 klassische klinischen Skalen, ohne „Maskulinität/Femininität“ und „soziale Introversion“) erfolgte eine Anreicherung mit weiteren Items aus dem MMPI-2. Auswahlkriterien waren eine ausreichend hohe Korrelation eines Items mit einer dieser 9 Keimskalen und niedrige Korrelationen mit den übrigen 11 Keimskalen.
3. Validierungsstudien gingen der endgültigen Aufnahme der klinischen Skalen in das MMPI-2-RF voraus. Details dazu finden sich im deutschen Manual nicht.

Klinische Skalen des MMPI-2 waren zu heterogen und korrelierten zu hoch

Konstruktionsschritte bei den neuen klinischen Skalen

Tab. 3.15 Skalen des MMPI-2-RF mit Angaben zur Itemzahl und Reliabilität

Kürzel	Name (ggf. Erläuterung)	Itemzahl	α	r_{tt}
<i>9 Validitätsskalen</i>				
VRIN-d	Variable Antwortinkonsistenz (zufälliges Antworten)	61 Paare	.70	.57
TRIN-r	Einseitige Antwortinkonsistenz	26 Paare	.47	.49
F-r	Seltenheitsskala (seltene Antworten)	32	.78	.88
Fp-r	Psychopathologische Seltenheitsskala (bei psychisch Kranken seltene Antworten)	21	.58	.78
Fa	Somatische Seltenheitsskala (körperliche Beschwerden, für körperlich Kranke jedoch selten)	16	.60	.85
FBS-r	Beschwerdenvalidität (somatische und kognitive Beschwerden, die auf Übertreibung hinweisen)	30	.71	.87
RBS	Antworttendenzskala (übersteigerte Gedächtnisbeschwerden)	28	.59	.85
L-r	Ungewöhnliche Tugenden	14	.51	.76
K-r	Ausgeglichenheitsvalidität (Angabe hoher emotionaler Ausgeglichenheit, mit Untertreibung assoziiert)	14	.59	.79
<i>3 Skalen höherer Ordnung (H0)</i>				
HOI	Internalisierungsstörungen (mit Stimmung und Affekt verknüpfte Probleme)	41	.87	.89
HOD	Denkstörungen	26	.75	.86
HOE	Externalisierungsstörungen (mit unkontrolliertem Verhalten verknüpfte Probleme)	23	.76	.83
<i>9 rekonstruierte klinische Skalen (RC)</i>				
RCd	Entmutigung	24	.88	.84
RC1	Körperbeschwerden (diffuse körperliche Beschwerden)	27	.82	.80
RC2	Mangel an positiven Emotionen	17	.74	.83
RC3	Zynismus	15	.77	.83
RC4	Antisoziales Verhalten	22	.79	.85
RC6 ^a	Verfolgungsgedanken	17	.66	.81
RC7	Dysfunktionale negative Emotionen (fehlangepasste Angst, Ärger, Sorgen)	24	.81	.85
RC8	Abweichende Erfahrungen (unübliche Wahrnehmungen oder Gedanken)	18	.72	.76
RC9	Hypomane Aktivierung (Hyperaktivität, Aggression, Impulsivität, Größenwahn)	28	.78	.80
<i>Spezialproblemskalen (SP)^b</i>				
<i>5 Somatisch-kognitive Skalen</i>				
UWS	Unwohlsein (allgemeines Gefühl einer angeschlagenen Gesundheit)	8	.66	.86
<i>9 Internalisierungsskalen</i>				
SUI	Suizidgedanken	5	.54	.77
<i>4 Externalisierungsskalen</i>				
JVP	Jugendliche Verhaltensprobleme	6	.65	.84
<i>5 Interpersonale Skalen</i>				
FML	Familiäre Probleme	10	.68	.75
<i>2 Interessenskalen</i>				
AES	Ästhetisch-literarisches Interesse	7	.61	.84
<i>5 Personality-Psychopathology-Five-Skalen (PSY-5)</i>				
AGG-r	Aggressivität	18	.76	.83
PSYCH-r	Psychotizismus	26	.74	.83
DISC-r	Unbeherrschtheit	20	.69	.78
NEGE-r	Negative Emotionalität	20	.75	.83
INTR-r	Introversion	20	.79	.86

Quelle: Auszüge aus Tab. 2.1 (S. 16 ff.) und Tab. 10.1 (S. 120 ff.) des Manuals (Engel 2019, © Hogrefe). Cronbachs α und Retest-Reliabilität r_{tt} aus der deutschen Normierungsstichprobe (Retest-Intervall: 10 Tage, $n=105$)

^aEine Skala RC5 gibt es im MMPI-2-RF nicht

^bDiese sind thematisch gruppiert; hier nur je ein Beispiel aufgeführt

In □ Tab. 3.15 sind die 9 rekonstruierten klinischen Skalen mit weiteren Angaben aufgeführt. Dort sind auch weitere Skalen aufgeführt, auf die nun kurz eingegangen wird.

Die *Validitätsskalen* dienen dazu, die Gültigkeit des Protokolls, d. h. die Gültigkeit der Angaben der Testperson insgesamt, auf unterschiedliche Weise zu überprüfen. Sie erfassen inhaltlich inkonsistente Antworten, Übertreibung sowie Untertreibung psychischer Störungen. Exemplarisch wird hier nur die Skala „variable Antwortinkonsistenz“ kurz erläutert. Die Skala besteht aus Paaren inhaltlich ähnlicher Items, die positiv korrelieren. Der Rohwert besteht aus der Anzahl von Itempaaren, die die Testperson übereinstimmend (konsistent) beantwortet hat. Bei zufälligem Ankreuzen ist dieser Wert niedrig.

Die *klinischen Skalen* sollten sowohl durch übergeordnete als auch durch spezifische ergänzt werden. Die übergeordneten Skalen konnten durch Faktorenanalysen der 9 klinischen Skalen, die 3 klinische Stichproben bearbeitet hatten, identifiziert werden. Es handelt sich um die 3 *Skalen höherer Ordnung* in □ Tab. 3.15. Die Items dieser Skala wurden faktorenanalytisch zunächst nur aus den Items ausgewählt, die hauptsächlich auf einer der übergeordneten Skalen laden. In einem 2. Schritt wurden alle 567 Items des MMPI-2 daraufhin getestet, ob sie sich diesen Skalen zuordnen ließen. Dazu mussten sie (ebenfalls in den 3 klinischen Stichproben) mit den Faktorwerten der 3 übergeordneten Skalen hinreichend hoch mit einem Faktor – und auch nur mit diesem – korrelieren.

Mit den *Spezialproblemskalen* sollten Aspekte der klinischen Skalen erfassen, die dort nicht hinlänglich abgebildet wurden. Beispielsweise sind in der Skala „Entmutigung“ keine Items zu Suizidgedanken enthalten. Für die Auswahl der Spezialproblemskalen waren auch Expertinnen- und Expertenratschläge und -urteile maßgeblich. Zur Identifikation passender Items wurde methodisch ähnlich vorgegangen wie bei der Konstruktion der klinischen Skalen. *Personality-Psychopathology-Five-Skalen* waren bereits in der amerikanischen Version des MMPI-2 enthalten. Sie sind an das Fünf-Faktoren-Modell der „normalen“ Persönlichkeit angelehnt. Wie der Zusatz „-r“ verrät, handelt es sich aber nun um revidierte Skalen. Von den 139 einschlägigen Items aus dem MMPI-2 wurden 65 eliminiert. Hinzu kamen 30 andere Items aus dem MMPI-2. Die Itemauswahl erfolgte in einem mehrstufigen Prozess, in dem sowohl die Korrelation eines Items mit einer Skala als auch die mit externen Kriterien eine wesentliche Rolle spielten. Auch beim MMPI-2-RF werden Items zum Teil für mehrere Skalen verrechnet, und zwar bis zu $6 \times$. Im Durchschnitt wird ein Item $2,1 \times$ verwendet (die Itempaare der Skalen für variable und einseitige Antwortinkonsistenz nicht mitgezählt).

Validitätsskalen

Klinische Skalen

Skalen höherer Ordnung, Spezialproblem- und Personality- Psychopathology-Five-Skalen

Durchführung und Auswertung Das MMPI-2-RF liegt in Papierform mit Antwortbogen und als Computerversion vor. Die Durchführung unterscheidet sich nicht von der des MMPI-2. Allerdings ist eine Auswertung mit Schablonen oder über das Fax-System nicht mehr vorgesehen. Bei Verwendung der Papierversion muss der Antwortbogen manuell in ein Auswertungsprogramm eingegeben werden. Die Durchführungszeit wird für die Computerversion mit 25–35 min und für die Papierversion mit 35–50 min angegeben.

Interpretation hoch standardisiert

Interpretation Das Computertestsystem bzw. das Auswertungsprogramm erzeugt einen umfangreichen Bericht, der Aussagen zur Gültigkeit der Testergebnisse anhand der Validitätsskalen, eine Profildarstellung aller Ergebnisse (in Normwerten) sowie ausformulierte Interpretationen für alle Ergebnisse sowie „diagnostische Überlegungen“ (welche Störungen sollten abgeklärt werden?). Im Testmanual finden sich zudem ausführliche Informationen zur inhaltlichen Bedeutung und Interpretation aller Skalen.

Überwiegend hohe Reliabilität der Skalen

Reliabilität In □ Tab. 3.15 finden sich Angaben zur internen Konsistenz (Cronbachs α) und zur Retest-Reliabilität nach 10 Tagen. Diese Angaben basieren auf den Daten der deutschen Normierungsstichprobe. Für die klinischen, die übergeordneten und die Personality-Psychopathology-Five-Skalen liegen die Werte fast ausnahmslos in einem für Persönlichkeitsfragebögen angemessenen Bereich. Cronbachs α reicht von (nur) .66 („Verfolgungsgedanken“ – r_{tt} aber .81) bis .88 („Entmutigung“). Die Restestwerte liegen mit Ausnahme der Skalen „abweichende Erfahrungen“ (.76) und „Unbeherrschtheit“ (.78) bei mindestens .80. Für die Spezialproblemskalen fallen die Reliabilitätskoeffizienten vergleichsweise niedrig aus, was vermutlich auf die meist kleine Itemzahl zurückzuführen ist.

Skala zu inkonsistentem Antworten spricht auf Zufallsantworten an

Validität Anders als beim deutschen MMPI-2 liegen beim deutschen MMPI-2-RF sehr viele Informationen zur Validität der Skalen vor. Im Manual werden auf 22 Seiten umfangreiche Validitätsbelege vorgelegt; auch Teile des Anhangs sind für die Beurteilung der Validität relevant. Insgesamt dürften mehrere Tausend Zahlenwerte vorliegen. Es ist uns aus Platzgründen nicht möglich, die Ergebnisse hier umfassend darzustellen und zu würdigen. Stattdessen gehen wir auf ausgewählte Skalen ein.

Zur Würdigung der *Gültigkeit des Protokolls* sind besonders die 2 Skalen zu inkonsistentem Antworten relevant. Protokolle von Personen, die die Items nicht (richtig) gelesen haben, sie nicht (richtig) verstanden haben oder nicht adäquat darauf geantwortet haben (z. B. teilweise zufällig geantwortet haben), sollten möglichst erkannt und nicht inhaltlich interpretiert werden. Im Abschnitt „Testinterpretation“ ist zu lesen, dass bei der Skala „variable Antwortinkonsistenz“ ein T-Wert von 80 oder höher bedeutet: „Das Testprotokoll ist nicht interpretierbar“ (Engel 2019, S. 42).

Als Validitätsbelege wenig aussagekräftig sind Korrelationen zwischen den Validitätsskalen des MMPI-2-RF mit vergleichbaren Skalen des MMPI-2 (Tab. 11.2 im Manual). Hohe Korrelationen können durch Itemüberlappung zustande gekommen sein. Ein starker Beleg ist dagegen der Effekt von (per Computerprogramm) eingestreuten Zufallsantworten in vorliegende Antwortbögen. Und hier sind die Ergebnisse beeindruckend: Wurden keine Zufallsantworten eingestreut, betrug der T-Wert für „variable Antwortinkonsistenz“ durchschnittlich 49,4, und nur 1,5 % der Protokolle galten als nicht interpretierbar ($T \geq 80$). Wurden 50 % der Antworten durch Zufallsantworten ersetzt, stieg der T-Wert auf 76,1. Das entspricht einer Effektstärke (Glass' Delta) von 2,90 (berechnet mit den Angaben für VRIN-d2 in Tab. 11.1 des Manuals). Immerhin 35,2 % dieser Protokolle galten damit als nicht interpretierbar ($T \geq 80$). Wir bewerten dies als einen sehr starken Validitätsbeleg für diese Skala.

Skala „Mangel an positiven Emotionen“ sollte bei depressiven Störungen erhöht sein

Ein hoher Wert auf der klinischen Skala „Mangel an positiven Emotionen“ gilt als Hinweis auf das Vorliegen einer depressiven Erkrankung, bei $T \geq 80$ sogar eventuell einer schweren Depression (Tab. 6.17 auf S. 56 im Manual). Patientinnen und Patienten mit einer diagnostizierten Depression sollten also auf dieser Skala erhöhte Werte aufweisen. Engel (2019, S. 135 f.) konnte 2676 Protokolle stationärer Patientinnen und Patienten auswerten. Viele davon hatten eine ICD-10-Diagnose für eine Form der Depression. Wir betrachten hier die Testwerte der Gruppe „mittelgradige depressive Episode“ F32.1 ($n=119$) und „schwere depressive Episode ohne psychotische Symptome“ F32.2. ($n=106$). Beide Gruppen sollten deutlich erhöhte Werte auf der Skala „Mangel an positiven Emotionen“ aufweisen, die F32.2- dabei höhere als F32.1-Gruppe. Die gleichen Hypothesen können auch für rezidivierende depressive Störung, mittelgradig (F33.1, $n=116$) vs. schwer (ohne psychotische Symptome, F33.2, $n=106$) aufgestellt werden. Anhang G des Manuals (Engel 2019, S. 238) zufolge liegen deren T-Werte im Durchschnitt bei 62 vs. 64 bzw. 67 vs. 69. Die Unterschiede zwischen einer mittelgradigen und schweren

depressiven Episode bzw. einer mittelgradigen und schweren rezidivierenden Störung sind nur marginal. Der Durchschnitt für alle Patientinnen und Patienten ($N=2676$) wird mit 60 angegeben. Hohe Skalenwerte finden sich auch bei Zwangsstörungen ($T=70$, $n=23$) und Borderline-Störungen ($T=67$, $n=41$); Mittelwerte von $T \geq 60$ finden sich bei weiteren 5 Störungen außerhalb des Depressionsbereichs. Erhöhte Werte sind also nicht sehr spezifisch für depressive Störungen.

Für die Personality-Psychopathology-Five-Skalen (die mit Bezug zu Persönlichkeitsstörungen entwickelt wurden) sind die Korrelationen mit den Big-Five-Skalen des NEO-PI-R (► Abschn. 3.3.3.5) aufschlussreich. Wir erwarten hohe Korrelationen zwischen den korrespondierenden Skalen (in □ Tab. 3.16 fett hervorgehoben) und vergleichsweise niedrige mit nicht korrespondierenden Skalen.

Die in □ Tab. 3.16 aufgeführten Ergebnisse können für die Skalen „negative Emotionalität“ und „Introversion“ als sehr starke Validitätsbelege gewertet werden. Für die Skalen „Aggressivität“ und „Unbeherrschtheit“ finden sich nur moderate Korrelationen mit Verträglichkeit, die aber konzeptuell nur entfernt ähnlich ist. Die Korrelationen mit nicht korrespondierenden Skalen fallen erwartungsgemäß relativ niedrig aus.

Korrelation der Personality-Psychopathology-Five-Skalen mit Big-Five-Skalen

Nur 1 Normgruppe trotz Alters- und Geschlechtsunterschieden

Normierung Für die Normierung wurden die Daten auf Itemebene aus der deutschen Normierungsstichprobe des MMPI-2 verwendet. Beide Geschlechter sind gleich stark vertreten ($N=916$; durch Entfernen nach Zufall ausgewählter Personen erreicht). Es gibt nur eine Normgruppe, obwohl bei einigen Skalen Geschlechts- und/oder Altersunterschiede bestehen, die in Tab. 7.7 auf S. 90 f. des Manuals dokumentiert sind (Engel 2019). Der Autor hat sich hier an dem amerikanischen Original orientiert, in dem aus rechtlichen Gründen (Verbot einer Berücksichtigung u. a. des Geschlechts bei der Personalauswahl) auf geschlechtsspezifische Normen verzichtet wurde. Die Rohwerte wurden in T-Werte (► Abschn. 2.6.4) transformiert. Weil die Rohwerte der meisten Skalen mehr oder weniger linksgipflig verteilt sind (niedrige Werte kommen häufiger vor als hohe), ist der Autor dem amerikanischen Original gefolgt und hat in diesen Fällen sog. „uniforme T-Werte“ berechnet. Diese differenzieren im klinisch relevanten oberen Bereich stärker als normale T-Werte.

Bewertung Mit dem MMPI-2-RF steht ein zeitökonomisches Verfahren mit einer sehr großen Informationsausbeute (51 Skalen) zur Verfügung, das überwiegend im klinischen Bereich Anwendung finden wird. Mithilfe der Validitätsskalen können wohl die meisten nicht korrekt bearbeiteten Protokolle

□ **Tab. 3.16** Korrelationen zwischen den Skalen der Personality Psychopathology Five und dem NEO-PI-R

NEO-PI-R	Personality Psychopathology Five				
	AGG-r	PSYCH-r	DISC-r	NEGE-r	INTR-r
Neurotizismus	-.18	.35	.09	.67	.04
Extraversion	.37	-.08	.20	-.21	-.61
Offenheit	.08	.00	.17	-.08	-.31
Verträglichkeit	-.38	-.13	-.34	-.15	.02
Gewissenhaftigkeit	.22	-.13	-.20	-.20	.00

Quelle: Auszug aus Tab. F-4 (Engel 2019, S. 236, © Hogrefe). $N=2620$ bis 2642
AGG-r = Aggressivität, PSYCH-r = Psychotizismus, DISC-r = Unbeherrschtheit, NEGE-r = negative Emotionalität, INTR-r = Introversion

erkannt werden. Die Testentwicklung ist gut dokumentiert und nachvollziehbar. Die Durchführung am Computer (bedingt auch die Bearbeitung eines Fragebogens) und die computerbasierte Auswertung gewährleisten eine sehr hohe Durchführungs- und Auswertungsobjektivität. Durch den umfangreichen schriftlichen Report mit standardisierten konkreten Aussagen zur Bedeutung der Testergebnisse ist auch die Interpretationsobjektivität gewährleistet. Die meisten Skalen sind hinreichend hoch reliabel. Zur Validität liegen außergewöhnlich viele Daten vor, die vom Testautor eher neutral beschreibend, aber angemessen kommentiert werden. Wir hätten uns hier ein stärker hypothesesgeleitetes Vorgehen gewünscht (was wird warum erwartet und wie passen die Ergebnisse zu den Erwartungen?). Die Normen sind schon älter. Da es nur eine Normgruppe gibt, werden die Ergebnisse jeder Person mit denen einer repräsentativen Stichprobe von über 900 anderen Menschen verglichen.

Bessere Bewertung als MMPI-2

Einsatz des IKP im klinischen Bereich

Itembeispiel aus dem IKP

Lange Tradition

Fazit Unser Fazit ist, dass es sich um ein ungewöhnlich aufwendig und sorgfältig konstruiertes Verfahren handelt, das bei relativ geringem Zeitaufwand sehr viele klinisch relevante Informationen über eine Testperson liefert.

Alternativen

Inventar Klinischer Persönlichkeitsakzentuierungen (IKP) Wenn man die Kombination von Messgegenstand und Konstruktionsprinzip des MMPI-2 bzw. dessen Nachfolger MMPI-2-RF als Maßstab nimmt, ist kein Konkurrenzverfahren in Sicht. Allerdings gibt es für den klinischen Bereich mit dem IKP (Andresen 2006) ein mehrdimensionales Persönlichkeitsinventar, das in etwa einen ähnlichen Messanspruch wie das MMPI hat. Persönlichkeitsakzentuierungen sind hohe Ausprägungen auf einer Persönlichkeitsdimension, die jedoch nicht zwangsläufig eine (klinisch relevante) Persönlichkeitsstörung anzeigen. Dem IKP liegt eine dimensionale Betrachtung von Merkmalen zugrunde, d. h., ein in Bezug zu einer Persönlichkeitsstörung stehendes Persönlichkeitsmerkmal kann mehr oder weniger ausgeprägt sein. Die erfassten Merkmale reichen also weit in den Normalbereich hinein. Damit unterscheidet sich das IKP von der in der Klinischen Psychologie und Psychiatrie häufig vorkommenden kategorialen Diagnostik (bei der eine bestimmte Störung als „vorliegend“ oder „nicht vorliegend“ bezeichnet wird).

An einem Itembeispiel lässt sich die Messintention und das Konstruktionsprinzip des IKP gut erläutern: „Ich bin wahnsinnig fantasievoll und romantisch, vor allem in Liebesangelegenheiten“. Das Item ist, wie auch alle anderen Items, auf einer 4-stufigen Skala von „völlig unzutreffend“ bis „völlig zutreffend“ zu beantworten. Auch gesunde Personen können hier eine starke Zustimmung zeigen. Auch diese würden im IKP dafür Punkte auf der Skala „histrionische Persönlichkeitsakzentuierung“ erhalten. Da jede Skala 10 Items umfasst, ist es sehr gut möglich, dass man am Ende auf dieser Skala im Normalbereich liegt, selbst wenn man bei einigen Items zustimmend antwortet.

Das IKP ist mit seinen 11 Skalen an den 11 offiziellen Diagnoseeinheiten nach dem „alten“ DSM-IV und der ICD-10 ausgerichtet. Ein Ergänzungsmodul enthält weitere Skalen.

3.3.3.3 Freiburger Persönlichkeitsinventar – revidierte Fassung (FPI-R)

Das FPI-R von Fahrenberg et al. (2020) ist ein in der Praxis viel verwendeter Persönlichkeitsfragebogen mit langer Tradition; die 1. Auflage des FPI erschien bereits 1970. Es handelt sich um eine eigenständige Entwicklung, die sich nicht dem Big-Five-Ansatz verpflichtet sieht. Die Autoren kritisieren im Vorwort zur 8. Auflage, dass die Frage, welches die wichtigsten Grundfaktoren der Persönlichkeit sind, in der Fachliteratur zu viel Raum eingenommen hätte; es sei wichtiger, die Auswahl geeigneter Eigenschaftskonzepte für bestimmte diagnostische Fragestellungen zu begründen.

Steckbrief FPI-R: Freiburger Persönlichkeitsinventar (9., vollständig überarbeitete Auflage mit neuer Normierung und Validitätshinweisen, Prinzipien der Testkonstruktion und modernen Assessmenttheorie; Fahrenberg et al. 2020)	
Zielsetzung und Testkonstruktion	
Messgegenstand	10 spezifische und 2 übergeordnete Persönlichkeitsmerkmale (Emotionalität und Extraversion)
Anwendungsbereich	Breiter Anwendungsbereich insbesondere im (sub-)klinischen Bereich
Theoretischer Hintergrund	Pragmatische Auswahl der Skalen; für Emotionalität und Extraversion war die damals einflussreiche Persönlichkeitstheorie von Eysenck maßgeblich
Testentwicklung	Weiterentwicklung des Vorgängers FPI; Itemauswahl nach faktorenanalytischen Ergebnissen und Trennschärfen; 2020 sprachliche Anpassung von einigen Items der Auflage von 2001 und zugleich 2010 sowie Überprüfung der Skalensstruktur
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche Instruktion; Hinweise für Umgang mit Fragen
Auswertung	Auszählen der Antworten mittels Schablone im Testheft; Einführung eines separaten Antwortbogens mit eigener Schablone angekündigt (Stand: Juni 2020)
Interpretation	Genaue Vorgaben zur Benennung der erfassten Merkmale und zur Verbalisierung ihrer Ausprägung sowie Fallbeispiele
Reliabilität	
Konsistenz	Cronbachs α der Skalen von .72 bis .82 (Normierungsstichprobe)
Retest	$r_{tt} = .63$ bis .85 ($N = 103$), Retest-Intervall: 4 Wochen
Validität	
Konstruktvalidität	Simultanfaktorisierung mit anderen Persönlichkeitsfragebögen; Korrelation mit Fremdbeurteilungen, Befinden im Alltag, biografischen Merkmalen
Kriteriumsvalidität	Angaben zu mehr oder weniger relevanten Kriterien im Manual enthalten
Normen	
Zusammensetzung	Normen ($N = 3450$), differenziert nach Alter und Geschlecht; für Deutschland repräsentative Stichprobe
Erhebungszeitraum	2018
Sonstiges	
Formen	Es gibt keine unterschiedlichen Testformen; Computerversion liegt vor
Testrezension	
Quelle	Rohrmann und Spinath (2011) zur 8. Auflage von 2010

Zum FPI-R und zu der Vorgängerversion FPI liegt für ein deutschsprachiges Verfahren sehr umfangreiche Forschung vor; einer Recherche der Testautoren (Fahrenberg et al. 2020) zufolge waren im Mai 2019 alleine in der Datenbank PSYNDEX-Tests zu „Freiburger Persönlichkeitsinventar“ 752 Publikationen verzeichnet. Für das FPI-R sind (Stand: Juni 2020) mit den Suchmaschinen PSYNDEX und MEDLINE über 550 Publikationen auffindbar.

Intensiv beforschtes Verfahren

Pragmatische Auswahl der Skalen

3

Konstruktionsprinzipien Die Entwicklung des Verfahrens orientierte sich nicht an einer spezifischen Persönlichkeitstheorie, sondern an den Interessen der Autoren an bestimmten Dimensionen des Verhaltens, und zwar teils im Hinblick auf die theoretischen Grundlagen (insbesondere Extraversion und Neurotizismus/Emotionalität), teils im Hinblick auf deren Implikationen für das soziale Zusammenleben (z. B. Aggressivität) und das subjektive Wohlbefinden oder Zurechtkommen mit Anforderungen (z. B. Lebenszufriedenheit, Beanspruchung). Die Skalenkonstruktion geschah zunächst deduktiv; die Autoren legten sich also zuerst auf bestimmte Konstrukte fest und entwickelten dann dazu passende Items. Die weitere Itemauswahl erfolgte teils nach faktorenanalytischen, teils nach Trennschärfeprinzipien; daneben spielten inhaltliche und praktische Erwägungen eine Rolle. Keinesfalls war eine Maximierung der internen Konsistenz das Ziel; die Autoren setzen sich sehr kritisch mit der Forderung nach einer hohen internen Konsistenz auseinander, die leicht zu Einbußen bei der Validität führen kann. Das FPI-R stellt eine Weiterentwicklung des Vorgängers FPI dar. Für die 9. Auflage (Fahrenberg et al. 2020) wurden lediglich einige Items sprachlich verbessert und die Zuordnung der Items zu den Skalen erfolgreich überprüft.

10 Standard- und 2 Zusatzskalen

Gliederung Der Test besteht aus 138 Aussagen in der Form „Ich (bin, fühle, würde usw.) ...“, die durch Ankreuzen von „stimmt“ oder „stimmt nicht“ zu beantworten sind. Das 1. Item „Ich habe die Anleitung gelesen und bin bereit, jeden Satz offen zu beantworten“ gehört zu keiner der Skalen. Alle weiteren Items verteilen sich auf 10 Standardskalen mit je 12 Items sowie 2 Zusatzskalen („Extraversion“ und „Emotionalität“, die dem damals einflussreichen Persönlichkeitsmodell von Eysenck entstammen) mit je 14 Items (von denen insgesamt 11 auch in den Standardskalen Verwendung finden). Die Skala „Offenheit“ dient auch als Kontrollskala. Bei Stanine-Werten von 1 bis 3 bestehen Zweifel, ob der Fragebogen offen und ehrlich bearbeitet wurde.

Skalen des FPI-R mit Itembeispielen

Zu jeder Skala sind die trennschärfsten Items (siehe Tab. 3.4 im Testmanual von Fahrenberg et al. 2010) aufgeführt:

1. **Lebenszufriedenheit:** „Alles in allem bin ich ausgesprochen zufrieden mit meinem bisherigen Leben.“
2. **Soziale Orientierung:** „Da der Staat schon für Sozialhilfe sorgt, brauche ich im Einzelnen nicht zu helfen.“ (Ablehnung spricht für soziale Orientierung)
3. **Leistungsorientierung:** „Ich habe gern mit Aufgaben zu tun, die schnelles Handeln verlangen.“
4. **Gehemmtheit:** „Ich werde ziemlich leicht verlegen.“
5. **Erregbarkeit:** „Oft rege ich mich zu rasch über jemanden auf.“
6. **Aggressivität:** „Wenn ich wirklich wütend werde, bin ich in der Lage, jemandem eine runterzuhaben.“
7. **Beanspruchung:** „Ich habe häufig das Gefühl, im Stress zu sein.“
8. **Körperliche Beschwerden:** „Mein Herz beginnt manchmal zu jagen oder unregelmäßig zu schlagen.“
9. **Gesundheitssorgen:** „Ich vermeide Zugluft, weil man sich zu leicht erkälten kann.“
10. **Offenheit:** „Ab und zu erzähle ich auch mal eine Lüge.“

E – **Extraversion:** „Ich kann in eine ziemlich langweilige Gesellschaft schnell Leben bringen.“

N – **Emotionalität:** „Ich bin oft nervös, weil zu viel auf mich einströmt.“

Durchführung Die Instruktion findet sich schriftlich und in leicht verständlicher Weise auf dem Fragebogen: Man soll nicht lange bei jedem Item nachdenken, sondern die Antwort geben, die einem unmittelbar in den Sinn kommt. In der Handanweisung werden zusätzliche Empfehlungen gegeben, wie den häufigsten Rückfragen und Einwänden vonseiten der Probandinnen bzw. Probanden zu begegnen ist. Die Dauer der Bearbeitung liegt zwischen 10 und 30 min.

Gut verständliche Instruktion

Auswertung und Interpretation Mithilfe einer Schablone, die auf die Antwortfelder der 4 Seiten des Fragebogens aufgelegt wird, werden skalenweise die Anzahl der Antworten im Sinne des Merkmals ausgezählt. Die Summenwerte werden auf dem Auswertebogen eingetragen und anschließend anhand von alters- und geschlechtsspezifischen Normen in Stanine-Werte transformiert. Die ermittelten Stanine-Werte ergeben ein Profil, nachdem sie auf dem Auswertebogen angekreuzt und miteinander verbunden wurden (Abb. 3.18).

Profilbogen

Reliabilität Die an der Normierungsstichprobe bestimmten Konsistenzkoeffizienten haben sich gegenüber der 8. Auflage nur geringfügig geändert ($\alpha = .72$ bis .82). Sie werden von den Autoren als „typisch für die Skalen von Persönlichkeitsfragebogen“ bezeichnet (Fahrenberg et al. 2020, S. 50). Retest-Reliabilitäts-Angaben liegen nur für eine Gruppe von Herz-Kreislauf-Patientinnen bzw. -Patienten ($N = 103$) vor. Die Testungen erfolgten zu Beginn und am Ende einer ca. 4-wöchigen Kur. Trotz der Homogenität dieser Stichprobe und des Treatments im Retest-Intervall fielen die Koeffizienten mit $r_{tt} = .63$ („soziale Orientierung“) bis .85 („Gehemmtheit“) zufriedenstellend aus. Der Mittelwert über alle 12 Skalen lag bei $r_{tt} = .73$.

Interne Konsistenz und Retest-Reliabilität angemessen hoch

Validität Unter „Validitätshinweise aus der Erhebung 2018“ finden sich im Manual zahlreiche Angaben. Sie werden ergänzt durch Befunde aus der Erhebung von 1999 (Fahrenberg et al. 2010). Beispielsweise fanden sich deutliche Zusammenhänge der Skalen Emotionalität, Lebenszufriedenheit und körperliche Beschwerden mit einem selbst als „gut“ bezeichneten Gesundheitszustand und „häufiger Tabletteneinnahme“. Die Interviewerinnen und Interviewer stuften die Testpersonen auf mehreren Skalen ein. Als „energisch vs. gar nicht energisch“ sowie als „selbstsicher, von sich überzeugt vs. ziemlich unsicher“ beurteilte Personen hatten höhere Werte auf den Skalen Leistungsorientierung und Extraversion sowie niedrigere Werte bei Gehemmtheit (Korrelationen um .30). Personen mit sichtbarer Tätowierung wiesen höhere Werte auf der Skala Aggressivität auf ($r = .23$).

Viele Validitätsbelege

Aus einer gemeinsamen Faktorenanalyse über mehrerer Testsysteme (Tab. 3.17) wird ersichtlich, dass nicht weniger als 4 FPI-Skalen (Lebenszufriedenheit, Erregbarkeit, Beanspruchung und körperliche Beschwerden) gemeinsam auf einem Neurotizismusfaktor laden, „Gehemmtheit“ (mit negativem Vorzeichen) relativ hoch auf einem Extraversionsfaktor lädt und „soziale Orientierung“ sowie „Aggressivität“ (negativ) mit Verträglichkeit in Beziehung stehen. Diese Resultate zeigen eine nicht vollständige Abdeckung des Fünf-Faktoren-Modells der Persönlichkeit durch das FPI-R; „Offenheit für Erfahrung“ wird mit dem FPI-R nicht erfasst.

Simultanfaktorisierung mit mehreren Persönlichkeitstests

In einer Studie von Schmidt und König (1986, s. Fahrenberg et al. 2020, S. 110 f.) wurden Personen, die das FPI-R ausgefüllt hatten, von sehr guten Bekannten direkt auf den Skalen des FPI-R eingestuft. Die mittlere Selbst-/Fremdeinschätzungskorrelation von $r_{tc} = .50$ kann als guter Validitätsbeleg angesehen werden; am geringsten war die Übereinstimmung bei sozialer Orientierung (.38) und Gehemmtheit (.36), am höchsten bei körperlichen Beschwerden (.61), Gesundheitssorgen (.60) und Lebenszufriedenheit (.58). Für die Kontrollskaala „Offenheit“ konnte auf diese Weise die Validität nicht belegt werden ($r_{tc} = .26$).

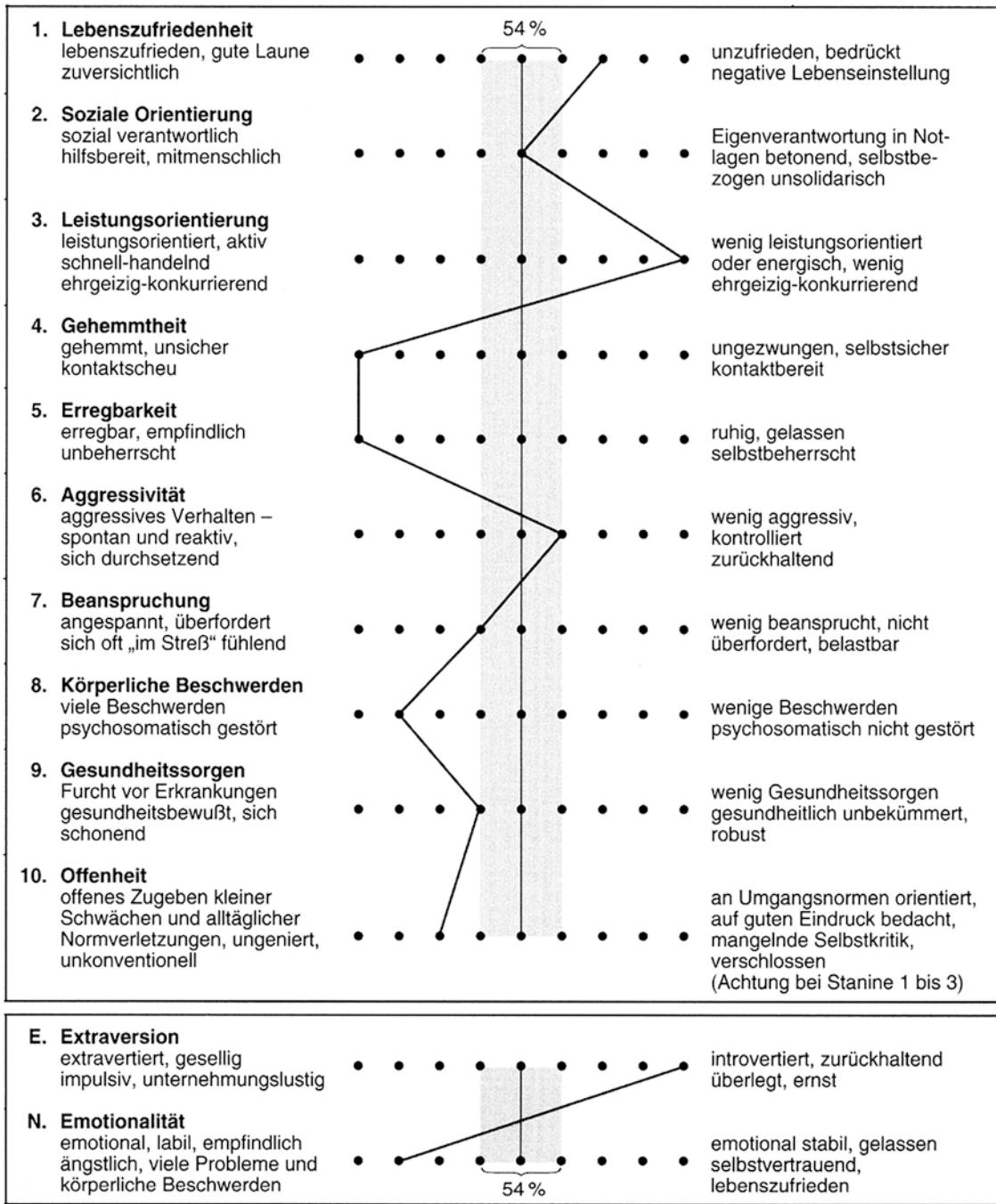
Korrelation mit Fremdbeurteilungen

Tab. 3.17 Rotierte Ladungsmatrix einer Simultanfaktorisierung von PRF, FPI, EPI und NEO-FFI

Skala	Faktoren ^a				
	I	II	III	IV	V
Personality Research Form (PRF)					
Leistungsstreben	.08	.06	.23	.12	.77
Geselligkeit	.01	.72	-.25	.33	.05
Aggressivität	.30	.40	.06	-.68	-.07
Dominanzstreben	-.19	.52	.16	-.44	.42
Ausdauer	-.24	-.13	.12	.08	.74
Bedürfnis nach Beachtung	.00	.76	.19	-.50	-.01
Risikomeidung	.25	-.46	-.44	.24	.10
Impulsivität	.26	.41	.30	-.11	-.57
Hilfsbereitschaft	.25	.29	.00	.65	.27
Ordnungsstreben	-.05	-.03	-.45	.10	.62
Spielerische Grundhaltung	-.01	.72	-.03	-.13	-.37
Soziales Anerkennungsbedürfnis	.38	.39	-.47	.10	.17
Anlehnungsbedürfnis	.56	.24	-.25	.28	-.33
Allgemeine Interessiertheit	.00	-.03	.75	.03	.19
Freiburger Persönlichkeitsinventar (FPI)					
Lebenszufriedenheit	-.61	.23	-.27	.13	.27
Soziale Orientierung	.28	.11	.33	.66	.13
Leistungsorientierung	-.15	.40	.06	-.18	.63
Gehemmtheit	.39	-.59	-.19	.11	-.24
Erregbarkeit	.24	.18	-.12	-.16	-.11
Aggressivität	.70	.31	.00	-.66	-.03
Beanspruchung	.69	.03	.04	.03	.18
Körperliche Beschwerden	.72	-.11	-.04	.04	-.03
Gesundheitssorgen	.25	-.17	-.38	-.09	.27
Eysenck-Persönlichkeits-Inventar (EPI)					
Extraversion	.02	.86	.03	-.18	-.11
Neurotizismus	.89	-.07	.04	.01	-.16
NEO-Fünf-Faktoren-Inventar (NEO-FFI)					
Neurotizismus	.79	-.10	.08	.07	-.27
Extraversion	-.05	.80	-.08	.12	.10
Offenheit für Erfahrung	.16	.02	.74	-.07	-.07
Verträglichkeit	.01	-.09	-.21	.75	-.06
Gewissenhaftigkeit	-.17	-.02	-.35	.09	.75

Quelle: Borkenau und Ostendorf (1993, S. 19, © Hogrefe)

^aLadungen ≥ 0.60 sind **fett** gedruckt



■ Abb. 3.18 Auswertungsbogen des FPI-R mit eingetragenem Profil. (Ausschnitt aus Fahrenberg et al. 2010, S. 99, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugsquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

In einer anderen Studie (Fahrenberg et al. 2020, S. 115 f.) stuften berufstätige Männer und Frauen 5× am Tag ihr momentanes Befinden ein. Dazu wurde ein kleiner tragbarer Computer verwendet. Zwischen einigen FPI-Skalen und den über 5 Messzeitpunkte aggregierten Befindensangaben konnten zum Teil relativ hohe Zusammenhänge festgestellt werden: Je höher die Lebenszufriedenheit der Probandinnen bzw. Probanden war, als desto ausgeglichener ($r=.51$) und weniger bedrückt (.49) beschrieben sie ihren aktuellen Zustand; zugleich stuften sie ihre Stimmung eher als gut ein (.51). Ähnliche

Korrelation mit Angaben zum Befinden im Alltag

Zusammenhänge fanden sich für Emotionalität. Die Gesundheitssorgen korrelierten zu .56 mit der Zustandsbeschreibung „aufgeregt, nervös“. Je größer die Beanspruchung laut FPI-R war, desto häufiger gaben die Teilnehmerinnen und Teilnehmer an, seit der letzten Eingabe „im Stress“ gewesen zu sein ($r = .52$). Es fanden sich weitere Validitätsbelege mit allerdings niedrigeren Korrelationen. Insgesamt sind das gute Belege dafür, dass die (meisten) FPI-Skalen das messen, was sie messen sollen.

Repräsentative Normstichprobe

Normierung Der 8. Auflage lagen noch die Normen der 7. Auflage zugrunde. Die Normierung für die 8. Auflage wurde im Herbst 1999 durch das Institut für Demoskopie in Allensbach an einer bevölkerungsrepräsentativen Stichprobe von 3740 Einwohnern aus allen Bundesländern durchgeführt. Dieses Institut war auch mit der Normierung für die 9. Auflage betraut. Von Juni bis November 2018 wurde eine für die deutsche Wohnbevölkerung im Alter ab 16 Jahren repräsentative Untersuchung von 3450 Personen durchgeführt. Anschließend erfolgte eine Gewichtung der Teilstichproben, um die Häufigkeit der Menschen im Osten und Westen der Bundesrepublik, der Geschlechter, der Altersgruppen, der Berufsgruppen, der Bildungsabschlüsse, des Familienstandes, der Konfessionen in der deutschen Bevölkerung nach Angaben des Statistischen Bundesamtes genau nachzubilden. Die Rohwerte wurden getrennt für die beiden Geschlechter und für jeweils 7 Altersgruppen in Stanine-Werte transformiert. Die einzelnen Gruppen sind mit 102 (Männer, 16–19 Jahre) bis 419 (Frauen, 70 Jahre und älter) Personen besetzt.

Bewährtes Verfahren mit guten Validitätsbelegen

Bewertung Das FPI-R ist ein bewährtes und in der Praxis sehr beliebtes Verfahren. Es liefert mit seinen Skalen offenbar über jene Merkmale Informationen, für die sich viele Anwenderinnen und Anwender interessieren; es bietet also eine für die Praxis nützliche Auswahl von Skalen. Die durch Konsistenzkoeffizienten belegte Messgenauigkeit reicht für gruppenstatistische Untersuchungen aus. Berechnet man anhand von Cronbachs α Konfidenzintervalle, fallen diese für einige Skalen recht breit aus, sodass ein „durchschnittlicher“ Stanine-Wert bei Berücksichtigung der Messgenauigkeit auch „hoch“ oder „niedrig“ bedeuten kann (s. dazu ▶ Abschn. 4.5.2) – eine unbefriedigende Erkenntnis. Deshalb wäre es wünschenswert, weitere Angaben zur Retest-Reliabilität zu bekommen. Die Retest-Reliabilität sollte an einer größeren Stichprobe, die zwischen beiden Messungen keiner Intervention ausgesetzt wird, nach wenigen Wochen sowie nach 1 Jahr oder später ermittelt werden. Das größere Intervall ist für Fragestellungen wichtig, die längerfristige Prognosen verlangen. Die Belege zur Validität sind insgesamt eindrucksvoll. Die Autoren haben viele Studien im Manual aufgegriffen. Besonders zu erwähnen ist, dass auch für die 9. Auflage eine große und bevölkerungsrepräsentative Eichstichprobe rekrutiert wurde. In einer Rezension (Rohrmann und Spinath 2011, S. 270) wird das FPI-R in der 8. Auflage von 2010, das sich mit der 9. Auflage aber kaum verändert hat, als ein Verfahren charakterisiert, „dass Forscher und Praktiker gleichermaßen überzeugt. Es ist zur fachgerechten und differenzierten Persönlichkeitsdiagnostik im anvisierten Konstruktbereich ausdrücklich zu empfehlen.“

Alternativen Der Hauptanwendungsbereich des FPI-R dürfte im subklinischen Bereich liegen, in dem es nicht vorrangig um eine Diagnostik psychischer Störungen nach den etablierten Diagnosesystemen ICD oder DSM (vgl. ▶ Abschn. 8.3) geht. Die folgenden Fragebögen stellen eher eine Ergänzung zum FPI-R als eine Alternative dar.

Trierer Integriertes Persönlichkeitsinventar

Trierer Integriertes Persönlichkeitsinventar (TIPI) Das TIPI (Becker 2003) kommt in diesem Bereich am ehesten als Alternative zum FPI-R infrage. Es erfasst die Persönlichkeit mit 34 Primärskalen, zusammengefasst in 4 Globalskalen

(Neurotizismus/seelische Gesundheit, Extraversion/Offenheit, Verträglichkeit und Gewissenhaftigkeit/Kontrolliertheit). Einige Skalen weisen eine konzeptuelle Ähnlichkeit mit Skalen im FPI-R auf, so „körperliche Beschwerden/Erschöpfung“ und „verbale Aggressivität“. Beispiele für Skalen, die keine Entsprechung im FPI-R haben, sind „Ungerechtigkeitsgefühl“, „Verlassensangst“, „Suizidalität“ und „magisches Denken“.

Gießen-Test – II (GT-II) Als weiteres Verfahren mit einer gewissen Ähnlichkeit zum FPI-R kann der GT-II (Beckmann et al. 2012) genannt werden. Die Skalen wurden vor einem psychoanalytischen Hintergrund konzipiert und nach den gängigen psychometrischen Prinzipien konstruiert. Die 6 Standardskalen lauten „soziale Resonanz“ (negativ sozial resonant vs. positiv sozial resonant), „Dominanz“ (dominant vs. gefügig), „Kontrolle“ (unterkontrolliert vs. zwanghaft), „Grundstimmung“ (hypomanisch vs. depressiv), „Durchlässigkeit“ (durchlässig vs. retentiv) und „soziale Potenz“ (sozial potent vs. sozial impotent). Der Fragebogen ist besonders für die Diagnostik von Paarbeziehungen geeignet, weil er mit den gleichen Items verschiedene Sichtweisen aufzeigen kann, die für Diagnostik und darauf aufbauende Beratung bei Partnerschaftsproblemen hilfreich sein können: Selbstbild, Fremdbild des Partners bzw. der Partnerin und Idealbild („Wie möchte ich gerne sein?“ bzw. „Wie wünsche ich mir meinen Partner bzw. meine Partnerin?“).

Gießen-Test – II

Fragebogen zur Partnerschaftsdiagnostik (FPD) Speziell für die Partnerschaftsdiagnostik steht mit dem FPD (Hahlweg 2016) ein nützliches Instrument zur Verfügung. Der „Partnerschaftsfragebogen“ dient mit den Skalen „Streitverhalten“, „Zärtlichkeit“ und „Gemeinsamkeit/Kommunikation“ zur Bestimmung der partnerschaftlichen oder Ehequalität. Dazu existiert auch eine Kurzform. Mit der sog. „Problemliste“ können wesentliche Konfliktbereiche in der Partnerschaft identifiziert werden. Weitere Informationen finden sich in einer Rezension von Meuwly et al. (2018).

Fragebogen zur Partnerschaftsdiagnostik

3.3.3.4 Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)

Das BIP von Hossiep und Paschen (2003, 2019) ist ein mehrdimensionaler Persönlichkeitsfragebogen speziell für die berufliche Eignungsdiagnostik. Es liegt in einer Selbst- und einer Fremdbeurteilungsform vor und soll 14 berufsrelevante Dimensionen der Persönlichkeit erfassen.

Verwendung von Persönlichkeitsfragebögen im Personalbereich

Eine Befragung von Personalverantwortlichen in großen deutschen Unternehmen (Hossiep et al. 2015) zeigt, dass vorzugsweise Verfahren eingesetzt werden, die aus konzeptioneller und psychometrischer Sicht problematisch sind. Am häufigsten verwendet werden der Befragung zufolge der Myers-Briggs-Typenindikator (MBTI; Bents und Blank 2003) und das persolog Persönlichkeits-Profil (Gay 2004), zu dem auch eine Testrezension vorliegt (König und Marcus 2013). Auf Platz 3 und 8 landen immerhin mit dem BIP bzw. BIP-6F und dem NEO-FFI bzw. NEO-PI-R (► Abschn. 3.3.3.5) 2 wissenschaftlich fundierte Verfahren. Etwa 24 % der befragten Unternehmen verwenden das BIP und etwa 63 % kennen es. Das NEO-FFI und das NEO-PI-R sind zwar relativ bekannt (etwa 36 % kennen die Verfahren), werden aber eher selten eingesetzt (etwa 7 %; alle Prozentwerte geschätzt nach Abb. 1 in Hossiep et al. 2015).

Für die berufliche Eignungsdiagnostik entwickelt

Das BIP liegt in mehr als einem Dutzend europäischer Sprachen vor; die Computerversion ist sogar als chinesische Version verfügbar. Die große Bekanntheit, Einsatzhäufigkeit und Verbreitung des BIP spricht dafür, das Verfahren hier vorzustellen. Im Anschluss wird auch auf das BIP-6 F kurz eingegangen.

3

Steckbrief BIP: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (3., durchgesehene Auflage; Hossiep und Paschen 2019)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Berufsrelevante Persönlichkeitsmerkmale
Anwendungsbereich	Personalauswahl, Platzierungsentscheidungen, Training, Coaching und Beratung
Theoretischer Hintergrund	Pragmatische Auswahl von Skalen, die nach Expertenmeinung relevante Persönlichkeitsmerkmale betreffen
Testentwicklung	Sammlung von Items zu den angezielten Persönlichkeitsmerkmalen, Reduktion nach inhaltlichen Gesichtspunkten und Itemkennwerten
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche Instruktion
Auswertung	Mit Schablonen
Interpretation	Normtabellen, kurze Erläuterung der Persönlichkeitsmerkmale im Profilbogen, Informationsbroschüre für Testpersonen
Reliabilität	
Konsistenz	Cronbachs $\alpha = .74$ bis $.91$
Retest	$r_{tt} = .77$ bis $.89$ für den Zeitraum von 8 bis 10 Wochen; für den Zeitraum von 5 Monaten bis 3 Jahren $r_{tt} = .69$ bis $.80$
Validität	
Konstruktvalidität	Korrelation mit inhaltlich ähnlichen Skalen anderer Persönlichkeitsfragebögen $r = .54$ bis $.84$
Kriteriumsvalidität	Korrelation mit selbstberichteten beruflichen Erfolgskriterien und Berufszufriedenheit (z. B. Führungsmotivation und erreichte Position: $r = .39$)
Normen	
Zusammensetzung	Über 100 spezifische Zielgruppen, insgesamt $N > 22.000$; Gesamtgruppe: $N = 9303$
Erhebungszeitraum	Bis 2018
Sonstiges	
Formen	Selbst- und Fremdbeurteilungsform; zusätzlich Computerverision
Testrezension	
Quelle	Marcus (2004); betrifft die 2. Auflage von 2003

Pragmatische Skalenauswahl

Theoretischer Hintergrund und Konstruktionsprinzipien Die Autoren suchten für die 1. Auflage (1998) Persönlichkeitsdimensionen, die für den beruflichen Erfolg relevant sind. Gespräche mit Psychologinnen und Psychologen, die in der Personalarbeit tätig sind, sowie Literaturrecherchen zur Beziehung zwischen Persönlichkeitsmerkmalen und Berufserfolg lieferten die

entscheidenden Hinweise. Die Auswahl der 14 Skalen erfolgte also (wie beim FPI-R; ▶ Abschn. 3.3.3.3) nach pragmatischen Gesichtspunkten. Dem Verfahren liegt keine bestimmte Persönlichkeitstheorie zugrunde. Zur Erfassung dieser Merkmale wurden anfangs sehr viele Items generiert, die dann nach inhaltlichen Gesichtspunkten und Itemkennwerten reduziert wurden, sodass am Ende 210 Items übrig geblieben sind. Die Items sind auf einer 6-stufigen Skala von „trifft voll zu“ bis „trifft überhaupt nicht zu“ zu beantworten.

Mit der 2. Auflage (2003) hat sich am Test selbst nichts verändert. Das Manual erfuhr eine Überarbeitung und vor allem eine Aktualisierung; ferner wurden neue Normen bereitgestellt. Für die 3. Auflage (2019) wurde das Manual überarbeitet. Dem Test ist nun ein neuer Normenband (Hossiep und Weiß 2018) beigelegt.

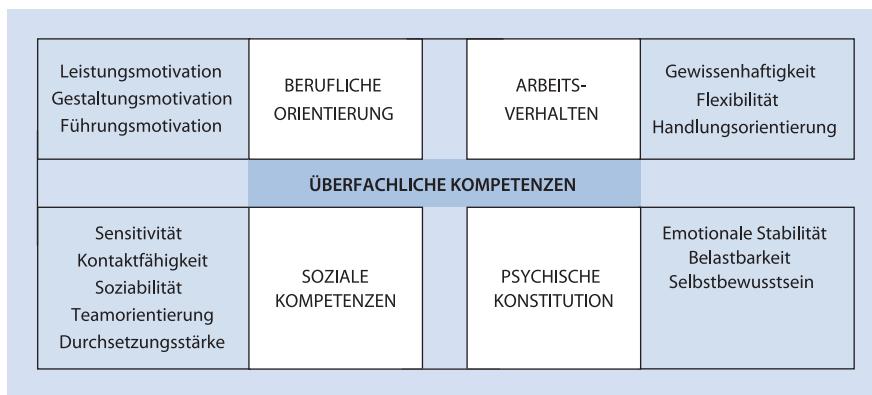
Gliederung Die Skalen des BIP lassen sich auf die 4 Bereiche „berufliche Orientierung“, „Arbeitsverhalten“, „soziale Kompetenzen“ und „psychische Konstitution“ verteilen, wie in □ Abb. 3.19 veranschaulicht. Diese Bereiche spielen allerdings weder bei der Auswertung noch der Interpretation eine Rolle und dürfen keinesfalls als Faktoren höherer Ordnung verstanden werden.

Jede der Skalen umfasst zwischen 12 und 16 Items in Form von Selbstbeschreibungen. Die Skalen des BIP sind zusammen mit Erläuterungen und Itembeispielen in □ Tab. 3.18 zusammengestellt.

Komplettiert wird das Instrumentarium durch einen Fremdeinschätzungsbo- gen. Jede der 14 Dimensionen wird durch 3 Items wie „Die von mir einzuschätzende Person ... ist motiviert, die eigene Arbeit kontinuierlich zu verbessern“ oder „... ist bestrebt, Missstände zu beseitigen“ abgedeckt. Hierbei muss der „Ausprägungsgrad des Verhaltens“ auf einer 9-stufigen Skala von „deutlich unterdurchschnittlich“ über „durchschnittlich“ bis zu „deutlich überdurchschnittlich“ beurteilt werden. Gedacht ist primär an einen Einsatz in Assessment-Centern, in Teamtrainings, als Stütze bei der Vermittlung von systematischem Feedback von Mitarbeiterinnen bzw. Mitarbeitern an ihre Vorgesetzten und in Forschungsarbeiten. Diese Fremdbeurteilungsversion hat eher noch den Charakter einer Forschungsversion (Schulz und Weiß 2018). Eine verbesserte und auch normierte Version folgt nach Auskunft des Projektteams zur Testentwicklung in absehbarer Zeit (Stand: Mai 2019).

14 Skalen zu 4 Bereichen

Fremdeinschätzungsbogen



□ Abb. 3.19 Die Struktur des BIP mit Bereichen (weiß) und Skalen. (© Hogrefe)

Tab. 3.18 Skalen des BIP mit Erläuterungen und Itembeispielen

Skala	Erläuterung ^a	Itembeispiel
Leistungsmotivation	<ul style="list-style-type: none"> – Stellt hohe Anforderungen an die eigene Leistung – Ist bereit, sich bei der Verfolgung seiner Ziele stark zu engagieren – Möchte die eigene Arbeit kontinuierlich verbessern 	Ich bin mit mir erst dann zufrieden, wenn ich außergewöhnliche Leistungen vollbringe.
Gestaltungsmotivation	<ul style="list-style-type: none"> – Verfügt über einen starken Willen, durch seine Tätigkeit gestaltend einzutreten – Ist motiviert, Missstände zu beseitigen – Möchte eigene Vorstellungen umsetzen 	Für einige bin ich ein unbequemer Querdenker.
Führungs motivation	<ul style="list-style-type: none"> – Möchte Führungsverantwortung wahrnehmen – Kann andere Personen überzeugen und für seine Auffassung gewinnen – Wirkt auf andere mitreißend und begeisternd 	Eine Spezialistentätigkeit ist mir lieber als eine Führungsaufgabe. (-)
Gewissenhaftigkeit	<ul style="list-style-type: none"> – Geht bei der Bearbeitung von Aufgaben sorgfältig vor – Hält sich zuverlässig an Vereinbarungen – Neigt zu Perfektionismus 	Ich nehme die Dinge ganz genau.
Flexibilität	<ul style="list-style-type: none"> – Stellt sich problemlos auf neue Situationen ein – Passt Methoden und Vorgehensweisen rasch an sich verändernde Bedingungen an – Kann uneindeutige Situationen gut tolerieren 	Wenn ich vor völlig unerwarteten Situationen stehe, fühle ich mich richtig in meinem Element.
Handlungsorientierung	<ul style="list-style-type: none"> – Beginnt nach der Entscheidungsfindung unverzüglich mit der Umsetzung – Lässt sich durch Ablenkungen und Schwierigkeiten bei der Arbeitsausführung nicht beeinträchtigen – Geht zielorientiert vor 	Wenn ich viele Aufgaben zu erledigen habe, weiß ich manchmal gar nicht, womit ich anfangen soll. (-)
Sensitivität	<ul style="list-style-type: none"> – Hat ein gutes Gespür für die Stimmungen anderer – Kann sich leicht auf verschiedenartige soziale Situationen einstellen – Kann die eigene Wirkung auf andere abschätzen 	Ich kann mich auf die unterschiedlichsten Menschen sehr gut einstellen.
Kontaktfähigkeit	<ul style="list-style-type: none"> – Kann auf andere Menschen zugehen und Kontakte knüpfen – Verfügt über vielfältige Beziehungen und Kontakte – Kommt gern mit anderen Menschen zusammen 	Ich brauche eine Weile, bis ich Bekanntschaften schließe. (-)
Soziabilität	<ul style="list-style-type: none"> – Tritt anderen Menschen freundlich und rücksichtsvoll gegenüber – Schätzt Harmonie im Umgang mit anderen – Hat eine hohe Bereitschaft, sich an unterschiedliche Personen anzupassen 	Ich gehe mit anderen rücksichtsvoll um.
Teamorientierung	<ul style="list-style-type: none"> – Arbeitet gern im Team – Sucht die Zusammenarbeit und den Austausch mit anderen – Ist bereit, Teamentscheidungen zu akzeptieren und mitzutragen 	Ich ziehe es vor, allein zu arbeiten. (-)
Durchsetzungsstärke	<ul style="list-style-type: none"> – Behält bei Auseinandersetzungen die Oberhand – Setzt eigene Vorstellungen durch – Vertritt eigene Auffassungen mit Nachdruck 	Bei Auseinandersetzungen gewinne ich andere leicht für meine Position.
Emotionale Stabilität	<ul style="list-style-type: none"> – Kommt schnell über Probleme und Misserfolge hinweg – Reagiert bei Schwierigkeiten gelassen – Lässt sich nicht entmutigen 	Mich wirft so leicht nichts aus der Bahn.
Belastbarkeit	<ul style="list-style-type: none"> – Ist resistent gegenüber Stress – Fühlt sich auch unter Druck noch leistungsfähig – Reagiert auch bei hoher Beanspruchung widerstandsfähig 	Bei gleichzeitigen Anforderungen von mehreren Seiten werde ich nervös. (-)
Selbstbewusstsein	<ul style="list-style-type: none"> – Ist selbstsicher im sozialen Umgang – Ist wenig besorgt über den Eindruck, der bei anderen hinterlassen wird – Bleibt gelassen in Situationen, in denen eine Bewertung der eigenen Person erfolgt (z. B. Bewerbungsgespräche) 	Vor Begegnungen mit wichtigen Personen werde ich nervös. (-)

Quelle: Adaptiert nach Hossiep und Paschen (2003, S. 22, © Hogrefe)

(-)=Item wird bei der Auswertung invertiert.

Durchführung, Auswertung und Interpretation Der Fragebogen liegt mit schriftlicher Instruktion und Erläuterungen anhand von Beispielitems vor. Die Bearbeitungsdauer beträgt etwa 30–45 min. Im Manual werden einige Einschränkungen bezüglich der Zielgruppe genannt, insbesondere ein Mindestalter von 21 Jahren und Erfahrung mit Tätigkeiten in der Privatwirtschaft (da einige Items entsprechende Situationen betreffen). Die Auswertung ist standardisiert; bei der Papier-und-Bleistift-Form werden Schablonen eingesetzt und die Ergebnisse in einen Profilbogen eingetragen. Der Interpretation dienen nicht nur die Normen sowie die kurzen Erläuterungen der Merkmale im Profilbogen, sondern auch eine Informationsbroschüre, die den Testpersonen ausgehändigt werden kann.

Informationsbroschüre für die Testpersonen

Psychometrische Gütekriterien Bei Fragebögen mit ausführlicher Instruktion und Schablonen sind Durchführungs- und Auswertungsobjektivität gegeben. Die Skalen haben überwiegend eine hohe interne Konsistenz; die meisten Skalen weisen für Cronbachs Alpha Werte im Bereich zwischen .80 und .90 auf (gesamter Streubereich: .74 bis .91). Auf ähnlichem Niveau liegen die Retest-Reliabilitäten nach 8–10 Wochen ($r_{tt} = .77$ bis .89; $N = 108$). Für die Konstruktvalidität sprechen erwartungsgemäße Korrelationen zwischen $r = .54$ und .84 mit konstruktkonvergenten Skalen des Eysenck-Persönlichkeits-Inventar (EPI), dem NEO-FFI (► Abschn. 3.3.3.5) und dem 16-Persönlichkeits-Faktoren-Test (16 PF-R; Schneewind und Graf 1998). Zur Kriteriumsvalidität finden sich Korrelationen mit selbstberichteten Erfolgskriterien, die auf großen Stichproben basieren. Wenig aussagekräftig sind die zahlreichen multiplen Regressionskoeffizienten für die Aufklärung von beruflichem Entgelt oder Berufserfolg.

Hohe interne Konsistenz und Retest-Reliabilität

Normen Der BIP-Normenband 2018 (Hossiep und Weiß 2018) enthält für alle Eichstichproben der 2. Auflage aktualisierte Normwerte, die bis 2018 an mehr als 22.000 Personen erhoben wurden. Normen liegen für Hochschulabsolventen, für verschiedene betriebliche Hierarchiestufen, für unterschiedliche Funktionsbereiche (z. B. Vertrieb) sowie für weibliche Fach- und Führungskräfte vor.

Viele und zudem große Normgruppen

Bewertung Die Items bilden intern konsistente und über die Zeit hinreichend stabile Skalen. Einige Skalen korrelieren sehr hoch miteinander („Führungs-motivation“ und „Durchsetzungsfähigkeit“: $r = .75$); es stellt sich daher die Frage, ob sie wirklich hinreichend unterscheidbare Merkmale erfassen. Leider werden im Manual der 2. Auflage des BIP keine detaillierten Ergebnisse von Faktorenanalysen berichtet. Es wird jedoch versichert:

- » In beinahe allen Befunden fanden sich jedoch Faktoren mit weitgehender Deckungsgleichheit zu den 14 Dimensionen des BIP. (Hossiep und Paschen 2003, S. 25).

Die Belege zur Konstruktvalidität in Form von Korrelationen mit bekannten Persönlichkeitsskalen sind insgesamt überzeugend; im Detail kann man jedoch kritisieren, dass etwa die Primärskala „Regelbewusstsein“ des 16 PF-R, die eng mit „Integrität“ verwandt ist, mit dem BIP nicht abgebildet wird (vgl. Marcus 2004). Die Bemühungen zur Validierung, so der Rezensent, sind weiter fortgeschritten als bei den meisten alternativen Verfahren. Besonders lobt Marcus (2004) die konsequente Berücksichtigung der Perspektive der Testpersonen durch die Hinweise im Manual zur Gestaltung der Rückmeldung.

Standardinstrument im Personalbereich

Kritisch merkt er an, dass noch immer eine Validierung an Leistungsbeurteilungen und für einzelne Skalen Belege für eine diskriminante Validität fehlen. Im Manual berichtete Befunde zur Akzeptanz sprechen dafür, dass das Verfahren in der Personalpsychologie gut angenommen wird. Die Eichstichprobe ist sehr groß und erlaubt es daher, diverse Untergruppen zu bilden. Insgesamt stellt das Verfahren eine gute Grundlage für Explorations-, Beratungs- und Rückmeldegespräche dar. Betrachtet man die eingangs erwähnte Befragung von Personalverantwortlichen, so sticht das BIP bezüglich der Anwendungshäufigkeit unter den wissenschaftlich fundierten Persönlichkeitsfragebögen deutlich heraus, sodass man es als wissenschaftlich fundiertes Standardinstrument im Personalbereich bezeichnen kann. Das BIP liegt inzwischen auch in sehr vielen Sprachen vor.

6 Faktoren überwiegend aus BIP-Items ermittelt

Sorgfältig entwickeltes Verfahren

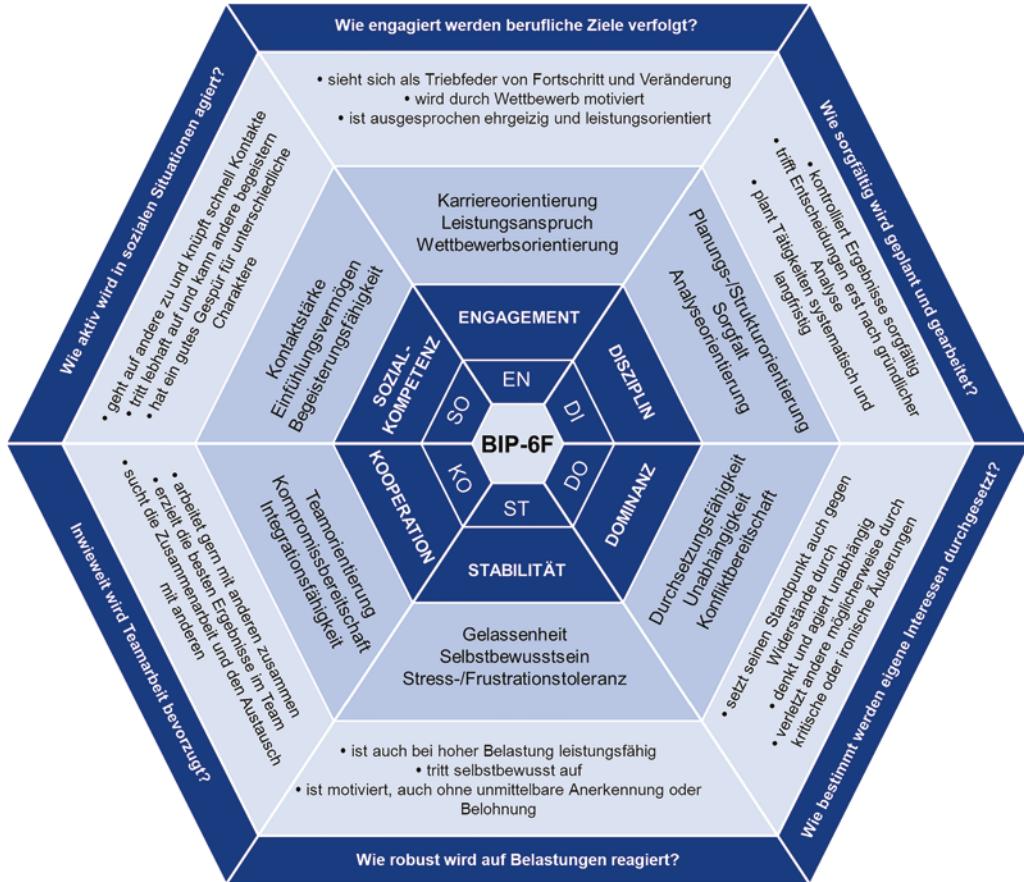
BIP und BIP-6 F können kombiniert werden

Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren (BIP-6 F) Hossiep und Krüger (2012) haben aus dem BIP eine kurze Version mit nur 48 Items abgeleitet. Mittels Faktorenanalysen des BIP sowie von Zusatzitems, die zu Forschungszwecken unter die Items des BIP gemischt worden waren, extrahierten sie 6 Faktoren (siehe □ Abb. 3.20). Es handelt sich also um einen induktiven und damit ergebnisoffenen Konstruktionsansatz. Die 6 Faktoren setzen sich aus 33 BIP-Items und neuen 15 Items aus Forschungsversionen zusammen. Anders als die BIP-Skalen sind die 6 Skalen des BIP-6 F weitgehend unabhängig voneinander. Die Skaleninterkorrelationen reichen von $r = -.01$ (Stabilität und Disziplin) bis $r = .33$ (Engagement und Sozialkompetenz). Das BIP-6 F erfasst mit jeweils 8 Items 6 berufsrelevante Persönlichkeitsmerkmale (□ Abb. 3.20). Die interne Konsistenz (Cronbachs α) der Skalen reicht von .74 (Stabilität) bis .85 (Kooperation). Die Re-test-Reliabilität liegt zwischen $r_{tt} = .80$ und .89.

Abrell-Vogel und Gerstenberg (2014, S. 196) kommen in einer Testrezenzierung zu einer sehr positiven Beurteilung:

- » Das BIP-6 F stellt ein ökonomisches Verfahren dar, das zur Beantwortung verschiedener Fragestellungen der Personalpsychologie herangezogen werden kann. Aufgrund seiner hohen Transparenz sowie Berufsbezogenheit erzeugt es Akzeptanz [...] Insgesamt handelt es sich beim BIP-6 F um ein nach wissenschaftlichen Kriterien sehr gut und sorgfältig entwickeltes Verfahren, das eine gleichsam ökonomische wie inhaltlich wertvolle Erfassung von 6 berufsbezogenen Persönlichkeitsfaktoren ermöglicht.

Vergleicht man BIP und BIP-6 F, so sind eine große Gemeinsamkeit und 2 wesentliche Unterschiede zu erkennen: Beide Verfahren decken mit ihren Items den Bereich der Persönlichkeit ab, der im Personalbereich relevant zu sein scheint. Das BIP-6 F ist mit nur 15 min Durchführungsduer einschließlich Instruktion wesentlich ökonomischer als das BIP. Da es „nur“ 6 Skalen hat, ist die Informationsausbeute zwangsläufig geringer als die des BIP mit seinen Skalen. Je nach Fragestellung ist es daher sinnvoll, sich auf eines der beiden Verfahren zu beschränken oder auch beide ggf. sukzessiv, beginnend mit dem kürzeren BIP-6 F, einzusetzen. Es soll jedoch auch betont werden, dass das BIP und das BIP-6 F kein „Allheilmittel“ für den Personalbereich sind. Die Relevanz der von diesen Verfahren abgedeckten Dimensionen muss im Einzelfall anforderungsanalytisch begründet sein. Mit dem BIP-AM und dem BIP-6 F-AM (► Abschn. 6.1.2.6) liegen dafür geeignete Instrumente vor.



■ Abb. 3.20 Skalen des BIP-6 F mit Erläuterungen. (Aus Hossiep und Krüger 2012, S. 20, mit freundlicher Genehmigung des Hogrefe Verlages. Bezugssquelle des Testverfahrens: Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, ► www.testzentrale.de)

■ Ergänzende Versionen

Wie bereits oben erwähnt, liegt das BIP auch als Fremdbeurteilungsversion vor. Diese existiert bislang nur als Forschungsversion. ■ Tab. 3.19 gibt einen Überblick über das Selbst- und Fremdbeschreibungsinventar sowie das Anforderungsmodul. Im Anforderungsmodul, d. h. bei der Beschreibung der Anforderungen an eine berufliche Tätigkeit (s. auch ► Abschn. 6.1.2.6), werden für jedes Persönlichkeitsmerkmal die Anforderungen durch Personen, die die Stelle gut kennen, beurteilt. Dazu dienen Items wie „Der Positionsnehmer sollte erst dann mit sich zufrieden sein, wenn er außergewöhnliche Leistungen vollbringt“ (Item zur Leistungsmotivation) oder „Der Positionsnehmer muss akribisch vorgehen“ (Item zur Gewissenhaftigkeit). Die Items zur Selbstbeschreibung sind denen der Anforderungsbeschreibung sehr ähnlich (hier: „Ich bin mit mir erst dann zufrieden, wenn ich außergewöhnliche Leistungen vollbringe“ bzw. „Ich nehme die Dinge ganz genau“; Hossiep und Weiß 2020). Die Beurteilung findet auf 6-stufigen Ratingskalen mit den Polen „trifft voll zu“ bis „trifft überhaupt nicht zu“ statt. Um die Skalenwerte

Anforderungsmodul,
Selbstbeschreibungs- und
Fremdbeschreibungsinventar

Tab. 3.19 Übersicht über zentrale Merkmale der BIP-Module

	Anforderungsbeschreibung	Selbstbeschreibung	Fremdbeschreibung
Gegenstand der Messung	Überfachliche Anforderungsbeschreibung(en) einer beruflichen Position	Berufsbezogenes Selbstbild der Persönlichkeit	Berufsbezogene(s) Fremdbild(er) der Persönlichkeit
Ausfüllende Person(en)	Eine oder mehrere Personen, die die infrage stehende berufliche Position gut kennen, z. B. Inhaber gleicher oder ähnlicher Stellen, Vorgesetzte, Kollegen, Personalexperten	Fokusperson selbst	Eine oder mehrere Personen, die die Fokusperson im beruflichen Kontext gut kennen, z. B. Kollegen, Vorgesetzte, Mitarbeiter, interne/externe Kunden
Aufbau	14 Skalen 98 Items 7 Items pro Skala	14 Skalen 210 Items 12–16 Items pro Skala	14 Skalen 98 Items 7 Items pro Skala
Hauptanwendungsfelder	– Personalauswahl u. Personalplatzierung – Personalentwicklung – Coaching	– Personalauswahl und Personalplatzierung – Personalentwicklung – Coaching – Berufs- und Karriereberatung	– Personalentwicklung – Coaching – Teamentwicklung – 360°-Beurteilungen – Platzierung interner Mitarbeiter

Quelle: Hossiep und Weiß (2020, S. 13, © Hogrefe)

Selbstbeschreibung,
Fremdbeurteilung und
Anforderungen in einem System
kombinierbar

mit denen der Selbst- oder Fremdbeschreibung vergleichbar zu machen, werden die Rohwerte anhand der Daten der Normierungsstichprobe für das Anforderungsmodul ebenfalls in Sten-Werte transformiert. So können die Anforderungen an eine Stelle direkt mit der Selbstbeschreibung der Person, die für die Stelle infrage kommt, und/oder der Fremdbeschreibung dieser Person in einer einheitlichen Metrik verglichen werden. Die verwendeten Sten-Werte stellen wie die Stanine-Werte (► Abschn. 2.6.4) Kategorien dar. Die Skala hat 10 Wertebereiche von 1 bis 10. Beispielsweise entspricht Sten 5 $z = -.5$ bis 0 und Sten 6 entsprechend $z = 0$ bis .5. Der Mittelwert der Sten-Werte beträgt 5,5 und die Standardabweichung 2.

Manchmal ist es sinnvoll, die Anforderungen oder die Fremdbeschreibungen von mehreren Personen zu erheben und diese dann zu mitteln. Auch das ist möglich, und so kann dann beispielsweise die Selbstbeschreibung einer Bewerberin oder eines Bewerbers mit einer durch mehrere Personen erstellten Beschreibung der Anforderungen kontrastiert werden. Dies geschieht in sehr anschaulicher Form, weil das Selbstbeschreibungs- und das Anforderungsprofil in einer Grafik dargestellt werden und Diskrepanzen und Übereinstimmungen leicht erkennbar sind. Die Beurteilungen können auch dazu genutzt werden, Anforderungskorridore festzulegen, d. h., eine Anforderung wird nicht als Punktwert auf der 10-stufigen Skala definiert, sondern als ein „grüner Bereich“ von mehreren Skalenstufen, in dem die einzustellende Person mit ihrer Selbstbeschreibung liegen soll.

Weiterführende Internetressourcen

Das Testentwicklungsteam um Dr. Hossiep veröffentlicht Informationen zu den aktuellen Forschungsversionen (► <https://www.testentwicklung.de/testverfahren/BIP/index.html.de> und ► <https://www.testentwicklung.de/testverfahren/BIP-6F/index.html.de>).

3.3.3.5 NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R)

Mit dem NEO-PI-R von Ostendorf und Angleitner (2004) liegt ein Verfahren vor, das dem Fünf-Faktoren-Modell der Persönlichkeit verpflichtet ist, aber zusätzlich die Big-Five-Dimensionen in Facetten unterteilt. Die 5 Globalskalen repräsentieren das Persönlichkeitsmodell und standen von Anfang an fest; sie wurden anschließend in Facetten unterteilt. Das Verfahren ist eng an das amerikanische Original angelehnt. Die Autoren legten dabei Wert auf eine sinngemäße und nicht wörtliche Übersetzung der Items. Auf eine Elimination oder Ersatz von Items mit niedrigen Trennschärfen wurde verzichtet. Das Verfahren liegt als Selbstbeurteilungs- und Fremdbeurteilungsversion (Form S und F) vor.

Fragebogen zum Fünf-Faktoren-Modell

Steckbrief NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (Ostendorf und Angleitner 2004)	
Zielsetzung und Testkonstruktion	
Messgegenstand	Globale Persönlichkeitsmerkmale (Neurotizismus, Extraversion, Offenheit für Erfahrung, Verträglichkeit und Gewissenhaftigkeit) und jeweils 6 Facetten
Anwendungsbereich	Insbesondere Grundlagen- und anwendungsbezogene Forschung, aber auch diagnostische Praxis (z. B. Personalpsychologie)
Theoretischer Hintergrund	Fünf-Faktoren-Modell der Persönlichkeit
Testentwicklung	Adaptation der amerikanischen Ausgabe
Maßnahmen zur Gewährleistung der Objektivität	
Durchführung	Standardisierte schriftliche Instruktion
Auswertung	Bei separatem Antwortblatt sind die Zahlen auf dem Durchschreibbogen skalenweise zu addieren; ansonsten Eingabe in Computerprogramm empfohlen
Interpretation	Normwerte; Erläuterungen zu den Global- und Subskalen im Beiheft „Persönlichkeitsbild“
Reliabilität	
Konsistenz	Globalskalen: $\alpha = .87$ bis $.91$ ($M = .89$); Facetten: $\alpha = .50$ bis $.82$ ($M = .71$); Angaben basierend auf repräsentativer Quotenstichprobe ($n = 871$)
Retest	Globalskalen: Median = $.90$ (nach 1–2 Monaten) bzw. $.75$ (nach 5 Jahren); Facetten: Median = $.82$ (nach 1–2 Monaten) bzw. $.68$ (nach 5 Jahren)
Validität	
Konstruktvalidität	Zusammenhang u. a. von Selbst- und Fremdbeurteilung bei $r = .47$ bis $.61$ (Facetten: $.27$ bis $.61$); Überprüfung der Faktorenanalyse
Kriteriumsvalidität	Keine Angaben
Normen	
Zusammensetzung	$N = 11.724$; bezüglich Alter, Geschlecht und Bildung nicht repräsentativ – daraus aber auch eine repräsentative sekundäre Quotenstichprobe ($n = 871$) gezogen
Erhebungszeitraum	Laut Testautoren nicht auf ein bestimmtes Jahr eingrenzbar (1999 wird als „Schätzung“ angegeben)
Sonstiges	
Formen	Selbst- und Fremdbeurteilungsversion; Computerversion verfügbar
Testrezension	
Quelle	Andresen und Beauducel (2008), Muck (2004)

5 Hauptskalen mit je 6 Facetten

Gliederung Jede der 5 globalen Persönlichkeitsdimensionen wird durch 6 Teilskalen mit je 8 Items näher beschrieben (► Tab. 3.20). Das NEO-PI-R besteht damit aus 30 Subskalen (Facetten) und 5 Hauptskalen mit insgesamt 240 Items. Zu jeder Facette liegen 8 Items vor; Itembeispiele sind: „Die meisten Menschen, die mir begegnen, sind mir wirklich sympathisch“ (E1 Herzlichkeit) und „Ich bin für meine Umsicht und meinen gesunden Menschenverstand bekannt“ (C1 Kompetenz; Ostendorf und Angleitner 2004). ► Tab. 3.20 zeigt die Skalen sowie Gütekriterien, die unten erläutert werden.

5-stufige Antwortskala

Durchführung, Auswertung und Interpretation Die 240 Items sind auf einer 5-stufigen Skala von „starke Ablehnung“ bis „starke Zustimmung“ zu beantworten. Es liegen 2 unterschiedliche Testhefte vor, die entweder ein Ankreuzen im Testheft oder auf einem separaten Antwortbogen vorsehen. Die Durchführung dauert etwa 30–40 min. Die Fremdbeurteilungsversion gleicht der Selbstbeurteilungsversion weitgehend. Die Items unterscheiden sich nur darin, dass sich die Aussage auf eine andere Person bezieht (z. B. statt „Ich bin leicht zu erschrecken“ „Er/Sie ist leicht zu erschrecken“). Bei den Fragebogenformen mit integriertem Antwortmodus empfehlen die Autoren,

► Tab. 3.20 Skalen des NEO-PI-R mit Angaben zur Reliabilität und Validität

Skala	α	r_{tc}	Skala	α	r_{tc}
Neurotizismus	.91	.53	Verträglichkeit	.87	.47
Ängstlichkeit	.79	.53	Vertrauen	.72	.38
Reizbarkeit	.72	.48	Freimütigkeit	.61	.27
Depression	.82	.52	Altruismus	.68	.33
Soziale Befangenheit	.67	.41	Entgegenkommen	.65	.49
Impulsivität	.60	.46	Bescheidenheit	.75	.37
Verletzlichkeit	.77	.49	Gutherzigkeit	.57	.46
Extraversión	.89	.61	Gewissenhaftigkeit	.88	.53
Herzlichkeit	.69	.49	Kompetenz	.62	.39
Geselligkeit	.77	.59	Ordnungsliebe	.63	.53
Durchsetzungsfähigkeit	.80	.60	Pflichtbewusstsein	.61	.39
Aktivität	.72	.51	Leistungsstreben	.67	.50
Erlebnishunger	.64	.61	Selbstdisziplin	.77	.50
Frohsinn	.79	.49	Besonnenheit	.74	.42
Offenheit für Erfahrungen	.89	.53			
Offenheit für Fantasie	.77	.37			
Offenheit für Ästhetik	.78	.57			
Offenheit für Gefühle	.75	.41			
Offenheit für Handlungen	.66	.50			
Offenheit für Ideen	.76	.50			
Offenheit des Normen- und Wertesystems	.50	.40			

Quelle: Ostendorf und Angleitner (2004, © Hogrefe). Hauptskalen **fett** gedruckt. Cronbachs α der Selbstbeurteilungsform ermittelt an einer repräsentativen Stichprobe ($N=871$; aus Tab. 38 und 39 im Manual); r_{tc} =Korrelation mit der Fremdbeurteilungsform (gemittelte Beurteilungen durch 2 Bekannte; $N=750$; aus Tab. 51 im Manual)

die Itembeantwortungen in einen Computer einzugeben und durch ein Programm auszuwerten (im Anhang des Manuals befindet sich eine Anweisung für die Auswertung mit der Statistiksoftware SPSS). Bei Verwendung der Testvariante mit separatem Antwortblatt mit Durchschreibform sind die Zahlenwerte (die für den Grad der Zustimmung zu einer Aussage stehen) für die angekreuzten Antwortalternativen für jede Skala aufzusummieren.

Die Transformation in Normwerte geschieht, indem auf einem zum Alter und Geschlecht der Testperson passenden Profilblatt lediglich die Rohwerte markiert werden. Die Rohwerte sind skalenweise so angeordnet, dass die Höhe des Rohwertes auf dem Profilblatt direkt dem Normwert entspricht. Mithilfe der am Rand des Profilblatts stehenden T-, Stanine- und Prozentrangwerte kann bei Bedarf eine Transformation in einen anderen Normwert erfolgen.

Zur Interpretation der Skalenwerte steht ein Beiheft mit dem Namen „Persönlichkeitssbild“ zur Verfügung. Darin finden sich in verständlicher Sprache Erläuterungen zu den Hauptskalen und den Facetten. Beispielsweise ist zur Skala „Vertrauen“ zu lesen:

- » Personen mit **hohen** Punktewerten neigen dazu, andere Menschen generell für ehrlich zu halten und ihnen gute Absichten zu unterstellen. Personen mit **niedrigen** Punktewerten beschreiben sich als eher skeptisch und misstrauisch gegenüber anderen Menschen. Sie unterstellen anderen schneller unredliche Absichten (Ostendorf und Angleitner 2004, S. 9).

Zur Verbalisierung der Ausprägung werden die T-Wert in 3 Bereiche von „sehr niedrig“ bis „sehr hoch“ unterteilt. Ferner sind u. a. Angaben zu den Konfidenzintervallen vorhanden.

Reliabilität Die internen Konsistenzen der Facettenskalen liegen im Durchschnitt nur bei $\alpha = .71$ (für Details s. □ Tab. 3.20). Bei einer separaten Auswertung für Männer und Frauen finden sich nur geringfügige Abweichungen von den Kennwerten der Gesamtgruppe. Die Items können jedoch auch über die Facetten hinweg zu den 5 Globalskalen verrechnet werden. Die großen Itemzahlen führen dazu, dass die internen Konsistenzen der Globalskalen mit durchschnittlich $\alpha = .89$ sehr hoch ausfallen (Details in □ Tab. 3.20). Die Retest-Reliabilitäten der Globalskalen liegen bei kurzen Zeitintervallen zwischen beiden Testungen (1–2 Monate) zwischen .82 und .91 (Median = .90) und bei längeren (5 Jahre) zwischen .74 und .78 (Median = .75). Für die Facetten fallen die Koeffizienten erwartungsgemäß etwas niedriger aus (Median = .82 bzw. .68).

Validität Die Ausführungen im Manual zur Konstruktvalidität umfassen 37 Seiten und weitere 7 Seiten zur Faktorenstruktur, die ebenfalls der Konstruktvalidität zuzurechnen ist. Von den zahlreichen Befunden können daher hier nur einige besonders erwähnenswerte vorgestellt werden. Die Faktorenstrukturen der 30 Skalen der Selbst- und Fremdbeurteilungsform stimmen sehr gut überein, ebenso die der Männer und Frauen sowie die von verschiedenen Altersgruppen. Die Zuordnung der Facetten zu den Hauptskalen wird durch Faktorenanalysen der Skalen überwiegend gut bestätigt. In einigen wenigen Fällen scheint jedoch Nachbesserungsbedarf bei der Zusammensetzung der Facettenskalen zu bestehen. So lädt die Impulsivitätsskala höher auf dem Extraversionsfaktor als auf dem Neurotizismusfaktor. Einige wenige Skalen

Profilblatt

Umfangreiche Interpretationshilfen

Interne Konsistenz und Retest-
Reliabilitäten der Globalskalen hoch

„Impulsivität“ passt nicht gut zu
„Neurotizismus“

(insbesondere Durchsetzungsfähigkeit) weisen beträchtliche Nebenladungen auf anderen Faktoren auf. Selbst- und Fremdberichte (gemittelte Beurteilung durch 2 Bekannte) korrelieren im Durchschnitt um .54 (Globalskalen) bzw. .47 (Facetten); die Variation ist jedoch beträchtlich (Tab. 3.20). Gemeinsame Faktorenanalysen der Globalskalen mit den Skalen anderer deutscher Persönlichkeitsskalen (u. a. Gießen-Test, FPI-R, BIP) informieren über die Einordnung dieser Skalen in das Fünf-Faktoren-Modell der Persönlichkeit.

Sehr große Normierungsstichprobe mit repräsentativer Teilstichprobe

Normierung Die Normierungsstichprobe für die Selbstbeurteilungsversion (Form S) umfasst 11.724 Personen, die an einer der zahlreichen Studien überwiegend in den Jahren 1999 und 2000 teilgenommen hatten. Daraus wurde zusätzlich nach den Angaben im Statistischen Jahrbuch für das Jahr 2001 eine sekundäre Quotenstichprobe ($N=871$) gezogen, die hinsichtlich Alter, Geschlecht und Bildungsstand als repräsentativ für Deutschland gelten kann. Normen mit separaten Profilblättern liegen für mehrere Gruppen vor: Gesamtstichprobe, jeweils Männer und Frauen von 16–20, von 21–24, von 25–29, von 30–49 und von über 49 Jahren, repräsentative Gesamtstichprobe (auch separat für Männer und Frauen). Für den Bereich „Offenheit“ stehen zusätzlich Normtabellen für Männer und Frauen zweier Altersgruppen und hohem vs. niedrigem Bildungsstand zur Verfügung. Die Normierung der Fremdbeurteilungsversion (Form F) erfolgte an 1547 Personen. Hier wird zwischen 4 Teilgruppen unterschieden, die nach Geschlecht und Alter in Jahren (16–29 sowie 30 und älter) gebildet wurden.

Sehr differenzierte Messung der Persönlichkeit

Bewertung Das NEO-PI-R erlaubt es, die großen 5 Persönlichkeitsdimensionen sehr zuverlässig zu messen. Die Facetten ermöglichen eine sehr differenzierte Beschreibung der Persönlichkeit – allerdings bei teilweise unbefriedigend niedrigen Skalenreliabilitäten. Dass auch eine normierte Fremdbeurteilungsversion zur Verfügung steht, ist für viele Anwendungen als ein großer Vorteil zu werten; durch eine Kombination der Selbst- und der Fremdbeurteilungsversion erschließen sich neue Anwendungsmöglichkeiten (z. B. Vergleich von Selbst- und Fremdbeurteilung zu Beratungszwecken). Das Verfahren wurde an einer großen Stichprobe normiert, zudem liegen auch differenzierte Normen vor.

Das NEO-PI-R ist international weitverbreitet; die amerikanische Originalversion wurde bislang in über 30 Sprachen übersetzt (Ostendorf und Angleitner 2004). Damit eröffnet sich die Chance, Forschungsarbeiten mit der deutschen Version international zu publizieren, was wiederum Forscherinnen und Forscher anregen wird, mit diesem Verfahren zu arbeiten. Zugeleich profitieren deutsche Anwenderinnen und Anwender von den nationalen und internationalen Forschungsarbeiten. Insgesamt ist das NEO-PI-R als ein theoretisch sehr gut fundiertes Verfahren zu bewerten. Zwischen der Übersetzung der Items und der Publikation des Verfahrens liegen 13 Jahre. Die Autoren haben die Zeit für eine sorgfältige Konstruktion, Evaluierung und Normierung genutzt. Dieses umsichtige Vorgehen hat Vorbildcharakter!

In einer Testrezension nach den Standards des Testkuratoriums fassen die Rezensenten ihr Gesamтурteil wie folgt zusammen:

- » Es handelt sich insgesamt um die gelungene Adaptierung eines solide konstruierten amerikanischen Persönlichkeitstests mit grundwissenschaftlichem Schwerpunkt, sehr großer Verbreitung und weltweiter Marktführerschaft. Die fünf Domänenskalen sind hoch konsistent und hinreichend zeitstabil gemessen, aber teilweise in den Faktorenpaaren E/O, N/C und N/E stärker korreliert

Grundwissenschaftlicher Schwerpunkt

als in konkurrierenden Verfahren, die bei den entsprechenden Paarungen angenähert Null-Korrelationen erzielen. Hier könnten Rekonstruktionen sinnvoll sein. Auf der Facettenebene besteht bei einzelnen Facetten, vor allem O6, Optimierungsbedarf hinsichtlich der Reliabilität. [...] Trotz der wünschenswerten Weiterentwicklungen handelt es sich beim NEO-PI-R aber um ein – gemessen am Konstruktionsziel – sehr gut konstruiertes Verfahren, dessen Einsatz eine fachgerechte und differenzierte Persönlichkeitsdiagnostik nach dem FFM [Fünf-Faktoren-Modell der Persönlichkeit] gestattet (Andresen und Beauducel 2008, S. 544).

Auch ein sorgfältig konstruiertes Verfahren kann kleine Mängel aufweisen. Im Einsatz des amerikanischen Originals bei Jugendlichen hatten sich einige Items als schwer verständlich erwiesen. McCrae et al. (2005) haben daher eine als NEO-PI-3 bezeichnete Version entwickelt, in der insgesamt 37 Items des in den USA weiter verwendeten NEO-PI-R ausgetauscht wurden. Da die deutsche Version eng an das amerikanische Original angelehnt ist, sind vermutlich auch hier einige Items verbesserungswürdig.

Im amerikanischen Original einige Items schwer verständlich

Alternativen

NEO-Fünf-Faktoren-Inventar (NEO-FFI) Anwenderinnen und Anwender, denen das NEO-PI-R mit 240 Items zu zeitaufwendig ist, können auf die revidierte Form des Fünf-Faktoren-Inventars, das NEO-FFI (Borkenau und Ostendorf 2008), zurückgreifen, das mit insgesamt 60 Items nur die Big-Five abbildet.

Auch andere Big-Five-Skalen verfügbar

Big-Five-Inventory-10 (BFI-10) Für Forschungszwecke stehen auch im Vergleich zum NEO-PI-R oder auch NEO-FFI deutlich kürzere Skalen zur Verfügung. So haben Rammstedt et al. (2013) auf der Basis einer amerikanischen Kurzversion ein deutschsprachiges Big-Five-Instrument mit nur 10 Items, das BFI-10, vorgelegt und evaluiert.

Weiterführende Internetressourcen

Zur Beschreibung der Persönlichkeit liegt eine frei verfügbare Itemsammlung vor, der *International Personality Item Pool* (► <https://ipip.ori.org/>). Für nähere Informationen kann auf Goldberg et al. (2006) verwiesen werden. Aus den über 3.320 englischsprachigen Items wurden über 250 Skalen in vielen Sprachen konstruiert (s. ► <https://ipip.ori.org/newitemtranslations.htm>). Für Forscherinnen und Forscher dürfte es interessant sein, dass aus dem Itempool auch deutschsprachige Instrumente hervorgegangen sind. Auch das amerikanische NEO-PI bzw. NEO-PI-R gehört übrigens dazu.

3.3.3.6 Persönlichkeitsfragebögen für Kinder

■ Persönlichkeitsfragebogen für Kinder zwischen 9 und 14 Jahren (PFK 9–14)

Einsatzbereich des PFK 9–14

Bei Kindern kommen Selbstbeurteilungsverfahren erst ab einem Alter in Frage, in dem sie hinreichend gut lesen können und über die nötige Selbsteinsicht verfügen. Im deutschen Sprachraum gibt es nur sehr wenige entsprechende Verfahren, die zudem 10 Jahre oder älter sind, aber auch ein mehrdimensionales Verfahren für Kinder, das schon lange auf dem Markt ist und in den Hitlisten der am häufigsten eingesetzten Fragebögen (► Tab. 3.12) auftaucht: der PFK 9–14 von Seitz und Rausche (2019). Er liegt in der 5., aktualisierten Auflage vor und wurde neu normiert. Als Einsatzbereiche werden Erziehungs- und schulpsychologische Beratung, Kinder- und Jugendpsychiatrie, Früherkennung von potenziell verhaltensauffälligen Kindern, forensisch-psychologische Begutachtung, Therapieverlaufskontrolle und Forschung genannt.

Der PFK 9–14 umfasst 15 Primärdimensionen, die sich auf die Kategorien „Verhaltensstile“, „Motive“ und „Selbstbild-Aspekte“ verteilen. Dazu existieren separate Testhefte, die eine getrennte Messung erlauben. Die Skalen sind hier mit ihrer Zuordnung zu den 3 Kategorien aufgeführt.

3

Skalen des PFK 9–14

- Verhaltensstile:
 - Emotionale Erregbarkeit
 - Fehlende Willenskontrolle
 - Extravertierte Aktivität
 - Zurückhaltung und Scheu im Sozialkontakt
- Motive:
 - Bedürfnis nach Ich-Durchsetzung, Aggression und Opposition
 - Bedürfnis nach Alleinsein und Selbstgenügsamkeit
 - Schulischer Ehrgeiz
 - Bereitschaft zu sozialem Engagement
 - Neigung zu Gehorsam und Abhängigkeit gegenüber Erwachsenen
 - Maskulinität der Einstellung
- Selbstbild-Aspekte:
 - Selbsterleben von allgemeiner (existenzieller) Angst
 - Selbstüberzeugung von eigenen Meinungen, Entscheidungen und Planungen
 - Selbsterleben von Impulsivität
 - Egozentrische Selbstgefälligkeit
 - Selbsterleben von Unterlegenheit gegenüber anderen

Fragebogen zur differenzierten Beschreibung der Persönlichkeit

Die internen Konsistenzen sind mit denen von Erwachsenenfragebögen vergleichbar ($\alpha = .63$ bis $.79$; Sekundärfaktoren: $\alpha = .80$ bis $.89$). Die 15 Primärdimensionen können faktorenanalytisch auf folgende 4 Sekundärfaktoren reduziert werden: „derb-draufgängerische Ich-Durchsetzung“, „Emotionalität (Angst)“, „aktives Engagement“ sowie „selbstgenügsame soziale Isolierung“.

Bisher wurden Fragebögen vorgestellt, die zur Messung von Persönlichkeitsmerkmalen im engeren Sinne dienen. Interessen und Motive können als Persönlichkeitsmerkmale im weiteren Sinne verstanden werden. Im Folgenden werden solche Verfahren exemplarisch vorgestellt.

3.3.3.7 Fragebögen zur Erfassung von Interessen

Interessentests dienen vor allem der Beratung bei der Berufswahl. Deshalb liegt es nahe, die Skalen nach Merkmalen der beruflichen Tätigkeit oder (wie beim EXPLORIX, s. u.) nach einer Theorie der Berufswahl auszuwählen. Ein Interessentest kann ein gesamtes Spektrum von Berufen abdecken oder auch nur auf einen Bereich wie etwa Sozialberufe fokussieren.

Die Möglichkeiten, Interessen zu messen, sind vielfältig. Ein eher konventionelles Vorgehen besteht darin, konkrete Tätigkeiten (z. B. sich draußen in der Natur aufzuhalten) zu benennen und die Testpersonen einstufen zu lassen, wie gerne sie das machen oder auch machen würden. Anstelle von verbalen Beschreibungen können auch Bilder verwendet werden, die Menschen bei (beruflichen) Tätigkeiten zeigen. In der Schweiz wird mit dem Foto-Interessen-Test F-I-T Serie 2020 (Jungo und Toggweiler 2019) dieser Ansatz verfolgt. Vermutlich sind ökonomische Gründe dafür verantwortlich, dass bei den meisten Interessen-Fragebögen verbale Items verwendet werden. Erstellung, Selektion und Modifikation professioneller Fotos sind aufwendiger und teurer, als Items zu schreiben und ggf. umzuformulieren.

Auch Fotos als Items möglich

Das Feld der Interessentests ist sehr überschaubar, wenn man ältere Verfahren ausklammert. Wir stellen 2 Verfahren vor, die den gleichen theoretischen Hintergrund haben, und zwar die Berufswahltheorie von John Holland (1997).

EXPLORIX – das Werkzeug zur Berufswahl und Laufbahnplanung

Der deutschsprachige EXPLORIX von Joerin Fux et al. (2012) stellt eine Adaptation und Weiterentwicklung des Instruments Self-directed Search (SDS) von Holland (1970) dar und wurde zur Unterstützung bei der Berufswahl und der Laufbahnplanung entwickelt. Eine Besonderheit ist, dass der Fragebogen auch online zur Selbstdiagnostik mit anschließendem Ergebnisbericht angeboten wird. Theoretischer Hintergrund ist die Berufswahltheorie von John Holland, die erstmals 1959 vorgestellt und seitdem bis zur letzten Fassung von 1997 weiterentwickelt worden ist. Holland postulierte, dass sich 6 Interessens- bzw. Persönlichkeitstypen unterscheiden lassen und analog dazu 6 Typen von Arbeitsumgebungen existieren, weil die Umwelten von den Menschen geprägt werden, die in ihnen tätig sind. In □ Tab. 3.21 werden die 6 „Typen“ kurz charakterisiert. Das Modell wird auch in ▶ Abschn. 6.3.3 skizziert.

Für Berufswahl und Laufbahnplanung entwickelt

Gliederung Im Anschluss an einige Fragen mit freier Beantwortung, darunter eine Auflistung von Berufen, die eine Person schon in Betracht gezogen hat („Berufsträume, Wünsche und Ideen“), folgen 4 Untertests, in denen die Items blockweise nach den Holland-Typen aufgeführt sind (in Klammern ist die Zuordnung zum Typ aufgeführt; Erläuterung der Abkürzungen in □ Tab. 3.21; Joerin et al. 2012):

4 Subtests zu unterschiedlichen Themen

- **Tätigkeiten:** 11 Items pro Typ; Wie gerne würde man Tätigkeiten wie „aus Holz ein Büchergestell zimmern“ (R) oder „kunstvolle Fotos machen“ (A) ausführen?
- **Fähigkeiten:** 11 Items pro Typ; Welche Tätigkeiten wie „gut vor Leuten sprechen“ (E) oder „mit großer Ausdauer sorgfältig arbeiten“ (C) kann man gut oder kompetent ausführen?

□ **Tab. 3.21** Die RIASEC-Typen der Berufswahltheorie von Holland

Typ	Charakterisierung	Werte, Ziele	Berufsbeispiele
R (Realistic)	Realistisch, handwerklich-technisch	Gesunder Menschenverstand	Zimmermann, Landwirtin
I (Investigative)	Intellektuell, untersuchend-forschend	Wissen/Lernen	Physikerin, Forscherin
A (Artistic)	Kreativ, künstlerisch, sprachlich, gestalterisch	Künstlerischer Ausdruck, Kultur	Musikerin, Schauspieler
S (Social)	Sozial, erziehend-pflegend	Helfen, Beziehungen	Lehrer, Psychotherapeutin
E (Enterprising)	Unternehmerisch, führend-organisierend-verkaufend	Finanzieller Erfolg, Verantwortung	Verkäufer, Politikerin
C (Conventional)	Konventionell, ordnend-verwaltet	Anpassung, gesellschaftliche Normen	Kaufmännischer Angestellte, Kassierer

Quelle: Nach Joerin et al. (2012, © Hogrefe)

- *Berufe*: 14 Items pro Typ; Welche Berufe wie „Wissenschaftsjournalist/-in“ (I) oder „Gerichtsbeamter/-beamtin“ (C) interessieren eine Person oder sprechen sie an?
- *Selbsteinschätzung*: 2 Items pro Typ; Wie schätzt man seine Fähigkeiten wie „Verkaufsgeschick“ (E) oder „Einfühlungsvermögen“ (S) ein?

3

Codes führen zu Berufen

Durchführung, Auswertung und Interpretation Der EXPLORIX kann selbstständig durchgeführt werden. Auch eine Gruppenuntersuchung ist möglich. Für die Bearbeitung der 228 Items sind etwa 20 min zu veranschlagen. Bei der Papierversion wertet die Testperson den Test im Regelfall selbst aus, indem sie lediglich für jeden Interessenstyp die zustimmenden Antworten auszählen hat. Bei der (kostenpflichtigen) Onlinetestung entfällt die selbst vorzunehmende Auswertung natürlich. Normen existieren nicht. Der höchste, zweithöchste und dritthöchste Wert ergibt den Holland-Code (Beispiel: R=40, I=35, A=20, S=18, E=30, C=20 ergibt RIE). In einem länder-spezifischen Berufsregister, das für Deutschland (ebenso für Österreich und für die Schweiz) über 1000 Berufe und Funktionen umfasst, sind für alle Holland-Codes passende Berufe mit Angabe des notwendigen Bildungswegs aufgeführt (z. B. für RIE ca. 40 Berufe von Biolandwirt/-wirtin bis Werkstoffingenieur/-ingenieurin). Die Autorinnen und Autoren raten dazu, für alle 6 Permutationen des Dreiercodes (im Beispiel also auch REI, IER, IRE, EIR, ERI) die Berufe nachzuschlagen. Wenn bereits ein Berufswunsch geäußert wurde, kann der Holland-Code für diesen Beruf mit dem Code der Probandin bzw. des Probanden verglichen werden. In einer Testrezension kritisieren Köller und Zettler (2017), dass die meisten Codes nur durch Expertinnen- bzw. Expertenratings bestimmt und nur sehr wenige empirisch überprüft wurden. Die Interpretation endet also damit, dass eine mehr oder weniger lange Liste von Berufen gefunden wird, zu denen die Testperson aufgrund ihrer 3 am stärksten ausgeprägten Interessensrichtungen passt.

Hohe interne Konsistenz und Retest-Reliabilität

Reliabilität Die internen Konsistenzen (Cronbachs α) für die 6 Interessenskalen liegen zwischen $\alpha=.86$ und $\alpha=.91$ und damit in einem hohen Bereich. Für eine Kurzform, bestehend aus den Skalen „Tätigkeiten“ und „Fähigkeiten“, beträgt die Retest-Reliabilität bei einem Zeitintervall von 15 bis 18 Monaten im Durchschnitt .80 (von .63 für C bis .87 für A; N=70 Berufstätige).

6 schwach korrelierende Faktoren

Validität Die 6 Typenskalen sind relativ unabhängig voneinander (höchste Korrelation: $r_{S-A}=.48$). Faktorenanalysen mit schiefwinkliger Rotation der 24 Subskalen (Tätigkeiten, Berufe, Selbsteinschätzung und Fähigkeiten für R, I, A, S, E und C) ergeben 6 schwach korrelierte Faktoren, die den 6 Typen entsprechen. Erwartungsgemäß treten erhebliche Geschlechtsunterschiede auf: Für „Realistic“ weisen Männer im Mittel höhere Werte auf als Frauen ($M=29$ vs. 18). In „Social“ und „Artistic“ zeigen Frauen durchschnittlich höhere Werte ($M=29$ und 30 vs. 21). Mit den Skalen des NEO-FFI finden sich einige plausible Zusammenhänge. So korreliert Offenheit mit „Artistic“ und „Investigative“ zu .47 bzw. .37, Extraversion mit „Enterprising“ zu .44 und Gewissenhaftigkeit mit „Enterprising“ und „Conventional“ zu .32 bzw. .27.

Keine Normen

Normen Die Interpretation basiert ausschließlich auf Rohwerten. Die Autorinnen und Autoren argumentieren, dass die 6 Typenskalen ungefähr gleich attraktiv seien; die Mittelwerte lägen bei etwa 25 Punkten. Dem Anhang des Manuals ist zu entnehmen, dass die Mittelwerte tatsächlich aber zwischen 21,6 (R) und 28,1 (S) liegen und die Streuungen ebenfalls uneinheitlich ausfallen ($SD=7,6\text{--}10,8$).

Bewertung Der EXPLORIX stellt auf dem deutschsprachigen Testmarkt seit dem Erscheinen der 1. Auflage im Jahr 2002 eine interessante und vielversprechende Innovation dar. Das Verfahren ist theoretisch gut fundiert und empirisch vergleichsweise intensiv untersucht. Die Skalen messen mit hoher Zuverlässigkeit 6 gut unterscheidbare Interessensrichtungen. Befunde zur Kriteriumsvalidität fehlen jedoch noch immer; die von den Autorinnen und Autoren berichteten Korrelationen mit einem anderen Interessentest und dem NEO-FFI dienen als Belege der Konstruktvalidität. Wünschenswert wäre der Nachweis, dass eine Beratung unter Zuhilfenahme von EXPLORIX zu einer größeren späteren Berufszufriedenheit führt als eine Beratung ohne dieses Instrument. Ferner sollten zufriedene Stelleninhaberinnen und Stelleninhaber häufiger den zu ihrem Beruf passenden Holland-Code aufweisen als unzufriedene. Dies sind wichtige, aber nicht erschöpfende Anforderungen an die Kriteriumsvalidität des Verfahrens. Die empirische Grundlage für den Verzicht auf Normen überzeugt nicht; von gleichen Mittelwerten und Streuungen der 6 Skalen ist, wie bereits erwähnt, nicht auszugehen. Bei einer Revision des Verfahrens könnten die Skalen durch eine veränderte Itemauswahl und/oder Ergänzung um weitere Items auf gleiche Mittelwerte und Streuungen eingestellt werden. In der Testrezension von Köller und Zettler (2017) erfährt die Validität nur die Bewertung „Anforderungen teilweise erfüllt“. Die vorgelegten Validitätsbelege „lassen zwar auf eine gute Validität der Interpretation der EXPLORIX-Ergebnisse hoffen, robuste, dies unterstützende empirische Befunde stehen allerdings noch aus“ (Köller und Zettler 2017, S. 100).

Verzicht auf Normen problematisch,
Kriteriumsvalidität nicht belegt

Exkurs Mit dem Fragebogen EXPLOJOB (Joerin Fux und Stoll 2006) steht ein Verfahren zur Beschreibung von Berufsansforderungen und -tätigkeiten zur Verfügung, das für bestimmte Fragestellungen zusammen mit dem EXPLORIX eingesetzt werden kann. EXPLOJOB beschreibt den Arbeitsplatz, EXPLOJOB den Menschen – und zwar anhand der gleichen 6 Dimensionen. Zusammen eingesetzt können die beiden Verfahren also Informationen dazu liefern, wie gut etwa Ratsuchende aufgrund ihrer Interessen zu einem ins Auge gefassten Beruf passen. Bei Unzufriedenheit mit dem Arbeitsplatz lässt sich prüfen, ob eine Passung zwischen den berufsrelevanten Interessen der Stelleninhaberin oder des Stelleninhabers und den Anforderungen der ausgeübten Tätigkeit vorliegt oder nicht (s. auch ▶ Abschn. 6.1.2.5).

EXPLOJOB zur Beschreibung von
Berufen

Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-3) – Version 3 (AIST-3)

Ebenfalls auf dem Holland-Modell aufgebaut ist der AIST-3 von Bergmann und Eder (2018). Der Umwelt-Struktur-Test entspricht konzeptuell dem EXPLOJOB. Eine weitere Gemeinsamkeit mit dem EXPLORIX besteht darin, dass über die Holland-Codes von den Testergebnissen auf die Passung zu bestimmten Berufen und Ausbildungen geschlossen werden kann. Der AIST besteht nur aus 60 Items, die in 10–15 min bearbeitet werden können. Er ist damit ökonomischer als der EXPLORIX. Trotzdem weisen die 6 Skalen ähnlich hohe interne Konsistenzen auf wie die korrespondierenden Skalen des EXPLORIX. Schließlich wurde das Verfahren an $N=4321$ Schülerinnen und Schülern sowie Studierenden im Alter von 14 bis 20 Jahren normiert. Mithilfe der alters- und geschlechtsspezifischen Normwerte ist eine wesentlich differenziertere Ergebnisrückmeldung möglich als beim EXPLORIX. Ein weiterer Vorteil besteht darin, dass die Ergebnisse des AIST-3 und des UST-3 in einen gemeinsamen Auswertungsbogen eingetragen werden können.

AIST-3 mit gleichem Messanspruch
und identischen Normen

3.3.3.8 Fragebögen zur Erfassung von Motiven

Fragebögen zur Leistungsmotivation

Für eine „breite“ Erfassung von Motiven liegt ein älteres Verfahren vor, die Deutsche Personality Research Form (PRF; Stumpf et al. 1985). Die PRF enthält 14 Skalen, von denen die meisten explizit ein Motiv erfassen (vgl. □ Tab. 3.17). Ein neueres Verfahren zur breiten Messung von Motiven, das in der Forschung viel Beachtung findet, liegt mit den Unified Motive Scales vor (Schönbrodt und Gerstenberg 2012). Von den zahlreichen Motiven, die postuliert wurden, hat das Leistungsmotiv in der Forschung und bei der Entwicklung von Messinstrumenten mit Abstand die größte Aufmerksamkeit gefunden. Zur *Leistungsmotivation* liegen im deutschen Sprachraum mehrere normierte Fragebögen vor, die in □ Tab. 3.22 aufgelistet sind. Drei Verfahren wurden primär für den schulischen Bereich entwickelt, eines für den Sportbereich und eines, das Leistungsmotivationsinventar (LMI; Schuler und Prochaska 2001), primär für die Personalpsychologie. Das LMI soll als das differenzierteste Verfahren ausführlich vorgestellt werden (s. u.). Es stand auch bei der Entwicklung des Sportbezogenen Leistungsmotivationstest (Frintrip und Schuler 2007) Pate.

Auch einige der bereits vorgestellten Persönlichkeitsinventare (► Abschn. 3.3.3) enthalten Skalen zur Leistungsmotivation. Die PRF wurde in ► Abschn. 3.3.3.7 bereits erwähnt. Im BIP kommt eine Skala „Leistungsmotivation“ vor (► Abschn. 3.3.3.4). Das FPI-R enthält mit „Leistungsorientierung“ ebenfalls eine einschlägige Skala (► Abschn. 3.3.3.3). Im NEO-PI-R trägt eine Facette im Bereich Gewissenhaftigkeit die Bezeichnung „Leistungsstreben“ (► Abschn. 3.3.3.5).

□ Tab. 3.22 Deutschsprachige Fragebögen zur Leistungsmotivation

Testname (Autorinnen und Autoren)	Skalen	Anwendungsbereich/Zielgruppe
LMI: Leistungsmotivationsinventar (Schuler und Prochaska 2001)	17 Skalen (ausführliche Beschreibung im Text)	Personalpsychologie, Sportpsychologie
SMT: Sportbezogener Leistungsmotivationstest (Frintrip und Schuler 2007)	Anspruchsniveau, Ausdauer, Dominanz, Einsatzbereitschaft, Flexibilität, Flow, Furchtlosigkeit, kompensatorische Anstrengung, Lernmotivation, Leistungsstolz, Selbstdisziplin, Selbstverantwortlichkeit, Statusstreben, Unabhängigkeit, Wettbewerbshaltung, Zielsetzung, Zuversicht	Sportpsychologie
FLM 7–13: Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13. Klasse (Petermann und Winkel 2015)	Leistungsstreben, Ausdauer und Fleiß, Anstrengungsvermeidung, Angst vor [tatsächlich „Erfolg“], Prüfungsangst	Schule; Klasse 7 bis 13
SELLMO: Skalen zur Erfassung der Lern- und Leistungsmotivation (Spinath et al. 2012)	Lernziele, Annäherungs-Leistungsziele, Vermeidungs-Leistungsziele, Tendenz zur Arbeitsvermeidung	Schule, Hochschule; Klasse 3 bis 10 sowie Studierende
FLM 3–6 R: Fragebogen zur Leistungsmotivation für Schüler der 3. bis 6. Klasse – Revision (Lohbeck und Petermann 2019)	Analog FLM 7–13 (s. o.)	Schule, Klasse 3 bis 6

Geordnet nach Publikationsjahr

■ Leistungsmotivationsinventar (LMI)

Schuler und Prochaska (2001) verfolgten bei der Entwicklung des LMI das Ziel, ein Verfahren zur „breiten“ Messung der berufsbezogenen Leistungsmotivation bereitzustellen. Sie kamen bei ihren Vorarbeiten zu der Erkenntnis, dass es sich bei der Leistungsmotivation um ein breites Konzept ohne scharfe Grenzen zu handeln scheint.

Diesen Zustand charakterisieren sie durch ihr „Zwiebelmodell“: Einige Merkmale sind zentral für die Leistungsmotivation und bilden die „Kernfacette“. Dazu gehören etwa Beharrlichkeit und Erfolgshoffnung. Andere wie Selbstständigkeit und Statusorientierung liegen weiter in der Peripherie („Randfacetten“). Noch weiter in der Peripherie liegen Merkmale wie Selbstvertrauen und Kontrollüberzeugung, die mit der Leistungsmotivation zumindest theoretisch verbunden sind. Im äußeren Randbereich schließlich sind Merkmale wie Gewissenhaftigkeit und Neurotizismus angesiedelt, die als „Hintergrundmerkmale“ einen Einfluss auf die Leistungsmotivation ausüben. Vor dem Hintergrund dieser Konzeption ist es nicht verwunderlich, dass die Autoren insgesamt 728 Items in die Vorauswahl genommen haben. Diese Zahl reduzierte sich allerdings in mehreren Auswahlsschritten. Die Autoren beschreiben ihr Vorgehen als „Wechsel von phänomenologisch-rationaler und empirischer Strategie“ (Schuler und Prochaska 2001, S. 12). Am Ende sahen sie 17 Dimensionen als angemessen für eine Beschreibung der beruflichen Leistungsmotivation an. Die Skalen des LMI wurden also induktiv entwickelt.

Gliederung Das LMI umfasst 17 Skalen mit je 10 Items (► Tab. 3.23). Die Items können jedoch auch zu einem Gesamtwert verrechnet werden. Anwenderinnen und Anwender, die nicht an einem differenzierten Persönlichkeitsbild ihrer Probandinnen bzw. Probanden interessiert sind, sondern nur „die“ Leistungsmotivation messen wollen, können die Kurzskala mit 30 Items einsetzen. Die Items wurden nach ihren Trennschärfen für den Gesamtwert aus der Langform herausgefiltert.

Breite Messung der berufsbezogenen Leistungsmotivation

Durchführung und Auswertung Das LMI kann einzeln oder in Gruppensitzungen durchgeführt werden. Die Bearbeitung der 170 Items, die auf einer Skala von 1 („trifft gar nicht zu“) bis 7 („trifft vollständig zu“) zu beantworten sind, nimmt etwa 30–40 min in Anspruch; für die Kurzform genügen etwa 10 min. Die Auswertung der Langform mit Schablonen ist außerordentlich mühsam, weil 10 Seiten des Testhefts durchzusehen, bei einigen Items Invertierungen vorzunehmen, die 170 Zahlenwerte auf einem Auswertungsbogen einzutragen und zu addieren sind.

„Zwiebelmodell“ mit induktiver Skalenentwicklung

Auch als Kurzskala mit 30 Items verfügbar

Auswertung mit Schablonen mühsam

Hohe interne Konsistenz und Retest-Reliabilität

Reliabilität Die interne Konsistenz (α) der Skalen liegt überwiegend im Bereich von .80. Solche Werte sind für eng umschriebene Merkmale bei 10 Items pro Skala völlig angemessen. Für die Kurzversion mit 30 Items beträgt $\alpha = .94$. Die Retest-Reliabilität nach einem Intervall von etwa 3 Monaten wird mit .66 (Flow) bis .82 (Furchtlosigkeit und Statusorientierung) sowie für die Kurzform mit .78 angegeben. Offenbar erfasst das LMI stabile Merkmale.

Einordnung in das Fünf-Faktoren-Modell

Validität Bei einem Persönlichkeitsfragebogen mit 17 Skalen stellt sich die Frage nach der Übereinstimmung oder auch Nichtübereinstimmung mit den großen 5 Persönlichkeitsdimensionen. Die Autoren berichten Korrelationen zu den 5 Skalen des NEO-FFI. Inzwischen liegt eine Untersuchung an 121 Sportstudenten vor, in der die Skalen des LMI zusammen mit denen des

Tab. 3.23 Skalen des LMI

Skala	Ladung ^a	Itembeispiel
Beharrlichkeit	-.57 (N), .56 (G)	Es fällt mir schwer, mich lange zu konzentrieren, ohne müde zu werden. (-)
Dominanz	-.50 (N), -.64 (V)	Wenn ich mit anderen zusammenarbeite, übernehme ich gewöhnlich die Initiative.
Engagement	.69 (G)	Ich arbeite mehr als die meisten anderen Leute, die ich kenne.
Erfolgzuversicht	-.60 (N)	Auch wenn ich vor schwierigen Aufgaben stehe, bin ich immer guten Mutes.
Flexibilität	-.72 (N)	Um etwas Neues auszuprobieren, gehe ich schon einmal ein Risiko ein.
Flow	.76 (O)	Es bereitet mir Freude, mich ganz in eine Aufgabe zu vertiefen.
Furchtlosigkeit	-.82 (N)	Wenn ich vor anderen etwas vorführen soll, habe ich Angst, mich zu blamieren. (-)
Internalität	.53 (E)	Wie weit man es beruflich bringt, ist zu einem guten Teil Glückssache. (-)
Kompensatorische Anstrengung	.51 (G)	Wenn ich fürchte, Fehler zu machen, strenge ich mich besonders an.
Leistungsstolz	.53 (E)	Für meine Selbstachtung ist es sehr wichtig, was ich geleistet habe.
Lernbereitschaft	.60 (O)	Einen großen Teil meiner Zeit verbringe ich damit, Neues zu lernen.
Schwierigkeitspräferenz	.60 (O)	Schwierige Probleme reizen mich mehr als einfache.
Selbstständigkeit	-.78 (N)	Manchmal ist es mir lieber, anderen die Entscheidung zu überlassen. (-)
Selbstkontrolle	.87 (G)	Häufig verschiebe ich Dinge auf morgen, die ich besser heute erledigen sollte. (-)
Statusorientierung	-.79 (V)	Es ist mir sehr wichtig, eine verantwortungsvolle Position zu erreichen.
Wettbewerbsorientierung	-.74 (V)	Der Wunsch, besser zu sein als andere, ist ein großer Ansporn für mich.
Zielsetzung	-.59 (V)	Im Allgemeinen bin ich stark auf die Zukunft ausgerichtet.

Quelle: Schuler und Prochaska (2001, © Hogrefe)

Items mit hohen Trennschärfen als Beispiele ausgewählt. (-)=Item wird invertiert.

^aLadungen ab .50 in der gemeinsamen Faktorisierung von NEO-PI-R und LMI (Ostendorf und Angleitner 2004, S. 153). Faktorenberechnungen: N = Neurotizismus, E = Extraversion, O = Offenheit für Erfahrungen, V = Verträglichkeit, G = Gewissenhaftigkeit.

NEO-PI-R (► Abschn. 3.3.3.5) faktorierte wurden (Ostendorf und Angleitner 2004). Die Ergebnisse sind in □ Tab. 3.23 aufgeführt. Sie belegen, dass sich die LMI-Skalen sehr gut in das Fünf-Faktoren-Modell der Persönlichkeit einordnen lassen, also offenbar verschiedene Facetten von Neurotizismus, Extraversion etc. erfassen. Lediglich 3 Skalen (Engagement, Internalität und Selbstkontrolle) laden niedriger als .50 auf einem der 5 Faktoren. Vor diesem Hintergrund interessiert die Korrelation der Gesamtskala mit den großen Persönlichkeitsfaktoren. Schuler und Prochaska (2001) berichten signifikante Korrelationen von .57 mit Gewissenhaftigkeit, -.40 mit Neurotizismus und .23 mit Extraversion. Das LMI kann also mit seinen Skalen, die auch im Gesamtwert mehr oder weniger gut repräsentiert sind, zu einem Großteil mit den Big Five in Zusammenhang gebracht werden. Es bleibt aber Erklärungsbedarf, was mit der restlichen systematischen Testvarianz gemessen wird.

Die Befunde zur Kriteriumsvalidität sind relativ unübersichtlich, weil zahlreiche Korrelationen mit unterschiedlichen Leistungsindikatoren wie Abiturnoten, Note des Ausbildungsabschlusses und Jahresgehalt berichtet werden. Viele Korrelationen sind insignifikant, und nur wenige liegen über $r=.30$. Einzelne herausragend hohe Korrelationen erscheinen nachträglich plausibel, so die zwischen „Dominanz“ und Stellung in der Hierarchie ($r=.43$) oder die zwischen „Lernbereitschaft“ und Bildungsniveau ($r=.35$).

Befunde zur Kriteriumsvalidität
unübersichtlich

Normen Es liegen (zum Teil geschlechtsspezifische) Normen für Studierende der Wirtschaftswissenschaften ($N=259$), Berufsschüler/-innen in kaufmännischen Ausbildungsberufen ($N=1008$), Schüler/-innen eines Wirtschaftsgymnasiums ($N=160$), Berufstätige in Finanzdienstleistungsunternehmen ($N=166$) und Hochleistungssportler/-innen ($N=78$) vor. Daraus konstruieren die Autoren zusätzlich eine nicht repräsentative „Gesamtnorm“ (auch getrennt für Männer und Frauen).

Heterogene Normierungsstichprobe

Bewertung Das LMI ist ein objektives und hinreichend reliables Verfahren zur Messung der Leistungsmotivation. Die Kriteriumsvalidität in den vorgeesehenen Anwendungsbereichen Personalauswahl und -entwicklung, Schul-, Studien- und Berufsberatung sowie Sportpsychologie ist noch umfassender zu belegen. „Die starke Differenzierung in 17 Dimensionen erweist sich zugleich als Chance und Problem“ (Schmidt-Atzert 2001, S. 144): Die vielen Skalen eröffnen die Chance, in bestimmten Anwendungsfeldern gute Einzelprädiktoren, beispielsweise für berufliche Leistungen, zu finden. Für eine Profilauswertung ist die große Zahl von zum Teil erheblich korrelierten Skalen hinderlich.

17 Dimensionen als Chance und Problem

3.3.4 Verfahren zur Erfassung aktueller Zustände

Die bisher vorgestellten Instrumente dienten der Erfassung von individuellen Unterschieden in habituellen Eigenschaften, also relativ breiten und zeitlich stabilen Dispositionen zu bestimmten Verhaltensweisen, die relativ konsistent in verschiedenen Situationen auftreten. Von diesen Traits sind die zeitlich viel stärker fluktuierenden States oder Zustände zu unterscheiden. Es handelt sich hierbei um temporäre Zustände von Aktiviertheit, Entspannung, guter oder schlechter Stimmung, Freude, Angst, Ärger etc. Die Aussage „Ich bin ein ängstlicher Mensch“ bezieht sich auf eine überdauernde Eigenschaft, während die Aussage „Ich habe Angst“ einen Zustand beschreibt.

Über Zeit und Situationen fluktuierende Zustände

Im Grunde genommen handelt es sich bei der Unterscheidung zwischen Eigenschaft und Zustand (State vs. Trait) nicht um eine echte Dichotomie. Die Begriffe markieren vielmehr Bereiche eines *Kontinuums*, das von „sehr stabil“ bis „sehr variabel“ reicht. Eigenschaften sind unterschiedlich stabil, und Zustände sind mehr oder weniger andauernd. Emotionale Zustände wie Überraschung oder Erschrecken können wenige Sekunden andauern oder auch – wie Angst, Traurigkeit oder Freude – über Stunden oder gar Tage fortbestehen. Stimmungen sind per Definition länger andauernde Zustände. Dennoch lässt sich auch hier eine große Spanne erkennen: So kann sich eine depressive Stimmung über wenige Stunden, aber auch über mehrere Wochen erstrecken.

Unterschiedlich stabile Merkmale

Das *momentane Befinden* kann thematisch in mindestens 3 Bereiche unterteilt werden (vgl. Hüppe et al. 2000):

3 Bereiche des Befindens

- Emotionale Befindlichkeit (z. B. Freude, gute Stimmung, Angst, Gereiztheit)
- Leistungsbezogene Befindlichkeit (z. B. Konzentriertheit, Müdigkeit)
- Körperliche Befindlichkeit (z. B. Schmerz, körperliches Unwohlsein)

Für die Erfassung von aktuellen Zuständen sind verschiedene Instrumente entwickelt worden (für eine Übersicht s. Hüppe et al. 2000; Schmidt-Atzert et al. 2014). Verfahren wie das nachfolgend beschriebene State-Trait-Angs-tinventar (STAII; Laux et al. 1981) erfassen nur ein einziges Merkmal (hier

Ein- und mehrdimensionale Verfahren

Einige Gemeinsamkeiten mit Persönlichkeitsfragebögen

Angst), andere Verfahren wie die Eigenschaftswörterliste (EWL; Janke und Debus 1978), die ebenfalls kurz vorgestellt wird, sind mehrdimensionale Verfahren, die verschiedene Aspekte des Befindens messen sollen. Meist handelt es sich dabei um Listen von Eigenschaftswörtern, Substantiven oder kurzen Erlebnisbeschreibungen („Ich bin ...“, „Ich fühle mich ...“), zu denen Stellung genommen werden muss, ob bzw. wie gut sie den momentanen Zustand beschreiben.

Verfahren zur Zustandsmessung weisen einige Gemeinsamkeiten mit Persönlichkeitsfragebögen auf: Sie verlangen, dass sich Menschen durch Ankreuzen von Items selbst beurteilen. Auch Fremdbeurteilungsvarianten sind möglich. Die Durchführungs- und Auswertungsobjektivität lässt sich durch Standardisierungsmaßnahmen gewährleisten. Wenn mehrere Items pro Merkmal vorliegen, kann die interne Konsistenz der Skala bestimmt werden.

Allerdings existieren auch deutliche Unterschiede zwischen Persönlichkeits- und Zustandsfragebögen: Während Persönlichkeitsfragebögen situationsübergreifende Merkmale erfassen sollen, wird allgemein anerkannt, dass das Befinden sehr stark durch die jeweilige Situation bedingt ist. Deshalb muss die Retest-Reliabilität von Zustandsmaßen niedriger ausfallen als die von Persönlichkeitstests. Weiterhin ist eine Normierung von Fragebögen zur Erfassung des momentanen Befindens nicht sinnvoll – es sei denn, mit einem Verfahren soll nur das Befinden in einer ganz bestimmten Situation erfasst werden.

Beispiel Angstmessung

Zustandsmessungen fallen situationsspezifisch aus

Wie kann ein Fragebogen zur Erfassung von Angst validiert werden? Dazu können wir den Fragebogen in Situationen bearbeiten lassen, die wir vorab für die Testpersonen als angstauslösend (vor einer Zahnbehandlung, einem Fallschirmsprung, einer schweren Prüfung, ...) oder als neutral klassifiziert haben. Treten große Mittelwertunterschiede zwischen beiden Situationsklassen auf, spricht das für die Konstruktvalidität des Fragebogens. Eine härtere Prüfung bestände darin, nicht neutrale, sondern z. B. Ärger oder Traurigkeit auslösende Situationen zum Vergleich heranzuziehen. Große Mittelwertunterschiede sprächen dafür, dass der Fragebogen nicht einfach nur negatives Befinden erfasst, sondern spezifisch Angst.

Wie hoch dürfen die Angstwerte der Testpersonen über die Situationen hinweg korrelieren? Wenn die Korrelationen zwischen den angstauslösenden Situationen sehr hoch ausfallen, erfasst der Fragebogen nicht nur Angst als Zustand, sondern auch Ängstlichkeit als Persönlichkeitsmerkmal. Das Gleiche gilt, wenn eine zumindest moderate Korrelation zwischen den angstauslösenden und den neutralen Situationen auftreten. Ängstliche Menschen empfinden auch in harmlosen Situationen mehr Angst als angstfreie Menschen.

Sind Normen nützlich? Ja, aber nur, wenn die Messungen in einer weitgehend standardisierten Situation erfolgen. Beispielsweise könnten Narkoseärzte daran interessiert sein, wie viel Angst ihre Patienten bzw. Patientinnen direkt vor einer schweren Operation haben (die Schwere könnte etwa über die Wahrscheinlichkeit schwerer Komplikationen definiert werden). Der Normwert zeigt an, wie stark die Angst im Vergleich zu der Angst anderer Personen in der gleichen Situation ist. Die Dosierung eines Beruhigungsmittels vor der Narkose könnte sich nach diesem Normwert richten.

3.3.4.1 State-Trait-Angstinventar (STAI)

Beim STAI von Laux et al. (1981) handelt es sich um die deutschsprachige Adaptation des von Spielberger et al. (1968) entwickelten „State-Trait Anxiety Inventory“. Das STAI ist ein international sehr bekannter Fragebogen mit einer Skala zur Messung von Angst als Zustand und einer für Angst als Eigenschaft (Ängstlichkeit). Wir stellen es hier vor, weil auch die deutsche Version immer noch gerne für Forschungszwecke verwendet wird. Fragt man die Autorinnen und Autoren, warum sie das Verfahren trotz der bekannten Unzulänglichkeiten (s. u.) einsetzen, erhält man oft die Antwort, weil es international eingesetzt wird und den Gutachterinnen und Gutachtern der Fachzeitschriften bekannt ist. Tatsächlich handelt es sich auch um ein oft zitiertes Verfahren. Für die diagnostische Praxis ist das STAI nicht brauchbar, weil die Normen für die Trait-Skala völlig veraltet sind und die State-Skala ohnehin nicht normiert ist.

Angst als Zustand und als Eigenschaft

Angstkonzept Spielberger et al. (1968) verstehen unter Angst etwas anderes, als in der Umgangssprache unter Angst verstanden wird. Nicht nur Laien ist es schwer zu vermitteln, dass Angst auch durch das Fehlen von positiven Gefühlen gemessen werden soll; das trennschärfste Item für Männer lautet in der State-Skala „Ich fühle mich wohl“ und in der Trait-Skala „Ich bin ausgelassen“. In einem grundlegenden Beitrag zum Angstkonzept hatte Spielberger (1966) stark auf Freud Bezug genommen, der Angst vor allem als einen unangenehmen Zustand charakterisiert. Dazu ist festzustellen, dass es auch andere negative Zustände wie Ärger, Ekel oder Scham gibt. Diese zeichnen sich ebenfalls durch die Abwesenheit von positiven Gefühlen aus. Mit anderen Worten: Das Fehlen positiver Gefühle ist nicht spezifisch für Angst.

Angst als negativer Zustand

Das Angstkonzept resultiert in Items wie „Ich neige dazu, alles schwer zu nehmen“ und „Unwichtige Gedanken gehen mir durch den Kopf und bedrücken mich“), die ebenso in einem Depressions- oder einem Neurotizismusfragebogen stehen könnten.

Gliederung Das STAI (Laux et al. 1981) enthält 2 Skalen mit je 20 (teilweise identischen) Items wie „Ich bin ruhig“, „Ich fühle mich wohl“ oder „Mir ist zum Weinen zumute“. Die Items sind auf einer 4-stufigen Skala von „überhaupt nicht“ bis „sehr“ (State) bzw. „fast nie“ bis „immer“ (Trait) zu beantworten. Ein Teil der Items ist in Richtung Angst, ein anderer in Richtung Angstfreiheit formuliert.

2 Skalen mit je 20 Items

Durchführung und Auswertung Bei gemeinsamer Anwendung der Skalen soll die State- stets vor der Trait-Variante bearbeitet werden. Die Instruktion für den State-Teil verlangt von den Probandinnen und Probanden, so zu antworten, „wie Sie sich jetzt, d. h. in diesem Moment fühlen ... (und) diejenige Antwort auszuwählen, die Ihren augenblicklichen Gefühlszustand am besten beschreibt“ (Spielberger et al. 1968). Die entsprechenden Passagen in der Trait-Instruktion lauten, so anzukreuzen, „wie Sie sich im Allgemeinen fühlen“. Die Bearbeitung und auch die mit einer Schablone vorgenommene Auswertung dauern nur wenige Minuten.

„Wie fühlen Sie sich jetzt?“ – bzw. „... im Allgemeinen?“

Reliabilität Sowohl für die State- als auch für die Trait-Skala liegen die Konsistenzen bei .90 und leicht darüber (in der Gesamtstichprobe und allen Normierungssubgruppen; eine Ausnahme bilden lediglich die 15- bis 29-jährigen Männer mit $\alpha = .89$). In 2 unterschiedlichen Stichproben von Studierenden fiel die Retest-Reliabilität (längstes Intervall: 73 Tage) für die State-Skala erwartungsgemäß deutlich niedriger aus als die der Trait-Skala (arithmetische Mittel $r_{tt} = .43$ bzw. .86).

Hohe interne Konsistenz

Höhere Trait-Skalenwerte bei klinischen Gruppen

Validität Die mitgeteilten Hinweise auf die Gültigkeit sind mannigfaltig. Sowohl in der State- als auch der Trait-Skala weisen Frauen etwas höhere Mittelwerte auf als Männer – ein Trend, der sich mit zunehmendem Alter verstärkt. Klinische Gruppen (Neurotiker/-innen, Alkoholiker/-innen und Schizophrene) zeigten durchschnittlich höhere Trait-Angstwerte als „unauffällig-normale“ Kontrollpersonen. Besonders hohe Mittelwerte finden sich, was ebenfalls den Erwartungen entspricht, bei Patientinnen und Patienten mit spezifischen Phobien und solchen mit generalisierten Ängsten. Darüber hinaus variierten die Mittelwerte der Trait-Skala zwischen neutralen und Klau-sursituationen nur unbedeutend, während die State-Skala erhebliche Schwankungen erkennen ließ.

State- und Trait-Skala korrelieren miteinander um $r=.60$. Trait-Angst steht mit Skalen eines ähnlichen Gültigkeitsanspruchs (s. Angstkonzept oben) in Beziehung (z. B. EPI-Neurotizismus $r=.77$, FPI-Nervosität $r=.74$, FPI-Depressivität $r=.72$, FPI-Gelassenheit $r=-.77$, FPI-Gehemmtheit $r=.67$, FPI-Emotionale Labilität $r=.70$). Bei einer Bearbeitung des STAI und der EWL von Janke und Debus (1978; ► Abschn. 3.3.4.2) durch eine Stichprobe von 136 Testpersonen fiel die Korrelationen der State-Skala mit der EWL-Skala für Ängstlichkeit ($r=.62$) ähnlich hoch aus wie die für Depressivität (.68), Ärger (.66), Erregtheit (.69) und Selbstsicherheit (-.65).

Normierung Da Alters- und Geschlechtseffekte bestanden, wurden getrennte Normen (T-Werte) für je 3 Altersgruppen von Männern und Frauen berechnet. Die Normierungsstichprobe bestand aus insgesamt $N=2385$ repräsentativ ausgewählten Personen. Für die State-Skala liegen keine Normen vor.

Ökonomisch, messgenau und unspezifisch

Bewertung Sowohl zur Messung von habitueller Ängstlichkeit als auch von Angst als Zustand liegen mehrere andere Verfahren vor. Das STAI zeichnet sich dadurch aus, dass es beide Aspekte isoliert, also nicht eingebettet in ein mehrdimensionales Verfahren, erfasst. Die beiden Skalen des STAI sind sehr ökonomisch und messgenau. Sie basieren auf einer international bekannten Angsttheorie, die Angst als relativ unspezifischen negativen Zustand konzipiert.

3.3.4.2 Die Eigenschaftswörterliste (EWL)

Mehrdimensionales Befindensmaß

Bei der EWL (Janke und Debus 1978; derweil nicht mehr im Verlagsprogramm gelistet) handelt es sich um ein mehrdimensionales Verfahren zur quantitativen Beschreibung des aktuellen Befindens. Wir führen sie hier auf, weil sie zumindest in den Testbibliotheken der psychologischen Institute oft vertreten sein wird und wohl das im deutschen Sprachraum bekannteste mehrdimensionale Instrument zur Erfassung des Befindens ist. Die Autoren der EWL sehen das Verfahren als „ein Forschungsinstrument zur Erfassung der Wirkung [von] Interventionen bei Gruppen“ (Janke und Debus 1978, S. 9). Bei diesen Interventionen ist an die Einflüsse von Umweltbedingungen (Lärm, Temperatur), Therapien, Psychopharmaka und Programmen mit motivational-emotionalen Auswirkungen zu denken. Als diagnostisches Instrument für individuelle Merkmalsausprägungen sei die EWL nur bei wiederholter Anwendung geeignet (Ermittlung der „durchschnittlichen Befindlichkeit“). Aus der EWL wurden verschiedene Kurzformen abgeleitet, so eine 40-Item-Version für Kinder mit 10 Skalen (Janke und Janke 2005).

6 Befindensbereiche, 15 Skalen

Gliederung Die EWL liegt in 2 Formen vor: Die „Normalversion“ (EWL-N) enthält 161 Items, eine kürzere (EWL-K) mit einer Teilmenge der Items aus der längeren Form beinhaltet 123 Items. Nachfolgend wird nur die EWL-N vorgestellt. Sie erfasst 6 Befindlichkeitsbereiche, von denen jeder durch

Diagnostische Verfahren

mehrere Skalen abgedeckt wird. Die 15 Skalen und die Bereiche, zu denen sie sich gruppieren lassen, sind mit Angaben zu den Items und zur Reliabilität in **Tab. 3.24** aufgeführt.

Durchführung und Auswertung Die Beantwortung der Items erfolgt durch Ankreuzen von „trifft zu“ oder „trifft nicht zu“, was anhand einiger Beispiele im Aufgabenheft eingeübt wird. Dafür brauchen die Probandinnen und Probanden zwischen 10 und 30 min. Für die Auswertung stehen Schablonen zur Verfügung.

Reliabilität Wie aus **Tab. 3.24** ersichtlich wird, variieren die α -Koeffizienten beträchtlich. Bei relativ großer Itemzahl (um 20 pro Skala) wird der Wert von .90 überschritten. Insgesamt liegen die Konsistenzkoeffizienten im gleichen Bereich wie die entsprechenden Kennwerte von Persönlichkeitsfragebögen. Wider Erwarten hoch sind die Retest-Koeffizienten. Der Mittelwert aller Skalen aus 2 Stichproben liegt bei ca. $r_{tt} = .78$. Bei Alkoholkranken und bei in psychiatrischer Behandlung befindlichen Probanden/Probandinnen sowie bei längeren Retest-Intervallen fallen die Stabilitäten allerdings niedriger aus.

Validität Neben korrelativen Studien unter Einbezug anderer State- und auch Trait-Maße, faktorenanalytischen Prüfungen sowie dem Vergleich von Selbst- mit Fremdeinschätzungen (Letztere durch den Arzt/die Ärztin; Koeffizienten in mittlerer Höhe bis zu $r_{tc} = .60$ bei Deprimiertheit, Ängstlichkeit, gehobene Stimmung) steht die Änderungssensitivität im Vordergrund. Diese ist in verschiedenster Weise belegt worden. Unter dem Einfluss von Lärm, Psychopharmaka, Androhung elektrischer Schläge, Teilnahme an Therapieverfahren etc. treten sehr unterschiedliche Effekte auf, die sich auf den einzelnen Subskalen abbilden. In gesonderten Experimenten stellten sich zudem (allerdings

Ankreuzen, ob Adjektiv zutrifft oder nicht

Retest-Reliabilität zum Teil hoch

Änderungssensitiv

Tab. 3.24 Skalen der EWL (Janke und Debus 1978, © Hogrefe)

Bereich	Subskala	Beispielitems ^a	Itemzahl	Reliabilität ^b
Leistungsbezogene Aktivität	A Aktiviertheit B Konzentriertheit	Energisch Aufmerksam	19 6	.93 .78
Allgemeine Desaktivität	C Desaktiviertheit D Müdigkeit E Benommenheit	Energielos Lahm Dösig	16 7 9	.91 .87 .76
Extraversion/Introversion	F Extravertiertheit G Introvertiertheit	Gesprächig Ungesellig	9 8	.81 .86
Allgemeines Wohlbefinden	H Selbstsicherheit I Gehobene Stimmung	Unbekümmert Heiter	8 16	.81 .94
Emotionale Gereiztheit	J Erregtheit K Empfindlichkeit L Ärger	Aufgeregter Kribbelig Gereizt	15 4 7	.88 .75 .78
Angst	M Ängstlichkeit N Deprimiertheit O Verträumtheit	Beklommen Traurig Gedankenverlorene	7 20 10	.77 .93 .81

^aDie Adjektivform der Skalenbezeichnung, bei „Konzentriertheit“ also „konzentriert“, gehört immer zu den Items. Zur Vermeidung von Redundanzen wurden diese Items hier nicht aufgeführt

^bCronbachs α aus der Analysenstichprobe II: $N=937$ unausgelesene männliche und weibliche Personen aller Bildungsstufen im Alter von 18 bis 65 Jahren

niedrige) Korrelationen zwischen Aktivierungsvariablen der EWL und physiologischen Variablen wie der Herzfrequenz und der elektrischen Hautleitfähigkeit heraus.

3

Keine Normen

Normierung Auf die Erstellung von Normen wurde verzichtet, da diese bei Verfahren zur Erfassung aktueller Zustände in sehr unterschiedlichen Situationen nicht sinnvoll sind.

Primär ein Forschungsinstrument

Bewertung Die EWL stellt in erster Linie ein Forschungsinstrument dar. Für Fragestellungen, in denen die Auswirkungen systematischer Beeinflussungen auf die aktuelle Befindlichkeit von Interesse sind, kann mit der EWL ein breites Spektrum an aktuellen Zuständen gemessen werden.

Weiterführende Literatur

Persönlichkeitsfragebögen sind auch in dem Kompendium *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (Brähler et al. 2002) verzeichnet. Ausgewählte Persönlichkeitstestsysteme sowie Fragebögen für einzelne Persönlichkeitsmerkmale werden in einem Übersichtsbeitrag von Borkenau et al. (2011) vorgestellt. Verfahren, die im Personalbereich Verwendung finden, werden von Hossiep und Mühlhaus (2005) systematisch dargestellt. Über Fragebögen, die für den klinischen Bereich relevant sind, informieren Hoyer et al. (2009).

?

Übungsfragen

— Abschn. 3.3:

- Welche Vor- und Nachteile haben Persönlichkeitsfragebögen?
- Wie kann Täuschung in Persönlichkeitsfragebögen eventuell verhindern und wie kontrollieren (erkennen)?
- Wie wurden die Items des Minnesota Multiphasic Personality Inventory-2 (MMPI-2) ausgewählt?
- Woher stammen die Items des MMPI-2-RF?
- Wie viele Skalen hat das MMPI-2-RF und in welche Kategorien lassen sie sich einteilen (möglichst jeweils auch ein Beispiel nennen)?
- Welche Funktion hat die Skala „Entmutigung“ für die Konstruktion der Skalen des MMPI-2-RF?
- Nennen Sie einige Validitätsskalen im MMPI-2-RF!
- Nach welchem Prinzip wurden die Skalen des Freiburger Persönlichkeitssinventars (FPI-R) zusammengestellt?
- Welche Erkenntnisse ergeben sich aus einer Simultanfaktorisierung mehrerer Testsysteme (neben dem FPI-R u. a. das NEO-Fünf-Faktoren-Inventar, NEO-FFI) für die Skalen des FPI-R?
- Wie ist das NEO-PI-R strukturell aufgebaut?
- Für welche Anwendungen wurde das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) entwickelt, und welche Art von Merkmalen soll es erfassen?
- Welche Gemeinsamkeiten haben Fragebögen zur Zustandsmessung (Befinden) mit denen zur Persönlichkeit, und worin unterscheiden sie sich?
- Wie ist das State-Trait-Angstinventar (STAII) aufgebaut, und was soll es messen?
- Wie ist die Eigenschaftswörterliste (EWL) aufgebaut, und was soll damit gemessen werden? Nennen Sie auch Skalenbeispiele!
- Für welche Zwecke wurde der EXPLORIX entwickelt, und welche Theorie liegt dem Verfahren zugrunde?
- Welche Theorie bzw. welches Modell liegt dem Leistungsmotivationsinventar (LMI) zugrunde? Beschreiben Sie den Aufbau des LMI!

3.4 Objektive Persönlichkeitstests

Die in ▶ Abschn. 3.3 vorgestellten Verfahren funktionieren nach dem Prinzip, dass ein Item vorgegeben wird und die Testperson darauf antworten soll. Die Antworten der befragten Personen sind auch immer von deren Selbst-einsicht und Auskunfts-bereitschaft abhängig. Sofern diese jedoch nicht gegeben sind, muss eine andere Lösung gesucht werden. Es existieren hierzu 3 Lösungsansätze: die Verwendung projektiver Tests (▶ Abschn. 3.5), der Einsatz sog. „objektiver Persönlichkeitstests“ und die Verwendung künstlicher Intelligenz zur Vorhersage von Persönlichkeitsmerkmalen. Letzteren Ansatz stellen wir hier im Kontext objektiver Tests vor, wenngleich es sich hierbei nicht um psychometrische Tests im eigentlichen Sinne handeln muss.

Persönlichkeitsdiagnostik
ohne Selbsteinsicht und
Auskunfts-bereitschaft

Objektive Persönlichkeitstests wirken zumeist wie ein Leistungstest: Die Testpersonen sollen auf bestimmte Aufgaben reagieren und haben dabei nur einen kleinen Verhaltensspielraum, können also nur zwischen wenigen Optionen wählen. Dadurch werden die Durchführung und die Auswertung standarisert. Letztere besteht beispielsweise darin, dass ein Kennwert für die Häufigkeit risikoreicher Entscheidungen oder für die Ausdauer bei einer gleichförmigen Aufgabe berechnet wird. Die Testperson ist sich dabei in der Regel nicht über die Bedeutung ihres Verhaltens im Klaren, sodass es ihr schwerfallen würde, sich sozial erwünscht darzustellen.

Wirkt zumeist wie Leistungstest

Definition

Die folgende Definition stammt von Schmidt (1975, S. 19): „**Objektive Tests (T-Daten)** zur Messung der Persönlichkeit und Motivation sind Verfahren, die unmittelbar das Verhalten eines Individuums in einer standardisierten Situation erfassen, ohne dass dieses sich in der Regel selbst beurteilen muss. Die Verfahren sollen für den Probanden keine mit der Messintention übereinstimmende Augenscheininvalidität haben. Das kann durch die Aufgabenauswahl oder bestimmte Auswertungsmethoden erreicht werden. Um als Test zu gelten, müssen auch die Objektiven Verfahren den üblichen Gütekriterien psychologischer Tests genügen.“

Objektive Persönlichkeitstests sind Tests, die Persönlichkeitsmerkmale aus dem Verhalten von Testpersonen erschließen. Verhalten wird unter hoch standardisierten Bedingungen „provoziert“ und durch eine systematische Verhaltensbeobachtung (s. dazu ▶ Abschn. 3.6.2) erfasst. Heute handelt es sich in der Regel um Computertests, in denen Aufgaben vorgegeben und die Antworten/Reaktionen vom Computer registriert und ausgewertet werden. Das Prinzip der objektiven Persönlichkeitstests wird durch den Vorschlag von Kubinger (2006), diese Verfahren künftig „experimentalpsychologische Verhaltensdiagnostik“ zu nennen, verdeutlicht.

Experimentalpsychologische
Verhaltensdiagnostik

Objektive Persönlichkeitstests haben eine lange Forschungstradition. Cattell (s. Pawlik 2006a, b) hat schon vor langer Zeit mit seiner Arbeitsgruppe sehr viele objektive Persönlichkeitstests entwickelt. Cattell und Warburton (1967) berichteten von über 400 solcher Tests. Die Faktorenstruktur – Cattell hatte 21 Faktoren angenommen – ließ sich allerdings nicht durch andere Forscher replizieren (Schmidt 2006). In Deutschland haben Häcker et al. (1975)

Objektive Testbatterie OA-TB 75

mit der Objektiven Testbatterie OA-TB 75 eine umfangreiche Serie solcher Tests vorgelegt, die sich an Arbeiten von Cattell orientierte. Die Autoren sahen die von ihnen vorgelegten Versionen allerdings nicht als „Endprodukt einer im herkömmlichen Sinne verstandenen Testkonstruktion [...], sondern als experimentelle Version, auf deren Basis eine standardisierte Testbatterie erstellt wird“ (Häcker et al. 1975, S. 9).

Im Testheft der Objektiven Testbatterie OA-TB 75 sind 50 Subtests, die zum Teil Leistungscharakter aufweisen, zusammengestellt. Ihre Auswahl erfolgte u. a. unter Ökonomie- und Kulturspezifitätsgesichtspunkten und teils danach, inwieweit sich ein Faktor in früheren Untersuchungen im angloamerikanischen Raum als replizierbar erwiesen hatte. Die Tests sollen insgesamt 21 Faktoren treffen, u. a. Stärke vs. mangelnde Selbstbehauptung, skeptische Zurückhaltung vs. Engagiertheit und ganzheitliches Verständnis vs. Willensschwäche. Darüber hinaus finden sich auch Faktoren, die bekanntere Dimensionen betreffen, z. B. Extraversion/Introversion, Angst, Realismus und Impulsivität. Beispiele für einige Items sind im Folgenden aufgeführt.

Beispielaufgaben aus der OA-TB 75

T 197 Was würden Sie lieber machen? [T steht hier für Testaufgabe].

- Mit anderen Bekannten einen Wettkampf machen
- Alleine laufen

Ausgewertet wird, ob Wettbewerbssituationen aufgesucht oder gemieden werden.

T 45 Beurteilung der Längen von Linien.

Jeweils 2 waagerechte oder etwas schräg nebeneinanderstehende Linien werden vorgegeben. Die Versuchsperson muss ankreuzen, ob:

- die linke Linie länger ist als die rechte,
- beide Linien gleich lang sind,
- die rechte Linie länger als die linke ist.

Abhängige Variable ist die Zahl der in der verfügbaren Zeit bearbeiteten Aufgaben.

T 43 Geschichten.

Die Versuchspersonen müssen die 2 Sätze „Als der Fahrer die Herrschaft über das Auto verlor...“, „Es war Herbst, und die Blätter fielen von den Bäumen...“, zu möglichst langen Geschichten fortsetzen. Dafür steht jeweils 1 min zur Verfügung.

Gemessen wird die Zahl der geschriebenen Wörter.

Die OA-TB 75 findet in der diagnostischen Praxis keine Verwendung, da sie nicht normiert ist. Sie hat aus heutiger Sicht Modellcharakter, indem sie Anregungen dafür gibt, wie man objektive Persönlichkeitstests grundsätzlich entwickeln kann. Mit dem Einzug von leistungsfähigen Computern in die Diagnostik haben sich völlig neue Möglichkeiten für die Konstruktion von objektiven Persönlichkeitstests ergeben, die auch genutzt wurden (Ortner et al. 2006). Einige ausgewählte Verfahren, die sich dieser Möglichkeiten bedienen, werden im Folgenden vorgestellt.

3.4.1 Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität

Die Testautoren, Kubinger und Ebenhöh (1996), orientierten sich bei der Konstruktion der kurzen Testbatterie zur Erfassung von Arbeitshaltungen eng an der OA-TB 75. Dazu wurden 3 Skalen in modifizierter Form auf dem Computer implementiert. Aus der Bearbeitung dieser 3 Skalen durch die Testpersonen leiten sich zahlreiche Kennwerte ab, die „Arbeitstugenden“ in Form von kognitiven Stilen und motivationalen Konzepten erfassen sollen. Diese Kennwerte basieren auf einer Faktorisierung der Daten von (nur) 60 Testpersonen, wo sie jeweils einen Faktor markierten. Für 5 der Kennwerte wird zudem „Konstruktvalidität im Hinblick auf die mittlerweile berühmten Big Five“ beansprucht (Kubinger und Ebenhöh 1996, S. 16). Alle Tests haben Leistungscharakter, die berechneten Kennwerte sollen jedoch Persönlichkeitseigenschaften erfassen.

„Arbeitshaltungen“ aus 3 Skalen der OA-TB 75 entstanden

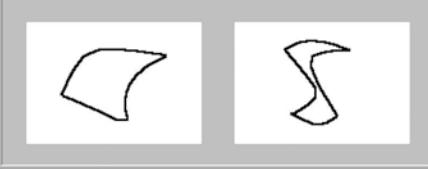
Gliederung Die Arbeitshaltungen bestehen aus 3 Untertests. Im Untertest „Figuren vergleichen“ soll die Testperson wiederholt beurteilen, welche von 2 unregelmäßig verlaufenden Linien eine größere Fläche umschließt (Abb. 3.21). In 30 s werden dabei maximal 20 Items präsentiert. Beim Untertest „Symbole kodieren“ erfolgt in 5 Durchgängen à 50 s fortlaufend die Darbietung von je einer von 4 abstrakten Schwarz-weiß-Figuren. Jede dieser Figuren ist einer anderen ebenfalls abstrakten, aber farbigen Figur fest zugeordnet. Diese farbige Figur muss mit der Maus angeklickt werden. Die Zuordnung von farbigen zu Schwarz-weiß-Figuren ist dabei ständig in der oberen Hälfte des Bildschirms als Legende dargestellt. Nach jedem Durchgang erhält die Testperson Rückmeldung über ihre Leistung; anknüpfend daran soll sie einschätzen, wie viele Symbole sie im nächsten Durchgang richtig bearbeiten wird. Außer bei der ersten Rückmeldung wird zusätzlich mitgeteilt,

3 Aufgabengruppen

Anleitung...

Ihre Aufgabe: Entscheiden Sie, welche der beiden Flächen den größeren Flächeninhalt besitzt.

Wählen Sie nun die größere der beiden Figuren aus.



 Zurück
Weiter 

Abb. 3.21 Aufgabe aus dem Test „Arbeitshaltungen“. (Mit freundlicher Genehmigung der SCHUHFRIED GmbH)

Bearbeitungsdauer hängt von Testperson ab

dass andere Personen durchschnittlich um 10 % besser sind. Beim Untertest „Figuren unterscheiden“ gilt es, aus jeweils 4 geometrischen Figuren die unpassende herauszufinden. Die Testperson erhält wiederholt und in Abhängigkeit von Bearbeitungsfehlern Rückmeldung.

Verschiedene Kennwerte

Durchführung und Auswertung Alle Instruktionen werden am Bildschirm präsentiert, sodass der Aufwand für die Testleiterin bzw. den Testleiter minimal ist. Die Eingabe erfolgt grundsätzlich mit der Maus, wobei entweder die fraglichen Figuren und Symbole oder beschriftete Buttons anzuklicken sind. Da die Bearbeitungsdauer beim letzten Untertest lediglich von der Ausdauer der Testperson abhängt, variiert die für die Durchführung benötigte Zeit zwischen 20 und 45 min.

Die Auswertung erfolgt automatisch. Es werden folgende Kennwerte ermittelt:

- Figuren vergleichen:
 - Exaktheit (Anteil richtiger Antworten)
 - Entschlussfreudigkeit (Anzahl der Antworten)
 - Impulsivität vs. Reflexivität ($\text{Fehler} \times 10.000 - \text{Richtig} \times 100 + \text{Weiß nicht} \times 1$)
- Symbole kodieren:
 - Frustrationstoleranz: (Differenz zwischen 5. und 2. Prognose) / (2. Prognose)
 - Anspruchsniveau: (1. Prognoseleistung im 2. Durchgang) / (Leistung im 2. Durchgang)
 - Leistungsniveau: richtige Kodierungen im 2. Durchgang
 - Zeitpunkt der Leistungsspitze: bester Durchgang
 - Zieldiskrepanz: mittlere Abweichung zwischen Prognose und darauffolgender Leistung
- Figuren unterscheiden:
 - Leistungsmotivation: Anzahl der bearbeiteten Items

Vorläufige Normen

Normierung Derzeit liegen nur vorläufige Normwerte von $N=314$ Personen vor. Die Normdaten stammen aus mehreren Untersuchungen und wurden vor allem an Psychologiestudierenden erhoben.

Keine Reliabilitätsbestimmung

Reliabilität Die Testautoren führen für jeden der berechneten Kennwerte bestimmte Gründe an, die einer empirischen Überprüfung der Reliabilität entgegenstehen (z. B. verhindern Ein-Punkt-Messungen die Bestimmung der internen Konsistenz; massive Übungs- und Gedächtniseffekte beeinträchtigen Retest-Stabilitäten; die relativen Differenzwerte zeigen die für Veränderungsmessungen generellen Reliabilitätsmängel usw.).

Validität unklar

Validität Die Testautoren beanspruchen für die erhobenen Kennwerte nicht nur Inhaltsvalidität bei der Erfassung von kognitiven Stilen und motivationalen Variablen, sondern auch Konstruktvalidität hinsichtlich der Big Five (E=Impulsivität vs. Reflexivität, N=Anspruchsniveau, O=Zieldiskrepanz, V=Frustrationstoleranz, G=Leistungsmotivation; vgl. ▶ Abschn. 3.3.3.5).

Angaben über die Korrelationen zwischen den einzelnen Kennwerten sowie mit den Big Five fehlen. Hinsichtlich der Kriteriumsvalidität verweisen die Testautoren auf 2 eigene Studien, in denen erfolgreiche von nicht erfolgreichen Mitarbeitern und Mitarbeiterinnen signifikant unterschieden werden konnten. Nur für einige der Kennwerte wurden Zusammenhänge mit dem Erfolgskriterium gefunden, deren Richtung je nach Validierungsstichprobe und -kriterium variierte, sodass Post-hoc-Interpretationen notwendig wurden.

Bewertung Die „Arbeitshaltungen“ stellen zumindest im deutschen Sprachraum den ersten computerbasierten objektiven Persönlichkeitstest dar. Sie sind aber auch ein gutes Beispiel dafür, wie die Erfüllung der herkömmlichen psychometrischen Gütekriterien doch deutlich hinter den ebenso attraktiven wie faszinierenden Darbietungs- und Auswertungsmöglichkeiten computerbasierter Tests zurückbleibt. Bei einer Revision sollten weitere Erkenntnisse zur Reliabilität und Validität generiert werden.

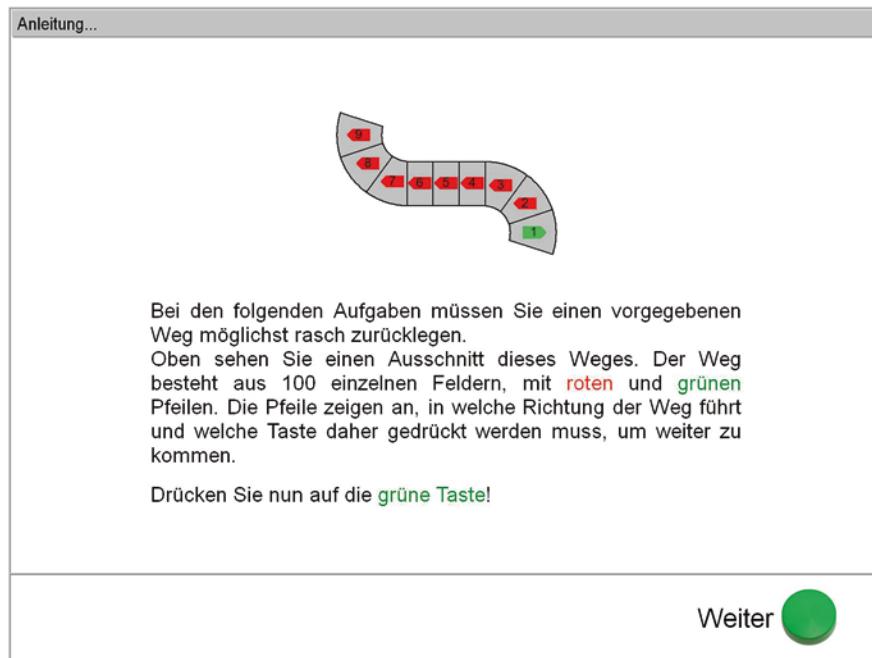
Erster computerbasiert
deutschsprachiger objektiver
Persönlichkeitstest

3.4.2 Objektiver Leistungsmotivations-Test (OLMT)

Der OLMT von Schmidt-Atzert (2007) soll, wie der Subtest „Figuren unterscheiden“ (► Abschn. 3.4.1), die Leistungsmotivation über eine kognitiv wenig anspruchsvolle Aufgabe messen. Die Aufgabe der Testpersonen besteht darin, durch Drücken von 2 Tasten eine „Straße“ auf dem Bildschirm abzufahren (► Abb. 3.22).

Die 100 Felder lange Straße führt abwechselnd nach rechts und links. Drückt die Testperson die richtige Taste (im Original: rot für links und grün für rechts), legt sie immer ein weiteres Feld zurück. Betätigt sie die falsche Taste, erfolgt eine optische und akustische Warnung, ohne dass ein Feld vorgerückt wird. Die Bearbeitungszeit ist vorgegeben; sie beträgt für jeden der insgesamt 30 Durchgänge genau 10 s. Erfasst wird die Schnelligkeit bzw. die Anzahl der zurückgelegten Felder, die exakt der Schnelligkeit des Tastendrückens entspricht, wenn keine Fehler gemacht werden. Die Ausdauer spielt insofern eine Rolle, als $30 \times$ die gleiche Aufgabe zu bewältigen ist, was bei maximaler Testleistung (Zurücklegen aller 100 Felder) immerhin 3000 Tastenanschlägen entspricht. Je mehr sich eine Testperson anstrengt, desto mehr Felder wird sie zurücklegen.

Kognitiv wenig anspruchsvolle
Aufgabe



► Abb. 3.22 Aufgabe im OLMT. (Aus Schmidt-Atzert 2007, mit freundlicher Genehmigung der SCHUHFRIED GmbH)

Forschungsergebnisse zur Leistungsmotivationsforschung umgesetzt

3

Aufgabe als Anreiz

Bei der Testkonstruktion fanden relevante Befunde der Leistungsmotivationsforschung Berücksichtigung. Leistungsmotiviertes Verhalten setzt voraus, dass die Testperson eine klare Zielsetzung hat (wird durch die Aufgabenstellung realisiert, möglichst viele Felder in 10 s zurückzulegen), alleine für das Ergebnis verantwortlich ist (wie viele Felder sie zurücklegt, liegt alleine an ihr) und Feedback über die erzielte Leistung erhält. Feedback erfolgt sowohl kontinuierlich (zurückgelegte Felder färben sich grau) als auch jeweils schriftlich am Ende eines Durchgangs (z. B. „Sie haben 67 Felder zurückgelegt“). Da die individuelle Leistungsmotivation durch Ziele, die man sich selbst setzt, sowie durch Konkurrenz angeregt werden kann, wurden 2 Subtests konzipiert, die erfassen sollen, wie stark die Person auf diese Anreizbedingungen anspricht.

Eigene Ziele setzen

Gliederung Der OLMT besteht aus 3 Subtests. Gemessen wird immer, wie viele Felder zurückgelegt werden. Der 1. Subtest „aufgabenbezogene Anstrengung“ erfasst die Leistung ohne andere Anreize als die Testaufgabe selbst. Allerdings wird nur der 8. bis 10. Durchgang ausgewertet, weil die Leistung normalerweise, vermutlich durch Übung bedingt, in den ersten Durchgängen ansteigt.

Im 2. Subtest „Motivation durch Ziele“ wird eine zusätzliche Anreizbedingung eingeführt: Die Testperson soll vor jedem Durchgang angeben, wie viele Felder sie nun schaffen will. Auf dem Bildschirm erscheint die Angabe, wie viele Felder sie zuletzt zurückgelegt hat. Sie soll über die Tastatur eingeben, wie viele Felder sie jetzt erreichen will. Erfasst wird hier nicht nur die Leistungsveränderung gegenüber Subtest 1, sondern auch das Anspruchsniveau, das aus der Abweichung der Ziele von den tatsächlichen Leistungen errechnet wird. Mit diesem Subtest soll erfassst werden, wie stark die Testperson dadurch motiviert wird, dass sie sich selbst Ziele für ihre Arbeitsergebnisse setzt, und wie hoch ihr Anspruchsniveau ist.

Im 3. Subtest „Motivation durch Konkurrenz“ tritt die Testperson gegen einen Konkurrenten an, der eine Straße parallel zu ihrer eigenen durchläuft. Der Konkurrent wurde angeblich vom Computer passend zur Testperson ausgewählt. Tatsächlich richtet sich dessen Leistung nach der der Testperson; allerdings ist er immer 10 % schneller als die Testperson in den letzten 3 Durchgängen. Damit soll gemessen werden, wie stark die Testperson sich davon motivieren lässt, dass sie ihre Leistungen mit denen eines Konkurrenten vergleicht. In □ Tab. 3.25 sind die Subtests mit ihren Kennwerten sowie Angaben zur Reliabilität aufgeführt.

Neben den Hauptkennwerten (□ Tab. 3.25) werden für jeden Subtest Fehlerquoten und die intraindividuelle Streuung der Leistungen berechnet, und der Leistungsverlauf über die 3 Subtests wird grafisch dargestellt. Die Hilfskennwerte dienen lediglich der Beurteilung der Hauptkennwerte. Beispielsweise kann eine große Leistungsschwankung auf Störungen oder Probleme während der Durchführung hinweisen. Eine hohe Fehlerrate spricht für große Anstrengung (die Fehlerrate korreliert um .30 mit der Anzahl zurückgelegter Felder).

Haupt- und Hilfskennwerte

Hohe interne Konsistenz

Objektivität und Reliabilität Die Durchführungsobjektivität ist durch die standardisierte Instruktion und Testvorgabe gegeben. Die Auswertungsobjektivität wird durch die automatische Berechnung der Testergebnisse gewährleistet. Die Interpretationsobjektivität ist gegeben, weil es sich um ein normiertes Testverfahren handelt und das Manual präzise Hinweise zur Interpretation der Kennwerte enthält. Die interne Konsistenz (□ Tab. 3.25) liegt in einem Bereich, der eher für Leistungstests typisch ist.

■ Tab. 3.25 Subtests und Hauptkennwerte des OLMT

Nummer	Subtest und Kennwert	Motivationaler Anreiz	Operationalisierung	α
1	Aufgabenbezogene Anstrengung	Aufgabe selbst	Anzahl zurückgelegter Felder in Subtest 1, Durchgänge 8 bis 10	.95 bis .96
2	a. Motivation durch Ziele	Eigenes Ziel	Anzahl zurückgelegter Felder in Subtest 2 im Vergleich zu Subtest 1	.88 bis .92
	b. Anspruchsniveau	Eigenes Ziel	Zielsetzung im Vergleich zur Anzahl tatsächlich zurückgelegter Felder in Subtest 2	.83 bis .94
3	Motivation durch Konkurrenz	Leistung des Gegners	Anzahl zurückgelegter Felder in Subtest 3 im Vergleich zu Subtest 1	.88 bis .92

Jeder Subtest besteht aus 10 Durchgängen von je 10 s. Angaben zu α für die 3 Altersgruppen der Normierungsstichprobe ($N=170, 72$ und 124).

(Schmidt-Atzert 2007, mit freundlicher Genehmigung der SCHUHFRIED GmbH)

Validität Positive Zusammenhänge in der Größenordnung um $r=.30$ fanden sich in mehreren Untersuchungen zwischen den Kennwerten des OLMT und Leistungen in verschiedenen kognitiven Leistungstests und Abiturnoten. Korrelationen in dieser Höhe bestehen nicht nur mit Speedtests, sondern auch mit einem Intelligenztest (SPM plus), der ohne Zeitbegrenzung bearbeitet wird. Meist erwies sich die „aufgabenbezogene Anstrengung“ als der Kennwert mit der höchsten Korrelation zu relevanten anderen Tests. In einer prospektiven Studie korrelierte dieser Kennwert zu $-.24$ mit der durchschnittlichen Vordiplomnote von Psychologiestudierenden und wies sogar eine inkrementelle Prognosegüte über die Abiturnote hinweg auf (Schmidt-Atzert 2005). Mit Fragebogenskalen zur Leistungsmotivation aus dem BIP (► Abschn. 3.3.3.4) konnte nur ein schwacher Zusammenhang festgestellt werden (Anspruchsniveau und Skala „Leistungsmotivation“ bzw. „Wettbewerbsorientierung“: $r=.29$ bzw. $.21$).

In einer experimentellen Untersuchung zur Verfälschbarkeit des Tests sollten die Testpersonen ihr Testergebnis nach oben bzw. nach unten verfälschen (Ziegler et al. 2007). Eine Verfälschung nach oben gelang nicht, wie der Vergleich mit einer neutralen Kontrollgruppe ergab. Die Testpersonen konnten eine niedrige Leistung vortäuschen, indem sie langsamer arbeiteten. Allerdings waren die Ergebnisse überwiegend so schlecht, dass eine Verfälschungen nach unten erkannt werden konnte.

Normierung Der OLMT wurde an einer nach Alter und Bildungsniveau repräsentativen Stichprobe normiert. Wegen der Altersabhängigkeit der Kennwerte wurden 3 Altersgruppen gebildet: 18–49;11 ($N=170$), 50–64;11 ($N=72$) und 65–80 Jahre ($N=124$); für die Seniorinnen- bzw. Seniorenstichprobe besteht kein Anspruch auf Repräsentativität.

Moderate Korrelation mit Leistungsmaßen

Verfälschung nach oben gelingt nicht

Repräsentative Normstichprobe

„Interessanter und
entwicklungsfähiger Versuch“

Bewertung In einer Testrezension schreibt Brandstätter (2005, S. 136 f.):

- » Es handelt sich um einen interessanten Versuch objektiver Messung der Leistungsmotivation, der eine nützliche Ergänzung der bisherigen Zugänge über projektive Verfahren oder Fragebogen verspricht.[...] Die Möglichkeiten der Computerpräsentation werden voll genutzt und machen die Testteilnahme interessant und anregend.[...] Der im Wiener Testsystem verfügbare OLMT ist als interessanter und Entwicklungsfähiger Versuch der objektiven Messung von Leistungsmotivation zu werten, der allerdings noch weitere Untersuchungen zur Konstruktvalidität [...] einschließen sollte.

Und in einer weiteren Rezension wird resümiert:

- » Der OLMT ist ein gutes Beispiel dafür, wie aktuelle theoretische Erkenntnisse unter Nutzung der Möglichkeiten des Computers optimal in ein innovatives Verfahren umgesetzt werden können: Leistungsorientiert und auf Grundlage von bestehenden theoretischen Erkenntnissen werden die Testwerte in computerisierter Simulation unverschiedenen Motivationsbedingungen für leistungsmotiviertes Verhalten ermittelt (Ortner und Sokolowski 2008, S. 306).

Bemängelt wird, dass Validierungsstudien aus Echtsituationen fehlen.

Nur wenige objektive
Persönlichkeitstests bei Testverlagen
erschienen

Anmerkung In der 5. Auflage dieses Lehrbuchs war an dieser Stelle noch zu lesen: „Das Interesse an objektiven Persönlichkeitstests ist wieder erwacht, nachdem in den Jahren nach Veröffentlichung der OATB-75 (s. o.) lange Zeit kein neues Verfahren mehr vorgestellt wurde“. Heute müssen wir feststellen, dass seitdem bei den Testverlagen im deutschen Sprachraum kein neuer objektiver Persönlichkeitstest mehr erschienen ist und von den damals aufgeführten Tests einige nicht mehr verfügbar sind. □ Tab. 3.26 gibt einen Überblick über computerbasierte objektive Persönlichkeitstests, die derzeit (Stand: Juni 2020) bei Testverlagen im deutschsprachigen Raum erhältlich sind; es handelt sich ausschließlich um computergestützte Verfahren.

□ Tab. 3.26 Übersicht über computerbasierte objektive Persönlichkeitstests

Testname	Quelle	Messgegenstand
Arbeitshaltungen – Kurze Testbatterie	Kubinger und Ebenhöh (1996)	Impulsivität/Reflexivität, Anspruchsniveau, Leistungsmotivation und Frustrationstoleranz
BaCO Belastbarkeits-Assessment bzw. BaCO-D Belastbarkeits-Assessment: computerisierte objektive Persönlichkeitstestbatterie	Ortner et al. (2007) ^a	Verschiedene Formen der Belastbarkeit
Multimethodische Objektive Interessensbatterie (MOI)	Proyer und Häusler (2008)	Berufsbezogene Interessen
Objektiver Leistungsmotivations-Test (OLMT)	Schmidt-Atzert (2007)	Leistungsmotivation
RISIKO – Risikowahlverhalten	Guttmann und Bauer (2004)	Risikobereitschaft
Wiener Risikobereitschaftstest Verkehr (WRBTV)	Hergovich et al. (2005)	Risikobereitschaft in Verkehrssituationen

^aVerlag existiert nicht mehr; Test jetzt bei Schuhfried erhältlich

3.4.3 Implizite Assoziationstests (IAT)

Auch Reaktionszeitbasierte („implizite“) Persönlichkeitstests, kurz IAT genannt, können den objektiven Persönlichkeitstests zugerechnet werden (s. Schmukle und Egloff 2011). Sie finden vor allem in der sozialpsychologischen Grundlagenforschung Verwendung, etwa um Einstellungen indirekt zu messen. Wir stellen diese interessante Messmethode kurz vor.

Einstellungen indirekt messen

Grundannahme Die Grundannahme von IAT ist, dass Gedächtnisinhalte in einem semantischen Netzwerk miteinander verbunden sind. Dabei spielt die Valenz, also ob die Testperson ein Element positiv oder negativ bewertet, oft eine zentrale Rolle. Mit einem IAT soll die Stärke der Assoziation zwischen 2 Konzepten gemessen werden. Beispielsweise wollen wir feststellen, wie positiv oder negativ der implizite Selbstwert von Menschen ist. Für das Konzept „Selbstwert“ können Wörter wie „ich“, „selbst“ oder [eigener Name] und zur Kontrolle Wörter wie „andere“, „sie“ oder „fremd“ verwendet werden. Da die Valenz des Selbstwertes von Interesse ist, bieten sich als Referenzkonzept Objekte an, deren positive oder negative Valenz allgemein bekannt ist, also beispielsweise „Kakerlake“ oder „Regenbogen“ (das Beispiel stammt von Grawronski 2006).

Assoziationsstärke im semantischen Netzwerk messen

Methodisches Vorgehen Das methodische Vorgehen wird nun kurz skizziert. Zunächst lernt die Testperson in Übungsdurchgängen eine einfache Diskriminationsaufgabe. Sie soll so schnell wie möglich eine Taste drücken, wenn sich das Wort auf das Selbst bezieht, und eine andere Taste, wenn es sich auf andere Menschen bezieht. Analog wird mit den Objekten (Kakerlake, Regenbogen etc.) verfahren. Bei der entscheidenden doppelten Diskriminationsaufgabe wird nun jede der beiden Tasten doppelt belegt:

- Die Testperson soll im ersten Durchgang z. B. die rechte Taste drücken, wenn das Objekt positiv ist (z. B. „Regenbogen“) oder wenn sich das Wort auf das Selbst bezieht (z. B. „ich“). Die andere (linke) Taste ist zu drücken, wenn sich das Wort auf andere bezieht oder das Objekt negativ ist.
- In einem weiteren Durchgang wird die Kopplung umgedreht. Die Selbstbegriffe sind nun mit der gleichen Taste zu beantworten wie negative Objekte, und die Begriffe, die andere betreffen, sind mit positiven Objekten verbunden.

Die Stärke der Assoziation des Selbstwertes mit positiv und negativ valenten Wörtern wird durch einen Differenzwert der Reaktionszeiten unter Bedingung A und B ermittelt. Bei einem positiven Selbstwert sollte die Assoziation mit positiven Wörtern stärker sein als die mit negativen Wörtern. Die Reaktion sollte also schneller erfolgen, wenn Selbstbegriffe und positive Wörter mit der gleichen Taste zu beantworten sind als wenn Selbstbegriffe und negative Wörter gepaart sind.

Implizite Assoziationstests liegen auch in anderen Varianten vor, als hier beschrieben (s. Schmukle und Egloff 2011). Für eine Verwendung für diagnostische Zwecke sind folgende Fragen relevant: Sind IAT-Maße hinreichend reliabel? Und gibt es gute Belege für ihre Validität?

Reliabilität Für die interne Konsistenz wurde in einer Metaanalyse (Hofmann et al. 2005) mit .79 ein Wert ermittelt, der etwa dem von Persönlichkeitsfragebögen entspricht. Die Retest-Reliabilität ist mit .51 allerdings eher niedrig.

Niedrige Retest-Reliabilität

Im Vergleich zu Fragebögen weniger valide

Validität In einer Metaanalyse von Hofmann et al. (2005) wurde nur eine durchschnittliche Korrelation von .24 zwischen IAT- und Fragebogenergebnissen festgestellt. Diese Korrelation ist doppelt minderungskorrigiert, beschreibt also den um den Messfehler bereinigten Zusammenhang. Moderatorvariablen wie die Art der IAT bzw. der Selbstberichte oder der Messgegenstand spielen eine Rolle. Für das Selbstkonzept (s. o.) betrug der korrigierte Zusammenhang $r = .21$. In einer weiteren Metaanalyse von Greenwald et al. (2009) wurde für die Kategorie „Persönlichkeitsmerkmale“ eine unkorrigierte Korrelation zwischen Fragebögen und IAT von $r = .17$ berichtet. Aber ist die Kriteriumsvalidität von IAT vielleicht der von Fragebögen ebenbürtig oder ihr sogar überlegen? Eine Antwort findet sich in der Metaanalyse von Greenwald et al. (2009). Bei Persönlichkeitsmerkmalen betrug die (unkorrigierte) Korrelation mit diversen Kriteriumswerten (subjektive Maße und Verhaltensmaße) $r = .35$ für Fragebogenmaße und $r = .28$ für IAT. Die Validitätskoeffizienten der Fragebogenmaße waren signifikant höher als die der impliziten Maße.

IAT für die Forschung nützlich

Fazit Fazit ist, dass IAT als Methode zur Messung von vielen Merkmalen, darunter auch Persönlichkeitsvariablen, „funktionieren“. Wegen ihrer niedrigen Retest-Reliabilität sind sie für viele Fragestellungen in der Einzelfallagnostik problematisch. Sie sind Fragebogenmaßen nicht bei der Validität überlegen. Eventuell weisen sie eine inkrementelle Validität auf, klären also zusätzlich zu einem Fragebogen Kriteriumsvarianz auf. Das ist jedoch im konkreten Fall zu prüfen. Ein weiteres Problem besteht darin, dass unserem Wissen nach zumindest im deutschen Sprachraum keine normierten IAT zur Verfügung stehen. Die Ausprägung eines Persönlichkeitsmerkmals kann nicht anhand von Rohwerten beurteilt werden. Für Forschungszwecke stellen IAT ein nützliches Werkzeug dar. Das gilt besonders für Forschungsgebiete wie die Forensische Psychologie, in denen es oft fraglich ist, ob die Testpersonen über die nötige Selbsteinsicht verfügen und bereit sind, offen und ehrlich zu antworten.

3.4.4 Weitere Forschung zu objektiven Persönlichkeitstests und impliziten Assoziationsstests

Es bleiben viele Fragen offen

Zum Thema objektive Persönlichkeitstests (inklusive IAT) ist weiter Forschung und Publikationstätigkeit zu verzeichnen. Besonders zu erwähnen ist eine aufwendige Studie von Koch et al. (2014), in der mehrere Fragen beantwortet werden sollten. Das Autorinnen- und Autorenteam stellt zunächst fest, dass die Konstruktvalidität der objektiven Persönlichkeitstests unbefriedigend ist; die Korrelationen mit Fragebögen zum gleichen Messgegenstand sind zu niedrig. Sie setzten 2 objektive Persönlichkeitstests, einen IAT (s. o.) und einen Fragebogen zur Gewissenhaftigkeit ein. Zum Vergleich verwendeten sie auch 2 Intelligenztests (die als objektive Persönlichkeitstests zur Intelligenz angesehen wurden), einen IAT und ein Selbstbeurteilungsverfahren zur Intelligenz ein. Die Testpersonen, 367 Studierende, bearbeiteten alle Verfahren zu 3 Messzeitpunkten. Der 1. objektive Persönlichkeitstest zur Gewissenhaftigkeit bestand aus Items des Test d2 (► Abschn. 3.2.2.2), die aber

ohne Zeitdruck korrekt zu bearbeiten waren. Der 2. objektive Persönlichkeitstest geht auf Cattell und Warburton (1967) zurück und findet sich in ähnlicher Form in den „Arbeitshaltungen“ (► Abschn. 3.4.1). Die Testpersonen müssen hier wiederholt und ohne Zeitdruck die Größe von 2 Figuren vergleichen. Der Forschungsansatz folgt also der Logik des Multitrait-Multimethod-Ansatzes (► Abschn. 2.6.3.4) und kombiniert diesen mit Messwiederholung. Somit kann auch geklärt werden, ob die Tests Zustände oder Persönlichkeitseigenschaften erfassen. Die Ergebnisse sind sehr komplex. Wir geben hier Antworten auf ausgewählte Fragen, die objektive Persönlichkeitstests zur Gewissenhaftigkeit betreffen:

1. Werden mit den beiden objektiven Persönlichkeitstests über die Messzeitpunkte hinweg stabile Merkmale erfasst? Ja, beim 1. objektiven Persönlichkeitstest (mit d2-Items) werden durchschnittlich 78 % und beim 2. Test (mit Items aus „Arbeitshaltungen“) 86 % der Varianz durch ein stabiles Merkmal erklärt (IAT und Fragebogen zum Vergleich: nur 41 % bzw. 95 %).
2. Wie hoch ist die konvergente Validität der objektiven Persönlichkeitstests? Die konvergenten Trait-Korrelationen der beiden objektiven Persönlichkeitstests zur Gewissenhaftigkeit waren niedrig ($r=.10$ bis $.24$; $M=.14$). Die entsprechenden Korrelationen zwischen den objektiven Persönlichkeitstests und dem IAT lagen zwischen $r=.14$ und $.20$ ($M=.14$) und die mit dem Fragebogen bei 0. Die objektiven Persönlichkeitstests zur Gewissenhaftigkeit messen demnach unterschiedliche Merkmale und haben zudem kaum oder sogar gar keine gemeinsame Varianz mit dem IAT und dem Fragebogen.
3. Wie hoch ist die diskriminante Validität? Mit Intelligenztests (als objektive Persönlichkeitstests für Intelligenz) fand sich kein signifikanter Zusammenhang ($r=-.08$ bis $.10$), was für die diskriminante Validität spricht. Für die IAT ergibt sich ein ganz anderes Bild: Die IAT zu Intelligenz und Gewissenhaftigkeit korrelierten sehr hoch miteinander ($r=.61$ und $.96$).
4. Können die objektiven Persönlichkeitstests ein Verhaltensmaß der Gewissenhaftigkeit vorhersagen? Die Pünktlichkeit, mit der die Testpersonen zur Untersuchung erschienen, diente als Verhaltensmaß. Ja, ein Zusammenhang zwischen objektiven Persönlichkeitstests und Pünktlichkeit war nachzuweisen. Allerdings leistete auch der Fragebogen einen unabhängigen Beitrag zur Erklärung der Pünktlichkeit.

3.4.5 Weitere digitale Ansätze – Machine Learning und künstliche Intelligenz

Seit einigen Jahren wird intensiv daran geforscht, u. a. Persönlichkeitsmerkmale mithilfe von digitalen Fußabdrücken, die Menschen im Internet hinterlassen (also objektiven Verhaltensdaten), zu messen. Der Definition objektiver Tests (► Abschn. 3.4) zufolge handelt es sich bei diesen Verfahren nicht um einen Test in engerem Sinne, weil das Verhalten nicht in einer standardisierten Situation erfasst wird. Vielmehr wird auf Verhaltensdaten zurückgegriffen, die Menschen in einer Vielzahl von natürlichen Situationen erzeugen.

Messung von
Persönlichkeitsmerkmalen anhand
digitaler Fußabdrücke

Die Definitionsmerkmale der Objektivität im Sinne des Verzichts auf Selbstbeurteilungen sowie der Undurchschaubarkeit (keine Augenscheininvalidität) treffen jedoch zu. Konzeptionell könnte man die Verfahren auch der Verhaltensbeobachtung zuordnen. Da es grundsätzlich aber auch möglich ist, die Verhaltensdaten in einer standardisierten Situation zu erheben, wie das Beispiel von PRECIRE JobFit (s. u.) zeigt, gehen wir im Kontext objektiver Persönlichkeitstests auf das Thema ein.

Das Konstruktionsprinzip lässt sich in groben Zügen leicht beschreiben: Viele Menschen bearbeiten zunächst psychologische Tests und stellen diese Daten für die Forschung zur Verfügung. Mithilfe von Algorithmen – der Ansatz wird als *Machine Learning* oder schlicht als *künstliche Intelligenz* bezeichnet – wird versucht, beispielsweise den Neurotizismuswert aus verfügbaren digitalen Fußabdrücken zu berechnen. Zwischen 2007 und 2012 hatten sich fast 7,5 Mio. Facebook-Nutzende an dem Projekt „myPersonality“ beteiligt und davon stimmten über 2 Mio. der Verwendung ihrer Daten zu. Mithilfe der Likes dieser Personen oder etwa ihrer Sprachmerkmale war es möglich, breite Persönlichkeitsmerkmale (Big Five) vorherzusagen (Bleidorn und Hopwood 2019). Das Potenzial dieses Ansatzes ist enorm, wenn erstens sehr viele Daten (Likes, Aufenthaltsorte, Nutzerverhalten bezüglich Musik und Filmen, Art und Anzahl sozialer Kontakte und vieles anderes mehr) von zweitens sehr vielen Menschen zur Verfügung stehen und drittens von ihnen ein Kriteriumswert (das Ergebnis in einem Persönlichkeitsfragebogen oder etwa im Beruf erzielter Umsatz) verfügbar ist. Die Algorithmen werden so lange an einem großen Datensatz trainiert, bis sie gelernt haben, beispielsweise den erzielten Umsatz vorherzusagen. Dabei gilt es, einige methodologische Aspekte zu beachten (Stachl et al. 2019):

- Ein Modell zur Vorhersage eines Persönlichkeitsmerkmals muss optimal an die Wirklichkeit angepasst werden. Es besteht u. a. die Gefahr des Overfitting; nur zufällig mit dem Merkmal assoziierte Verhaltensdaten können Eingang in das Modell finden.
- Deshalb ist es erforderlich, ein Modell immer wieder an einem neuen Datensatz zu validieren. Dadurch versteht man es zwar nicht, stellt aber sicher, dass es weiterhin funktioniert. Mit dem Begriff „model degradation“ wird der Umstand beschrieben, dass ein Modell an Vorhersagekraft verliert. So wurde spekuliert, dass Likes in sozialen Netzwerken für die Fernsehserie „Game of Thrones“ im Jahr 2017 schon 2019 eine andere Bedeutung für die Vorhersage von Offenheit für Erfahrung haben kann, weil die Serie inzwischen Mainstream-Charakter hat und in der Fangemeinde anderes bewertet wird.
- Ein Modell kann gut zur Vorhersage geeignet sein, ohne dass man es versteht, also erklären kann, wie es funktioniert. Die Algorithmen stellen dann eine Blackbox dar.

Stachl et al. (2019) äußern sich vorsichtig optimistisch, was den Einsatz von Machine Learning angeht. Bei korrektem Gebrauch führe die Methode wohl zu besseren Beurteilungen der Persönlichkeit. Nach dem Interview mit Richard Justenhoven zur künstlichen Intelligenz stellen wir einen Test vor, der diesen Ansatz verwendet, um Persönlichkeit objektiv zu messen.

Interview mit Richard Justenhoven

Richard Justenhoven

Herr Justenhoven, als Direktor Produktentwicklung bei Aon Assessment Solutions beschäftigen Sie sich schon lange mit künstlicher Intelligenz (KI). Haben Sie eine für unsere Zwecke, also Vorhersage von berufsrelevantem Verhalten und Erleben, einfache Definition von KI?

Der Begriff künstliche Intelligenz beschreibt ein Forschungsfeld der Informatik und subsumiert unterschiedliche Technologien dieses Teilgebiets.

Grundlegende Gemeinsamkeiten dieser Technologien sind die Verarbeitung großer Mengen komplexer Datenstrukturen, ein hohes Maß an Automatisierung in der Datenverarbeitung, sowie Fähigkeiten in der Mustererkennung, die eigenständiges Lernen (im Sinne einer Vorhersage- bzw. Verarbeitungsverbesserung) ermöglichen.

Die Kombination von Mustererkennung, Lernfähigkeit und Integration von unterschiedlichen Daten zur Entscheidungsfindung ist, was diese „intelligenten“ Technologien von traditionellen Ansätzen unterscheidet. Ein häufig beschriebenes Ziel von KI ist es, menschliche Datenverarbeitungsmuster und -fähigkeiten nachzuzeigen.

Welche Art von Daten stehen KI-Anwendungen im Bereich der Psychologischen Diagnostik zur Verfügung?

Es ist zunächst wichtig, zwischen öffentlich zugänglichen Informationen (beispielsweise Inhalte von Social-Media-Profilen) und spezifisch erhobenen Informationen (beispielsweise Daten psychometrischer Tests) zu unterscheiden.

Grundsätzlich gilt: Zweckgebundene Entscheidungen, wie z. B. in der Personalwahl, sollten stets auf spezifisch dafür erhobenen relevanten Daten basieren.

Innerhalb der spezifisch erhobenen Daten gibt es neben den tatsächlichen Testwerten noch weitere Daten, sog. „Paradaten“. Dazu zählen die Bearbeitungszeit insgesamt oder pro Item, Antwortverzögerungen sowie in manchen Fällen die Bearbeitungsreihenfolge von Items. Auch die Interaktion von Bewerberinnen und Bewerbern mit einer Online-Testoberfläche wie das wiederholte Aufrufen einer Hilfe-Seite oder Übersicht ist in bestimmten Fällen verfügbar. Paradaten werden in klassischen Onlinetests nicht zur Bestimmung des Testwertes genutzt, können allerdings von vielen KI-Modellen genau hierfür herangezogen werden.

Offenere Testformate wie Videointerviews oder auch frei geschriebene Texte liefern weitere, durch KI in der Psychologischen Diagnostik nutzbare Informationen.

Allerdings muss Bewerberinnen und Bewerbern zu Beginn einer Datenerhebung deutlich gemacht werden, welche Art von Daten erhoben und für die Auswertung herangezogen werden. Nur bei einer Einwilligung der Bewerberin bzw. des Bewerbers ist es möglich, Paradaten und Daten aus offenen Testformaten mithilfe von KI-Modellen auszuwerten.

Der klassische psychologisch-diagnostische Ansatz besteht darin, Verhalten (z. B. berufliche Leistungen) und Erleben (z. B. Berufszufriedenheit) durch

psychologische Konstrukte (z. B. Intelligenz, Persönlichkeitsmerkmale) vorherzusagen. Bei der Operationalisierung der Konstrukte greifen wir auf Verhalten (z. B. Lösen von Testaufgaben) und Erleben (z. B. Angaben in Fragebögen) zurück. Braucht man bei Anwendung der KI diesen „Umweg“ über Konstrukte? Wenn ja/oder nein, warum (nicht)

Auch wenn es verlockend erscheint, ein KI-System direkt von beobachtetem Verhalten auf Kriterien schließen zu lassen, ersetzt nicht die Nutzung von Konstrukten.

Ein gutes Beispiel hierfür ist der Befund von Kosinski et al. (2013). Sie konnten zeigen, dass sich Intelligenz u. a. durch Facebook-Likes für Curly Fries vorhersagen ließ. Allerdings ist es unwahrscheinlich, dass dieser Zusammenhang über die Zeit und über Stichproben hinweg stabil ist. Im Jahr 2019 sind Curly Fries für die meisten Menschen keine Neuigkeit mehr, das Facebook-Nutzungsverhalten vieler Menschen hat sich verändert und immer weniger junge Menschen frequentieren die Plattform. Ob sich dieser Zusammenhang auch heute noch derart findet, ist daher fragwürdig.

Die Nutzung von Konstrukten in KI-Anwendungen vereinfacht es, wie auch in traditionellen eignungsdiagnostischen Szenarien, Befunde zu generalisieren und Zusammenhänge in verschiedenen Kontexten zu spezifizieren und anzupassen.

Auch für die Transparenz von Entscheidungen ist der Konstruktansatz zentral. Für Bewerberinnen und Bewerber sowie für Personalpsychologinnen und -psychologen sollte stets nachvollziehbar sein, wie ein KI-System eine bestimmte Entscheidung generiert hat, welche Indikatoren als Indikativ für ein Konstrukt bewertet wurden und wie dieses Konstrukt mit einem Kriterium zusammenhängt. Ist diese Transparenz nicht gegeben, kann auch das Ergebnis einer KI-basierten Analyse nicht vollständig nachvollzogen werden.

Welche ethischen Probleme sehen Sie bei der Anwendung der KI zur

Vorhersage von berufsrelevantem Verhalten und Erleben?

Im Kontext ethischer Bedenken können drei übergeordnete Bereiche unterschieden werden: Entscheidungsfindung, Transparenz und Einwilligung.

Diese Bereiche sind nicht nur relevant für die Akzeptanz der Nutzung von KI-Systemen durch Bewerberinnen und Bewerber und ethisch vertretbare Personalauswahlpraktiken, sondern auch für die Sicherstellung legaler Entscheidungen.

In allen Anwendungen, in denen Entscheidungen mit weitreichenden Konsequenzen getroffen werden, z. B. in der Personalauswahl, sollte KI nicht genutzt werden, um die Menschen zu ersetzen, die für die Eignungsdiagnostik verantwortlich sind, sondern um sie zu unterstützen. Es bleibt der Anspruch, nachweisen zu können, wie eine (Auswahl-)Entscheidung zustande gekommen ist. Insbesondere im Training von KI-Systemen sind Aufsicht und wiederholte Evaluationen und Korrekturen unverzichtbar. KI-Modelle sind (noch) keine Selbstläufer und bedürfen weiterhin kritischer Überprüfung und Hinterfragung unserer Expertise. Sich voll und ganz auf ein einmalig validiertes KI-Modell zu verlassen, kann fatale Folgen haben.

Besonders der Aspekt der Transparenz ist für die etische Nutzung von KI-Modellen relevant. In allen Datenverarbeitungs- und Entscheidungsschritten muss nachvollziehbar sein, welche Datenpunkte für Schlussfolgerungen der Modelle genutzt wurden. Die Zuhilfenahme des Konstruktansatzes unter Verwendung nachvollziehbarer Modelle ist der Transparenz zuträglich.

Werden KI-Systeme genutzt, sollten Bewerberinnen und Bewerber darüber informiert werden. Dies gilt umso mehr unter Berücksichtigung der Datenschutzgrundverordnung (DSGVO) der EU. Abzuwägen ist, inwieweit eine explizite Zustimmung zur Erfassung und Nutzung bestimmter Informationen erforderlich ist. Möglicherweise verändertes Verhalten der Bewerbenden oder auch die Ablehnung eines Verfahrens sollten hierbei berücksichtigt werden.

3.4.6 Objektive sprachbasierte Eignungsdiagnostik

Mit JobFit von PRECIRE (2016) ist ein diagnostisches Verfahren auf den Markt gekommen, das verspricht, 11 Persönlichkeitsmerkmale und 6 Skills wie etwa Durchsetzungsfähigkeit sowie einige „Kommunikationskennwerte“ objektiv (im Sinne objektiver Tests) und unverfälschbar über eine Sprachprobe zu messen. Die Testpersonen müssen sich lediglich einem automatischen Telefoninterview mit standardisierten Fragen wie „Bitte beschreiben Sie den Ablauf eines typischen Sonntags“ unterziehen. Sie dürfen dabei auch einzelne Fragen auslassen, weil die Inhalte der Antworten überhaupt keine Rolle spielen.

PRECIRE JobFit

Die Software sucht in den transkribierten Antworten und in den Sprachaufnahmen nach Text- und Sprachmerkmalen sowie deren Kombinationen, die die Ausprägung der oben genannten Merkmale „verraten“. Bei der Testentwicklung hatten über 5000 Personen u. a. diverse Persönlichkeitsfragebögen bearbeitet und eine Sprachprobe abgeliefert. Die Software ist laut Manual in der Lage, pro Skala bis zu 10.000 „Features“ (Sprachmerkmale) zu messen. In einer Validitätsstudie konnten relativ hohe Übereinstimmungen ($r = .49$ bis $.66$) zwischen 5 maschinell erfassten Persönlichkeitsmerkmalen und konstruktkonvergenten Big-Five-Skalen in einem Persönlichkeitsfragebogen nach dem HEXACO-Modell (► Abschn. 3.3.1) festgestellt werden. Dieser Fragebogen war bei der Testentwicklung nicht eingesetzt worden. Zur Kriteriumsvalidität (Berufserfolg) liegen keine aussagekräftigen Studien vor.

Testentwicklung und Validität

Der Preis für eine Einzeltestung im Top Executive Management kostet bis zu 695 €. JobFit hat nicht nur in Fachkreisen, sondern auch in der Öffentlichkeit große Beachtung erfahren. Das Spektrum der Resonanz reicht von Faszination für die neue Technik bis Skepsis. Beispielsweise erschien am 4. November 2017 in der *Süddeutschen Zeitung* der Beitrag „Bewerbungsgespräch bei einem Computer“ (Hoffmeyer 2017). Wichtige Fragen sind, ob sich der Einsatz des Verfahrens lohnt und ob es bedenkenlos eingesetzt werden kann. In einer Testrezension (Schmidt-Atzert et al. 2019, S. 20) wird dazu festgestellt:

Skepsis ist angebracht

» „PRECIRE“ erfasst bestimmte Persönlichkeitsmerkmale, die ansonsten mit Fragebogen gemessen werden. Es bleibt aber offen, ob damit die berufliche Eignung besser erkannt werden kann. Es fehlt sogar der Nachweis, dass „PRECIRE“ diesbezüglich den Fragebogen ebenbürtig ist.

Mit anderen Worten: Über die inkrementelle Validität gegenüber den sehr kostengünstigen Fragebögen wissen wir noch nichts. Zum anderen wird auf ethische und rechtliche Probleme hingewiesen. Im Manual wird der Datenschutz angesprochen, aber nur in Bezug auf die Weitergabe der gewonnenen Daten.

Ethische und rechtliche Probleme

» Zum Datenschutz gehört aber auch die informationelle Selbstbestimmung. Wer eine Sprachprobe ab liefert, weiß nicht, was er damit über sich verrät (Schmidt-Atzert et al. 2019, S. 20).

Weiterführende Literatur und Internetressourcen

Über die theoretischen Grundlagen objektiver Persönlichkeitstests sowie über die Entwicklung mehrerer konkreter Verfahren informiert ein von Ortner et al. (2006) herausgegebenes Buch. Einen etwas aktuelleren Überblick geben Ortner und Proyer (2015). Sehr informativ ist auch der Beitrag von Schmukle und Egloff (2011) in der Enzyklopädie der Psychologie.

Sehr empfehlenswert ist ein etwa 15-minütiger Videoclip, den Prof. Dr. Uwe Kanning zum Thema „Was verrät die Sprache über einen Menschen?“ erstellt hat (► <https://www.youtube.com/watch?v=t3xTmqd26bo>). Er setzt sich darin kritisch mit dem Einsatz der Sprachanalyse zur beruflichen Eignungsdiagnostik auseinander.

?

Übungsfragen

— Abschn. 3.4:

- Was versteht man unter objektiven Persönlichkeitstests?
- Nennen Sie 3 Beispiele für objektive Persönlichkeitstests!
- Warum sind implizite AssoziationsTests für die Einzelfalldiagnostik nicht brauchbar?
- Nennen Sie ein standardisiertes diagnostisches Verfahren, das durch Anwendung von künstlicher Intelligenz Persönlichkeitsmerkmale messen soll und beschreiben Sie dessen Durchführung!
- Warum ist bei dem Test, der künstliche Intelligenz zur Diagnostik von Persönlichkeitsmerkmalen verwendet, der Nachweis von inkrementeller Validität gegenüber Fragebögen wichtig?

3.5 Projektive Verfahren

Ähnlichkeit mit objektiven Persönlichkeitstests

Zu den wohl umstrittensten diagnostischen Verfahren gehören die sog. „projektiven Tests“. Für die einen stellen sie einen einzigartigen und ergiebigen Zugang zur Persönlichkeit eines Menschen, seinen Motiven sowie dessen Wünschen dar – für andere sind sie ein psychometrischer Albtraum, also Verfahren mit völlig unzulänglichen Gütekriterien. Zur letztgenannten Position ist zu sagen, dass es nicht „die“ projektiven Tests gibt, sondern sehr unterschiedliche – auch mit unterschiedlichen Gütekriterien. Projektive Verfahren gleichen in mehreren Aspekten den in ▶ Abschn. 3.4 vorgestellten objektiven Persönlichkeitstests:

- Auf Selbstberichte wie bei Fragebögen wird verzichtet.
- Aus dem Verhalten der Testperson in einer standardisierten Testsituation wird auf deren Persönlichkeitseigenschaften geschlossen.
- Der Test ist für die Testpersonen weitgehend undurchschaubar – sie wissen nicht, was genau gemessen wird.

Insofern könnte man projektive Verfahren auch den objektiven Persönlichkeitstests zuordnen. Sie heben sich von diesen jedoch durch eine ihnen zugrunde liegende Theorie ab: Die Reaktionen auf das Testmaterial kommen durch Projektion eigener Eigenschaften in das mehr oder weniger diffuse Testmaterial zustande.

Klassischer Projektionsbegriff

Projektion Der Begriff der Projektion geht auf Freud zurück und meint ursprünglich die Verlegung einer Eigenschaft, die das Ich bedroht und an der eigenen Person nicht wahrgenommen wird, auf eine Person der Außenwelt. Beispielsweise mag jemand sehr geizig sein, dies aber in einem Interview oder bei Bearbeitung eines Fragebogens von sich weisen. Dieselbe Person meint nun, bei einer anderen Person – die vielleicht gar nicht geizig ist – genau diese Eigenschaft zu „beobachten“. Die bei sich selbst nicht akzeptierte und wahrgenommene Eigenschaft wird in eine andere Person projiziert. Im Grunde ist es der eigene Geiz, der in der anderen Person wahrgenommen wird. Der klassische Projektionsbegriff zeichnet sich durch 3 Merkmale aus:

- Projektion besteht darin, dass man anderen Menschen Eigenschaften, Gefühle und/oder Wünsche unterstellt, die man selbst hat, aber sich nicht eingestehst, weil sie gewöhnlich negativ bewertet werden.
- Projektion ist ein unbewusster Vorgang.
- Es handelt sich um einen Abwehrmechanismus.

Diagnostische Verfahren

Würde man dem klassischen Projektionsbegriff folgen, müssten projektive Tests Personen als „Projektionsfläche“ verwenden. Tatsächlich gibt es auch projektive Tests, die zumindest Bilder von Personen als Testmaterial verwenden, so der Thematische Apperceptionstest (TAT; ▶ Abschn. 3.5.1.2). In einer verallgemeinerten Variante versteht man unter Projektion, dass sich eigene Interessen, Gewohnheiten, Zustände, Wünsche etc. auf die Wahrnehmung bzw. Interpretation von mehrdeutigem Material auswirken. Damit kommen als „Projektionsfläche“ auch Tintenkleckse, andere mehrdeutige Objekte und sogar ein weißes Blatt Papier infrage. Die Annahme, dass es sich um einen unbewussten Vorgang handelt, der zudem der Abwehr dient, ist nicht mehr zwingend erforderlich. Für die Konstruktion von projektiven Tests folgt daraus, dass mehrdeutige Reize als Testmaterial gut geeignet sind. Wenn man ein bestimmtes Merkmal messen will, sollten die Reize einen spezifischen Aufrufcharakter für dieses Merkmal haben, also dazu anregen, merkmalsbezogene Antworten zu geben.

Projektive Verfahren werden somit nach der Theorie klassifiziert, die ihnen zugrunde liegt. Das ist ein ungewöhnlicher Kategorisierungsgesichtspunkt, da die Gruppierung üblicherweise nach dem Messgegenstand (z. B. Intelligenz, Konzentration, Persönlichkeit) oder äußereren formalen Kriterien (z. B. Papier-und-Bleistift-Test, Einzel- oder Gruppentest usw.) erfolgt. Die Bereichsbildung nach „der“ (einen) Theorie ist jedoch problematisch, weil es divergierende Darstellungen darüber gibt, was genau eine Projektion ist und welche Variante im konkreten Fall vorliegt.

Einsatzbereiche Projektive Verfahren werden in der diagnostischen Praxis primär im klinischen Bereich verwendet. In einer Befragung in 92 stationären und ambulanten Einrichtungen für Kinder und Jugendliche (Bölte et al. 2000) gaben 31 % der Befragten an, „immer“ projektive Verfahren einzusetzen – für Persönlichkeitsfragebögen lag der vergleichbare Wert bei 29 %. Weitere 34 % der Befragten berichteten, diese Verfahren „oft“ einzusetzen. Auf die Frage, welche Verfahren sie einsetzen, gaben viele auch den „Baumtest“ oder „Familie in Tieren“ an – Verfahren, die den üblichen psychometrischen Kriterien absolut nicht genügen. Wie nachfolgend deutlich wird, bleibt nur zu hoffen, dass diese Verfahren nur zum Einstieg in ein Gespräch mit Kindern und Jugendlichen verwendet werden und jede Auswertung unterlassen wird.

Einteilung Projektive Verfahren sind so vielfältig, dass man sie nach verschiedenen Kriterien einteilen kann:

- Formdeuteverfahren (Beispiel: Rorschach-Test; Aufgabe: Tintenkleckse deuten)
- Verbal-thematische Verfahren (Beispiel: TAT; Aufgabe: Geschichten zu Bildern erfinden)
- Zeichnerische und Gestaltungsverfahren (Beispiel: Familie in Tieren; Aufgabe: die eigene Familie als Tiere zeichnen)

Eine andere Einteilung, die Lilienfeld et al. (2000) zufolge auf Lindzey (1959) zurückgeht, orientiert sich alleine am Antwortmodus:

- Konstruktion (Beispiel: TAT; Aufgabe: etwas frei konstruieren oder beantworten)
- Ergänzung (Beispiel: Rosenzweig Picture-Frustration-Test; Aufgabe: Antworten in eine Sprechblase einfügen)
- Anordnung/Selektion (Beispiel: Szondi-Test; Aufgabe: Bilder von Menschen danach auswählen, ob sie einem gefallen oder nicht gefallen)
- Ausdruck (Beispiel: Analyse der Handschrift)

Verallgemeinerter Projektionsbegriff

Theorie gibt den Tests ihren Namen

Verwendung primär im klinischen Bereich

Auswahl der vorgestellten projektiven Tests

Arten projektiver Tests Zu den weltweit am häufigsten verwendeten projektiven Verfahren gehören der Rorschach-Test und der Thematische Apperzeptionstest (TAT). Bei diesen Verfahren handelt es sich um Klassiker, die schon sehr lange auf dem Markt sind und zu denen inzwischen fast unüberschaubar viele Publikationen vorliegen. Zum TAT wurden Varianten wie der Leistungsmotivations-TAT, die auch eine hohe Standardisierung der Auswertung erlauben, und als semiprojektives Verfahren das Multi-Motiv-Gitter (MMG) entwickelt. Aus der Gruppe der zeichnerischen und Gestaltungsverfahren wird im Folgenden die „Familie in Tieren“ vorgestellt und kritisch erörtert.

Wie bereits erwähnt sind projektive Verfahren für Testpersonen schwerer zu durchschauen als Fragebögen. Dennoch sind sie nicht gegen Verfälschung resistent, wie mehrere Untersuchungen belegen. Exemplarisch soll eine Untersuchung zum TAT (► Abschn. 3.5.1.2) beschrieben werden: Holmes (1974) zeigte Studierenden, die keine Vorkenntnisse über projektive Verfahren hatten, 2 TAT-Bilder. Sie sollten, wie beim TAT vorgesehen, jeweils eine Geschichte dazu verfassen. Die Hälfte der Testpersonen sollte dabei ehrlich sein, die andere Hälfte sollte sich als leistungsmotiviert darstellen. Drei Wochen später wurde der Test wiederholt; die Teilnehmerinnen und Teilnehmer wurden nun der jeweils anderen Bedingung zugewiesen. In der 1. Sitzung unterschieden sich die beiden Gruppen sehr deutlich in ihrer Leistungsmotivation (nur dafür wurde ein Kennwert bestimmt). Würde man die Leistungsmotivation in Standardwerten ausdrücken (mit der Kontrollgruppe als „Normgruppe“, einem Mittelwert von 100 und einer Standardabweichung von 10), hätten die Testpersonen in der Faking-Bedingung einen Wert von 112 erreicht. In der 2. Sitzung, in der nun ein wenig Testerfahrung bestand, war der Unterschied noch größer; in der Faking-Gruppe entsprach der Testwert einem Standardwert von 116.

Auch das bereits erwähnte semiprojektive MMG (► Abschn. 3.5.2.3) erwies sich in einer experimentellen Studie (Ziegler et al. 2007) als verfälschbar. Die Testpersonen sollten sich vorstellen, arbeitslos zu sein und nun an einer Eignungsuntersuchung teilzunehmen. In der Faking-good-Bedingung wurde ihnen mitgeteilt, die Stelle sei sehr attraktiv. In der Faking-bad-Bedingung wurde die Stelle als sehr unattraktiv beschrieben. Sie sollten sich dementsprechend gut oder schlecht darstellen, aber so geschickt vorgehen, dass Expertinnen bzw. Experten die Verfälschung nicht entdecken. Nur das Leistungsmotiv wurde untersucht. Gegenüber einer neutralen Kontrollgruppe fiel die Furcht vor Misserfolg in der Faking-good-Gruppe um 7 Standardwertpunkte niedriger aus und die Hoffnung auf Erfolg um 6 Punkte höher. Bei faking bad zeigte sich das umgekehrte Muster: Der Standardwert für Furcht vor Misserfolg lag 4 Punkte höher und der für Hoffnung auf Erfolg 7 Punkte niedriger als in der Kontrollbedingung.

3.5.1 Klassische projektive Tests

Wie bereits erwähnt, gehören der Rorschach-Test, ein Formdeuteverfahren, sowie der TAT, ein verbal-thematisches Verfahren, zu den am häufigsten verwendeten projektiven Verfahren. Wir stellen zunächst beide eingehender vor, um im Anschluss aus dem TAT abgeleitete Testprinzipien und semiprojektive Verfahren aufzugreifen.

3.5.1.1 Rorschach-Test

Der Rorschach-Test (Rorschach 1949) wurde nach seinem Erfinder Hermann Rorschach, einem Schweizer Psychiater, benannt. Rorschach (Abb. 3.23) arbeitete seit 1918 mit einer von ihm entwickelten Serie von 40 Tintenklecksbildern. Er hatte die Karten 117 gesunden Kontrollpersonen und 188 Patientinnen und Patienten mit Schizophrenie mit der Frage „Was könnte das sein?“ vorgelegt.

„Die Geschichte der Tintenkleckse“

Wie kam Rorschach auf die aus heutiger Sicht ungewöhnliche Idee, Zufallsformen für diagnostische Zwecke deuten zu lassen? Rorschach hatte schon in seiner Kindheit Spaß daran, „Bilder“ herzustellen, indem er Tinte oder Farbe auf ein Blatt Papier gab und dieses zusammenfaltete; von seinen Mitschülern erhielt er dafür den Spitznamen „Klecks“. Das spielerische Herstellen von Klecksbildern war früher unter Kindern verbreitet. Der schwäbische Arzt Justinus Kerner hatte bereits 1857 ein Buch mit dem Namen *Klecksografien – Gedichte zu Dintenklecksen* veröffentlicht. Kerner hatte sich von selbst hergestellten Tintenklecksen zu Gedichten anregen lassen. Tintenklecksbilder wurden bereits vor dem Rorschach-Test für wissenschaftliche Zwecke verwendet: Alfred Binet, der Schöpfer des modernen Intelligenztests, hatte Tintenkleckse als Testmaterial zur Messung von Kreativität eingesetzt. Bereits 1917 hatte Szymon Hens eine Doktorarbeit über einen Tintenkleckstest bei dem berühmten Schweizer Psychiater Bleuler verfasst, der auch später der Doktorvater von Rorschach war.

Tintenkleckstest mit Vorläufern

Ein Buchmanuskript mit 15 ausgewählten Bildern wurde von mehreren Verlegern abgelehnt. Der Bircher-Verlag, der das Buch *Psychodiagnostik: Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments (Deutlassen von Zufallsformen)* schließlich veröffentlichte, ging bald bankrott. Hans Huber gründete 1927 in Bern einen Verlag und kaufte die Rechte von Bircher. Das Buch und die 10 dazugehörigen Tafeln trugen ganz wesentlich zum wirtschaftlichen Erfolg des Verlages bei. Rorschach konnte den Erfolg nicht miterleben; er verstarb 1922 im Alter von 37 Jahren an den Folgen einer Blinddarmentzündung.

„Wahrnehmungsdiagnostisches Experiment“



Abb. 3.23 Hermann Rorschach. (© akg-images/picture-alliance)

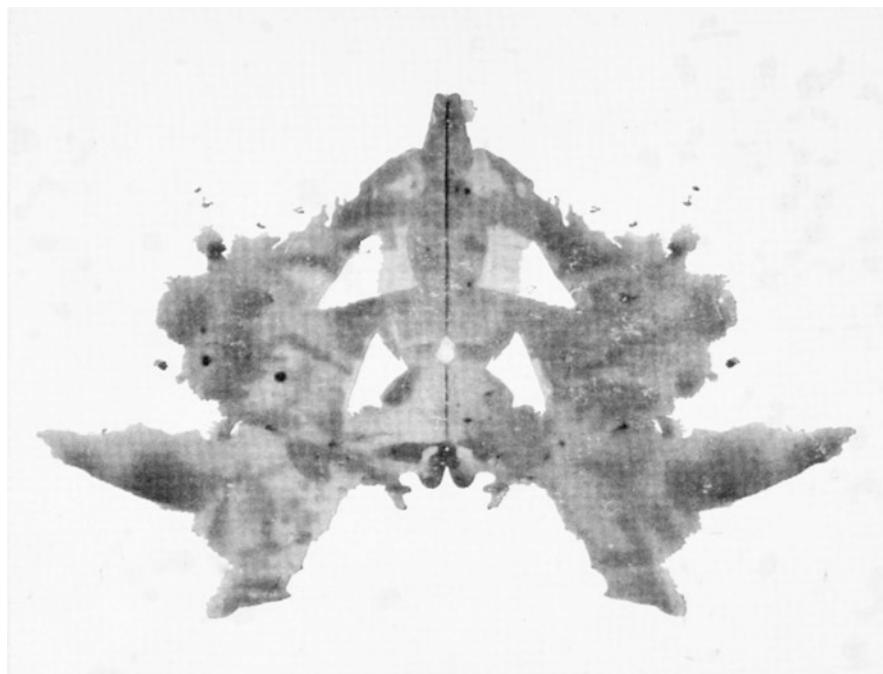
Erst später als projektives Verfahren eingeordnet

Erfassungsmodus, Determinanten, Inhalt, Originalität

Rorschach hatte das Verfahren zur Diagnostik der damals neu definierten Störung „Schizophrenie“ entwickelt und großen Wert auf formale Prinzipien der Wahrnehmung gelegt (s. Schmidt-Atzert 2006). Die theoretische Einordnung des Rorschach-Tests als projektives Verfahren wurde erst 1939 vorgenommen (Mattlar 2004). Der Test wurde in den 1990er-Jahren rund 6 Mio. Mal pro Jahr angewendet (Sutherland 1992; zitiert nach Lilienfeld et al. 2000). Im Jahr 2009 kam es zu einem Eklat: Der kanadische Arzt James Heilman lud die 10 Farbkleckstafeln auf die Online-Enzyklopädie Wikipedia – inklusive gängiger Interpretationen (Schneibel 2009). Die Bilder waren zwar schon vorher im Internet abrufbar, aber deren Verbreitung samt Interpretationshinweisen über eine bekannte Internetadresse bewirken, dass Testpersonen nun eventuell andere Deutungen liefern, wenn sie sich vorher informiert haben.

Durchführung und Auswertung Der Rorschach-Test ist ein Formdeuteverfahren, bei dem der Testperson nacheinander 10 Klecksbilder (für ein Beispiel s. □ Abb. 3.24) vorgelegt werden. Die Frage dazu lautet immer: „Was könnte das sein?“ Die Tafeln dürfen gedreht werden, die Zahl der Antworten ist beliebig. Jede Antwort wird protokolliert und anschließend nach bestimmten Kategorien mehrfach signiert.

Die klassische Auswertungsmethode von Rorschach sieht eine Beurteilung der Antworten nach 4 Gesichtspunkten vor: Erfassungsmodus (Ganz- oder Detaildeutung), Determinanten (Form, Farbe, Bewegung), Inhalt (z. B. Tier, Mensch, Anatomie) und Grad der Originalität (z. B. „Vulgärantwort“ für häufig vorkommende Deutungen – etwa „Fledermaus“ bei der in □ Abb. 3.24 gezeigten Tafel). Weitere Indikatoren für das „Psychogramm“ sind die Antwortzahlen, die Reaktionszeiten, die Sukzession der Erfassungsmodi, der Erfassungs- und Erlebnistyp und verschiedene Prozentwerte (Menschen, Tier-, Anatomiendeutungen usw.). Eine inhaltliche Analyse nach tiefenpsychologischen Prinzipien ist nicht obligatorisch. Ist schon die Signierung



□ Abb. 3.24 Tafel aus dem Rorschach-Test. (Hermann Rorschach, Psychodiagnostik. Der Rorschach®-Test (Tafel I). © Verlag Hans Huber, Hogrefe AG, Bern/Switzerland)

der Antworten nicht einfach, so bedarf deren Interpretation eingehender Schulung und langjähriger Erfahrung, weil die Ergebniskategorien nur im Zusammenhang bewertet werden sollen.

Im deutschen Sprachraum stellte lange Zeit ein erstmals 1951 erschienenes umfangreiches Buch (Bohm 2004) die Grundlage für die Auswertung und Interpretation des Rorschach-Tests dar. In den USA steht mit dem Comprehensive System (Exner 1974, 2003) ein Handbuch für den Rorschach-Test zur Verfügung, das detaillierte Anweisungen für die Durchführung und Auswertung sowie Normen für Kinder und Erwachsene enthält. Eine deutschsprachige Fassung der 5. Auflage des Arbeitsbuchs für das Comprehensive System ist 2010 erschienen (Exner 2010).

Gütekriterien Trotz des detaillierten Comprehensive System ergeben sich für die Auswertungsobjektivität erhebliche Probleme. So korrelieren einige Kategorien deutlich mit der nicht festgelegten Antwortzahl, und die Übereinstimmung zwischen verschiedenen Auswertenden variiert bei unterschiedlichen Stichproben und Kategorien zwischen 52 und 98 %. Die Auswertungsübereinstimmung (Kappa oder andere Koeffizienten) für das Comprehensive System liegt Mattlar (2004) zufolge zwischen .85 und .94 und ist damit relativ hoch.

Auch die Reliabilitätsbestimmung ist problematisch. Ein Halbierungskoeffizient lässt sich kaum berechnen, weil die 10 Testtafeln sehr unterschiedlich sind. Die Bestimmung einer Paralleltest-Reliabilität scheitert am Fehlen einer Parallelform. Aber auch die Ermittlung der Retest-Reliabilität ist schwierig, da die gegebenen Antworten leicht erinnert werden können und häufig zu Kontrastreaktionen bei der Testwiederholung führen. Dennoch liegen Befunde zur Retest-Reliabilität vor. Wood et al. (2015) berichten Werte zwischen .34 und .69 bei einem Retest-Intervall von 3 Monaten.

Zur Validität des Rorschach-Tests liegen viele Hundert Publikationen unterschiedlicher Qualität vor. Einschlägige Metaanalysen erlauben jedoch eine Beurteilung der Befunde. Hiller et al. (1999) haben eine Metaanalyse zur Validität von MMPI (► Abschn. 3.3.3.1) und Rorschach-Test durchgeführt; die Daten wurden nach einer Kritik an methodischen Details teilweise reanalyisiert (Rosenthal et al. 2001). Rosenthal et al. (2001) zufolge beträgt die durchschnittliche Korrelation (gewichteter Median) des Rorschach-Tests mit diversen Validitätskriterien $r = .29$, die des MMPI im Vergleich dazu .35. Allerdings fielen die Korrelationen für den Rorschach-Test in jüngerer Zeit höher aus als früher, was von der Autorin und den Autoren mit der Einführung des Comprehensive System in Verbindung gebracht wurde. Die Korrelation mit objektiven Kriterien wie „Gefängnisinsasse oder nicht“ fällt jedoch für den Rorschach-Test höher aus als für das MMPI ($r_{tc} = .37$ vs. .20), während das MMPI höher mit psychiatrischen Diagnosen korreliert als der Rorschach-Test ($r_{tc} = .37$ vs. .18; Hiller et al. 1999, Tab. 9).

In einer weiteren Metaanalyse sind Wood et al. (2010) der Frage nachgegangen, wie gut mit dem Rorschach-Test Psychopathen bzw. Psychopathinnen erkannt werden können. Dazu suchten sie Studien, in denen Personen mit einer antisozialen Persönlichkeitsstörung (diagnostiziert mit einer Psychopathie-Checkliste) mit Kontrollpersonen verglichen wurden. Zu 37 Kennwerten aus dem Comprehensive System lagen Ergebnisse vor, die aber insgesamt enttäuschend ausfielen: Nur für einen einzigen Kennwert, der sich auf das Aggressionspotenzial bezieht, lag die Korrelation über .20 ($r_{tc} = .23$). Damit wurde der Befund, dass der Rorschach-Test mit psychiatrischen Diagnosen schlecht übereinstimmt, erneut bestätigt.

Comprehensive System

Objektivität und Reliabilität nicht optimal

Geringe Übereinstimmung mit psychiatrischen Diagnosen

Metaanalyse zur Validität von 65 Kennwerten

3

Eine Metaanalyse von Mihura et al. (2013) befasst sich mit der Validität von 65 Kennwerten („Variablen“) nach dem Comprehensive System. Die Daten stammen aus 215 unabhängigen Studien mit insgesamt über 25.000 Testpersonen. Die Kennwerte wurden entweder durch Selbstberichte (insbesondere Fragebögen) oder „extern“, z. B. durch psychiatrische Diagnosen oder Beobachtung, validiert. Die wichtigsten Ergebnisse lassen sich wie folgt zusammenfassen:

- Wer mit dem Comprehensive System arbeitet, kann sich in Tab. 2 der Publikation über die Validitätsbelege aller 65 Kennwerte informieren. Diese Informationen sind nützlich für die Schlussfolgerungen, die man aus einem Rorschach-Protokoll zieht.
- Aus wissenschaftlicher Sicht ist es bemerkenswert, dass für immerhin 13 Kennwerte „exzellente“ ($r > .32$) und für weitere 17 „gute“ Validitätsbelege ($r > .20$) vorliegen.
- Die Korrelation der Rorschach-Variablen mit Kriterien war höher, wenn externe statt subjektiver Kriterien herangezogen wurden ($r = .27$ vs. $.08$).
- Einschränkend ist zu sagen, dass bei einigen Kennwerten keine eindeutige Aussage möglich ist; der Messgegenstand ist eher vage definiert. So spricht „menschliche Bewegung“, also wenn die Testperson eine Aktivität von Menschen sieht (z. B. „Zwei Menschen sitzen an einem Tisch“) für geistige Fähigkeiten einschließlich Planung, Fantasie und Empathie. Die Zusammenhänge zu externen Kriterien sind deutlich ($r = .33$) und gut belegt ($N = 9576$). Es gibt auch Kennwerte mit klarer Interpretation, für die gute Validitätsbelege vorliegen. Als Beispiel sei Textur genannt ($r = .24$, $N = 1648$). Feine Abstufung von hell und dunkel wie in der Antwort „Ein weicher, pelzartiger Vorleger“ sollen dafür sprechen, dass die Person einen Wunsch nach zwischenmenschlicher Nähe hat.

Kritik an der Metaanalyse

Die Studie von Mihura et al. (2013) löste 2 kritische Stellungnahmen aus. Wood et al. (2015) stimmten den Ergebnissen und deren Interpretation weitgehend zu. Sie überprüften gezielt 4 Kennwerte des Comprehensive System aus dem nichtkognitiven Bereich, die Mihura et al. (2013) zufolge die höchsten Validitätskoeffizienten aufweisen. Dazu führten sie eine eigene Metaanalyse durch, in der sie unveröffentlichte Dissertationen berücksichtigten. Das methodologische Vorgehen (z. B. fehlende Korrektur für die Anzahl der Antworten, die Verwendung von Dissertationen, die offenbar Fehler enthalten), das bei den untersuchten Variablen zu leicht anderen Validitätskoeffizienten führt, wird von Mihura et al. (2015) in einer Replik als unangemessen zurückgewiesen.

Czopp und Zeligman (2016) interessierten sich dagegen für 13 Kennwerte, die sich in der Metaanalyse als nicht oder wenig valide erwiesen hatten. Ihre Kritik besteht im Wesentlichen darin, dass Mihura et al. (2013) einige dieser Variablen etwas anders interpretieren als im Comprehensive System vorgeschlagen. Beispielsweise wird die Variable „Tierbewegung“ (ein Tier, das eine artgemäße Aktivität ausführt) im Comprehensive System interpretiert als „disconcerting awareness of unmet needs“, während Mihura et al. (2013) dies als „pressing primary needs“ verstanden (Czopp und Zeligman 2016, Tab. 1). Die Frage, was mit einer Variablen erfasst wird, wirkt sich natürlich auf die Auswahl geeigneter Validitätskriterien aus und damit auf das Ergebnis der Validierung. Die Kritik bleibt auf der konzeptuellen Ebene; eine Reanalyse der 13 Variablen wird nicht vorgenommen. Mihura et al. (2016) wehren sie gegen die Kritik, indem sie auf ihre fachliche Expertise und die Konsultation

von Comprehensive-System-Experten sowie ihr sorgfältiges Vorgehen bei der Auswahl der Validitätsstudien hinweisen. Eine Fußnote (Nr. 9) macht deutlich, wie schwer es offenbar ist, eine unangreifbare Metaanalyse zur Validität von Rorschach-Variablen vorzulegen. Mihura et al. (2013) hatten 2.467 Artikel unter die Lupe genommen. In jedem Beitrag wurden meist Dutzende von Ergebnisse berichtet; insgesamt fanden sie in den Publikationen 2.468 postulierte Zusammenhänge zwischen einem Prädiktor und einem Kriterium. Bei mindestens 25 Rorschach-Ergebnissen pro Artikel und 2468 zu prüfenden Prädiktor-Kriterium-Beziehungen kommt man auf über 30 Mio. zu treffende Entscheidungen.

Insgesamt erscheint uns die Kritik nicht geeignet, die wesentlichen Ergebnisse der Metaanalyse von Mihura et al. (2013) in Zweifel zu ziehen.

Kritik zurückgewiesen

Fazit Fazit ist, dass der Rorschach-Test nicht pauschal mit dem Argument abgelehnt werden kann, er sei nicht hinreichend reliabel und valide. Seine Schwächen sind eher der große Aufwand für die Auswertung und Interpretation und der für viele Fragestellungen fehlende Nachweis einer inkrementellen Validität gegenüber ökonomischeren Verfahren.

Im Übrigen scheint es 2 Lager zu geben, die sich in ihrer Einstellung gegenüber dem Rorschach-Test unterscheiden. Eine Umfrage richtete sich an amerikanische Psychologinnen und Psychologen, die als Mitglieder der Society for Personality Assessment (SPA) und/oder der American Psychological Association (APA) registriert waren (Musewicz et al. 2009). Die meisten gaben an, im klinischen Bereich tätig zu sein (92 %), 65 % waren in einer privaten Praxis tätig; Diagnostik gehörte zu ihren beruflichen Aufgaben (98 %). In der Befragung ging es auch um die Einstellungen gegenüber dem Rorschach-Test. Dazu dienten u. a. die folgenden beiden Fragen: „I believe that the Rorschach is an effective personality test“ und „Based on my interpretation of the current research regarding the Rorschach's statistical reliability and validity, the test should continue to be used in clinical practice“. Diese waren auf einer Skala von 1 („strongly disagree“) bis 7 („strongly agree“) zu beantworten. Die beiden Fragen wurden von den SPA-Mitgliedern mit sehr starker Zustimmung beantwortet ($M=6,3$ und 6,2). Die APA-Mitglieder antworteten deutlich zurückhaltender ($M=4,8$ und 4,9). Nun muss man wissen, dass es sich bei der SPA um eine traditionsreiche Gesellschaft handelt, die historisch stark mit dem Rorschach-Test und auch anderen projektiven Verfahren verbunden ist. Im Jahr 1950 wählte sie den Namen „Society of Projective Techniques“ und die von ihr herausgegebene Fachzeitschrift war das *Journal of Projective Techniques*. Namen können sich ändern: Die Zeitschrift wurde in *Journal of Projective Techniques and Personality Assessment* umbenannt und trägt seit 1971 den Namen *Journal of Personality Assessment* (Weiner 2018).

Einstellung zum Rorschach-Test
lagerabhängig

3.5.1.2 Thematischer Apperceptionstest (TAT)

Der TAT von Murray (1991) wurde erstmals im Jahr 1943 publiziert. Der Test soll es geübten Diagnostikerinnen und Diagnostikern ermöglichen, einige der vorherrschenden Triebe, Gefühle, Gesinnungen, Komplexe und Konflikte der Testperson zu erkennen. Theoretischer Hintergrund ist die Persönlichkeitstheorie von Murray, der zufolge das Verhalten stark durch Motive (needs) und Zwänge der Umwelt (presses) determiniert wird. Der TAT besteht aus 31 Bildtafeln, die grundlegende menschliche Problemsituatien ansprechen (Abb. 3.25). Eine Tafel ist völlig weiß, also ohne jedes

Theoretischer Hintergrund:
Persönlichkeitstheorie von Murray



Abb. 3.25 Tafel aus dem TAT. (Aus Murray 1936, THEMATIC APPERCEPTION TEST by Henry A. Murray, Card 12 F, Cambridge, Mass.: Harvard University Press, Copyright © 1943 by the President and Fellows of Harvard College, Copyright © renewed 1971 by Henry A. Murray)

Bild. Es handelt sich dabei noch immer um die alten Schwarz-weiß-Originale; allerdings wurden auch diverse TAT-Varianten mit zum Teil anderen Bildern entwickelt. Im deutschen Sprachraum wurde der Test vor allem durch Revers (1958) bekannt.

Dramatische Geschichte zu jedem Bild erzählen

Durchführung und Auswertung Der Test wird in 2 etwa 1-stündigen Sitzungen appliziert, in denen je etwa 10 Tafeln vorgelegt werden, die nach der Fragestellung unter Berücksichtigung von Alter und Geschlecht der untersuchten Person auszuwählen sind. Auf der Rückseite der Tafel ist vermerkt, ob das Bild für Männer, Frauen, Jungen oder Mädchen geeignet ist. Die Testperson wird aufgefordert, zu jedem Bild eine möglichst dramatische Geschichte zu erzählen. Darin soll enthalten sein: Was führte zu der gezeigten Situation? Was geschieht gerade? Was fühlen und denken die Personen? Wie geht die Geschichte aus? Die Testleiterin oder der Testleiter hat die Antworten zu protokollieren und bei Bedarf an die Instruktion zu erinnern. In einer Nachbefragung sollen die angesprochenen Themen und Konflikte, deren Vorgeschichte sowie ihre weitere Entwicklung herausgearbeitet werden. Für Murray stand eine Satz-für-Satz-Auswertung im Mittelpunkt, die nach den Kräften und Aktivitäten fragte, die entweder von „Heldinnen“ bzw. „Helden“ der Geschichte ausgehen oder auf diese wirken („needs“ bzw. „presses“). Es gibt jedoch verschiedene Auswertungsmethoden, deren Verwendung von der verfügbaren Zeit, dem Quantifizierungsanspruch der Testleiterin bzw. des Testleiters, den unterlegten Persönlichkeitstheorien, Rahmenbedingungen und anderen Faktoren abhängt.

Objektivität problematisch

Objektivität Ist die Objektivität der Testdurchführung schon wegen der Freiheit bei der Auswahl der Tafeln infrage gestellt, so birgt die für die Testperson ungewohnte Aufgabenstellung und die Enge des Kontakts die Gefahr, dass sie versucht, aus dem verbalen und nonverbalen Verhalten der Testleiterin oder des Testleiters Hinweise für ihre Antworten zu erlangen. Dadurch kann der Testleitereffekt verstärkt werden. Orientiert man sich an den Werken

von Murray (1991) oder Revers (1958), so gibt es keine hinreichend standariserte Auswertung. In der Praxis wird der TAT meist intuitiv ausgewertet. Normen, die helfen würden, die Häufigkeit bestimmter Aussagen oder Themen, die in den Geschichten anzutreffen sind, einzuordnen, sind nicht vorhanden. Damit ist die Interpretationsobjektivität fraglich.

Allerdings wurden für Forschungszwecke Auswertungsmethoden zur Erfassung des Leistungs-, Gesellungs- und Machtmotivs entwickelt (für eine Anwendung und Literaturhinweise s. Langan-Fox und Grant 2006). Zur Auswertung des TAT steht ein detailliertes Manual von Winter (1994) zur Verfügung, anhand dessen Kodiererinnen und Kodierer bei entsprechendem Training zu übereinstimmenden Auswertungen gelangen können (z. B. Schultheiss und Brunstein 2001).

Reliabilität Für die Berechnung der Reliabilität ergeben sich infolge der Heterogenität der Tafeln und wegen des Fehlens einer Parallelserie ganz ähnliche Probleme wie beim Rorschach-Test. Das Gleiche gilt auch für die Erinnerungseinflüsse bei einer Testwiederholung. Der TAT wurde gerne in der Leistungsmotivationsforschung eingesetzt. In diesem Kontext wurden Reliabilitätschätzungen für Kennwerte der Leistungsmotivation durchgeführt. Die Werte für die interne Konsistenz lagen selten über .30 oder .40 (Entwistle 1972). Allerdings ergab eine Berechnung der internen Konsistenz anhand von Testmodellen, die einen dynamischen Verlauf von Motiven über die Testdauer annehmen, durchaus zufriedenstellende Kennwerte (Lang 2014). Der Autor nahm einen dynamischen Verlauf von Motiven an, da mit einem Testitem ein Motiv zunächst befriedigt werden kann und dann für nachfolgende Items so lange nicht zum Tragen kommt, bis sich das Motivpotenzial wieder aufgebaut hat. Für die Retest-Reliabilität ermittelte Fineman (1977) einen mittleren Wert (Median) von .32.

Für Leistungsmotivation niedrige Reliabilität ermittelt

Validität Zur Leistungsmotivation liegen ebenfalls viele Validitätsuntersuchungen (meist älteren Datums) vor. Spangler (1992) führte eine Metaanalyse über 105 Validitätsstudien durch und verglich die Validität des TAT mit der von Fragebögen zur Leistungsmotivation. TAT und Fragebögen korrelieren extrem niedrig ($r=.09$), was zu der Annahme passt, dass mit dem TAT eher implizite (unbewusste) Motive und mit Fragebögen explizite Motive erfasst werden, die unkorreliert sein können (McClelland et al. 1989). Natürlich ist angesichts dieser Ergebnisse auch die Schlussfolgerung möglich, dass die Konstruktvalidität nicht belegt ist. Bei den Kriterien, die Spangler (1992) zur Validierung heranzog, handelte es sich zum Teil um objektive Indikatoren von Leistung im Beruf wie etwa den Umsatz, den Farmer erzielen. Während Fragebögen durchschnittlich zu $r=.13$ mit solchen „harten“ Kriterien korrelierten, betrug die vergleichbare Korrelation des TAT immerhin $r=.22$.

Metaanalyse zur Messung des Leistungsmotivs

Auf den ersten Blick ist dies ein Beleg für die Kriteriumsvalidität des TAT. Allerdings ist zu bedenken, dass der Leistungsmotivationskennwert mit dem Umfang der erzählten Geschichten zunimmt und schwach mit Intelligenz korreliert. Deshalb wäre es sinnvoll, den Zusammenhang zwischen TAT und Berufserfolg für Intelligenz zu kontrollieren, was aber nicht geschehen ist. Entwistle (1972) hat bereits 1972 zahlreiche Belege dafür gesammelt, dass der Leistungsmotivwert deutlich mit der Länge der produzierten Geschichten zusammenhängt und dass die verbale Produktivität höher mit Schulleistungen korreliert als der eigentliche Motivwert. Die verbale Produktivität hängt

Länge der Geschichten kann Validitätsbefunde erklären

weiterhin mit Intelligenz zusammen; wird der Zusammenhang zwischen Motivwert und Schulleistung für Intelligenz kontrolliert, sagt die Motivstärke die Schulleistung nicht mehr vorher.

Weitere Metaanalyse

3

Eine Berufsgruppe, bei der Leistungsmotivation besonders zum beruflichen Erfolg beitragen sollte, sind Entrepreneurs. In einer Metaanalyse von Collins et al. (2004) wurden nach einer engen Definition darunter Menschen verstanden, die ein Unternehmen gegründet haben. Die breitere Definition schließt dagegen auch Menschen ein, die eigenständig ein Unternehmen führen. Allerdings erwies sich die Definition nicht als relevanter Moderator, sodass die Ergebnisse zusammengefasst wurden. Wenn also jemand einen Betrieb führt, spielt es demnach für die Bedeutung der Leistungsmotivation keine Rolle, ob er oder sie den Betrieb etwa geerbt oder selbst gegründet hat. Die Metaanalyse stellt eine wichtige Ergänzung zu der oben vorgestellten Metaanalyse von Spangler (1992) dar, in der nur etwa 4 % der Studien Entrepreneurs einschlossen. Collins et al. (2004) untersuchten den Zusammenhang zwischen TAT bzw. Fragebögen zur Leistungsmotivation und zum Berufserfolg sowie zur Karriereentscheidung (selbstständig oder angestellt). Ihre Hypothese, dass der TAT diesbezüglich überlegen ist, konnten sie nicht bestätigen. Die Korrelation zwischen TAT und Berufserfolg lag bei $r=.16$ (8 Studien, $N=915$), die entsprechende Korrelation für Fragebögen lag bei $r=.19$ (5 Studien, $N=803$). Das Gleiche gilt für die Karriereentscheidung; TAT und Fragebögen wiesen mit $r=.20$ die gleiche Korrelation auf (10 bzw. 26 Studien, $N=907$ bzw. 4424). Damit wird erneut deutlich, dass der TAT als Maß der Leistungsmotivation eine niedrige Kriteriumsvalidität hat. Allerdings gilt diese Aussage auch für Fragebögen.

Metaanalyse zur Validität bei Messung des Macht- und Affiliationsmotivs

Der TAT wurde nicht nur zur Messung des Leistungsmotivs eingesetzt, sondern auch für das Macht- und das Affiliationsmotiv. In einer Metaanalyse zum TAT und verwandten Verfahren haben Köllner und Schultheiss (2014) den Zusammenhang mit Fragebögen ermittelt, die die gleichen Merkmale messen sollen. Zunächst konnten sie den Befund von Spangler (1992) zum Leistungsmotiv replizieren, dass der Zusammenhang zwischen TAT und Fragebögen extrem schwach ist ($r=.07$ bzw. .12 nach Korrektur für Mess- und Stichprobenfehler). Für das Affiliationsmotiv ($r=.08$ bzw. korrigiert .12) und für das Machtmotiv ($r=.06$ bzw. korrigiert .04) fanden sie ebenfalls sehr niedrige Korrelationen mit entsprechenden Fragebögen.

Angesichts dieser Werte sind, wie bereits zuvor erwähnt, folgende Schlussfolgerungen möglich: Entweder ist die Konstruktvalidität des TAT, zumindest bei der Messung dieser 3 Motive, nicht belegt (für Fragebögen liegen Belege zur Konstruktvalidität vor; ► Abschn. 3.3.3). Oder TAT und Fragebögen messen unterschiedliche Phänomene. Letzteres wird von vielen Forscherinnen und Forschern, die mit dem TAT arbeiten, angenommen. Ihre Erklärung klingt ganz einfach: Mit den Fragebögen wird das gemessen, was den befragten Menschen leicht zugänglich ist und was sie leicht erinnern können. Mit dem TAT dagegen – so die Erklärung – werden implizite, dem Bewusstsein nicht zugängliche Motive erfasst (vgl. McClelland et al. 1989).

Annahme impliziter Motive

TAT ökonomischen Verfahren überlegen?

Auch wenn sich der TAT für eine bestimmte Fragestellung als valide erweisen sollte, stellt sich die Frage, ob das Gleiche nicht auch mit einem einfach handhabbaren, ökonomischen Test zu erreichen wäre. Eine diesbezüglich eindrucksvolle Untersuchung stammt von Wildman und Wildman (1975). Klinischen Psychologinnen und Psychologen wurde die Aufgabe gestellt, u. a. anhand von MMPI- und/oder TAT-Ergebnissen festzustellen, ob es sich

bei der Testperson um eine psychiatrische Patientin oder um eine Krankenschwesternschülerin handelt. Mit dem TAT alleine erzielten sie eine Trefferquote von 57 %, während sie mithilfe des MMPI 88 % richtig klassifizierten – bei blindem Raten wären 50 % Treffer zu erwarten gewesen. Wurden den Expertinnen und Experten TAT und MMPI zur Verfügung gestellt, verbesserte sich die Trefferquote nicht, sondern sie wurde schlechter; sie betrug nur noch 80 %.

Fazit Der TAT stellt in Kombination mit dem Handbuch von Murray oder von Revers (1958) kein hinreichend objektives Verfahren dar; die Anweisungen zur Auswertung und Interpretation sind nicht eindeutig und präzise genug, um zu gewährleisten, dass verschiedene Auswerterinnen und Auswerter zum gleichen Resultat gelangen. Er kann bei aufwendigem Training der Kodiererinnen und Kodierer unter Verwendung des Manuals von Winter (1994) hinreichend objektiv ausgewertet werden. Die Befunde zur Reliabilität variieren deutlich und bedürfen weiterer Klärung. Der stärkste Validitätsbeleg ist die Metaanalyse von Spangler (1992), in der sich der TAT Persönlichkeitsfragebögen als überlegen erwies. Die Ergebnisse werden aber durch die Metaanalyse von Collins et al. (2004) relativiert, die einen „Gleichstand“ von TAT und Fragebögen dokumentiert.

Psychometrisch problematisches Verfahren

3.5.2 Abgeleitete Testprinzipien und semiprojektive Tests

Nach der bisherigen Kritik an den mangelnden Testeigenschaften des TAT stellt sich die Frage, ob das Verfahren so verbessert werden kann, dass es testdiagnostischen Mindestkriterien genügt. Folgende Strategien bieten sich an:

- Begrenzung auf ein Motiv in Verbindung mit einem präzisen Auswertungsschema. Diesen Ansatz finden wir beim Leistungsmotivations-TAT (► Abschn. 3.5.2.1).
- Modularer Aufbau mit Bildern, die für bestimmte Motive gut geeignet sind; die Auswahl erfolgt nach Bedarf. Auch hier ist ein präzises Auswertungsschema erforderlich. Manche Picture Story Exercises (PSE; ► Abschn. 3.5.2.2) entsprechen diesen Anforderungen.
- Anstatt Geschichten frei erzählen zu lassen, gibt es eine Auswahl von vorgegeben Antwortmöglichkeiten. Die Auswertung wird dadurch wesentlich vereinfacht, weil feststeht, welche Antwort für welches Motiv steht. Dieses Vorgehen wird als semiprojektive Tests bezeichnet. Wir stellen dazu kurz das Multi-Motiv-Gitter (MMG) vor (► Abschn. 3.5.2.3).

3.5.2.1 Leistungsmotivations-TAT

Anstelle der Messung eines sehr breiten Spektrums an Persönlichkeitseigenschaften könnte man sich auf ein einziges Merkmal konzentrieren und dafür geeignete Bilder auswählen. Die Auswertung, die sich nun auf ein Merkmal beschränkt, wäre durch präzise Handlungsanweisungen leichter zu standardisieren und einfacher zu handhaben als bei einem Test mit vielen Motiven. Heckhausen (1963) hat sich daher auf die Messung des Leistungsmotivs konzentriert. Dieses Verfahren soll hier exemplarisch dargestellt werden, um das Potenzial, das der TAT hat, genauer zu betrachten. Beim Leistungsmotivations-TAT werden statt vieler verschiedener menschlicher Problemsituationen nur 6 Leistungssituationen gezeigt (► Abb. 3.26); die Instruktion ist der des TAT sehr ähnlich.

Leistungsmotivations-TAT

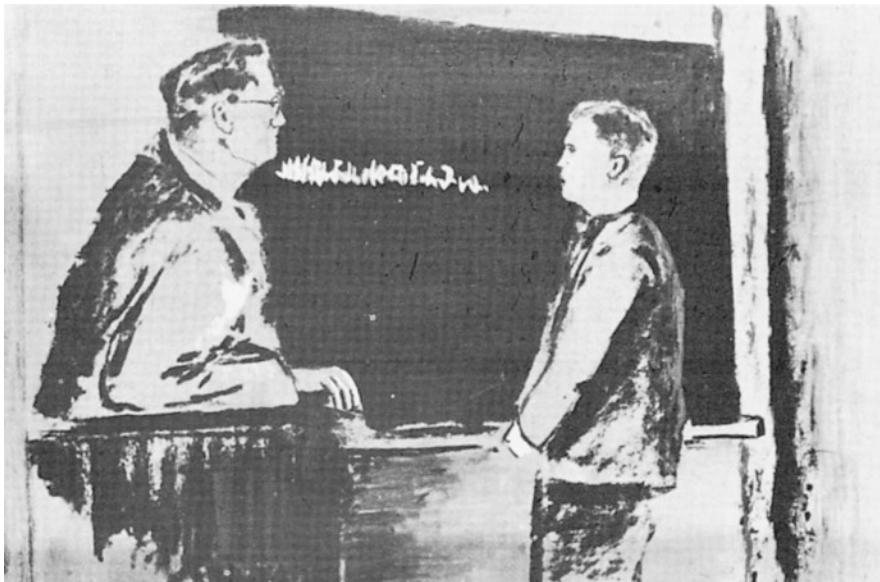


Abb. 3.26 Bild aus dem Leistungsmotivations-TAT (Heckhausen 1963)

Standardisierte inhaltsanalytische Auswertung

Bei der Auswertung werden jedoch die Differenzen deutlich. Die Signierung erfolgt anhand eines detaillierten Inhaltsschlüssels, der eine Zuordnung der bei offener Beantwortung vielfältigen verbalen Reaktionen und deren Quantifizierung ermöglicht. Als Ergebnis werden Kennwerte für „Hoffnung auf Erfolg“ und „Furcht vor Misserfolg“ ermittelt, deren Summe einen Wert für Gesamtmotivation und deren Differenz einen Kennwert für Nettohoffnung ergeben. An 236 Schülerinnen und Schülern an Grund- und Handelschülern sowie 251 Studierenden wurden Normwerte erhoben.

Moderate Reliabilität

Die Übereinstimmung von 2 geschulten Auswerterinnen bzw. Auswertern erreichte einen Wert von .84. Die interne Konsistenz der Bilderserie war mit .66 relativ niedrig, was aber bei nur 6 Items nicht überraschen kann. Bei der Testwiederholung nach 5 Wochen ergaben sich Retest-Koeffizienten von .42 für Hoffnung auf Erfolg und .59 für Furcht vor Misserfolg, die die Problematik einer Testwiederholung bei derartigen Verfahren noch einmal verdeutlichen. Die Studienleistungen von Studierenden korrelierten mit ihrer Erfolgszuversicht in Höhe von .56. Dieser Validitätskoeffizient ist ungewöhnlich hoch – in vielen anderen Studien zeigte Leistungsmotivation keinen derart starken Zusammenhang mit Studienerfolg.

3.5.2.2 Picture Story Exercises (PSE)

Die allgemeinere Methode, zu der auch der TAT gezählt wird, ist als PSE bekannt. Sie ist kein standardisiertes diagnostisches Verfahren und schon gar nicht für die Einzelfalldiagnostik geeignet, sondern ein Testprinzip, das unterschiedlich ausgestaltet sein kann. Das Prinzip besteht darin, den Testpersonen wie beim TAT ausgewählte Bilder zu zeigen (Abb. 3.27) und diese dazu Geschichten schreiben zu lassen. Die Antworten werden inhaltsanalytisch ausgewertet.

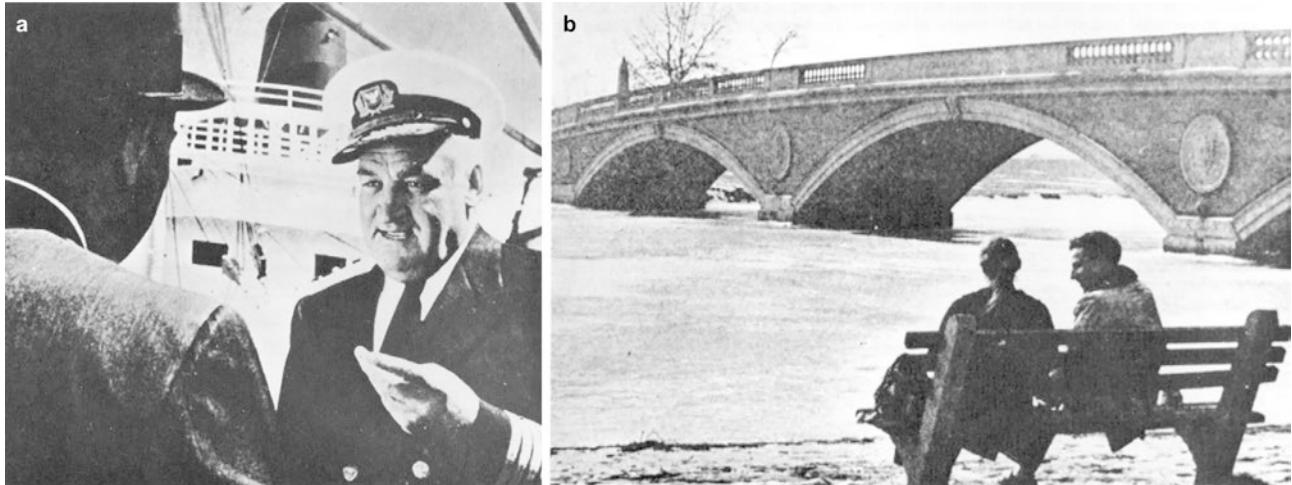


Abb. 3.27 Ausgewählte Bilder, bei denen deutsche Testpersonen ($n=428$) die höchsten Ausprägungen für ein Motiv zeigten (Kapitänen: Macht; Paar vor Brücke: Affiliation). (Nach Pang 2010, © Oxford University Press)

Durchführung der PSE

Pang (2010) zufolge sind folgende Schritte obligatorisch:

- Auswahl von idealerweise 5–8 Bildern nach verschiedenen Kriterien: Bei Pang (2010) finden sich in tabellarischer Form Angaben dazu, wie stark bestimmte Bilder bei verschiedenen Personenstichproben die Motive „Leistung“, „Macht“ und „Affiliation“ anregen.
- Abfolge der Bilder festlegen: Idealerweise wird diese für jede Testperson randomisiert.
- Präsentation am besten auf einem Computerbildschirm: Die Bilder werden für 10–15 s eingeblendet, dann folgt ein dunkler Bildschirm.
- Standardisierte Instruktion: Pang (2010) macht einen Vorschlag dazu; wichtige Elemente: Jedes Bild ist 10 s lang zu sehen, dann hat die Testperson 5 min Zeit, eine imaginäre Geschichte dazu zu schreiben, die einen Anfang, eine Mitte und ein Ende hat. Die Leute auf dem Bild sollen beschrieben werden – was fühlen, denken und wünschen sie? Was führte zu der Situation und wie endet alles?
- Die Geschichten können handschriftlich oder auch über die Computertastatur aufgezeichnet werden.
- Eventuell sind unbrauchbare Geschichten (z. B. weniger als 30 Wörter) auszuschließen.
- In den Protokollen müssen alle Informationen eliminiert werden, die Rückschlüsse auf die Person oder das jeweilige Bild zulassen.
- Unten den zahlreichen Kodierungssystemen ist ein passendes auszuwählen; dieses wird von zuvor trainierten Beurteilerinnen und Beurteilern angewendet. Diese benötigen etwa 2–5 min pro Geschichte, die normalerweise aus etwa 100 Wörtern besteht.

Es ist nicht möglich, hier die gesamte Literatur zu den Gütekriterien aufzuarbeiten. Exemplarisch wird kurz eine Studie zur Reliabilität vorgestellt (eine weitere zur Validität folgt im Anschluss an das MMG; ▶ Abschn. 3.5.2.3). Schultheiss et al. (2008) haben die Auswerterübereinstimmung und die Reliabilität einer PSE-Variante mit 8 Bildern untersucht. Die Bilder wurden nach

Objektivität und Reliabilität der PSE

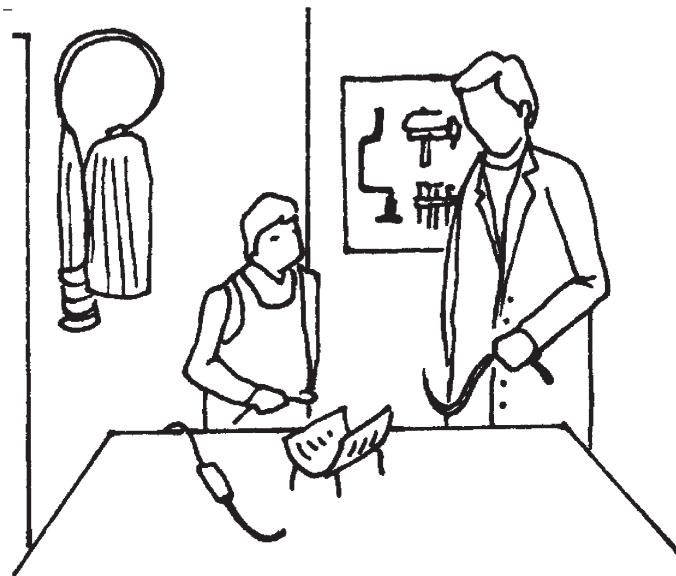
den oben vorgestellten Standards je 10 s auf einem Bildschirm gezeigt. Nach 4 min erschien die Aufforderung auf dem Bildschirm, nun fertig zu werden. Das Prozedere wurde nach 2 Wochen wiederholt. Verwertbar waren am Ende die Daten von 90 studentischen Versuchspersonen. Zwei trainierte Personen werteten die Geschichten für die Motive „Macht“, „Affiliation“ und „Leistung“ aus. Dabei kam das Auswertungsschema von Winter (1994) zum Einsatz. Affiliation wurde etwa kodiert, wenn sich die Hauptperson der Geschichte um die Herstellung, die Aufrechterhaltung oder die Wiederherstellung einer freundschaftlichen Beziehung kümmert. Dies kann sich durch das Zeigen positiver Gefühle anderen gegenüber, Traurigkeit über eine Trennung, gesellige Aktivitäten oder fürsorgliche Verhaltensweisen äußern.

Die Auswertungsübereinstimmung lag bei beiden Sitzungen für alle 3 Motive zwischen .70 und .87. Die interne Konsistenz fiel erwartungsgemäß sehr niedrig aus, weil nur wenige und zudem sehr heterogene Items verwendet wurden ($\alpha = -.02$ bis .43). In der in ▶ Abschn. 3.5.2.3 vorgestellten Validitätsstudie von Schüler et al. (2015) wurden ebenfalls niedrige Werte ermittelt ($\alpha = .35$ bis .44). Die Retest-Reliabilität lag bei .37 (Leistung), .39 (Macht) bzw. .61 (Affiliation).

3.5.2.3 Multi-Motiv-Gitter (MMG)

MMG als Beispiel für ein semiprojektives Verfahren

Eine weitere Alternative sind semiprojektive Verfahren. Die Antworten auf mehrdeutige Reize werden hier hoch standardisiert erfasst und ausgewertet. So werden beim MMG (Schmalt et al. 2000) Situationen skizzenhaft dargestellt (► Abb. 3.28), und die Testperson muss lediglich ankreuzen, welche der vorgegebenen Aussagen auf diese Situation zutrifft. Mit dem Verfahren soll die Ausprägung des Macht-, Leistungs- und des Anschlussmotivs erfasst werden. Die möglichen Antworten passen zum Teil auf eines oder mehrere dieser Motive. Die Auswertung besteht darin, die Anzahl der zu jedem der Motive passenden Antworten auszuzählen. Durchführung und Auswertung sind also genauso hoch standardisiert wie bei Persönlichkeitsfragebögen; das Verfahren ist sogar in einer Computerversion verfügbar. Gegenüber den meisten anderen projektiven Verfahren hat es den Vorteil, dass es normiert ist.



■ Abb. 3.28 Item aus dem MMG. (Aus Schmalt et al. 2000, © 2011, 2., unveränderter Nachdruck Pearson Assessment & Information GmbH, Frankfurt am Main)

Diagnostische Verfahren

Bereits in ▶ Abschn. 2.6.3.2 haben wir eine Studie dargestellt, die der Frage nachgegangen ist, ob das MMG auch ohne Bilder zu ähnlichen Ergebnissen führt wie mit Bildern (in ▶ Abschn. 2.6.3.2 haben wir diese Frage größtenteils bejaht; vgl. Krumm et al. 2016).

Mit der Validität der PSE und des MMG befasste sich auch eine Untersuchung von Schüler et al. (2015). Diese Studie zeichnet sich dadurch aus, dass neben den beiden (semi-)projektiven Verfahren noch ein weiteres, der Operante Multi-Motive-Test (OMT; Kuhl und Scheffer 1999), zum Einsatz kam, sodass alle 3 Verfahren die gleichen Motive messen. Zudem bearbeiteten die über 200 studentischen Versuchspersonen 3 Fragebögen zu diesen Motiven, von denen die (deutsche) Personality Research Form (PRF; Stumpf et al. 1985) das bekannteste und am besten erforschte Fragebogenverfahren ist.

Die 6 Bilder der PSE wurden je 15 s dargeboten, die Bearbeitungszeit und auch die Auswertung wurde wie bei Schultheiss et al. (2008; ▶ Abschn. 3.5.2.2) gehandhabt. Im OMT (Kuhl und Scheffer 1999) werden mehrdeutige Strichzeichnungen als Stimuli verwendet. Die Testpersonen sollen zu jedem der verwendeten 15 Bilder auf 4 Fragen stichpunktartig antworten. Die Fragen lauten: „Was ist für diese Person in dieser Situation wichtig und was tut sie? Wie fühlt sich diese Person? Warum fühlt die Person sich so? Wie geht die Geschichte aus?“ Das MMG wurde oben bereits beschrieben. Zu sämtlichen Verfahren liegen mehrere Validitätsstudien vor (s. dazu Schüler et al. 2015).

In □ Tab. 3.27 sind die Ergebnisse exemplarisch für das Leistungsmotiv dargestellt. Der □ Tab. 3.27 ist im oberen Dreieck zu entnehmen, dass die 3 (semi-)projektiven Verfahren um 0 und zudem nicht signifikant miteinander korrelieren. Mit dem Fragebogen (untere Zeile) finden sich erwartungsgemäß ebenfalls nur extrem niedrige und nicht signifikante Zusammenhänge. Die konvergente Validität der 3 hier zur Diskussion stehenden Tests wurde damit nicht einmal annäherungsweise bestätigt.

Betrachten wir nun die diskriminante Validität. Die 3 Motivskalen sollten innerhalb der einzelnen Verfahren niedrig miteinander korrelieren, weil sie unterschiedliche Motive erfassen sollen. Die Skaleninterkorrelationen sollten somit niedriger ausfallen als die zur konvergenten Validität. Diese Annahme lässt sich nicht bestätigen. Insbesondere beim MMG waren moderate und signifikante Skaleninterkorrelationen zu verzeichnen ($r=.40$ bis $.44$). Für die beiden anderen Verfahren lagen die Werte deutlich niedriger ($r=-.04$ bis $.24$). Nun bleibt noch die Frage zu klären, ob sich auch bei den anderen Motiven so niedrige konvergente Validitäten ergaben. Die Korrelationen (die denen im oberen Dreieck von □ Tab. 3.27 entsprechen), lagen beim Affiliationsmotiv zwischen $-.02$ und $.14$ und beim Machtmotiv zwischen $.06$ und $.15$. Der Zusammenhang mit den entsprechenden Skalen der PRF lagen mit $r=.02$ bis $.16$ im gleichen extrem niedrigen Bereich. Damit liefert diese Studie eine

3 (semi-)projektive Verfahren im Vergleich

Durchführung

Ergebnisse zum Leistungsmotiv

Unerwartet schlechte Ergebnisse zur Konstruktvalidität von 3 (semi-)projektiven Verfahren

□ Tab. 3.27 Interkorrelation von 3 (semi-)projektiven Verfahren und einem Fragebogen zum Leistungsmotiv bei Schüler et al. (2015)

Verfahren	PSE	OMT	MMG
Picture Story Exercise (PSE)			
Operanter Multi-Motive-Test (OMT)	.12		
Multi-Motiv-Gitter (MMG)	-.12	-.04	
Personality Research Form (PRF)	.01	-.05	-.03

Quelle: Auszug aus Tab. 2 von Schüler et al. (2015); die Korrelationen sind nicht signifikant

wichtige Erkenntnis. Wurden bisher die sehr niedrigen Korrelationen zwischen „impliziten“ und per Fragebogen erfassten Motiven als theoretisch begründet angesehen, so muss nun festgestellt werden, dass auch die impliziten Motivtests nur in der gleichen Größenordnung korrelieren. Die Autorinnen und Autoren der Studie sprechen von unerwarteten Ergebnissen, betonen aber, dass es sich nur um die Befunde einer einzigen Studie handelt.

3.5.3 Zeichnerische und Gestaltungsverfahren

Aus der Gruppe der zeichnerischen- und Gestaltungsverfahren soll mit der „Familie in Tieren“ ein in der Praxis beliebtes Verfahren (Bölte et al. 2000) vorgestellt werden, das sich aber als psychometrisch unzulänglich erweist.

Eigene Familie als Tiere malen

Familie in Tieren Das Verfahren „Familie in Tieren“ (Brem-Gräser 2001) wurde für Kinder konzipiert. Zur Durchführung benötigt man lediglich ein Blatt Papier (keine Angabe zur Größe) und einen Stift (keine Angabe ob Bleistift, Buntstifte – und wenn ja, welche Farben und welche Art). Das Kind wird aufgefordert, sich seine Familie als Tiere vorzustellen und sie zu malen. Schon die Durchführung ist nicht standardisiert, da wichtige Angaben zum Testmaterial fehlen (welche Größe des Papiers, welche Stifte in welcher Farbe?).

Zweifelhafter theoretischer Hintergrund

Der theoretische Hintergrund ist deutlich zu kritisieren, beispielsweise werden Mythen und Redewendungen zur Interpretation der Tierfiguren bemüht. Über die Schlange (die Geschwister werden sehr oft als Schlange dargestellt, aber auch für Vater und Mutter verwenden Kinder gerne die Schlange) sei bekannt, dass sie das Prinzip des Bösen verkörpert; es heißt aber auch, „seid klug wie eine Schlange“; auch ein höheres Wissen und die Kraft zur Heilung sei ihr eigen etc. Immerhin findet man im Manual Angaben, wie häufig bestimmte Tiere für Vater, Mutter, Bruder und Schwester gewählt werden; die Daten stammen aus der Analyse von 2000 Familiendarstellungen. So erfahren wir, dass die Schlange insgesamt am häufigsten vorkommt, der Vater aber auch oft als Pferd oder Elefant und die Mutter oft als Vogel oder Hase gemalt wird. Diese Angaben dürfen nicht als Normtabellen verstanden werden; sie sagen lediglich aus, wie gewöhnlich oder ungewöhnlich die Wahl eines bestimmten Tieres ist. Für die Interpretation des Charakters eines Tieres und damit der Person, die es verkörpert, finden sich vage Hinweise (der Vogel sei z. B. „das eigentliche Luftwesen, der Götterbote“), aber auch konkrete Charaktereigenschaften (z. B. beschwingt, lustig, rege, schwankend, kleinmütig, frech). Eine Hilfestellung, wie man die passende Eigenschaft auswählt, sucht man vergebens. Die Auswertungsobjektivität ist nicht gegeben, da es keine genaue Anleitung gibt.

Die Zeichnungen sollen inhaltlich und formal gedeutet werden. Zur inhaltlichen Auswertung gehört es etwa, herauszufinden, welche Eigenschaften die einzelnen Tiere für das Kind verkörpern. Ein Kind malt sich beispielsweise als Katze und den Bruder als Rabe. Die Autorin deutet dies in einer Fallbesprechung (Nr. 20) als Feindschaft, da die Katze den Vogel jagt und frisst. Neben den Charakteren der Tiere können auch die Größenverhältnisse und die räumliche Anordnung der Familienmitglieder oder etwa die Beobachtung, dass jemand vergessen wurde (manchmal das Kind selbst – aber hat es die Instruktion vielleicht falsch verstanden?) interpretiert werden. Je- denfalls kann man den Fallbeispielen entnehmen, dass solche Interpretationen offenbar zulässig sind. Zur formalen Analyse gehört etwa die Beurteilung

der Strichstärke und die Gestaltung der Flächen – Schraffierung weise darauf hin, dass sich die Verstands- und die Antriebsbeteiligung die Waage halten. Angaben zur Reliabilität und Validität fehlen.

Die eigene Familie als Tiere zeichnen zu lassen, kann bei Kindern, die gerne malen, ein guter Einstieg in ein diagnostisches Interview sein. Zeichnungen der Familie als Tiere bergen möglicherweise diagnostische Informationen über die Eigenschaften, die das Kind den einzelnen Mitgliedern zuschreibt, über das Verhältnis der Familienmitglieder untereinander, über Konflikte etc. Damit solche Informationen jedoch verwertbar sind, müssen sie aber unter besser standardisierten Bedingungen erhoben, nach genauen Anweisungen ausgewertet, in Kennwerte überführt und schließlich mithilfe von Vergleichswerten (Normen) interpretiert werden. Schließlich sind die Gütekriterien des Verfahrens zu bestimmen. Die „Familie in Tieren“ erfüllt diese Voraussetzungen nicht.

Psychometrisch unzulängliches Verfahren

Weiterführende Literatur

Lilienfeld et al. (2000) haben eine sehr lesenswerte, kritische Auseinandersetzung mit projektiven Verfahren, insbesondere dem Rorschach-Test und dem TAT, vorgelegt, in der sie sich auch ausführlich mit der Reliabilität und Validität dieser Verfahren befassen. Dieser viel zitierte Beitrag hat auch Kritik hervorgerufen, so etwa von Hibbard (2003). In einer Replik setzen sich Garb et al. (2005) sehr differenziert mit der Kritik auseinander. Einen eher pro Rorschach verfassten Übersichtsbeitrag hat Mattlar (2004) vorgelegt.

Wir wagen es, zum Rorschach-Test auch ein populärwissenschaftliches Buch aufzuführen, das immerhin in der *New York Times* rezensiert worden ist (Whippman 2017). Das Buch von Searls (2017) ist 2019 unter dem Titel *Im Auge des Betrachters: Hermann Rorschach und sein bahnbrechender Test* auch als deutsche Ausgabe bei Kindler erschienen ist.

?

Übungsfragen

— Abschn. 3.5:

- Welche 3 Gemeinsamkeiten weisen projektive Tests mit objektiven Persönlichkeitstests auf?
- Nennen Sie die 3 Merkmale des klassischen Projektionsbegriffs!
- Was versteht man in einer verallgemeinerten Version unter Projektion, und welche Konsequenzen ergeben sich daraus für das Testmaterial?
- Nennen Sie Kategorien, in die man projektive Tests unterteilen kann!
Nennen Sie jeweils ein Verfahren als Beispiel!
- Wodurch unterscheiden sich semiprojektive von projektiven Tests?
- Als was wurde der Rorschach-Test ursprünglich entwickelt?
- Nach welchen 4 Aspekten wird der Rorschach-Test in der klassischen Variante ausgewertet?
- Wie gut ist der Rorschach-Test geeignet, psychiatrische Störungen zu erkennen?
- Wie wird der TAT durchgeführt (Testmaterial, Instruktion)?
- Was soll mit dem TAT gemessen werden?
- Welche Ergebnisse brachte eine Metaanalyse zur Validität des TAT im Kontext von Leistungsmotivation (Spangler 1992)? Warum sind die Ergebnisse kritisch zu hinterfragen?
- Beschreiben Sie das Multi-Motiv-Gitter (MMG) und beurteilen Sie Befunde zu dessen Validität!
- In der Studie von Schüler et al. (2015) bearbeiteten die Testpersonen ein projektives (PSE), 2 semiprojektive (OMT, MMG) Tests und einen Fragebogen (PRF). Welche Ergebnisse fanden Schüler et al. zum Leistungsmotiv und wie interpretieren sie diese?
- Wie wird der Test „Familie in Tieren“ durchgeführt, und wie ist dessen psychometrische Qualität zu bewerten?

3.6 Verhaltensbeobachtung und -beurteilung

Die Verhaltensbeobachtung dient der Beschreibung des Verhaltens einer oder mehrerer Personen. Sie wird in der diagnostischen Praxis sehr oft eingesetzt – und sei es nur, um das Verhalten einer Person bei der Durchführung eines Tests oder während eines diagnostischen Interviews zu beschreiben. Sie kann bei der Beobachtung des Verhaltens bei der Testbearbeitung ergänzende Informationen liefern, die eventuell das Testergebnis relativieren. Manchmal hat sie eine herausragende Bedeutung, weil andere Informationsquellen nicht verfügbar oder unergiebig sind, etwa wenn ein Kindergartenkind ein gestörtes Sozialverhalten zeigt, aber keine der Bezugspersonen im Interview dazu differenzierte Angaben machen kann. Eine Verhaltensbeobachtung liefert im Idealfall Erkenntnisse, die weitgehend frei von subjektiven Bewertungen sind.

Verhalten beschreiben vs. beurteilen

Eine Verhaltensbeurteilung beinhaltet immer eine Beurteilung, eine Interpretation. Sie setzt eine Verhaltensbeobachtung voraus, geht aber über das Beobachtete hinaus. Beispielsweise schreibt eine Psychologin oder ein Psychotherapeut in einem Bericht, dass sie bzw. er das untersuchte Kind beim freien Spiel im Kindergarten beobachtet hat: „Das Kind nahm einem gleichaltrigen Mädchen zweimal das Spielzeug weg und schubste einen ein Jahr jüngeren Jungen so stark, dass dieser hinfiel.“ Dies ist ein Auszug aus dem Protokoll einer Verhaltensbeobachtung. Die Aussage, dass sich das beobachtete Kind aggressiv verhielt, wäre dagegen eine Verhaltensbeurteilung.

Verhaltensbeurteilung mit und ohne Begründung

Eine Verhaltensbeurteilung kann sich an eine Beobachtung anschließen. Die im Beobachtungsprotokoll gesammelten Fakten werden zwecks Beantwortung einer Fragestellung interpretiert. In diesem Fall muss die Interpretation nachvollziehbar begründet sein. Die Verhaltensbeurteilung wird aber auch als eigenständige Methode angesehen. Sie kann ohne explizite Begründung der Schlussfolgerungen stattfinden. Typischerweise geschieht dies mithilfe von Ratingskalen. Die Beurteilerin oder der Beurteiler kreuzt etwa an, wie kooperativ sich die beobachtete Person in einer Assessment-Center-Übung verhalten hat.

Definition

Eine **Verhaltensbeobachtung** ist die freie oder systematische Beobachtung des Verhaltens einer Person oder mehrerer Personen in einer näher bestimmten Situation. Die Beobachtungen werden so weit wie möglich frei von Wertungen und Schlussfolgerungen in geeigneter Form protokolliert.

Bei einer **Verhaltensbeurteilung** handelt es sich um Schlussfolgerungen, die aus der Beobachtung von Verhalten in einer näher bestimmten Situation gezogen werden. Diese können sich auf Eigenschaften oder Verhaltensmuster der beobachteten Person(en) beziehen.

Erläuterungen

Zu diesen Definitionen sind einige kurze Erläuterungen angebracht.

- Wir beziehen uns auf *Personen*, obwohl es unstrittig ist, dass auch andere Lebewesen Verhalten zeigen. Die Formulierung „Lebewesen“ würde hier etwas befremdlich wirken (wäre aber auch richtig), weil sich die Psychologische Diagnostik ganz überwiegend mit Menschen befasst.
- Verhalten findet immer in einer bestimmten *Situation* statt. Eine Verhaltensbeschreibung und auch eine Verhaltensbeurteilung kann erst richtig

interpretiert werden, wenn der situative Kontext bekannt ist. Deshalb wird die Situation in die Definitionen aufgenommen.

- Beobachtungen können *Wertungen und Interpretationen* enthalten. In der Definition wird nicht verlangt, grundsätzlich und vollständig darauf zu verzichten. Formulierungen wie „läuft weg“ oder „versteckt sich“ implizieren etwa eine Absicht. Flüchtet die Person wirklich vor etwas oder läuft sie zu etwas hin, das wir gerade nicht sehen können? Setzt sie sich hinter den Baum, damit sie nicht von anderen gesehen wird oder will sie sich vielleicht nur im Schatten ausruhen? Exakte und wertungsfreie Beschreibungen sind in der Regel umfangreicher und trotzdem weniger informativ als Verkürzungen wie „weglaufen“ oder „verstecken“. Beobachterinnen und Beobachter verfügen oft über Hintergrundwissen, das ihnen implizite Interpretationen ermöglicht. Nehmen wir beispielsweise an, die beobachtete Person hat große Angst vor Hunden. Nun läuft ein großer Hund direkt auf sie zu. Da liegt es nahe, das beobachtete Verhalten als „weglaufen“ zu beschreiben. Und wenn Kinder Verstecken spielen, liegt es nahe, warum sich das beobachtete Kind gerade hinter den Baum setzt. Ist man sich nicht sicher, kann man immer noch als Beobachtung formulieren „scheint vor dem Hund wegzulaufen“ oder „versteckt sich offenbar“.
- Zu einer Verhaltensbeobachtung gehört, dass das Ergebnis – in der Regel durch ein *Beobachtungsprotokoll* – festgehalten wird. Ohne dies wäre es so, als würde man einen Test durchführen, aber nicht auswerten.

Wir stellen zunächst die Verhaltensbeobachtung als diagnostische Methode vor und kommen im Anschluss daran auf die Methode der Verhaltensbeurteilung zu sprechen.

3.6.1 Arten der Verhaltensbeobachtung

Die Verhaltensbeobachtung lässt sich anhand mehrerer Aspekte, die miteinander kombinierbar sind, näher charakterisieren:

- Frei oder systematisch (gebunden)
- Direkt oder indirekt (anhand von Aufzeichnungen)
- In natürlicher Umgebung („im Feld“) oder in einer Situation, die von Beobachterinnen bzw. Beobachtern geschaffen wurde
- Verdeckt oder offen
- Wenn offen, dann teilnehmend oder nicht teilnehmend
- Selbst- oder Fremdbeobachtung

Varianten der Verhaltensbeobachtung

Von einer *freien Verhaltensbeobachtung* spricht man, wenn die Beobachterin oder der Beobachter selbst entscheidet, welche Verhaltensweisen sie oder er beobachtet. Für die Beobachtung gibt es in der Regel einen Anlass, und sie dient meist einem bestimmten Zweck. Sie kann der Überprüfung konkreter Hypothesen bzw. der Beantwortung von Fragen dienen. Selbst wenn sie explorativ ist, wird nicht jedes beliebige Verhalten beobachtet, sondern die Aufmerksamkeit gilt zumeist bestimmten Bereichen wie Arbeits-, Sozial-, Zwangs- oder Spielverhalten oder der Vermeidung von angstauslösenden Reizen. Das Ergebnis ist ein mehr oder weniger detaillierter schriftlicher Bericht.

Freie Verhaltensbeobachtung:
Beobachtende entscheiden, was ihnen wichtig ist

Empfehlungen für die freie Verhaltensbeobachtung

- Beobachtete Person, Anlass und Zweck der Beobachtung nennen
- Angaben zu Ort, Umgebungsbedingungen (u. a. anwesende Personen) und Zeit machen
- Verhalten so konkret und in Zusammenhängen beschreiben (z. B. was ging dem Verhalten voraus?), dass eine Leserin oder ein Leser eine lebhafte Vorstellung davon bekommt
- Verhalten neutral (nicht wertend) und so weit wie möglich ohne Interpretationen beschreiben
- Interpretationen sind aber oft unvermeidbar. Sie sollten daher als solche kenntlich gemacht und durch Verhaltensweisen belegt werden. Beispiel: „Wirkt teilnahmslos (nimmt keinen Blickkontakt auf, kaut an Bleistift)“.
- Nicht nur das Verhalten beschreiben, sondern auch dessen Auslöser und Konsequenzen

Systematische Verhaltensbeobachtung:
Beobachtende handeln nach Vorgaben

Direkte Beobachtung oder Videoaufnahme

Bei einer *systematischen Verhaltensbeobachtung* wird den Beobachtenden dagegen genau vorgegeben, worauf sie zu achten haben und wie das Beobachtete zu protokollieren ist. Man muss also bereits Hypothesen darüber haben, was in der Beobachtungssituation wichtig ist. Die systematische Verhaltensbeobachtung erfordert einen erheblichen Aufwand bei der Vorbereitung: Meist erstellt man eine Liste mit relevanten Verhaltensweisen („Indexsystem“; ▶ Abschn. 3.6.2), und in der Beobachtungsphase wird beispielsweise beim Auftreten einer dieser Verhaltensweisen ein Strich in der Liste gemacht. Die Beobachtung wird also auf vorher festgelegte Verhaltensweisen eingeengt und die Art der Registrierung wird vorgeschrieben. Das Ergebnis liegt beispielsweise in Form einer Strichliste vor. Das genaue Vorgehen bei einer systematischen Verhaltensbeobachtung wird in ▶ Abschn. 3.6.2 näher erläutert.

Die *direkte Verhaltensbeobachtung* hat gegenüber der Verwendung einer Kamera den Vorteil, dass man seinen Blick dahin richten kann, wo das Geschehen interessant ist. Die Beobachtenden können ihren Standpunkt so verändern, dass sie ungehindert das beobachten können, was sie sehen wollen. Ein Nachteil besteht darin, dass es nicht möglich ist, gleichzeitig zu beobachten und zu registrieren. Wenn man das Beobachtete aufzeichnet (und sei es auch nur durch einen Strich in einer Liste), lässt die Aufmerksamkeit für das Geschehen zwangsläufig nach. Konzentriert man sich hingegen ganz auf die Beobachtung, ist man gezwungen, den Bericht anschließend aus dem Gedächtnis zu verfassen – mit der Gefahr, etwas Wichtiges zu vergessen oder Fakten falsch wiederzugeben (in ▶ Abschn. 3.6.2 wird jedoch eine Lösung für dieses Problem vorgestellt). Die Verwendung einer oder mehrerer Videokameras bietet dagegen den großen Vorteil, die Beobachtung ganz vom Registrieren trennen zu können: Man hält die Aufzeichnung einfach an, wenn man etwas aufschreiben will. Bei Bedarf ist es möglich, sich eine bestimmte Szene noch einmal anzuschauen. Videoaufnahmen können zudem mit speziellen Softwarelösungen analysiert werden, etwa um die Dauer eines Verhaltens exakt zu bestimmen. Es liegt auf der Hand, dass die größere Informationsausbeute und Exaktheit auch mit einem höheren Aufwand verbunden ist.

Eine *Verhaltensbeobachtung im Feld*, beispielsweise die Beobachtung eines Kindes beim Spiel mit den Eltern in häuslicher Umgebung, hat den Vorteil, dass auch die Kontextbedingungen erfasst werden, die das interessierende Verhalten beeinflussen könnten. Die Frage, ob die Beobachtungsergebnisse, die in einer künstlichen Laborsituation gewonnen wurden, auf den Alltag der Klientin bzw. des Klienten generalisierbar sind, stellt sich dann nicht. Eine von der Beobachterin oder dem Beobachter geschaffene oder speziell ausgewählte *laborähnliche Situation* hat dagegen den Vorteil, dass sie standardisiert werden kann. Immer wenn Vergleiche angestellt werden sollen, muss die Situation konstant gehalten werden. So kann man eine Patientin oder einen Patienten vor, während und nach einer Intervention beobachten; die Situation könnte in der Darbietung bestimmter angstauslösender Reize bestehen. In vergleichbarer Weise wird beispielsweise in einem Assessment-Center (► Abschn. 6.2.1.1) das Verhalten von Bewerberinnen oder Bewerbern in einer weitgehend standardisierten Situation beobachtet.

Beobachtung im Feld oder unter laborähnlichen Bedingungen

Der Fremde-Situations-Test

Der Fremde-Situations-Test (Ainsworth et al. 1978) ist eine häufig durchgeführte, systematische Form der Verhaltensbeobachtung, die im Labor stattfindet. Sie dient der Beurteilung der Beziehungsqualität des Kindes zu einem Elternteil (Zimmermann 2020). Dazu erfährt das Kind in festgelegter Abfolge und für eine genau spezifizierte Dauer die Trennung und die Rückkehr des Elternteils sowie die Begegnung und das Spiel mit einer fremden Person. Aus dem Verhalten des Kindes in den jeweiligen Sequenzen (z. B. Weinen bei Trennung vom Elternteil) wird auf die Bindungsqualität geschlossen.



(© buritora/stock.adobe.com)

Verdeckte Beobachtung soll Reaktivität verhindern

Bei einer *verdeckten Verhaltensbeobachtung* ist die beobachtende Person nicht sichtbar. Eine verdeckte Verhaltensbeobachtung sollte aus ethischen Gründen nur mit Zustimmung der beteiligten Personen durchgeführt werden. Verdeckt beobachten kann man mit einer Kamera, die fest installiert ist und nur zu bestimmten Zeiten aufzeichnet. Mit der Zeit gewöhnt sich die zu beobachtende Person an die Kamera, beachtet sie nicht weiter und verhält sich weitgehend natürlich. Eine verdeckte Beobachtung setzt immer voraus, dass entsprechende räumliche und technische Voraussetzungen gegeben sind, also beispielsweise eine Kamera installiert werden kann oder ein Beobachtungsraum vorhanden ist, der durch eine Einwegscheibe von einem Nachbarraum getrennt ist. Der wichtigste Grund, sich für eine verdeckte Beobachtung zu entscheiden, ist die Vermeidung von Reaktivität. Darunter versteht man die Beeinflussung des Messgegenstands durch den Messvorgang: Der Vorgang des Beobachtens führt dazu, dass sich das zu beobachtende Verhalten ändert.

Erfolgt die Verhaltensbeobachtung offen, sind 2 Varianten zu unterscheiden: *teilnehmend oder nicht teilnehmend*. Bei der teilnehmenden Beobachtung ist die Beobachterin oder der Beobachter in der Beobachtungssituation nicht nur anwesend, sondern nimmt selbst am Geschehen teil. Beispielsweise führt eine Lehrkraft den Unterricht durch und beobachtet dabei „nebenbei“ ein Kind. Ein Beispiel für eine nichtteilnehmende Verhaltensbeobachtung wäre, wenn sich die Lehrkraft in ein Klassenzimmer setzt und sich ganz der Beobachtung eines Kindes widmet, während eine andere Lehrerin oder ein anderer Lehrer den Unterricht durchführt.

Für die teilnehmende Beobachtung gibt es ökonomische und pragmatische Gründe: Ein ökonomischer Grund wäre, wenn die Hauptaufgabe und die Beobachtung ohne Weiteres von einer Person bewältigt werden kann. So ist es möglich, die routinierte Administration eines Tests, die wenig Kapazität erfordert, mit der Beobachtung des Verhaltens der Testperson zu verbinden. Ein pragmatischer Grund für eine teilnehmende Beobachtung wäre, wenn eine passive Teilnahme als störend oder unpassend empfunden wird. Man stelle sich vor, in einer geselligen Runde würde eine Beobachterin oder ein Beobachter sitzen und weder mit den anderen reden noch mit ihnen essen und trinken. Ein derart unnatürliches Verhalten hätte vermutlich einen negativen Effekt auf die Akzeptanz der Beobachtung und könnte darüber hinaus zu Reaktivität führen. Es ist aber grundsätzlich zu beachten: Wenn die mit der Teilnahme verbundene Aufgabe anstrengend ist – wie im Beispiel die Unterrichtsgestaltung –, wird die Beobachtungskapazität eingeschränkt.

Auch wenn man bei der Verhaltensbeobachtung zunächst daran denkt, dass es vorteilhaft ist, wenn eine andere, neutrale und unvoreingenommene Person jemanden beobachtet, gibt es manchmal gute Gründe für eine *Selbstbeobachtung*. So wird beispielsweise aus ethischen Gründen in der Regel darauf verzichtet, das Sexualverhalten eines Paares zu beobachten, das wegen Eheproblemen Rat sucht. Und aus ökonomischen Gründen verzichtet man darauf, etwa den Zigaretten- oder Alkoholkonsum oder das Ess- oder Zwangsverhalten im Alltag durch Fremdbeobachtung zu erfassen. Eine Beobachterin oder ein Beobachter müsste dazu schließlich mehrere Tage oder sogar Wochen ständig die zu beobachtende Person begleiten. In den genannten Fällen wird man die Klientin oder den Klienten zumeist dafür gewinnen können, selbst Aufzeichnungen vorzunehmen. Für die praktische Umsetzung kommen das Führen eines Tagebuchs oder das Protokollieren von Verhaltensweisen zu festen Zeitpunkten infrage. Das Führen eines Tagebuchs kann als eine Form der freien Verhaltensbeobachtung betrachtet werden, während Angaben in Verhaltenslisten der systematischen Verhaltensbeobachtung (► Abschn. 3.6.2) zuzuordnen sind. Solche Verhaltenslisten können in Form eines Heftchens mitgeführt oder über ein Smartphone administriert werden; die

Nichtteilnehmende Beobachtung kann unpassend sein

Gute Gründe für Selbstbeobachtung

Klientinnen und Klienten kreuzen beispielsweise zu bestimmten Zeitpunkten an, welche der genannten Verhaltensweisen (z. B. Zigarette geraucht) sie in der letzten Stunde wie oft ausgeführt haben.

Selbstbeobachtung im Alltag

Es gibt verschiedene Methoden, mit denen man das Verhalten (und Erleben) im Alltag durch Selbstbeobachtung erfassen kann. Als Oberbegriff werden dafür häufig „ambulantes (auch: ambulatorisches) Assessment“ und „Ecological Momentary Assessment (EMA)“ verwendet. Damit wird eine Erhebungsstrategie bezeichnet, die relevante Phänomene unmittelbar und unverzerrt durch Gedächtniseffekte in „natürlicher“ Ökologie erfassen soll. Die folgenden Ansätze sind zu unterscheiden: Die in der Fachliteratur als „Experience Sampling Method (ESM)“ bekannte Methode (Csikszentmihalyi und Larson 1987), auch manchmal „Event Sampling Method“ genannt, zeichnet sich durch die Erfassung von Verhalten (oder Erleben) direkt im Alltag aus. Während anfangs Papierversionen in Verbindung mit einem kleinen Signalgeber verwendet wurden, steht heute Software zur Verfügung, die auf einem Smartphone installiert wird. Die Probandinnen und Probanden werden zu bestimmten Zeitpunkten durch ein Signal aufgefordert, Einträge vorzunehmen. Welche Fragen zu beantworten sind, richtet sich nach dem Untersuchungsziel. Beispielsweise kann die momentane Aktivität erfragt werden, wobei Antwortalternativen angeklickt werden können, aber auch freie Antworten möglich sind. Die Angaben können sich auf die momentane oder auf die letzte abgeschlossene Aktivität beziehen („Was machen Sie gerade?“ bzw. „Was haben Sie zuletzt gemacht?“). Kontextvariablen (Aufenthaltsort, Anwesenheit anderer Personen etc.) können miterfasst werden. Die Methode ist sehr flexibel bezüglich Abfrageraten, Messgegenstand und Umfang der Protokolle. Die elektronische Version hat mehrere Vorteile: Es ist sichergestellt, dass die Aufzeichnung auch zum vorgesehenen Zeitpunkt stattfindet und keine nachträglichen Korrekturen vorgenommen werden und dass die vorausgegangenen Einträge nicht sichtbar sind. Nachteilig ist, dass die Menschen eventuell bei ihren momentanen Aktivitäten gestört werden. Man kann jedoch vorsehen, dass in solchen Fällen auf die Protokollierung verzichtet wird und eventuell später eine erneute Aufforderung kommt.

Eine andere Methode zielt darauf ab, den abgelaufenen Tag zu rekonstruieren, um dann Fragen zum Verhalten (oder Erleben) zu beantworten. Sie ist unter dem Namen „Day Reconstruction Method (DRM)“ (Kahneman et al. 2004) bekannt. Die teilnehmende Person rekonstruiert zunächst den abgelaufenen Tag schriftlich und unter Beachtung von vorgegebenen Leitfragen. Dies dient dazu, ein klares und vollständiges Bild des Tagesablaufs zu bekommen. Diese Aufzeichnungen verbleiben bei der betroffenen Person. Erst dann bearbeitet sie eine Liste mit Fragen. Kahneman et al. (2004) geben den Aufwand mit 45–75 min pro Tag an. Die Methode führt zu sehr ähnlichen Ergebnissen wie die zuvor beschriebene ESM. Der Vorteil gegenüber der ESM liegt vor allem darin, dass der Tagesablauf nicht gestört wird. Einschränkend ist zu sagen, dass eine sehr gute Mitarbeitbereitschaft nötig ist, nicht zuletzt wegen des relativ großen Aufwands.

Mit der „Event Reconstruction Method (ERM)“ steht zusätzlich eine Methode zur Verfügung, bei der Personen nicht den abgelaufenen Tag, sondern ein bestimmtes Ereignis rekonstruieren sollen (Grube et al. 2008). Solche Ereignisse können die letzte Begegnung mit kritischen Kundinnen bzw. Kunden oder das letzte Gespräch mit der Führungskraft sein. Der Vorteil gegenüber der DRM liegt nach Grube et al. (2008) darin, dass auf bestimmte Ereignisse fokussiert und damit Zeit beim Protokollieren der Beobachtungen gespart werden kann.

Stichprobenartige Aufzeichnung der akustischen Umwelt

Während sich die oben beschriebenen Methoden durch eine aktive Protokollierung des Verhaltens auszeichnen, eröffnen technische Entwicklungen auch neue Formen der Alltagsbeobachtung. Eine automatische Aufzeichnung findet ohne aktive Mitarbeit der beobachteten Person statt. Hierbei kommt beispielsweise die Sprache infrage. Mehl (2017) stellt unter der Bezeichnung „Electronically Activated Recorder (EAR)“ eine Methode vor, die mithilfe einer speziellen Software und eines Smartphones über längere Zeit stichprobenartig die akustische Umwelt aufzeichnet. So kann insbesondere die Sprechaktivität analysiert werden; die Aufzeichnungen werden dazu transkribiert. Die Methode wirft rechtliche und ethische Probleme auf, weil in die Privatsphäre der betroffenen Person und deren Interaktionspartner/-innen eingedrungen wird. Mehl (2017) erklärt dazu u. a., dass die teilnehmende Person sich die Aufzeichnungen vor Weitergabe anhören und zensieren kann. Ob davon Gebrauch gemacht wird, ist eine andere Frage. Wird beispielsweise 5 % der Zeit aufgezeichnet (alle 12 min je 30 s lang; das Beispiel stammt vom Autor), müsste die Person nach 12 h eine halbe Stunde lang akustische Stichproben aufmerksam abhören. Interaktionspartnerinnen bzw. -partner und unbeteiligte Personen sollen durch ein „Warndreieck“ auf dem Smartphone informiert werden, dass eine Aufzeichnung stattfinden kann. Diese Schutzmaßnahme kann nur funktionieren, wenn die untersuchte Person das Smartphone auch tatsächlich immer gut sichtbar vor sich legt und alle Menschen, die in die Nähe kommen, dies auch bemerken. Mittlerweile können Smartphones und andere sog. „Wearables“ (allgemein für tragbare Computer; für ein Beispiel s. □ Abb. 3.29) vielerlei Verhaltensweisen sowie physiologische Daten ihrer Benutzerinnen und Benutzer aufzeichnen. Dadurch lässt sich beispielsweise das Gesundheitsverhalten (z. B. körperliche Aktivität, Schlaf) sowie die aktuelle Belastung messen und durch entsprechende Rückmeldungen optimieren. Eine kritische Auseinandersetzung mit diesen Technologien findet sich bei Peake et al. (2018).



□ Abb. 3.29 Beispiel für tragbare Bewegungssensoren. (© sitthiphong stock.adobe.com)

3.6.2 Systematische Verhaltensbeobachtung

„Verhalten“ existiert vor allem in den Köpfen der Menschen, die beobachten. Es handelt sich dabei keineswegs um ein Abbild der physikalischen Welt, sondern um eine Auswahl von Ereignissen aus dem ständigen Fluss von Verhalten. Vieles ist nicht wichtig und wird daher nicht beachtet; Beobachten ist also immer mit einer **Selektion** verbunden. Was schließlich im Verhaltensstrom als relevant entdeckt wird, grenzen wir voneinander ab und benennen es meist nach seiner vermuteten Bedeutung. Diesen Vorgang kann man als **Segmentierung** bezeichnen. Lachen, Weinen, Antworten und Schimpfen sind Beispiele für solche **Verhaltenssegmente**. Schließlich werden die so bestimmten Verhaltensweisen **quantifiziert**, indem sie mit Aussagen über die **Intensität**, **Dauer** oder **Häufigkeit** versehen werden. Dazu finden Begriffe wie „sehr stark“, „heftig“ (Intensität), „anhaltend“, „ständig“ (Dauer), „oft“ oder „hin und wieder“ (Häufigkeit) Verwendung. Auch die Wahl eines verhaltensbeschreibenden Verbs oder Substantivs kann zumindest implizit eine Aussage über die Intensität, Dauer oder Häufigkeit einer Verhaltensweise gemacht werden. Ein Verhalten kann beispielsweise als „Weinen“, „Heulen“, oder „Schluchzen“ beschrieben werden. Diese kurzen Ausführungen erklären, warum freie Verhaltensbeschreibungen verschiedener Beobachter oder Beobachterinnen praktisch nie identisch ausfallen. Mit einer **systematischen Verhaltensbeobachtung** sollen **Selektion**, **Segmentierung** und **Quantifizierung** so weit wie möglich **standardisiert** werden.

Selektion, Segmentierung und Quantifizierung standardisieren

Bei einer systematischen Verhaltensbeobachtung wird keine vollständige Beschreibung des Verhaltens einer Person angestrebt, sondern es soll immer nur ein bestimmter **Teilaspekt** des Verhaltens erfasst werden, beispielsweise Aggressivität, Kooperations- oder Dominanzverhalten. Die Vielfalt der Verhaltensweisen, die einen solchen Teilbereich des Verhaltens ausmachen, soll auf wenige Aussagen reduziert und zudem quantifiziert werden. **Zeichensysteme** dienen dazu, ausgewählte Teile des interessierenden Verhaltens zu erfassen. Bei **Kategoriensystemen** wird dagegen versucht, jeden Verhaltensakt zu erfassen (beispielsweise alle Aussagen im Rahmen einer Verhandlungssituation). Alle wesentlichen Details des Verhaltensstroms werden einer begrenzten und damit überschaubaren Zahl von Oberbegriffen zugeordnet. Eine weitere Gruppe stellen die Ratingverfahren dar, die in ▶ Abschn. 3.6.3 beschrieben werden. Alle Systeme dienen dazu, **Verhalten zu quantifizieren**.

Reduktion auf wenige Aussagen

Index- bzw. Zeichensysteme Die Begriffe „Indexsystem“ und „Zeichensystem“ sind als synonym zu verstehen. Bei ihnen geht es darum, ausgewählte Verhaltensweisen, die als **Indikatoren** oder Anzeichen für den zu messenden **Verhaltensbereich** gelten, zu registrieren. Die zu beobachtenden Verhaltensweisen werden auch **Beobachtungseinheiten** genannt. Beispielsweise kann der Verhaltensbereich „Aggression“ durch konkrete Verhaltensweisen wie Schlagen, Treten, Umstoßen, Ausreißen von Haaren, Beißen, Kratzen, Anspucken, Anschreien und Beleidigen operationalisiert werden. Allerdings ist es sinnvoll, nur solche Verhaltensweisen aufzunehmen, die in der Beobachtungssituation auch zu erwarten sind.

Einzelne Verhaltensweisen als Anzeichen für einen Verhaltensbereich

Wie findet man die „richtigen“ Anzeichen für einen Verhaltensbereich? Diese Frage lässt sich im Prinzip einfach beantworten. Die grundsätzliche Antwort lautet, dass die Verhaltensweisen repräsentativ für das zu messende Merkmal in dem vorgesehenen Anwendungsbereich sein sollen. Hier wird also Inhaltsvalidität gefordert – die Verhaltensweisen sollen repräsentativ sein und zwar für etwas ganz Bestimmtes.

Die Umsetzung der Antwort bereitet aber eventuell Schwierigkeiten. Nehmen wir als Beispiel Aggression. Beginnen müssen wir mit einer definitiven Klärung des Konzepts „Aggression“. Die praktisch relevante Frage lautet dann: In welchen Verhaltensweisen kann sich im vorgesehenen Beobachtungskontext Aggression zeigen? Bei Kindern im Vorschulalter, die einen Kindergarten besuchen, kommen ganz andere Verhaltensweisen infrage als bei erwachsenen Straftäterinnen und Straftätern, die sich wegen Körperverletzung, Mord, Totschlag oder anderer Gewaltdelikte im geschlossenen Strafvollzug befinden und die auf dem Pausenhof eines Gefängnisses beobachtet werden sollen. Es kann hilfreich sein, die einschlägige Fachliteratur zu Aggression bei Kindern im Kindergarten bzw. bei Strafgefangenen im Strafvollzug zu sichten. Auch die Befragung von Personen, die mit entsprechenden Situationen vertraut sind, also das Personal in Kindergärten bzw. im Strafvollzug, kann zu guten Ergebnissen führen. Ebenso kommt eine freie Verhaltensbeobachtung infrage. Man sollte sich darüber im Klaren sein, dass solche Recherchen erst einmal nur zu Hypothesen führen, welche Verhaltensweisen relevant für die geplante Beobachtung sind.

Zwei Arten von Fehlern spielen eine Rolle: Erstens kann die Liste Verhaltensweisen enthalten, die kein Indikator für Aggression (so wie sie zuvor definiert wurde) sind. Dies lässt sich am besten mithilfe von Expertinnen und Experten feststellen, die mit der zugrunde gelegten Definition vertraut sind und jede Verhaltensweise danach beurteilen, ob sie ein Anzeichen für Aggression sind oder nicht (hierzu kann der in ▶ Abschn. 2.6.3.1 genannte Ansatz der quantitativen Inhaltsanalyse genutzt werden). Der andere Fehler besteht darin, wichtige Verhaltensweisen übersehen zu haben. Dieses Problem zeigt sich erst in der probeweisen Anwendung des Beobachtungssystems: Es treten Verhaltensweisen auf, die offenbar Anzeichen für Aggression sind, aber in der Liste fehlen. In diesem Fall muss die Liste erweitert werden. Um in diesem Schritt dann nicht doch den zuerst genannten Fehler zu machen, sollte noch einmal ein Expertinnen- bzw. Expertenrating durchgeführt werden. Aus pragmatischen Gründen wird die Liste der Anzeichen noch überarbeitet werden. Sie muss eventuell auf einen gut handhabbaren Umfang reduziert werden. Eine Maßnahme kann sein, selten vorkommende Verhaltensweisen zu eliminieren. Die Liste verkürzt sich, wenn die Beobachtungseinheiten breiter gefasst werden. So könnte man schlagen, treten, Haare ausreißen, beißen, kratzen und ggf. weitere Verhaltensweisen zu einer breiten Einheit „körperlich verletzen“ zusammenfassen. Die detailliertere Liste der Verhaltensweisen dient eventuell in Klammern als Beispiel. Welchen Umfang die Liste haben soll, ergibt sich vor allem aus dem Zweck der Untersuchung, der Expertise und Kapazität der Beobachterinnen und Beobachter sowie der Beobachtungsdauer (► Abb. 3.30).

Einschlägige Anzeichen suchen

Striche in Beobachtungs-Checkliste

Die Häufigkeit von Verhaltensweisen ermitteln Ziel einer systematischen Verhaltensbeobachtung ist es, die Häufigkeit oder die Dauer der zu beobachteten Verhaltensweisen zu ermitteln. Wir betrachten zunächst die Registrierung der Häufigkeit. Das Auftreten eines definierten Zeichens kann erstens in Form einer Checkliste festgehalten werden, in der etwa 10 Verhaltensweisen aufgeführt sein können. Die Beobachterinnen und Beobachter sollen jedes Mal, wenn sie eine dieser Verhaltensweisen bemerken, einen Strich in die entsprechende Zeile des Beobachtungsbogens machen. Diese Methode hat den Nachteil, dass die Beobachtung immer kurz aussetzt, wenn ein Eintrag vorgenommen wird. Die richtige Zeile zu finden und einen Strich am Ende zu machen, lenkt kurzfristig von der Beobachtung ab. Als Vorteil ist zu werten, dass die Protokollierung so zeitnah erfolgt, dass Erinnerungseffekte keine Rolle spielen.



■ Abb. 3.30 Wutausbruch während einer Pressekonferenz. Enthält das Zeichensystem das Item „greift andere verbal an“, wird hier ein Strich in der Beobachtungsliste gemacht. (© Douliery Olivier/ABACA / picture alliance)

Zweitens kann die Beobachtungs-Checkliste anders bearbeitet werden. Die Beobachterinnen und Beobachter sind mit der Checkliste vertraut, widmen sich aber erst einmal nur der Beobachtung. Am Ende der Beobachtungsphase stufen sie ein, wie häufig jede der Verhaltensweisen aufgetreten ist. Von Vorteil ist, dass die ganze Zeit über beobachtet werden kann. Dieser Vorteil wird aber damit „bezahlt“, dass das Ergebnis durch Erinnerungsfehler und Ungenauigkeiten beim Einschätzen der Häufigkeit verzerrt werden kann; eine Einstufung liefert nicht so exakte Ergebnisse wie das Zählen von beobachteten Verhaltensweisen.

Bei der dritten Methode finden die gleichen Beobachtungseinheiten wie in der Beobachtungs-Checkliste Verwendung. Das ganze Beobachtungsintervall wird jedoch in viele kleine Zeitabschnitte unterteilt. Die Methode wird auch Zeitstichprobe (time sampling) genannt. Der Begriff ist leider etwas irreführend, weil zumindest in der Statistik unter der Stichprobe eine (zufällige oder systematische) Auswahl verstanden wird. Wie gleich gezeigt wird, kann time sampling auch eine Vollerhebung über alle Zeitabschnitte bedeuten.

Wir betrachten zunächst die Variante einer Zeitstichprobe, die den Zusatz „Stichprobe“ im Namen unzweifelhaft verdient, weil sich die Beobachtung auf Stichproben des Verhaltens bezieht. Dabei werden die Beobachtung und die Protokollierung arbeitstechnisch getrennt, was die Beobachterinnen und Beobachter entlastet. Quasi nebenbei fallen dabei Informationen zum Verlauf an. Konkret bedeutet dies, dass die gesamte Zeitstrecke in Beobachtungs- und Protokollierungsphasen unterteilt wird, die sich abwechseln. Die beiden Phasen können beispielsweise jeweils 10 s dauern (aber auch unterschiedlich lange Phasen sind möglich). Die Beobachterinnen und Beobachter beobachten in der Beobachtungsphase zunächst nur das Verhalten der Person (ohne zu protokollieren). Danach kommt die Protokollierungsphase, in der sie die Checkliste bearbeiten, aber nicht weiter beobachten. Sie machen bei allen Verhaltensweisen, die in der Beobachtungsphase vorkamen, einen (!) Strich. Es spielt also keine Rolle, wie oft das Verhalten in der Beobachtungsphase

Beobachtungs-Checkliste mit Rating der Häufigkeit

Time sampling

Geplante Pausen zum Registrieren

vorkam. Es folgt wieder eine Beobachtungsphase mit anschließender Protokollierungsphase. Wenn beide Phasen gleich lang sind, haben die Beobachterinnen und Beobachter am Ende genau die Hälfte der Zeit für die Verhaltensbeobachtung aufgewendet. □ Tab. 3.28 zeigt den Aufbau eines solchen Beobachtungsprotokolls.

3

□ Tab. 3.28 Aufbau eines Beobachtungsprotokollbogens (Zeichensystem)

Aggressives Verhalten	Zeitabschnitt										
	1	2	3	4	5	6	7	8	9	10	etc.
Schlagen											
Treten											
Beißen											
etc.											

Für alle Fragestellungen, die auf einem inter- oder auch intraindividuellen Vergleich basieren, reicht die verkürzte Beobachtungszeit aus. Beispielsweise wollen wir feststellen, ob und wie stark sich ganz kleine und etwas ältere Kinder in ihrem Spielverhalten im Kindergarten unterscheiden und setzen dazu ein eigens konstruiertes Beobachtungsinstrument ein. Die eine Gruppe hat im Durchschnitt 20× und die andere 25× Spielverhalten gezeigt.

Bei einer anderen Variante des time sampling wird auf eigene Registrierphasen verzichtet. Hier wird durchgearbeitet. Ein Signal, etwa ein Piepton, fordert die Beobachterinnen und Beobachter z. B. alle 10 s auf, zu protokollieren, welche der seit dem letzten Signal beobachteten Verhaltensweisen vorkamen. Es handelt sich also um eine Vollerhebung des Verhaltens. Wenn den Beobachterinnen und Beobachtern keine relevanten Verhaltensweisen durch die Doppelbelastung entgehen, ist die Anzahl der beobachteten Verhaltensweisen am Ende doppelt so hoch wie bei der zuvor beschriebenen Variante mit gleich langen Beobachtungs- und Registrierphasen.

Die Anzahl der Registrierungen unterscheidet sich bei den Time-Sampling-Methoden geringfügig gegenüber dem gleichen Kennwert bei Anwendung der Beobachtungs-Checkliste. Der Unterschied liegt darin, dass beim time sampling erfasst wird, ob ein Verhalten im Zeitintervall vorkam. Trat es zum Beispiel 3× auf, wird es dennoch nur 1× kodiert; zog es sich über 2 Zeitintervalle hin, wird es 2× kodiert. Bei der Beobachtungs-Checkliste wird ein 3-maliges (kurzes) Auftreten mit 3 Strichen kodiert; und ein Verhalten, dass sich über 2 Zeitintervalle hinzieht, zählt als ein einzelnes Ereignis, wird also mit 1 Strich kodiert. Der Unterschied wird in □ Abb. 3.31 veranschaulicht.

Allen hier beschriebenen Zeichensystemen, außer der Checkliste mit anschließendem Rating der Häufigkeiten, ist gemeinsam, dass für jede Verhaltensweise ein Summenwert gebildet wird. Die Summenwerte können zu einem Gesamtwert verrechnet werden, der dann beispielsweise als Maß für die Häufigkeit von aggressivem Verhalten im Beobachtungszeitraum gilt. Natürlich ist es auch möglich, Verhaltensweise zu gruppieren und dafür analog zu Subskalen eines Tests separate Häufigkeiten zu bestimmen. Beispielsweise könnte

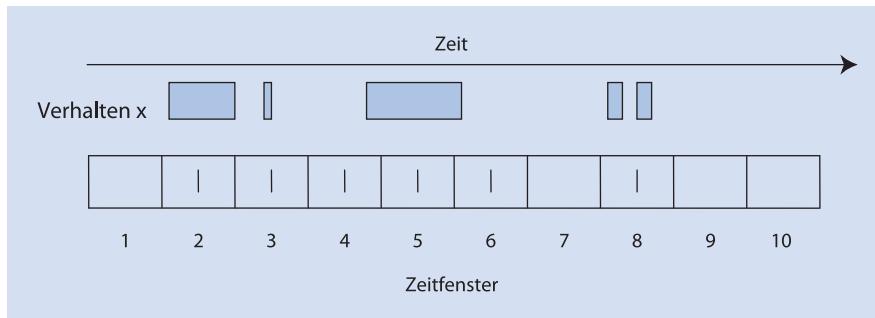


Abb. 3.31 Time sampling. In jedem Zeitabschnitt (von z. B. 20 s) wird eine Markierung vorgenommen, wenn die Verhaltensweise auftritt – unabhängig von ihrer Dauer

man für körperliche und verbale Aggression, Aggression durch Zerstörung von Dingen und Aggression durch Verweigerung von Hilfe eigene Kennwerte berechnen.

Die Dauer einer Verhaltensweise ermitteln Eine exakte Bestimmung der Dauer einer Verhaltensweise – der Begriff wird hier gleichbedeutend mit Ereignis verwendet – erfolgt nur beim *event sampling*. Dazu werden der **Anfang und das Ende der Verhaltenssequenz** zeitlich genau bestimmt, um daraus die Dauer zu berechnen. In der Regel ist dazu eine *Videoaufnahme* nötig, in die die Zeit eingebendet ist. Durch eventuell wiederholtes Betrachten der Aufnahme im Slow-Motion-Modus gelingt es, den Zeitpunkt des Beginns und des Endes sehr genau festzustellen. Ein entsprechendes Instrument zur Protokollierung von Verhaltensweisen kann eigens für geplante Forschungs- oder Anwendungszwecke entwickelt werden; spezielle Software ist aber auch kommerziell erhältlich. Alternativ können Videoaufzeichnungen manuell analysiert werden: Die Verhaltensweisen werden in einem Protokollbogen aufgelistet, und die Beobachterinnen und Beobachter protokollieren jeweils die Zeit für den Beginn und das Ende eines Verhaltens. Geschieht die Protokollierung etwa in einer Excel-Tabelle, kann mithilfe von hinterlegten Formeln die Dauer eines Ereignisses berechnet und auch über alle Ereignisse der gleichen Art summiert werden. Der Aufwand für die Protokollierung beim *event sampling* ist vergleichsweise groß und steigt natürlich mit der Anzahl der analysierten Verhaltensweisen an.

Event sampling

Die oben beschriebenen Index- bzw. Zeichensysteme mit unterschiedlichen Aufzeichnungsmethoden haben gegenüber der freien Verhaltensbeobachtung den großen Vorteil, dass sie quantitative Daten liefern. Sie sind dafür wesentlich aufwendiger: Das Beobachtungsinstrument muss in der Regel erst selbst konstruiert werden. Die Durchführung ist nicht selbsterklärend, sondern die Beobachterinnen und Beobachter müssen speziell für das Verfahren trainiert werden. Schließlich ist der Einsatz auf eine bestimmte Situation oder einen engen Anwendungsbereich beschränkt. Die freie Verhaltensbeobachtung kann dagegen in fast allen denkbaren Situationen durchgeführt werden. Wir stellen abschließend ein Beispiel für eine sehr sorgfältig geplante und durchgeführte Verhaltensbeobachtungsstudie vor, die man als einzigartig bezeichnen kann.

Vergleich von systematischer und freier Verhaltensbeobachtung

Systematische
Verhaltensbeobachtung zur
Erforschung von Unfallrisiken

100-Car Naturalistic Driving Study

Analysen von Unfällen im Straßenverkehr weisen darauf hin, dass oftmals die Fahrerin oder der Fahrer unaufmerksam war. Das US-Verkehrsministerium hat eine Studie in Auftrag gegeben, die die Rolle der Unaufmerksamkeit bei Verkehrsunfällen klären sollte (Klauer et al. 2006, 2010). Dazu wurden über 100 Pkw technisch aufwendig ausgestattet, damit sowohl das Verhalten der Fahrer/-innen und Fahrzeugdaten als auch die Umgebung erfasst werden konnten. Die Fahrzeuginsassen wurden 12–13 Monate lang dabei beobachtet, wie sie sich im Straßenverkehr verhielten. Die Liste von beobachteten Verhaltensweisen mit Definition und Beispielen umfasst allein 26 Seiten (Klauer et al. 2010, Anhang B). Unauffällig platzierte Kameras erfasssten das Gesicht der Fahrerinnen oder des Fahrers, den Blick über die Schultern auf Hände und Armaturen, die Fahrerseite im Auto, den Blick aus dem Auto nach vorne und nach hinten sowie die Beifahrerseite.

Unaufmerksamkeit wurde u. a. durch „sekundäre Tätigkeiten“ operationalisiert. Das sind Aktivitäten, die nicht zum Führen eines Pkw erforderlich sind. Sie wurden in komplexe, moderat komplexe und einfache Tätigkeiten unterteilt. „Komplexe“ sekundäre Tätigkeiten wurden beispielsweise als Tätigkeiten definiert, bei denen mehr als 2 Tasten zu betätigen sind oder der Blick mehr als 2× von der Straße vor dem Auto abgewendet werden muss. Beispiele sind das Wählen einer Nummer auf dem Handy, Lesen, ein Insekt im Auto verfolgen und sich schminken. Besonders ambitioniert war die Erfassung von Schläfrigkeit, weil hier viele Stufen unterschieden wurden. Verhaltensweisen wie langsames Schließen der Augen, schlaffe Gesichtsmuskulatur, verminderte Körperbewegungen, verringerte Blickbewegungen oder das Schließen der Augen für 2 bis fast 4 s („sehr schläfrig“) oder für 4 oder mehr Sekunden („extrem schläfrig“) wurden ebenso beachtet wie Reiben der Augen, Verziehen des Gesichts, Gähnen oder unruhiges Hin-und-her-Bewegen im Sitz als Indikatoren für eine „Bekämpfung“ der Müdigkeit.

Um die Nützlichkeit dieser systematischen Verhaltensbeobachtung zu demonstrieren, werden ausgewählte Ergebnisse und damit Erkenntnisse für die Verkehrssicherheit, basierend auf dem Abschlussbericht (Klauer et al. 2010), beschrieben. Dazu ist eine kurze Erläuterung zur Auswertung nötig. Während der gesamten Erhebungszeit passierten 69 Unfälle und 761 Beinaheunfälle, bei denen die Fahrerin oder der Fahrer entweder Schuld oder Teilschuld hatten. Es geht nun um die Frage, welche Verhaltensweisen bei einem (Beinahe-)Unfall gehäuft vorkommen. Es genügt nicht, einfach nur die Verhaltensweisen zu ermitteln, die einem (Beinahe-)Unfall vorausgegangen sind. Wenn beispielsweise oft das Rauchen einer Zigarette vorkommt, muss man prüfen, ob dieses Verhalten beim unfallfreien Autofahren vielleicht genauso häufig vorkommt. Wenn ja, ist es kein Risikofaktor für Unfälle. Dazu wurde eine sog. „Fall-Crossover-Analyse“ durchgeführt: Für jedes analysierte (Beinahe-)Unfallereignis („Fall“) wurden für die gleiche Person zum Vergleich mindestens

6 Baseline-Episoden mit unfallfreiem Fahren gesucht, die bezüglich mehrerer Variablen (z. B. Tageszeit, Verkehrsichte, Ort – soweit GPS-Daten vorlagen) mit dem Fall vergleichbar waren. Die Baseline-Episoden mussten vor dem Unfallereignis liegen, weil sich das Fahrerverhalten infolge des (Beinahe-)Unfalls geändert haben könnte. Insgesamt wurden rund 10,000 Baseline-Episoden in die Auswertung einbezogen. Verglichen wurden immer die letzten 15 s vor einem (Beinahe-)Unfall mit mehreren 30-sekündigen Baseline-Sequenzen. Das Ergebnis wird als Odds-Ratio dargestellt. Das ist das relative Risiko, dass auf ein bestimmtes Verhalten ein (Beinahe-)Unfall erfolgt. Ein Wert von 1 bedeutet, dass das Verhalten vor einem (Beinahe-)Unfall genauso wahrscheinlich ist wie beim Fahren ohne Unfall. Tatsächlich zeigte sich, dass bei „einfachen Tätigkeiten“, dazu gehören Verhaltensweisen wie Rauchen oder Trinken, die relative Wahrscheinlichkeit für (Beinahe-)Unfälle mit einer Odds-Ratio von .8 nicht signifikant erhöht war. Für moderate und komplexe sekundäre Tätigkeiten betrug die Odds-Ratio jedoch 1,3 bzw. 2,1 (Klauer et al. 2010, Tab. 14). Bei den entsprechenden Verhaltensweisen war das Unfallrisiko signifikant erhöht. Wer als Fahrerin oder Fahrer moderat komplexe Tätigkeiten ausführt, z. B. mit dem Handy telefonieren oder etwas essen, hat ein 1,3× größeres Risiko für einen (Beinahe-)Unfall, als wenn sie oder er dies unterlässt. Komplexe Tätigkeiten, z. B. eine Nummer auf dem Handy eingeben, etwas lesen oder sich schminken, erhöhen das Risiko um den Faktor 2,1. Einen dramatischen Effekt hatte das Auftreten von moderater bis extremer Schläfrigkeit mit einer Odds-Ratio von 38,7! Schläfrigkeit wurde in der Baseline selten beobachtet (in .5 % aller Episoden), kam aber bei einem (Beinahe-)Unfall relativ oft vor (16,3 %).



Auf einem Handy Eingaben zu machen, erhöht das Unfallrisiko beim Autofahren stark. (© Andrey Popov/stock.adobe.com)

Alle relevanten Verhaltensweisen werden einer Kategorie zugeordnet

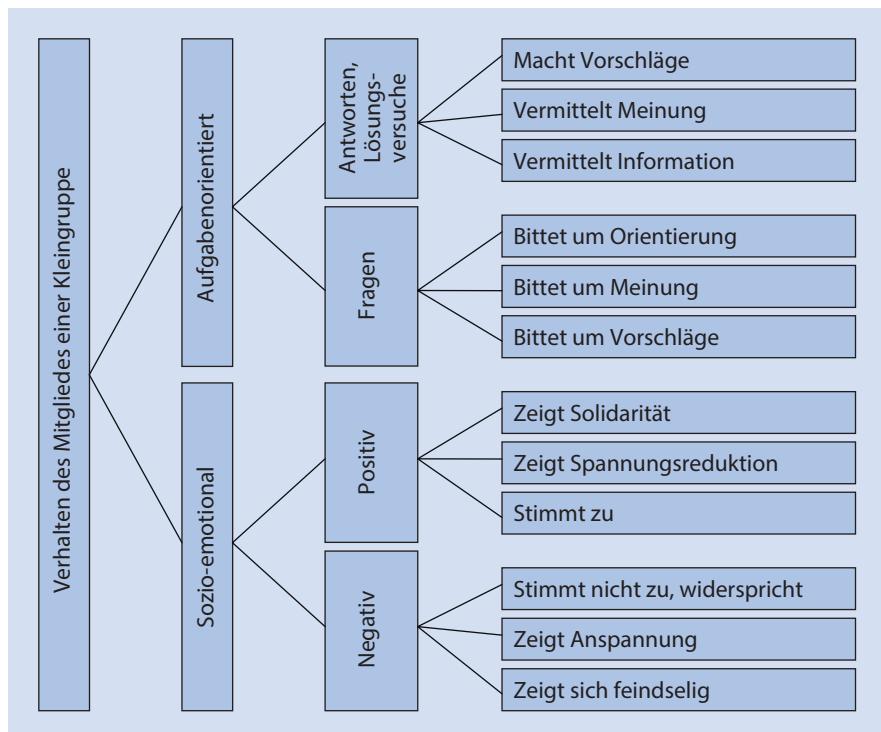
Interaktionsprozessanalyse

Instrument zur Kodierung von Diskussionen (IKD)

Kategoriensysteme Mithilfe von Kategoriensystemen will man ein Verhalten vollständig anstatt über ausgewählte Verhaltensweisen (s. o.) erfassen. Das Verhalten wird dazu in mehrere klar definierte und voneinander abgrenzbare Kategorien unterteilt. Die Kriterien Vollständigkeit, Eindeutigkeit und Überschneidungsfreiheit sind jedoch schwer zu erfüllen; Kategoriensysteme spielen daher in der diagnostischen Praxis praktisch keine Rolle. Ein Beispiel – allerdings von begrenztem Nutzen – wäre die Unterteilung der Reaktionen von Erzieherinnen und Erziehern in Belohnung, Bestrafung und neutrales Verhalten. Wie bei Zeichensystemen kann das bloße Auftreten eines Verhaltens bzw. hier einer Verhaltenskategorie über die Zeit hinweg registriert werden.

Zur Erforschung von Interaktionen in Kleingruppen hat Bales bereits 1950 die sog. Interaktionsprozessanalyse (Bales 1975) entwickelt (► Abb. 3.32). Die Systematik ist theoretisch fundiert und führt zu insgesamt 12 Kategorien des Verhaltens. Beobachterinnen und Beobachter, die mit einem solchen Kategoriensystem arbeiten, müssen jede Verhaltensweise einer dieser Kategorien zuordnen – eine Restkategorie gibt es nicht.

Mit dem Instrument zur Kodierung von Diskussion (IKD) von Schermuly et al. (2010) soll hier beispielhaft eines der wenigen neuen Kategoriensysteme kurz vorgestellt werden. Das IKD dient der standardisierten Beobachtung von Kommunikationsanlässen zwischen 2 oder mehr Personen. Dabei wird die gesamte Kommunikation zuerst in einzelne Sequenzen zerlegt. Die Autoren benennen konkrete Regeln zum Vorgehen bei der Sequenzierung. Danach wird das Verhalten aus jeder Sequenz hinsichtlich seiner interpersonalen und seiner funktionalen Bedeutung bewertet (in diesem Teil ist das IKD also eine Verhaltensbeurteilung; ► Abschn. 3.6.3). Zur interpersonalen Bewertung stehen 2 Dimensionen zur Verfügung: dominant-submissiv und freundlich-feindlich. Hinsichtlich seiner Funktion ist das Verhalten als Inhalts- oder



► Abb. 3.32 Kategoriensystem zur Beobachtung von Interaktionsprozessen in Kleingruppen. (Nach Bales 1975, © R. F. Bales)

Steuerungsaussage zu kategorisieren. Darüber hinaus kann es als Vorschlag und/oder Frage bewertet werden. Ein Kodierbogen ist in Abb. 3.33 dargestellt. Darüber hinaus steht eine Kodiersoftware zur Verfügung.

3.6.3 Verhaltensbeurteilung

Die Verhaltensbeobachtung liefert Daten über die Häufigkeit oder Dauer von konkreten Verhaltensweisen. Diese können von einer Diagnostikerin oder einem Diagnostiker, die nicht selbst Beobachtende gewesen sein müssen, als Ausprägung von Eigenschaften interpretiert werden: Wer sehr viele aggressive Verhaltensweisen im Vergleich zu anderen Personen in vergleichbaren Situationen zeigt, wird als sehr aggressiv beurteilt. Bei der Verhaltensbeurteilung nimmt die beobachtende Person diese Interpretation direkt vor: Sie sieht

Interpretation der Ausprägung direkt durch beobachtende Person

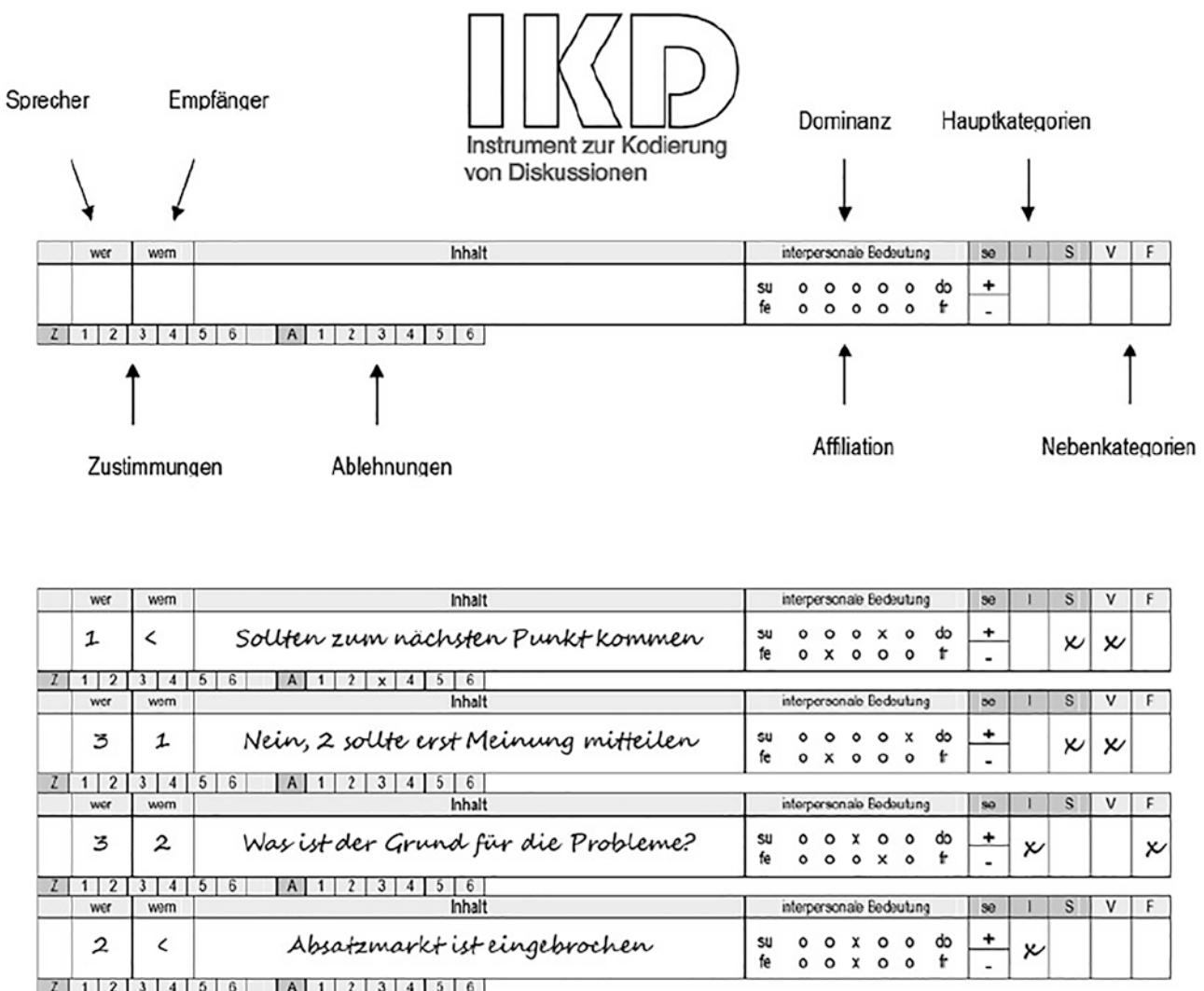


Abb. 3.33 Kodierbeispiel aus dem Instrument zur Kodierung von Diskussionen (IKD). Legende: wer zu wem: <=an alle; interpersonale Bedeutung: su=submissiv, do=dominant, fe=feindlich, fr=freundlich; funktionale Bedeutung: se=socio-emotionale Aussage, I=Inhaltsaussage, S=Steuerungsaussage, V=Vorschlag, F=Frage; Reaktionen: Z=Zustimmung, A=Ablehnung. (Aus Schermuly et al. 2010, mit freundlicher Genehmigung des Hogrefe Verlages)

bestimmte Verhaltensweisen und schließt daraus direkt auf die Eigenschaft. Die Registrierung während des Beobachtungsvorgangs entfällt; die beobachtende Person braucht ihr Urteil in der Regel nicht einmal zu begründen.

Ratingskalen

3

Die Beurteilung erfolgt in standardisierter Form: Die zu beurteilende Eigenschaft (z. B. Teamfähigkeit) wird vorgegeben, und das Urteil ist auf einer mehrstufigen Ratingskala durch Ankreuzen abzugeben. Die Zahlen symbolisieren die Ausprägung des beobachteten Verhaltens. Vorsicht ist geboten, wenn Häufigkeiten zu beurteilen sind; die Zahlen (z. B. „2“) könnten als absolute Häufigkeiten missverstanden werden (das Verhalten kam 2× vor). Werden mehrere Aspekte des Verhaltens beurteilt, beispielsweise verbale Aggression, Aggression gegen Sachen, körperliche Aggression, so dienen die Zahlen auch der Verrechnung zu einem Gesamtwert. Beliebt sind 5- bis 7-stufige numerische Skalen, deren Pole meist verbal verankert sind (z. B. „sehr niedrig“ und „sehr hoch“). Da es bei dieser Art von Skalen den beurteilenden Personen völlig überlassen bleibt, was sie beispielsweise unter einer sehr hohen Teamfähigkeit verstehen, werden die Skalen gerne auch zusätzlich durch typische Verhaltensweisen erläutert. Das nennt man „verhaltensverankerte Skalen“. Bei vielstufigen Skalen ist es kaum möglich, für alle Ausprägungen passende Verhaltensweisen zu finden. Deshalb werden meist nur die Pole, manchmal auch die Mitte, mit Verhaltensangaben erläutert. Es folgt ein Beispiel für eine verbal (oben) und verhaltensverankerte Skala (unten).

Beispiel für verhaltensverankerte Ratingskalen

Wie verbal aggressiv verhält sich das Kind?

1	2	3	4	5
gar nicht	schwach	mittel	stark	sehr stark
kein Schimpfen oder Drohen		schimpft laut, wenn etwas nicht klappt		brüllt bei geringen Anlässen, beleidigt oder beschimpft andere heftig

Verhaltensverankerung soll Beurteilendenübereinstimmung erhöhen

Brunswik'sches Linsenmodell zur Erklärung von Verhaltensbeurteilung

Der Wert einer Skala steht und fällt mit der Eindeutigkeit der Definition der einzelnen Skalenelemente. Aber erst eine gründliche Schulung kann sicherstellen, dass verschiedene Beurteilerinnen und Beurteiler zu einigermaßen übereinstimmenden Ergebnissen kommen. Je globaler das einzuschätzende Verhaltensmerkmal ist, desto schwieriger wird es, Übereinstimmung zu erzielen. Durch eine Verhaltensverankerung soll die Übereinstimmung verbessert werden.

Warum Urteile manchmal nicht übereinstimmen, lässt sich besser verstehen, wenn man sich mit dem Urteilsprozess befasst. Dazu ist das *Linsenmodell* von Brunswik (1952) hilfreich. Brunswik hat argumentiert, dass Menschen ihre Umwelt nicht direkt wahrnehmen, sondern sie erschließen. Dazu nutzen sie einen „Fächer“ von Hinweisreizen, die von Objekten der Umwelt ausgehen. Sie nehmen diese Hinweisreize wahr und bilden dann ein Urteil. In Abb. 3.34 wird die Logik dieses Modells auf die Verhaltensbeurteilung angewendet: Die beobachtete Person sendet willentlich und unwillentlich Hinweisreize aus; sie wird deshalb hier als „Sender“ bezeichnet. Bei

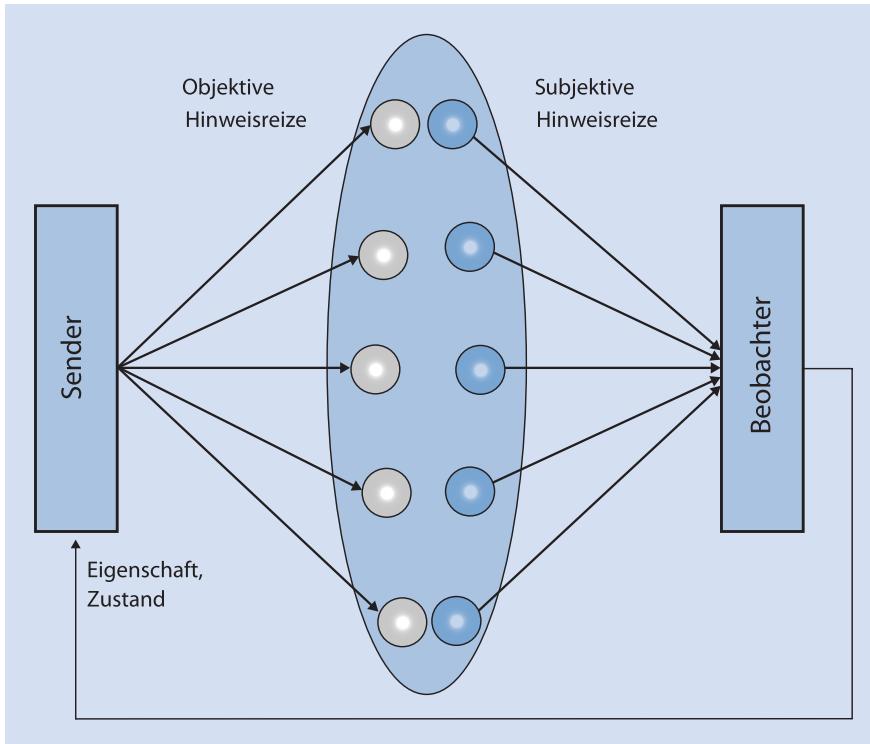


Abb. 3.34 Modifiziertes Brunswik'sches Linsenmodell zur Erklärung von Verhaltensbeurteilungen (Erläuterungen im Text)

diesen Hinweisreizen handelt es sich um konkrete Verhaltensweise wie Gähnen, Spielen mit einem Bleistift, verzögertes Reagieren auf direkte Ansprache oder eine schlaffe Körperhaltung. Diese Verhaltensweisen sind real vorhanden und werden deshalb in der Abbildung als „objektiv“ bezeichnet. Die beobachtende Person nimmt diese Hinweisreize wahr – allerdings nicht wie eine Kamera, sondern subjektiv – mit allen Eigentümlichkeiten und Fehlern, die für die menschliche Wahrnehmung typisch sind. Unterschiedliche Beobachterinnen und Beobachter würden das Geschehen deshalb auch teilweise unterschiedlich erfassen oder kodieren: Ein Hochziehen der Mundwinkel kann als Lächeln oder als Grinsen wahrgenommen werden, und ein 5-sekündiger Blick in Richtung der Augen einer anderen Person kann „Blickkontakt“ oder „Anstarren“ bedeuten.

Die Beobachterin oder der Beobachter verarbeitet die Einzelinformationen zu einem Urteil über die beobachtete Person und schreibt dieser nun eine Eigenschaft oder einen Zustand zu. Die oben genannten Hinweisreize könnten zu der Zuschreibung von Langeweile führen. Die Beobachterin oder der Beobachter werden sich dieses Prozesses in der Regel nicht bewusst sein; sie oder er wird beispielsweise argumentieren, doch gesehen zu haben, dass sich die andere Person gelangweilt hat. Für eine mangelnde Übereinstimmung zwischen 2 Beurteilenden kommen mehrere Gründe infrage: Unterschiede in der Wahrnehmung einzelner Hinweisreize und in deren Verarbeitung zu einem Urteil, beispielsweise aufgrund unterschiedlicher Gewichtungen der Hinweisreize.

Urteilsprozess mit Zuschreibung von Eigenschaften oder Zuständen

Die Brunswik'sche Linsengleichung

Zuvor haben wir das Brunswik'sche Linsenmodell im Kontext von Verhaltensbeobachtungen betrachtet. Es ist jedoch auch darüber hinaus anwendbar. Es beschreibt ganz allgemein, unter welchen Randbedingungen Urteilende eine zu beurteilende Gegebenheit (aus der Realität) korrekt beschreiben oder vorhersagen können. Wir erinnern uns: „Beschreiben“ und „Vorhersagen“ sind zentrale Ziele der Psychologischen Diagnostik (vgl. ▶ Kap. 1). Das Brunswik'sche Linsenmodell ist also für die Psychologische Diagnostik von zentraler Bedeutung. Unter welchen Randbedingungen es zu einem hohen Zusammenhang zwischen menschlichen (diagnostischen) Urteilen einerseits und der zu beurteilenden Gegebenheit (z. B. Persönlichkeitsmerkmale, Rückfälligkeit von Straftäterinnen und Straftätern) andererseits kommt, wird von der Brunswik'schen Linsengleichung spezifiziert. Dort wird der Zusammenhang zwischen menschlichen Urteilen und der Realität als „achievement“ (r_a) bezeichnet. Dieses achievement kann nur hoch sein, wenn 3 wesentliche Bedingungen gegeben sind:

1. Urteilende nutzen die zur Verfügung stehenden Hinweise (cues) auf eine systematische (und damit auch statistisch vorhersagbare) Art und Weise. Es erscheint logisch, dass Urteilende, die Hinweise immer wieder anders, ja geradezu beliebig nutzen, keine guten und mit der Realität korrespondierenden Urteile fällen können. In der Brunswik'schen Linsengleichung wird das Ausmaß, in dem die gefällten Urteile durch die vorliegenden Hinweisreize statistisch zu erklären sind, „response linearity“ (R_s) genannt.
2. Die vorliegenden Hinweisreize sind tatsächlich dazu geeignet, die zu beurteilende Gegebenheit zu beschreiben oder vorherzusagen. Es liegt auf der Hand, dass auch eine systematische Verwertung der Hinweisreize durch Urteilende zu keinem treffsicheren Urteil führt, wenn die Hinweise nichts mit der Realität zu tun haben. Das wäre so, als wolle man anhand der Schädelform, des Geburtsmonats und der Haarfarbe die Persönlichkeit von Menschen einschätzen. Das Ausmaß, in dem die zu beurteilende Gegebenheit tatsächlich durch die vorliegenden Hinweisreize beschreibbar oder vorhersagbar ist, wird „environmental predictability“ (R_e) genannt.
3. Urteilende gewichten Hinweisreize so, dass dies deren Relevanz für die Vorhersage der Realität entspricht. Anders ausgedrückt: Die tatsächlich wichtigsten Hinweisreize werden von Urteilenden auch am stärksten gewichtet, weniger wichtige Hinweise werden das Urteil nur marginal beeinflussen. Man könnte auch sagen, dass die Nutzung der Hinweise symmetrisch zu deren Relevanz sein sollte (vgl. ▶ Abschn. 2.6.3.4). Auch diese Forderung leuchtet ein, wenn man sich ein einfaches Beispiel überlegt. Die Prognose der Rückfälligkeit von Straftäterinnen und Straftätern wird nicht gut gelingen, wenn die damit betrauten forensische Psychologinnen und Psychologen den eigentlich wichtigsten Indikator (Anzahl an Vorstrafen) nur am Rande berücksichtigen. In der Brunswik'schen Linsengleichung wird das Ausmaß, in dem sich die Relevanz der Hinweisreize und deren Nutzung symmetrisch verhält, „matching index“ (G) genannt.

Das Ausmaß, in dem diese 3 Randbedingungen erfüllt sind, bestimmt – multiplikativ miteinander verknüpft –, wie gut Urteile die Realität beschreiben oder vorhersagen können. Die gesamte Linsengleichung lautet (Karelaia und Hogarth 2008):

$$r_a = R_s \times R_e \times G + C \times \sqrt{(1 - R_s^2) \times (1 - R_e^2)}$$

Zu der multiplikativen Verknüpfung der 3 benannten Randbedingungen wird in der Formel noch etwas addiert: das Ausmaß, in dem durch die Hinweisreize unerklärte Anteile des Urteilendenverhaltens ($1 - R_s^2$) und der Realität ($1 - R_e^2$)

miteinander korrelieren (C). So könnte es sein, dass Urteilende Hinweise nutzen, die bei der statistischen Prognose nicht entdeckt und berücksichtigt wurden, aber tatsächlich relevant sind.

Manche Leserinnen und Leser werden sich vielleicht fragen: Welchen praktischen Nutzen hat diese Linsengleichung? Wenn diagnostische Urteile vorliegen und die Realität ebenfalls bekannt ist, kann ich dann nicht beides miteinander korrelieren und kenne damit die Güte der Urteile, also r_a ? Im Fall der Prognose von Rückfällen im Strafvollzug könnte man ja einfach die Einschätzung des Rückfallrisikos durch forensische Psychologinnen und Psychologen mit den später tatsächlich beobachteten Rückfallraten vergleichen. Die Antwort lautet: ja, das kann man. Aber was tut man, wenn diese Korrelation (also das r_a) gering ist? Dann ist zunächst unklar, woran das liegt. Operiert man mit der Brunswik'schen Linsengleichung, so lässt sich feststellen, welche Randbedingung(en) suboptimal sind:

- Ist die Realität durch die verfügbaren Hinweisreize schlichtweg nur schlecht zu beschreiben (geringes R_e)?
- Nutzen Urteilende die Hinweise nicht systematisch (geringes R_s)?
- Oder nutzen Urteilende Hinweise besonders stark, die eigentlich nicht besonders wichtig sind (geringes G)?

Sobald man identifiziert hat, wodurch die geringe Übereinstimmung zwischen Urteilen und Realität begründet ist, lassen sich gezielte Maßnahmen ableiten. Im Falle einer geringen R_e müssen bessere Hinweisreize gefunden werden. Im Falle geringer R_s oder G müssen Urteilende besser geschult werden.

Manchmal ist es auch einfach nur interessant, herauszufinden, was Menschen tun, wenn sie eine bestimmte Gegebenheit beurteilen sollen. Gosling et al. (2002) wollten herausfinden, ob Fremde anhand eines Zimmers die Persönlichkeit der Bewohnerin bzw. des Bewohners beurteilen können. Dazu baten sie 83 Studierende bzw. „frische“ Absolventinnen und Absolventen, ihr eigenes Zimmer zu zeigen sowie einen Persönlichkeitsfragebogen auszufüllen. Drei geschulte Mitarbeiterinnen und Mitarbeiter inspizierten die Zimmer akribisch und notierten mögliche Hinweisreize. So wurde beispielsweise festgehalten, ob die Zimmer sauber waren oder nicht und ob sie modern eingerichtet waren oder nicht. Der eigentliche Teil der Studie bestand nun darin, dass bis zu 6 fremde Personen die Zimmer sichteten und ein Urteil über die Persönlichkeitsmerkmale (Big Five) der Bewohnerin bzw. des Bewohners abgaben. Dabei zeigte sich, dass es Fremden ziemlich gut gelang, die Offenheit für Erfahrungen der Zielpersonen zu beurteilen (Korrelation zwischen Fremdurteil und Selbstbeschreibung der Bewohnerinnen und Bewohner = .65). Besonders hilfreich war dabei der Hinweisreiz, ob die im Zimmer befindlichen Bücher thematisch divers oder einheitlich waren. Am schlechtesten gelang die Einschätzung der Verträglichkeit. Hier nutzten die Urteilenden Hinweise, die dafür irrelevant waren (z. B. oder der Raum einladend gestaltet war). Zudem fanden sich nur wenige Hinweisreize, die tatsächlich aussagekräftig für die Verträglichkeit der Bewohnerinnen und Bewohner waren.

3.6.4 Gütekriterien von Beobachtungs- und Beurteilungsverfahren

Der Haupteinwand gegen Beobachtungs- und Beurteilungsverfahren richtet sich gegen deren angeblich zu geringe *Objektivität*, also die Unabhängigkeit des Ergebnisses von der Person, die das Verfahren durchführt und auswertet. Bei Zeichen- und Kategoriensystemen kann die *Übereinstimmung* der Registrierungen ermittelt werden.

Beobachtenden- bzw.
Beurteilendenübereinstimmung

Cohens Kappa oder Intraklassenkorrelation

3

Haben die Daten Nominalskalenniveau – ein Verhalten liegt vor oder nicht vor –, ist Cohens Kappa (Cohen 1960) das geeignete statistische Verfahren, um die Übereinstimmung zweier Beobachterinnen bzw. Beobachter numerisch auszudrücken. Die prozentuale Übereinstimmung stellt dagegen keine geeignete Maßzahl dar, weil sie stark von der Auftretenshäufigkeit der einzelnen Verhaltensweisen abhängt. Ist die Auftretenshäufigkeit sehr hoch oder sehr niedrig, kommt es leicht zu scheinbar hohen Übereinstimmungen. Bei intervallskalierten Variablen wird die Intraklassenkorrelation berechnet.

Berechnung von Cohens Kappa

Zur Berechnung von Cohens Kappa benötigt man lediglich den Anteil der vorliegenden Übereinstimmungen (p_o) und den Anteil der Übereinstimmungen, der zufallsbedingt zu erwarten wäre (p_e)

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Zur Illustration verwenden wir ein Zahlenbeispiel von Wirtz und Kutschmann (2007, S. 4). Es zeigt die Übereinstimmung von 2 beurteilenden Personen in ihren dichotomen Urteilen („auffällig“ vs. „unauffällig“) für 200 beurteilte Personen.

		Beurteilende Person B:		
		auffällig	unauffällig	Σ
Beurteilende Person A:				
auffällig		74	25	99
unauffällig		24	77	101
Σ		98	102	200

Insgesamt stimmen die beiden urteilenden Personen in $74 + 77 = 151$ Urteilen überein. Damit liegt p_o bei $151/200 = 0,755$. Der Anteil der Übereinstimmungen, der zufallsbedingt zu erwarten wäre, berechnet sich wie folgt aus den 4 Randsummen:

$$p_e = \frac{99}{200} \times \frac{98}{200} + \frac{101}{200} \times \frac{102}{200} = 0,5001$$

Somit ergibt sich ein κ von:

$$\kappa = \frac{0,755 - 0,5001}{1 - 0,5001} = 0,51$$

Die Frage, zu welchem Gütekriterium die Übereinstimmung von Beobachtenden bzw. Beurteilenden gehört, ist nicht leicht zu beantworten. Die Definition „Objektivität“ bedeutet, dass die Ergebnisse eines diagnostischen Verfahrens unabhängig davon zustande kommen, wer die Untersuchung, die Auswertung und die Interpretation durchführt“ (► Abschn. 2.6.1), lässt sich auch auf die Beobachtung bzw. Beurteilung übertragen. Objektivität bedeutet dann, dass die Ergebnisse einer Verhaltensbeobachtung – dabei handelt es sich unzweifelhaft um ein diagnostisches Verfahren – unabhängig davon zustande kommen, wer die Beobachtung durchführt, auswertet und interpretiert. Die Übereinstimmung der Beobachtenden bzw. Beurteilenden wird in der Literatur häufig auch als sog. „Interrater-Reliabilität“ bezeichnet.

Gründe für die Übereinstimmung von Beobachtungsergebnissen

Wie kommt es, dass die Beobachtungen (und das gilt auch für die Beurteilungen) mehrerer Personen mehr oder weniger gut übereinstimmen? Folgende Faktoren wirken hier zusammen:

1. Die Beobachtungssituation
2. Das Beobachtungsinstrument
3. Die Personen, die mit dem Instrument arbeiten (die Beobachtenden)

Die Beobachtungssituation kann mehr oder weniger „schwer“ sein. Viele Akteure/Akteurinnen, schneller Ablauf, sehr kurze Verhaltenssequenzen, teilweise verdeckte Personen etc. machen es schwer, das Verhalten zu beobachten. Wir betrachten im Folgenden nur 2 Quellen der Übereinstimmung: das Messinstrument und die Menschen, die damit arbeiten. Beide Faktoren sind in der Praxis gemeinsam und untrennbar dafür verantwortlich, dass die Beobachtungsergebnisse (oder die Beurteilungen) hoch oder niedrig übereinstimmen.

Ein fiktives Beispiel soll dies erläutern: Eine Beobachtungsverfahren dient der Erfassung der Selbstsicherheit von Bewerberinnen und Bewerbern. Die Kurzform besteht aus 3 Items, und zwar „nimmt unaufgefordert Platz“, „bittet um ein Getränk“ und „eröffnet das Gespräch“ (Antwortmodus: ja – nein). Als Beobachtende fungieren mit Bewerbungsgesprächen gut vertraute Mitarbeiterinnen und Mitarbeiter einer Personalberatungsgesellschaft. Die Beobachtendenübereinstimmung (Interrater-Reliabilität) fällt bei allen 3 Items und damit für das ganze Instrument extrem hoch aus. Liegt dieses erfreuliche Ergebnis am Instrument (hohe Durchführungsobjektivität) oder an den Personen (hohe Interrater-Reliabilität)? Wir können nur feststellen: Die zur Beobachtung eingesetzten Personen konnten mit der 3-Item-Version sehr gut umgehen. Wollen wir beide Varianzquellen trennen, müssen wir das Instrument und die beurteilenden Personen variieren und beispielsweise auch die Langform mit 10 Items verwenden und zudem unerfahrene Praktikantinnen und Praktikanten zur Beobachtung einsetzen.

Ohne entsprechende Forschungsergebnisse ist es voreilig, das Instrument oder die Personen, die damit arbeiten, für eine hohe oder niedrige Übereinstimmung verantwortlich zu machen. Mit der Verwendung der Begriffe „Durchführungsobjektivität“ oder „Interrater-Reliabilität“ treffen wir aber solche Zuweisungen.

Kennwert der Objektivität oder Reliabilität?

Cohens Kappa beschreibt ebenso wie andere Übereinstimmungsmaße, wie gut Beobachtungen oder Beurteilungen durch andere Beobachter oder Beobachterinnen reproduzierbar sind. Das Ausmaß der Übereinstimmung wird durch mehrere Faktoren determiniert, die zum Teil dem Messinstrument, zum Teil den situativen Bedingungen, unter denen es eingesetzt wird, und zum Teil den beobachtenden Personen zuzuordnen sind. Eigenschaften des Instruments wie sprachlich eindeutige Benennungen des zu beobachteten Verhaltens und die Berücksichtigung der Kapazität der Beobachtenden durch eine angemessene Anzahl an zu beobachtenden Verhaltensweisen, ggf. eine zeitliche Trennung von Beobachtungs- und Protokollierungsphasen sowie ggf. die Verwendung von Videoaufnahmen, die wiederholt angeschaut werden können, fördern eine gute Übereinstimmung. An wichtigen vorteilhaften situativen Bedingungen sind zu nennen: Standardisierung der Beobachtungssituation, die zu beobachtenden Verhaltensweisen sollten in vorher bekannten Phasen mit einer gewissen Wahrscheinlichkeit auftreten, die Beobachtungssituation sollte so kurz sein, dass Ermüdung und nachlassende Aufmerksamkeit der Beobachterinnen und Beobachter auszuschließen sind. Seitens der beobachtenden Personen stellen ein erfolgreich absolviertes Training, Erfahrung (Übung) und Vertrautheit mit der Beobachtungssituation günstige Bedingungen dar.

Faktoren, die eine hohe Übereinstimmung fördern

Validität

Zur Beurteilung der *Konstruktvalidität* sind bei Verfahren, die mehrere Verhaltensmerkmale erfassen sollen, Angaben zur Interkorrelation der Skalen bzw. Faktorenanalysen der Skalen relevant. Korrelationen mit Fragebögen, die das gleiche Merkmal erfassen sollen, kommen eventuell infrage. Für die *Kriteriumsvalidität* können Gruppenvergleiche und Veränderungen durch eine bewährte Intervention aufschlussreich sein. Werden Beobachtungs- oder Beurteilungsdaten zur Vorhersage von Berufserfolg oder von anderen quantifizierbaren Kriterien verwendet, liefert die Korrelation zwischen Prädiktor und Kriterium einen gut interpretierbaren Kennwert. Besteht der Anspruch, alle für ein Merkmal in einer bestimmten Situation relevanten Verhaltensweisen zu erfassen, ist die *Inhaltsvalidität* relevant, die idealerweise über ein Rating von Expertinnen und Experten ermittelt wird.

Halo-Effekt

Fehlerquellen Sowohl für mangelnde Übereinstimmung wie auch Validitätsprobleme werden Urteilsfehler und Antworttendenzen verantwortlich gemacht. Urteilsfehler betreffen, wie ihr Name sagt, den Urteilsprozess. Am bedeutsamsten ist vermutlich der *Halo-Effekt*, auch als „Hofeffekt“ bezeichnet, der sich in unangemessen hohen Korrelationen der Urteile einer Beobachterin bzw. eines Beobachters zwischen verschiedenen Merkmalen einer Person äußert. Er soll dadurch zustande kommen, dass das Urteil über ein herausragendes Merkmal die Beurteilungen anderer Merkmale einer Person „überstrahlt“. Beispielsweise wirkt eine Person sehr freundlich; andere Merkmale werden daraufhin positiver beurteilt. Der Halo-Effekt wirkt sich vor allem negativ auf die Validität von Verhaltensbeurteilungen aus. Die Übereinstimmung zwischen 2 Beurteilerinnen oder Beurteilern kann durch den Halo-Effekt künstlich erhöht werden. Abschwächen lässt sich dieser Effekt, indem man nicht alle Merkmale einer einzelnen Person nacheinander beurteilen lässt, sondern zunächst nur den Ausprägungsgrad eines einzelnen Merkmals bei allen einzuschätzenden Personen erhebt.

Logischer Fehler

Ebenfalls durch unangemessen hohe Interkorrelationen mehrerer Merkmale ist der *logische Fehler* definiert. Bei ihm dominiert nicht ein vorherrschendes Merkmal die restlichen Urteile, sondern die implizite Annahme über die logische Zusammengehörigkeit bestimmter Merkmale. Anstatt den Ausprägungsgrad einzelner Verhaltensweisen unabhängig einzuschätzen, wird er aufgrund impliziter Zusammenhangsannahmen erschlossen. Beispielsweise hat eine beurteilende Person die implizite Theorie, dass dominante Menschen aggressiv und nicht teamfähig sind. Die Beobachtung von dominantem Verhalten kann dazu führen, dass die Person als aggressiver und weniger teamfähig eingestuft wird.

Primacy- und Recency-Effekt

Unter *Primacy- und Recency-Effekt* versteht man, dass sich Beurteilerinnen und Beurteiler übermäßig stark von den Beobachtungen beeinflussen lassen, die sie am Anfang bzw. am Ende der Beobachtungsphase gemacht haben. Beide Effekte können sich bei Verhaltensbeobachtungen insbesondere bemerkbar machen, wenn nicht kontinuierlich und zeitnah protokolliert wird, sondern erst am Ende einer Beobachtungsphase. Eine plausible Erklärung für das Zustandekommen eines Primacy-Effekts ist, dass sich schon sehr früh ein Gesamteindruck bildet und dieser durch weitere Beobachtungen zu bestätigen versucht wird. Ein Recency-Effekt kann entstehen, wenn die Beobachtungen lange Zeit ein unklares Gesamtbild ergeben; die Beurteilerinnen und Beurteiler stützen ihr Urteils dann zu stark auf die zuletzt beobachteten und damit sehr präsenten Verhaltensweisen.

Eine *Beobachterdrift* stellt vor allem bei der Verhaltensbeobachtung eine Fehlerquelle dar. Die Genauigkeit der Beobachtung lässt entweder über die Beobachtungsphase nach oder sie nimmt zu. Für ein Nachlassen kommen insbesondere Müdigkeit und mangelnde Motivation infrage; die Aufmerksamkeit der Beobachterinnen und Beobachter nimmt ab, und sie entdecken zunehmend weniger Verhaltensweisen, die sie eigentlich protokollieren müssten. Die Drift kann auch in die andere Richtung gehen. Manchmal steigern die Beobachterinnen und Beobachter ihre Leistung, weil sie mit dem Instrument vertrauter werden und deshalb zunehmend mehr relevante Verhaltensweisen entdecken. Dafür verantwortlich ist meist eine ungenügende Schulung.

Beobachterdrift

Reaktivität ist ein Fehler, bei dem sich das Verhalten der beobachtenden Person durch die Messung, also durch das Beobachtetwerden, verändert. Dabei kann das Verhalten der Beobachterinnen und Beobachter den Effekt verstärken, etwa durch die Wahl der Kleidung oder der Beobachtungsposition (im Blickfeld der beobachteten Person) oder deutlich sichtbares Protokollieren. Generell ist es hilfreich, der beobachteten Person Gelegenheit zu geben, sich an die Anwesenheit der Beobachterin oder des Beobachters oder auch einer Kamera zu gewöhnen.

Reaktivität

Antworttendenzen betreffen nicht den Urteilsprozess, sondern die Abbildung des Urteils auf einer Skala. Ein *Mildefehler* wird darin sichtbar, dass eine Beurteilerin oder ein Beurteiler überwiegend positive Urteile abgibt (Abb. 3.35a). Ein *Strengefehler* ist dagegen an einer Tendenz zu negativen Urteilen erkennbar (Abb. 3.35b). Eine *zentrale Tendenz*, d. h. eine Bevorzugung mittlerer Skalenpositionen, lässt sich an Urteilen im mittleren Skalenbereich mit eingeschränkter Varianz erkennen (Abb. 3.35c). Durch Verwendung von Skalen ohne exakt erkennbaren Mittelwert, also beispielsweise mit Stufen von 1 bis 4 oder 1 bis 6, kann diese Tendenz etwas gemildert werden. Bei einer *Tendenz zu Extremurteilen* ergibt sich dagegen eine erhöhte Varianz der Urteile (Abb. 3.35d).

Antworttendenzen

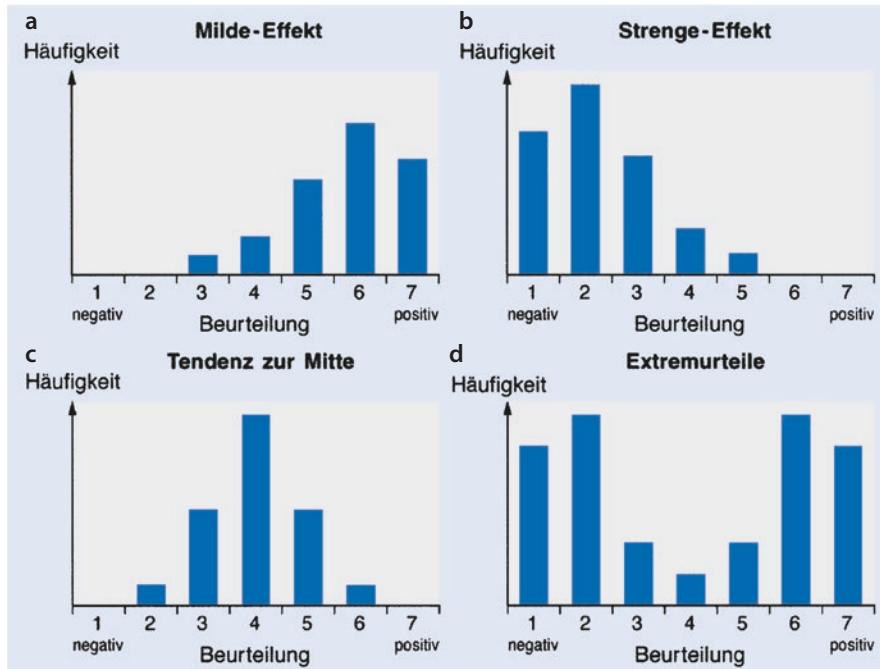


Abb. 3.35 a–d Illustration einiger Beurteilungsfehler bei der Einschätzung von Eigenschaftsausprägungen oder Verhaltenstendenzen

Instrument sorgfältig konstruieren und erproben

Maßnahmen zur Optimierung der Objektivität Maßnahmen zur Verhinderung oder zumindest Verringerung von Urteilsfehlern sowie zur Optimierung der Objektivität können am Beobachtungs- bzw. Beurteilungsinstrument und an den Menschen, die damit arbeiten, ansetzen. Ein Manual, in dem das Instrument verständlich erläutert und die Durchführung und Auswertung genau beschrieben werden, ist wünschenswert. Bei der systematischen Verhaltensbeobachtung ist es wichtig, dass die Beobachtungseinheiten operational definiert und durch Beispiele erläutert werden. Die Anzahl der Beobachtungseinheiten muss den Kapazitäten der Beobachterinnen und Beobachter in der Beobachtungssituation angemessen sein. Auch für Beurteilungsinstrumente ist zu fordern, dass die Merkmale operational definiert (Woran ist beispielsweise „Kooperation“ zu erkennen? Etwa: „Die Person arbeitet spontan oder auf eine Aufforderung hin mit anderen konstruktiv zusammen“) und durch konkrete Beispiele erläutert werden (z. B. bietet Hilfe an, reagiert positiv auf Bitte um Unterstützung, übernimmt Aufgaben). Eventuell ist es hilfreich, Skalenstufen verbal zu verankern, also kurz verhaltensnah zu beschreiben, was etwa typisch für eine niedrige und für eine hohe Ausprägung ist (z. B. „sehr niedrig“=arbeitet auf Bitte hin nicht mit, „sehr hoch“=übernimmt spontan Aufgaben, um andere zu entlasten).

! Ein Instrument, das am Schreibtisch entstanden ist, wird in der Regel noch nicht einsatzbereit sein. Erst bei einer Erprobung unter realistischen Bedingungen zeigt sich, wie gut es funktioniert und wo Änderungsbedarf besteht.

Training von Beobachtenden und Beurteilenden hilfreich

Zu den wichtigsten Maßnahmen, die an den Menschen ansetzen, die mit einem Instrument arbeiten sollen, ist ein Beobachtenden-/Beurteilendentraining. Dazu gibt es verschiedene Trainingskonzepte. Ein Ansatz besteht darin, die Beurteilerinnen und Beurteiler mit einem Bezugssystem vertraut zu machen, um damit zu arbeiten. Das Grundprinzip des *Frame-of-Reference-Trainings* besteht darin, verhaltensnah zu vermitteln, wie sich eine durchschnittliche, eine hohe und eine niedrige Ausprägung des zu beurteilenden Merkmals äußert. Damit werden alle Beurteilerinnen und Beurteiler sozusagen auf ein einheitliches Bezugssystem eingeschworen. Roch et al. (2012) haben eine Metaanalyse zur Effizienz von Frame-of-Reference-Trainings durchgeführt. Als durchschnittlicher Effekt des Trainings auf die Akkuratheit der Urteil wurde eine Effektstärke von $d=0,50$ ermittelt. Allerdings waren die meisten Studien mit Studierenden durchgeführt worden. In den wenigen Studien mit Angestellten erwiesen sich die Trainings als effektiver ($d=0,80$). Das Training ist nicht nur bei der Leistungsbeurteilung, sondern auch bei anderen Beurteilungen effizient. Das verwendete Trainingsmaterial spielt eine wichtige Rolle. Werden die Szenarien nur schriftlich beschrieben, so ist dies relativ ineffizient ($d=0,22$); besser ist es, zusätzlich Videomaterial zu verwenden ($d=0,50$).

Fazit Die Verhaltensbeobachtung ist eine Methode, die diagnostisch relevante Informationen direkt über das Verhalten liefert. Sie ergänzt Informationen, die mit anderen Methoden (Tests, Interview etc.) erhoben wurden. Die *freie Verhaltensbeobachtung* wird in fast allen diagnostischen Situationen mit eingesetzt, beispielsweise um das Verhalten bei der Testdurchführung zu beschreiben. Bei einer *systematischen Verhaltensbeobachtung* wird meist nur ein bestimmter Verhaltensbereich (z. B. Aggression) erfasst. Sie erfordert umfangreiche Vorbereitungen zur Erstellung eines Zeichen- oder Kategoriensystems. Zeichensysteme verlangen lediglich die Identifizierung von zumeist eindeutig definierten Verhaltensweisen, während Kategoriensysteme eine Einordnung von Verhaltensweisen zu relativ abstrakten Kategorien verlangen und damit weitaus stärker von Interpretationsleistungen abhängig sind. Beide

Varianten führen zu Quantifizierungen des Verhaltens, indem entweder die Häufigkeit oder die Dauer von Verhaltensweisen registriert werden. Am wenigsten aufwendig ist die *Verhaltensbeurteilung*, bei der die Beurteilerinnen und Beurteiler jedoch weitreichende Schlussfolgerungen über die „Bedeutung“ des Verhaltens anstellen müssen. Diese Methode ist daher besonders anfällig für Urteilsfehler.

Grundsätzlich ist bei der Verhaltensbeobachtung und bei der Verhaltensbeurteilung zu bedenken, dass sie zu Reaktivität führen können: Die beobachteten Personen verhalten sich anders als gewohnt, weil sie sich beobachtet fühlen. Deshalb ist immer abzuwägen, ob die Beobachtung verdeckt erfolgen soll oder zumindest eine Eingewöhnungsphase vorgeschaltet wird. Wenn der Aufwand einer Fremdbeobachtung zu groß ist oder wenn ethische Bedenken wegen der Verletzung der Privatsphäre bestehen, ist eine Selbstbeobachtung vorzuziehen. Dazu stehen erprobte Vorgehensweisen zur Erfassung von Verhalten im Alltag zur Verfügung. Alle Varianten der systematischen Verhaltensbeobachtung und die Verhaltensbeurteilung sollten erst nach einem gründlichen Training der Beobachterinnen und Beobachter eingesetzt werden.

Weiterführende Literatur

Eine umfangreiche und systematische, aber nicht immer leicht zu lesende deutschsprachige Darstellung zum Thema „systematische Verhaltensbeobachtung“ bietet das Buch von Faßnacht (1995). Für Unterrichtszwecke sowie die praktische Durchführung und Auswertung von Verhaltensbeobachtungen und -beurteilungen kann ein Buchbeitrag von Stemmler und Margraf-Stiksrud (2015) empfohlen werden. Über Methoden zur Berechnung der Beurteilendenübereinstimmung informieren auch Wirtz und Caspar (2002) sowie Wirtz und Kutschmann (2007). Eine sehr gute Übersicht über Methoden zur Erfassung von Verhalten im Alltag gibt das von Mehl und Conner (2012) herausgegebene *Handbook of research methods for studying daily life*. Ein Buchbeitrag von Höft und Kersting (2018) befasst sich mit der Anwendung der Verhaltensbeobachtung und -beurteilung im eignungsdiagnostischen Kontext.

Übungsfragen

Abschn. 3.6:

- Worin unterscheiden sich freie und systematische Verhaltensbeobachtung?
- Nach welchen 6 Aspekten kann Verhaltensbeobachtung näher charakterisiert werden?
- Welche Möglichkeiten der Selbstbeobachtung und der Registrierung von Verhalten und Ereignissen im Alltag kennen Sie? Beschreiben Sie diese Ansätze kurz!
- Was ist beim Einsatz einer verdeckten Beobachtung zu beachten?
- Welche Vorteile bieten elektronische Aufzeichnungen (z. B. mittels Smartphone oder Tablet) gegenüber der Verwendung von Verhaltensprotokollen in Papierform?
- Was bedeuten Selektion, Segmentierung und Quantifizierung?
- Was bedeuten time und event sampling?
- Wodurch zeichnet sich ein Kategoriensystem aus?
- Nennen Sie ein Beispiel für ein Kategoriensystem, und geben Sie an, was damit erfasst werden soll!
- Worin unterscheidet sich die Verhaltensbeurteilung von der Verhaltensbeobachtung?
- Beschreiben Sie mithilfe des Brunswik'schen Linsenmodells die Entstehung einer Verhaltensbeurteilung!
- Mit welchen statistischen Verfahren wird die Übereinstimmung zwischen Beobachtenden und Beurteilenden berechnet?
- Nennen Sie wichtige Fehler, die bei der Verhaltensbeobachtung und -beurteilung auftreten können!

3.7 Diagnostisches Interview

3

Informationserhebung mittels Gespräch

Anamnese, Exploration, Einstellungs- oder Auswahlgespräch

Interviews dienen generell der Erhebung von Informationen mittels Gespräch. Interviews werden beispielsweise von Journalistinnen und Journalisten dazu eingesetzt, politische Standpunkte in Erfahrung zu bringen, oder von Marktforscherinnen und Marktforschern zur Erkundung von Einstellungen gegenüber bestimmten Produkten. In anderen Kontexten nennt man Interviews beispielsweise Verhör oder Zeugenbefragung. Der Zusatz „diagnostisches“ Interview macht klar, dass es sich um ein Interview zu diagnostischen Zwecken handelt (Abb. 3.36). Damit wird ein ganz bestimmter Verwendungszweck impliziert.

Innerhalb der Psychologischen Diagnostik sollte der Begriff „diagnostisches Interview“ als Oberbegriff für alle Methoden zur Erhebung von diagnostisch relevanten Informationen mittels Gespräch verstanden werden. Je nach Art der zu erhebenden Informationen können diagnostische Interviews zusätzlich spezifiziert werden. Unter einer *Anamnese* oder einer Anamneseerhebung wird in Anlehnung an den Sprachgebrauch der Medizin die gesprächsweise Erkundung der Vorgeschichte einer Erkrankung oder Störung verstanden. Der Begriff „Exploration“ stammt ursprünglich aus der Psychiatrie und bezeichnet die Erkundung des subjektiven Lebensraums einer Person (vgl. Trost 1996). Anamnese und Exploration können daher Bestandteil, in bestimmten Fällen auch alleiniger Bestandteil, eines diagnostischen Interviews sein. Diagnostische Interviews können aber auch Funktionen übernehmen, die nicht unter die Begriffe „Anamnese“ und „Exploration“ fallen. Sie dienen u. a. der Erhebung von Informationen zur Eignung einer Person für



Abb. 3.36 Interviews in verschiedenen Kontexten. (a: © deagonez/stock.adobe.com, b: © Microgen/stock.adobe.com, c: © Photographee.eu/stock.adobe.com, d: © Syda Productions/stock.adobe.com)

einen bestimmten Beruf oder für ein bestimmtes Studium. Nach dem primären Verwendungszweck kann das diagnostische Interview dann auch näher als *Einstellungs- oder Auswahlgespräch* bezeichnet werden.

Definition

Diagnostisches Interview ist der Überbegriff für Methoden zur Erhebung von diagnostisch relevanten Informationen mittels Gespräch. Mit Begriffen wie Anamnese, Exploration, Einstellungsgespräch oder Auswahlgespräch kann der Verwendungszweck oder die Zielsetzung eines diagnostischen Interviews näher bestimmt werden. Diagnostische Interviews unterscheiden sich durch den Grad ihrer Standardisierung.

Die Definition macht keine Aussage darüber, wie das Gespräch geführt wird. Grundsätzlich bestehen hier folgende Möglichkeiten:

- Face-to-face Interview
- Telefoninterview
- Interview via Internet:
 - Synchron (mit oder ohne Blickkontakt)
 - Asynchron (Fragen per E-Mail schicken; Beantwortung, wenn Interviewpartner/-innen Zeit hat)

Telefoninterviews werden schon lange bei der Vorauswahl von Bewerberinnen und Bewerbern eingesetzt. Da heute Laptops verbreitet sind und diese meist mit Kamera und Mikrofon ausgestattet sind, kann das Interview oft auch als Videokonferenz durchgeführt werden. Bei beiden Varianten muss eine Interviewerin oder ein Interviewer mit jeder Bewerberin und jedem Bewerber individuell ein Interview führen. Bei einer neuen Form von Interviewtechnik ist dies nicht mehr erforderlich (s. dazu das Interview mit Sara Lindemann). Die Bewerberin oder der Bewerber wird per E-Mail aufgefordert, sich auf einer Videoplattform des Unternehmens einzuloggen. Das Interview kann in einem gesetzten Zeitrahmen von beispielsweise einer Woche an einem beliebigen Ort durchgeführt werden. Meist besteht die Möglichkeit, vorher ein kurzes Interview zu Übungszwecken durchführen, bei Bedarf auch mehrmals. Die Fragen werden meist in Textform auf dem Monitor eingeblendet.

Die Durchführung von Interviews über Telefon oder unverschlüsselt über das Internet wirft auch Fragen nach der Vertraulichkeit des Gesagten auf. Kann das, was man sagt, auch in falsche Hände geraten?

Wie vertraulich ist ein Interview über das Internet?

James und Busher (2012) weisen darauf hin, dass eine Überwachung der Internetaktivität verbreitet ist. Sie zitieren eine Warnung, die schon 2003 erfolgte: „The U.S. central government can read and track e-mails sent by people to anywhere in the world“ (Hessler et al. 2003).

Die Vertraulichkeit von Angaben in einem Interview, das über das Internet geführt wird oder das unverschlüsselt per E-Mail verschickt wird (etwa als Anhang zu einem Gutachten), ist also nicht automatisch gewährleistet und bedarf konkreter Schutzmaßnahmen. Das Bundesamt für Sicherheit in der Informationstechnik (2020) warnt: „Die meisten Menschen kommunizieren nahezu jeden Tag per E-Mail. Doch nur Wenigen ist bewusst, dass unverschlüsselte E-Mails mitgelesen oder verändert werden können. Ähnlich einer Postkarte werden hier private oder sensible Informationen nicht vor unerwünschten Mitlesern geschützt.“

Interview mit Dipl.-Psych. Sara Lindemann, Head of Customer Success und Mitgründerin der viasto GmbH, zum Thema „zeitversetzte Videointerviews“



Sara Lindemann, Head of Customer Success und Mitgründerin der viasto GmbH.

Frau Lindemann, Ihr Unternehmen bietet Lösungen für zeitversetzte Videointerviews an. Was versteht man unter zeitversetzten Videointerviews? Bei einem zeitversetzten Videointerview zeichnet ein Kandidat oder eine Kandidatin seine/ihrre Antworten auf Interviewfragen mithilfe einer Software eigenständig auf. Die Antworten werden dann automatisch in den Account des Recruiter bzw. der Recruiterin geladen, der/die dann ggf. gemeinsam mit Kolleginnen und Kollegen die Bewertung der Antworten vornimmt. Es gibt unterschiedliche Anbieter für diese Technologie. Wichtig ist natürlich, dass die Aufzeichnung des Interviews vollständig strukturiert abläuft, also dass alle Kandidaten die gleiche Zeit zur Vorbereitung und zur Beantwortung der Fragen haben.

Welche Vorteile bieten zeitversetzte Videointerviews gegenüber herkömmlichen Interviewformen?

Zum einen fällt natürlich zunächst die Effizienz und die Flexibilität aller Beteiligten ins Auge. Dadurch, dass Bewerbende und Unternehmen keinen gemeinsamen Termin finden müssen, wird die Auswahl beschleunigt. Was ich viel spannender finde ist, dass die Standardisierung in der Durchführung und der Auswertung (zumindest, wenn die Softwarelösung nach dieser Logik aufgebaut ist) dazu führt, dass weniger Beobachtungs- und Bewertungsfehler auftreten. Dieser Effekt entsteht nicht nur durch die komplette Strukturierung des Interviews, sondern auch dadurch, dass die Person, die die Auswertung der Antworten vornimmt, nur einen sehr geringen „cognitive load“ in der Bewertungssituation erlebt. Sie ist nicht in das Gespräch involviert und kann sich bei Bedarf die Antworten auch mehrmals ansehen. Je geringer dieser Faktor ist, desto unwahrscheinlicher treten Wahrnehmungsfehler wie Halo-Effekte oder andere Phänomene auf. So konnten wir in unterschiedlichen Studien beispielsweise auch herausfinden, dass Kandidatinnen und Kandidaten aufgrund des Geschlechts oder des kulturellen Hintergrunds weniger diskriminiert werden (Kroll und Ziegler 2016).

Ist mit Einbußen bei der psychometrischen Qualität zu rechnen?

Eher im Gegenteil. Wie gesagt ist der hohe Standardisierungsgrad eigentlich die optimale Umsetzung einer Interviewdiagnostik nach Lehrbuch, wie sie eigentlich in der Praxis nur sehr selten zum Einsatz kommt. Dreh- und Angelpunkt sind natürlich gute Fragen, die auch ohne Interaktion, also ohne vertiefendes

Nachfragen funktionieren. Und hier kommen wir an den Punkt, an dem (zumindest bei unserer Software) in Zukunft künstliche Intelligenz zum Zug kommen wird. Wir haben einen Algorithmus entwickelt, der die Anforderungskriterien der Stelle, aber auch Faktoren wie die Wettbewerbssituation um Talente und noch zahlreiche weitere Faktoren in die Erstellung eines Interviewleitfadens einbezieht. Dadurch nehmen wir den Unternehmen den absolut erfolgskritischen Schritt der Formulierung von guten Fragen ab. Nur die Bewertung wird aktuell dann noch, gestützt durch die Bewertungssystematik in der Software, durch die Menschen vorgenommen (zumindest noch ...).

Sicherlich gibt es auch Nachteile – was ist zu beachten?

Wie ich bereits ausgeführt habe, ist es wichtig, die eben genannten qualitätskritischen Aspekte im Hinterkopf zu behalten. Außerdem ist es sehr wichtig, für Akzeptanz bei den Kandidaten/Kandidatinnen zu sorgen, indem man ihnen den Mehrwert verdeutlicht (= fairere Chancen durch die Strukturierung der Durchführung und Beurteilung). Dieser Mehrwert ist für Bewerbende nicht immer gleich erkennbar. Hier hat es sich bewährt, vorab eine klare und transparente Kommunikation über die Karriereseite oder andere Kanäle zu pflegen, damit sich Kandidatinnen und Kandidaten der Vorteile bewusst werden. Viele vermissen das Feedback, das sie durch einen Gesprächspartner bzw. eine Gesprächspartnerin bekommen. Dass diese positive Bestätigung eben oft auch mit Beurteilungsfehlern zusammenhängt, ist einem Laien natürlich nicht bewusst.

Welche technologischen Entwicklungen hinsichtlich der Durchführung und Auswertung diagnostischer Interviews erwarten uns in näherer Zukunft?

Wenn man den ganz großen Sprung in die Zukunft macht, sagen wir 2030, dann kann man sich fragen, ob es überhaupt noch Interviews geben wird. Sie dienen ja letztlich der beiderseitigen Informationsgewinnung und der Bewertung, ob Kandidatin bzw. Kandidat und Stelle bzw. Unternehmen zusammenpassen. Ich bin überzeugt davon, dass das in Zukunft nicht mehr am besten in Form von Interviews erreicht werden kann (Stichwort: Matching-Algorithmen über digitale Spuren in Social-Media-Profilen oder digitale Footprints etc.). In den jetzt kommenden Jahren werden die Interviews, die wir noch führen, immer stärker technologieunterstützt ablaufen. Kein Protokollieren mehr, da wir dann speech-to-text-basierte Features haben, kein Entwickeln von Fragen, da das die künstliche Intelligenz übernimmt. Dynamische Anpassung von Fragen, weil die künstliche Intelligenz „mithört“ und Zwischenauswertungen macht und die besten Folgefragen vorschlägt. Und die Ergebnisformulierung und die Gutachtenerstellung wird ebenfalls parallel von der künstlichen Intelligenz erledigt. Und irgendwann ermöglichen uns die besseren Alternativen zum Interview dann wohl auch, dass die Personen, die dann später miteinander arbeiten werden, sich nur noch einmal auf einen Kaffee treffen. Die Entscheidung, dass man zueinander passt, ist dann bereits gefallen. Dann plaudert man einfach nur noch ein bisschen ... – klingt doch eigentlich auch ganz nett, oder?

Interviews sind in unterschiedlichem Ausmaß standardisiert. Völlig *unstandardisiert* ist ein Interview, wenn nur dessen Zweck feststeht (etwa eine klinische Diagnose stellen) und sich die Fragen im Laufe des Gesprächs erst ergeben. Zwei unstandardisierte Interviews könnten sich daher selbst bei identischer Fragestellung und gleicher Struktur sehr stark voneinander unterscheiden. Völlig *standardisiert* ist ein Interview, wenn jede Frage vorher genau festgelegt und immer im gleichen Wortlaut vorzutragen ist.

Standardisiert – strukturiert

Strukturiert vs. standardisiert

In der Fachliteratur trifft man häufig den Begriff „strukturiertes“ Interview an. „Strukturiert“ bedeutet eigentlich nur, dass ein Interview systematisch aufgebaut ist. Die inhaltliche Struktur ergibt sich bei klinischen Interviews zur Diagnostik nach DSM-5 oder ICD-10 durch die diagnostischen Kriterien für psychische Störungen bzw. Persönlichkeitsstörungen. Levashina et al. (2014) stellen für eignungsdiagnostische Interviews fest, dass „strukturierte“ Interviews sehr unterschiedlich definiert werden. Sie stellen verschiedene Kriterien für ein gutes „strukturiertes“ Interview auf. Die Struktur ergibt sich aus vorher durchgeführten Anforderungsanalysen. Es folgen Maßnahmen zur Standardisierung. So sollen etwa allen interviewten Personen die gleichen Frage gestellt werden.

Eine sachlich angemessene Struktur ist eine notwendige, aber noch keine hinreiche Voraussetzung für ein gutes Interview. Für eine gute psychometrische Qualität ist es unbedingt erforderlich, dass auch die Durchführung und die Auswertung standardisiert sind. Umgekehrt macht eine Standardisierung allein noch kein gutes Interview aus. Man stelle sich nur vor, dass die falschen Fragen hoch standardisiert vorgegeben und ausgewertet würden. Sowohl die Bezeichnung „strukturiert“ als auch „standardisiert“ greift also zu kurz. Eigentlich müsste die Bezeichnung „standardisierte strukturierte Interviews“ laut. Wir wollen hier aber keinen neuen Begriff einführen und verwenden die Begriffe standardisiert und strukturiert im Zusammenhang mit Interviews synonym. Der jeweils fehlende Teil muss also mitgedacht werden. Ein standardisiertes Interview ist auch strukturiert und umgekehrt.

Unterschiedliche Grade der Standardisierung

Zwischen unstandardisierten und standardisierten Interviews sind viele Abstufungen denkbar. Mit dem Begriff „halbstandardisiert“ wird zumeist zum Ausdruck gebracht, dass eine Standardisierung angestrebt wird, eine Festlegung auf exakte Wortlaute und Abfolgen von Fragen aber nicht gewollt oder nicht möglich ist. Die Vorlage für ein halbstandardisiertes Interview kann etwa aus einer Liste von Themen mit stichpunktartig charakterisierten Fragen bestehen. Auch Abstriche bei der Standardisierung der Auswertung können ein Merkmal von halbstandardisierten Interviews sein. Beispielsweise könnte ein aus festgelegten Fragen bestehendes Einstellungsgespräch in seiner Durchführung hoch standardisiert sein, und die Auswertung sieht am Ende so aus, dass „aus dem Bauch heraus“ entschieden wird, ob eine Bewerberin oder ein Bewerber nun geeignet ist oder nicht.

Standardisierung der Auswertung bedeutet, dass die Antworten nach festen Regeln verwertet werden. Meist müssen die Personen, die das Interview durchführen oder in beobachtender Funktion daran teilnehmen, bei jeder Interviewfrage die Antwort(en) bewerten. Dabei ist es unerlässlich, dass das Merkmal oder die Anforderung präzise und verständlich definiert oder beschrieben wird. Beispielsweise könnte das Merkmal „Machtstreben“ als

„versucht andere Menschen zu beeinflussen oder zu manipulieren“ erläutert werden. Beurteilt wird, in welchem Ausmaß das Merkmal, das mit der Frage angezielt wurde, ausgeprägt ist oder einer Anforderung entspricht. Das Spektrum möglicher Antworten reicht von „ja – nein“ bis zu mehrstufigen Ratingskalen. Da die meisten Menschen unterschiedliche Vorstellungen davon haben, was viel und wenig ist, sind Verhaltensverankerungen hilfreich. Für jede Antwortstufe oder zumindest die Extrempole werden typische Antworten oder Antwortelemente genannt. Im folgenden Beispiel wird gezeigt, wie die Antworten auf eine standardisierte Frage mithilfe einer verhaltensverankerten Skala standardisiert ausgewertet wird. Eine Besonderheit ist hier, dass sich Anforderungsmerkmale zu einem Merkmal auch addieren können, so dass im Beispiel bis zu 3 Punkten möglich sind.

Fragen (zur Führungsmotivation) mit standardisierter Auswertung

Haben Sie während Ihrer Schulzeit irgendwelche Führungsaufgaben übernommen, beispielsweise in der Schule oder in einem Verein? Waren Sie beispielsweise einmal Klassensprecher oder Jugendgruppenleiter? Haben Sie vielleicht für eine größere Gruppe alleine eine Aktivität, beispielsweise eine Theateraufführung oder eine Jugendfreizeit, organisiert? (Nachfrage: Wie lange haben Sie das gemacht?).

Bewertung der Antwort:

<input type="checkbox"/>	(2 Punkte)	Führungsaufgabe(n) mit konkreter Funktionsbezeichnung (Schülersprecher/-in, Leitung von Jugendgruppen im Verein etc.) für insgesamt mindestens 2 Jahre
<input type="checkbox"/>	(1 Punkt)	(Weitere) Führungsaufgabe(n) mit konkreter Funktionsbezeichnung für mindestens 3 Monate
<input type="checkbox"/>	(0 Punkte)	Keine oder nur unbedeutende Führungsaufgabe übernommen (z. B. Sammelbestellung organisiert)
<hr/>		Punkte insgesamt

Auch die *Verrechnung der Antworten* wird vorher festgelegt. Meist werden bei jeder Frage Punkte für die Antworten (die sich wie im Beispiel oben auch addieren können) vergeben. Bei Ratingskalen werden die Stufen meist von 0 bis z. B. 6 numerisch kodiert; der Skalenwert ergibt sich aus der angekreuzten Stufe. Es ist ratsam, innerhalb eines Interviews bei einem Antwortformat zu bleiben.

Am Ende sind lediglich die Punkte über alle Fragen, wenn erforderlich auch separat für alle Themenbereiche, zu addieren. Dabei sind vorab festgelegte Gewichtungen möglich. Wenn beispielsweise bei einem eignungsdiagnostischen Interview 4 Anforderungsmerkmale, z. B. Teamorientierung, Flexibilität, Belastbarkeit und Führungsmotivation, erfasst werden. Da die Führungsmotivation besonders wichtig ist, weil es sich um eine Leistungsfunktion handelt, könnten die auf der Dimension Führungsmotivation erzielten Punkte beispielsweise mit dem Faktor 1,5 gewichtet werden. Liegen für jeden Merkmalsbereich unterschiedlich viele Fragen vor, es sind aber alle Bereiche gleich wichtig, lässt sich die ungleiche Anzahl von Fragen ebenfalls durch eine entsprechende Gewichtung korrigieren. Es empfiehlt sich, die Gewichte so zu wählen, dass man das Ergebnis auf der verwendeten Skala ablesen kann. Dieses Vorgehen wird in □ Abb. 3.39 anhand eines Beispiels demonstriert. Darüber hinaus können auch Mindestwerte (Cut-off-Werte; ▶ Abschn. 5.1.1) für die einzelnen Bereiche festgelegt werden.

Antworten gewichten

3.7.1 Standardisierte strukturierte Interviews

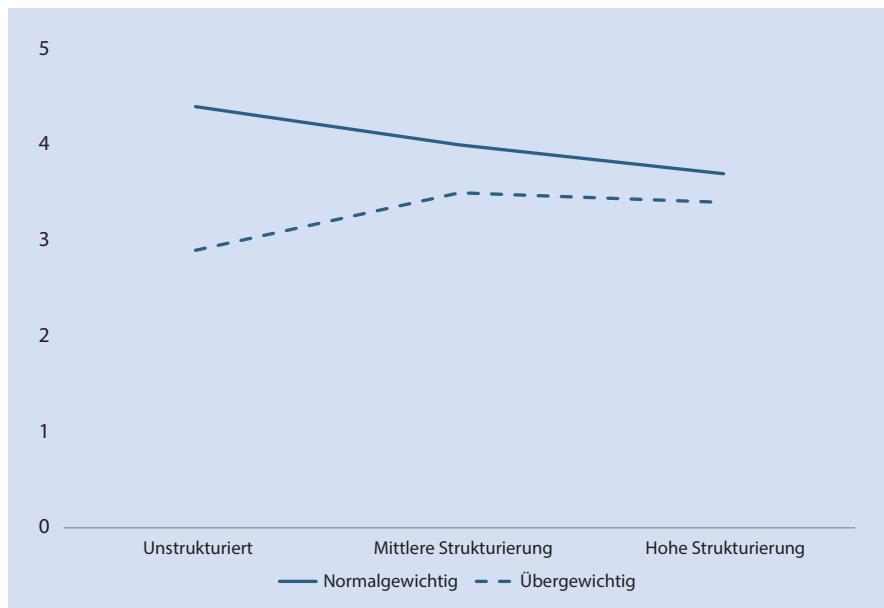
Die besten Belege für den Nutzen eines strukturierten und standardisierten Vorgehens liegen für Eignungsinterviews vor. Zahlreiche Einzelstudien und Metaanalysen zeigen die Überlegenheit strukturierter gegenüber unstrukturierten Interviews, wenn es um die (statistische) Vorhersage beruflicher Leistungen geht. Levashina et al. (2014) zufolge liegen insgesamt 12 Metaanalysen vor, die eine Überlegenheit „strukturierter“ Eignungsinterviews gegenüber unstrukturierten bezüglich der Validität und der Reliabilität belegen. Einen eindrucksvollen Beleg für den Vorteil strukturierter Eignungsinterviews liefert eine Studie von Kutcher und Bragger (2004). Der Autor und die Autorin ließen unstrukturierte, mittelstrukturierte und hoch strukturierte Interviews mit einer Bewerberin durchführen, die entweder als normalgewichtig auftrat oder einen Fatsuit trug, also übergewichtig erschien. Der Inhalt ihrer Antworten wurde über die unterschiedlichen Interviewformen und die beiden Erscheinungsbilder der Bewerberin konstant gehalten. Versuchspersonen sollten sich die Interviews ansehen und eine Kompetenzeinschätzung der Bewerberin abgeben (auf einer Skala von 0 bis 5). Die Ergebnisse sind in ▶ Abb. 3.37 dargestellt. Es zeigte sich eine deutliche Benachteiligung der scheinbar übergewichtigen Bewerberin bei unstrukturierten und mittelstrukturierten Interviews, nicht aber bei hoch strukturierten Interviews (für weitere Ausführungen zu Beurteilungsfehlern in unterschiedlich stark strukturierten Interviews s. auch ▶ Abschn. 3.7.1).

Einsatz im klinischen Bereich und zur Eignungsdiagnostik

Strukturierte Interviews kommen vor allem im klinischen Bereich und in der Eignungsdiagnostik zum Einsatz. Generell ist der Einsatz von standardisierten Interviews immer dann zu erwägen, wenn sich eine Fragestellung oft wiederholt oder wenn die Ergebnisse mit denen anderer Personen verglichen werden sollen. Im klinischen Bereich kommen bestimmte Fragestellungen immer wieder in gleicher Form vor.

Deshalb wurden speziell zur Diagnostik von psychischen Störungen Interviews entwickelt, die mit zugehörigem Manual über Testverlage vertrieben werden. Die Diagnostik von psychischen Störungen nach einem der

Diagnosekriterien



■ Abb. 3.37 Mittlere Kompetenzbewertung einer Bewerberin in einem Interview; Skala von 0 bis 5. (Aus Kutcher und Bragger 2004, mit freundlicher Genehmigung von John Wiley and Sons)

bekannten Klassifikationssysteme (DSM-5 oder ICD-10) folgt einem einfachen Prinzip: Für jede Störung wurde eine Liste möglicher Symptome per Konvention festgelegt. Diese Symptome stellen die Kriterien für das Vorliegen einer bestimmten psychischen Störung dar. Weiterhin wurde, ebenfalls per Konvention, festgelegt, wie viele Kriterien erfüllt sein müssen, damit eine Diagnose vergeben werden kann. Beispielsweise wird eine Major Depression (unipolare depressive Störung) nach DSM-5 durch 9 Kriterien definiert. Für eine Diagnose müssen mindestens 5 davon vorliegen. Diese müssen in derselben 2-wöchigen Periode fast jeden Tag auftreten. Eines dieser Kriterien lautet „deutlich verminderter Interesse oder Freude an allen oder fast allen Aktivitäten für die meiste Zeit des Tages“. Dieses Symptom ist übrigens eines von zweien, die unbedingt vorliegen müssen, damit die Diagnose überhaupt gestellt werden darf. Auch ohne eine Standardisierung des klinischen Interviews würde eine Psychologin oder ein Psychologe auch nach diesem Symptom fragen. Entsprechende Fragen könnten lauten:

- „Haben Sie längere Zeit keine Lust gehabt, die Dinge zu tun, die Ihnen sonst Spaß machen?“ Nachfrage: „Wie lange schon?“
- „Haben Sie die Freude an Dingen verloren, die Sie früher gerne gemacht haben? Wenn ja, wie lange geht es Ihnen schon so?“
- „Wofür interessieren Sie sich im Beruf und in der Freizeit? Worüber können Sie sich jeden Tag freuen?“ Nachfrage: „Nun vergleichen Sie den momentanen Zustand einmal damit, wie es Ihnen früher ging. Waren Sie früher stärker interessiert und haben Sie sich öfter über etwas gefreut?“
- „Ich möchte Ihnen zunächst einige Fragen zu Ihrer Stimmung stellen. Während der letzten 4 Wochen: Haben Sie das Interesse oder die Freude an fast allen Aktivitäten verloren, die Ihnen gewöhnlich Freude machen?“

Da von den Antworten letztlich die Diagnose abhängt, ist es naheliegend, die Fragen zu standardisieren, also immer auf die gleiche Weise zu stellen. Genauso das leisten standardisierte klinische Interviews. Alle Diagnostikerinnen und Diagnostiker, die das Strukturierte Klinische Interview für DSM-5-Störungen – Klinische Version (SKID-5-CV) durchführen, stellen die Frage so: „Hatten Sie während dieser Zeit [während des letzten Monats] weniger Interesse oder Freude an Aktivitäten, die Ihnen gewöhnlich Freude machten?“ Oder bei Nichtzutreffen der vorherigen Frage: „Gab es während des letzten Monats [...] eine Zeit, in der Sie das Interesse oder Freude an Aktivitäten verloren haben, die Ihnen gewöhnlich Freude machten?“ (Beesdo-Baum et al. 2019, S. 10).

Die Standardisierung des Interviews endet nicht mit der Festlegung der Fragen. Es werden auch Kategorien zur Bewertung einer Antwort vorgegeben, und die Verrechnung der Bewertungen ist geregelt. Die Diagnostikerin oder der Diagnostiker kreuzt anhand der Antwort an, ob das erfragte Kriterium „sicher vorhanden und kriteriumsgemäß ausgeprägt“ ist. Die Punkte werden für jede Störung addiert, und die Diagnose wird nur gestellt, wenn die geforderte Mindestzahl an Kriterien erfüllt ist und ggf. die obligatorischen Kriterien vorliegen. Andere standardisierte Interviews haben dagegen nur bestimmte Bereiche von Störungen, beispielsweise Essstörungen, zum Gegenstand (► Abschn. 8.4.1). Standardisierte klinische Interviews können auch andere Funktionen haben, als Diagnosen zu stellen. Beispielsweise mögen sie dazu dienen, die Ätiologie einer Störung abzuklären (► Abschn. 8.1) oder den Verlauf bzw. das Ergebnis einer Therapie zu evaluieren (► Abschn. 8.6.1). Ausführliche Informationen zur klinischen Diagnostik finden sich in ► Kap. 8.

Standardisierte Interviews zur Diagnostik psychischer Störungen

Kategorien zur Bewertung der Antworten

Warum nicht gleich schriftliche Fragen vorlegen?

An dieser Stelle werden sich manche Leserinnen und Leser fragen, warum man nicht einfach die ohnehin präzisen Fragen den Klientinnen und Klienten zur Beantwortung vorlegt. Das wäre doch die höchstmögliche Standardisierung einer Befragung.

Die mündliche Face-to-Face-Befragung hat 2 große Vorteile: Sie erlaubt ertens Nachfragen sowohl der interviewten Person (z. B. „Wie ist das gemeint?“) als auch der Interviewerin oder des Interviewers (z. B. bei einer unklaren Antwort). Zweitens erleichtert sie der interviewten Person die Durchführung: Sprechen ist weniger aufwendig als Schreiben. Viele Menschen haben Schwierigkeiten mit dem schriftlichen Ausdruck und der Rechtschreibung, was sich negativ auf die Ergiebigkeit der Antworten auswirken kann.

Mögliche Unterschiede zwischen schriftlicher und mündlicher Befragung wurden auch empirisch untersucht. Zu Essstörungen liegt mit der Eating Disorder Examination (EDE) ein teilstandardisiertes Interview vor, das 4 Bereiche abdeckt (Restraint, Eating Concern, Shape Concern, Weight Concern). Die Befragten sollen zusätzlich angeben, wie häufig bestimmte Verhaltensweisen (wie selbst herbeigeführtes Erbrechen oder Gebrauch von Abführmittel) bei ihnen in einem definierten Zeitraum vorkamen. Da das Interview relativ aufwendig ist, wurde auch eine Fragebogenversion (EDE-Q) dazu entwickelt. Berg et al. (2011) haben in einer Metaanalyse über 16 Studien, in denen jeweils beide Verfahren zum Einsatz kamen, Interview und Fragebogen verglichen. Die Korrelationen der 4 Skalen lagen zwischen $r = .68$ und $.76$ – ein Beleg dafür, dass beide Verfahren nicht genau das Gleiche messen. Der Fragebogen lieferte höhere Skalenwerte als das Interview. Die Effektstärken entsprechen umgerechnet einer Standardwertdifferenz von 3 bis 6 Punkten. Die Häufigkeitsangaben fielen dagegen im Interview geringfügig höher aus als im Fragebogen (umgerechnet 2 Punkte), und die Übereinstimmung war mit $r = .55$ (objektive) bzw. $.37$ (subjektive Items) sehr niedrig. Die Studie sagt nichts darüber aus, welche Methode besser ist. Sie zeigt aber, dass beide im Einzelfall zu unterschiedlichen Ergebnissen führen.

In der Eignungsdiagnostik sollen sich Interviewfragen immer nach den Anforderungen, die mit einer bestimmten Stelle oder einem Studienplatz verbunden sind, richten. Deshalb gibt es hier keine universell einsetzbaren Interviewverfahren. Es haben sich allerdings bestimmte Frageformate etabliert: situative und biografische Fragen.

Mit *biografischen Fragen* möchte man herausfinden, ob eine Bewerberin oder ein Bewerber früher schon unter Beweis gestellt hat, eine bestimmte Anforderung effizient erfüllt zu haben. Es geht also um entsprechendes reales Verhalten in der eigenen Biografie. Die Annahme ist, dass vergangenes Verhalten auch in Zukunft gezeigt wird und damit für die Entscheidung, ob eine Person eine zukünftige Stelle ausfüllen kann, sehr relevant ist. *Situative Fragen* dagegen sind fiktiv. Man schildert eine Situation, in der die Anforderung relevant ist und fragt die Bewerberinnen oder Bewerber, wie sie sich in dieser Situation verhalten würden. Die Annahme ist hierbei, dass die von Bewerberinnen geschilderten Handlungen, die sie in der hypothetischen Situation ausführen würden, auch für reale zukünftige Situationen Gültigkeit haben. Natürlich kann man bei einer passenden Antwort nicht sicher sein, ob die interviewte Person nur weiß, was hier das richtige Verhalten ist, oder ob sie sich später tatsächlich so verhält.

Biografische und situative Fragen

► Beispiel

Es soll überprüft werden, wie gut eine Bewerberin oder ein Bewerber bei Konflikten zwischen Mitarbeiterinnen und Mitarbeitern vermitteln kann.

Biografische Frage: „Schildern Sie bitte eine Situation, in der es zu einem ernsten Konflikt zwischen Ihren Mitarbeiterinnen oder Mitarbeitern kam.“ Nachfrage: „Was haben Sie unternommen, damit dieser Konflikt gelöst wird?“

Situative Frage: „Stellen Sie sich bitte vor, dass es zwischen zwei Mitarbeiterinnen oder Mitarbeitern zu einem ernsten Konflikt kommt. Ein wichtiger Arbeitsauftrag, für den beide gemeinsam verantwortlich waren, wurde nicht erledigt. Beide beschuldigen sich gegenseitig. Was würden Sie unternehmen, damit beide sich wieder vertragen, um künftig wieder gut zusammenarbeiten zu können?“

Kriterien einer guten Antwort: Lädt die Betroffenen zu einem Gespräch ein, gibt beiden Gelegenheit, den Konflikt kurz aus eigener Sicht darzustellen, fasst die zentralen Diskrepanzen nicht wertend zusammen, fordert beide auf, gemeinsam eine Lösung zu finden, begleitet den Prozess der Lösungsfindung moderierend, schlägt geeignete Regeln vor (z. B. die andere Person nicht herabsetzen oder beschuldigen, Lösungen finden, die für beide Seiten akzeptabel sind, konkrete Schritte vereinbaren) und achtet auf deren Einhaltung. ◀

In 2 Metaanalysen (Tab. 3.29) erwiesen sich Interviews mit biografischen Fragen denen mit situativen Fragen als überlegen. Ein Nebenergebnis war, dass die Verwendung von verhaltensverankerten Skalen günstig für die Validität ist. Bei Taylor und Small (2002) betrug die Validität der biografischen

Interviews mit biografischen Fragen
valider

■ Tab. 3.29 Ergebnisse von Metaanalysen zur Kriteriumsvalidität von biografischen und situativen Eignungsinterviews

Metaanalyse	k	N	Spezifikation	Validität ^a
Marchese und Muchinski (1993)	31	3960	Kriterium Berufserfolg	.38
	23	2290	Subjektive Kriterien	.37
	12	1875	Objektive Kriterien	.39
McDaniel et al. (1994)	160	25.244	Kriterium Berufserfolg	.37
	106	12.847	Strukturierte Interviews	.44
	39	9330	Unstrukturierte Interviews	.33
	75	59.844	Kriterium Trainingserfolg	.36
	26	3576	Strukturierte Interviews	.34
	30	47.576	Unstrukturierte Interviews	.36
	90	11.393	Berufserfolg, ein Interviewer	.43
	54	11.915	Berufserfolg, Interviewerteam	.32
Huffcutt et al. (2004)	32	2815	Situative Interviews	.43
	22	2721	Verhaltensbeschreibende Interviews	.51
Taylor und Small (2002)	30	2299	Situative Interviews	.45
	19	1855	Biografische Interviews (nur verhaltensverankerte Beurteilungsskalen)	.56
	29	2142	Situative Interviews	.47
	11	1119	Biografische Interviews	.63

k = Anzahl der Studien, N = Anzahl der Personen in den Studien insgesamt.

^aKorrigierte Validität (Korrekturfaktoren zwischen den Metaanalysen leicht verschieden).

Gute Gründe, auch situative Fragen zu verwenden

3

Fragen können bekannt werden

„Interrater-Reliabilität“

Interviews ohne Verwendung von verhaltensverankerten Skalen .47 und mit verhaltensverankerten Skalen .63. Bei situativen Fragen war der Unterschied nur gering.

Angesichts der Überlegenheit biografischer Interviews stellt sich die Frage, ob man in eignungsdiagnostischen Interviews überhaupt situative Fragen einsetzen sollte. Die Frage lässt sich eindeutig mit „Ja“ beantworten. Erstens sind biografische Fragen nur dann sinnvoll, wenn die interviewte Person eine berufliche Biografie vorzuweisen hat. So fällt es schwer, bei Bewerberinnen und Bewerbern für einen Ausbildungsplatz für bestimmte Anforderungsmerkmale gute Fragen zu finden. Zur Bereitschaft, Verantwortung zu übernehmen, zur Leistungsmotivation oder zur Teamfähigkeit lassen sich Fragen finden, die sich auf das Verhalten in der Schule, in Vereinen oder im Freundeskreis beziehen. Da Verhalten immer auch von situativen Bedingungen abhängt, ist es schwer, den Anteil festzustellen, der mit Eigenschaften der Person zu erklären ist. Beispielsweise kann sich ein junger Mensch sehr teamfähig verhalten, weil dies im Freundeskreis oder im Verein erwartet wird und andere sich auch so verhalten. In diesem Fall könnte es problematisch sein, von früherem Verhalten auf künftiges Verhalten in der Ausbildung zu schließen. Zweitens können sich situative und biografische Fragen gut ergänzen. Sie liefern unterschiedliche Sichtweisen auf künftiges Verhalten im Beruf, in einer Ausbildung oder auch in einer Fußballmannschaft. Eine biografische Frage liefert Erkenntnisse, wie sich jemand künftig in ähnlichen Situationen wie der geschilderten verhalten wird. Es können später aber auch deutlich andere Situationen auftreten, in denen ebenfalls beispielsweise Teamfähigkeit verlangt wird. Es bleibt bei biografischen Fragen oft unklar, wie stark man die Befragungsergebnisse auf das Verhalten in anderen Situationen übertragen kann. Situative Fragen können sehr gut an die bevorstehende berufliche Realität angepasst werden. Man kann die Fragen an kritischen Ereignissen ausrichten, die erfahrungsgemäß in dem Beruf, in dieser Ausbildung oder in dieser Fußballmannschaft vorkommen.

Allerdings kann eine Standardisierung auch Nachteile haben. Man stelle sich vor, dass ein großes Unternehmen zur Auswahl von Auszubildenden, eine Universität zur Auswahl von Studierenden oder eine medizinisch-psychologische Untersuchungsstelle bei der Begutachtung von Menschen, die im Straßenverkehr wegen Alkoholkonsums auffällig geworden sind, stets dieselben Fragen verwenden würden. Die Fragen könnten von Betroffenen weitergegeben werden, eventuell sogar für alle Interessierten in Internetforen. Der Effekt wäre, dass ein Teil der künftig interviewten Personen gut vorbereitet zur Untersuchung erscheint. Die Aussagekraft der Interviewergebnisse wäre damit eingeschränkt. Außerdem beständen berechtigte Zweifel an der Fairness des Verfahrens.

Beurteilung der Gütekriterien An die Gütekriterien von strukturierten Interviews sind die gleichen Anforderungen zu stellen wie bei Tests. Zur Beurteilung der *Objektivität* sind die Maßnahmen zur Standardisierung der Durchführung, Auswertung und Interpretation relevant. Bei der Auswertung, also der Bewertung von Antworten, stellen Maßnahmen zur Standardisierung eine notwendige, aber keine hinreichende Bedingung dar. Es ist nicht davon auszugehen, dass alle Personen, die mit der Auswertung betraut sind, ein und dieselbe Antwort gleich bewerten. Deshalb ist es erforderlich, die Auswertungsübereinstimmung empirisch zu bestimmen. In der Fachliteratur wird die Auswertungsübereinstimmung oftmals als Kennwert der Reliabilität behandelt („Interrater-Reliabilität“; s. dazu auch die Ausführungen in ▶ Abschn. 3.6.4).

Für eignungsdiagnostische Interviews liegen zahlreiche Befunde zur *Reliabilität* (Auswertungsübereinstimmung) vor, die von Huffcutt et al. (2013) in einer Metaanalyse unter die Lupe genommen wurden. Wenig überraschend

ist die Erkenntnis, dass die Beurteilendenübereinstimmung (Interrater-Reliabilität) mit dem Grad der Strukturierung/Standardisierung des Interviews zunimmt. Allerdings spielt dabei eine entscheidende Rolle, unter welchen Bedingungen die Übereinstimmung ermittelt wird. Wenn 2 oder mehr Beurteilerinnen oder Beurteiler gemeinsam ein und dasselbe Interview bewerten (Panel Interview), ist die Übereinstimmung deutlich höher, als wenn 2 Interviewer oder Interviewerinnen mit den gleichen Personen getrennte Interviews durchführen. Als mittlere Auswertungsübereinstimmung ermittelten Huffcutt et al. (2013) .74 bzw. .44. Für hoch strukturierte Interviews betragen die Übereinstimmungskoeffizienten .78 und .61. Dabei sei angemerkt, dass hier alle Studien ausgeschlossen wurden, in denen die Bewertung nach einer gemeinsamen Diskussion erfolgte. (Bei Bewertung nach gemeinsamer Diskussion steigt die Übereinstimmung dramatisch an und beträgt .99 bei mittel- und .98 bei hoch strukturierten Interviews – ein triviales Ergebnis.)

Was bedeutet die Ergebnisse? Dazu müssen wir überlegen, welche Faktoren zu einer hohen Übereinstimmung beitragen. Bei Panel-Interviews sehen die Auswerterinnen und Auswerter stets das gleiche Verhalten der interviewten Person und das zudem im gleichen situativen Kontext. Zudem nehmen sie die nonverbalen Reaktionen der anderen Auswerterin oder des anderen Auswerts wahr. Mit anderen Worten: Sie beeinflussen sich gegenseitig. Nur die separate Interviewdurchführung (gleiche interviewte Person, unterschiedliche Interviewerinnen und Interviewer) liefert eine angemessene Schätzung der Reliabilität des Interviews als Messinstrument. Wir können diese Schätzung auch als Retest-Reliabilität bezeichnen. Und die ist mit .61 keinesfalls als hoch, nicht einmal als zufriedenstellend zu bezeichnen! Darüber hinaus wurde deutlich, dass eine hohe Standardisierung vorteilhaft für die Reliabilität ist.

Die Reliabilität kann grundsätzlich auch als Konsistenzmaß bestimmt werden. Das ist aber nur angemessen, wenn das Merkmal homogen ist. Diese Voraussetzung wird allenfalls bei klinischen Interviews, die sich auf eine einzige Störung beziehen, erfüllt sein.

Bei der Beurteilung der *Validität* kommt der *Inhaltsvalidität* oft eine große Bedeutung bei. Decken die Fragen alle relevanten Aspekte der psychischen Störung ab, die mit dem Interview diagnostiziert werden soll? Die Frage kann bejaht werden, wenn das Interview direkt nach den „offiziellen“ Kriterien psychischer Störungen wie dem DSM-5 oder der ICD-10 konstruiert wurde. Bei einem eignungsdiagnostischen Interview liegt es nahe, zu überprüfen, ob es alle zuvor erhobenen Anforderungen abdeckt. Das ist aber zu kurz gedacht, weil ein Interview nicht bei allen Anforderungen die erste Wahl ist. Die Intelligenz wird besser mit einem Intelligenztest erfasst, und fachliche Qualifikationen lassen sich oft sehr viel objektiver und einfacher anhand von Dokumenten (Zeugnissen, Abschlüssen etc.) feststellen. Deshalb ist es angebracht, die Inhaltsvalidität nur für den Teil der Anforderungen zu beurteilen, die explizit mit dem Interview abgedeckt werden sollen.

Über die *Konstruktvalidität von klinischen Interviews* ist wenig bekannt. Grundsätzlich ist es möglich, die Übereinstimmung mit Fragebögen oder Verhaltensbeurteilungen zum gleichen Störungsbild zu überprüfen. Erstaunlicherweise führte eine Literatursuche nach Metaanalysen zur Validität von klinischen Interviews (im Juli 2019) zu keinem „richtigen“ Treffer. Eine Metaanalyse von Markon et al. (2011) befasst sich jedoch mit der Reliabilität und Validität von Instrumenten zur Psychopathologie mit Schwerpunkt auf einem Vergleich zwischen kategorialen und dimensionalen Ansätzen. Insgesamt 20 Studien mit klinischen Stichproben wurden gefunden. Allerdings werden nur in einem Teil dieser Studien strukturierte klinische Interviews eingesetzt worden sein. Die Messinstrumente wurden leider nicht als Moderatorvariable

Retest-Reliabilität relativ niedrig

Inhaltsvalidität bei klinischen Interviews wichtig

Wenige Studien zur Konstruktvalidität von klinischen Interviews

berücksichtigt. Auch liegen keine Informationen dazu vor, woran die Verfahren validiert wurden. Die in Tab. 7 berichtete Validität von .32 für kategoriale Instrumente in klinischen Stichproben liefert nur eine vage Vorstellung davon, in welchem Bereich die Konstruktvalidität von klinischen Interviews liegen könnte. Man kann hier aber unabhängig von der schwachen Befundlage einwenden, dass standardisierte Interviews bereits den „Goldstandard“ darstellen. Zu niedrige Korrelationen mit einem „schwachen“ Referenzverfahren wird man daher eher nicht dem Interview anlasten.

Sind psychische Störungen Konstrukte?

Sind psychische Störungen reale, direkt beobachtbar Gegebenheiten in der Natur (auch wenn sie mit zum Teil vielen Symptomen komplex sein können) oder handelt es sich um Konstrukte? Für die 2. Interpretation spricht, dass psychische Störungen neu in ein Klassifikationssystem wie dem DSM oder der IDC hinzukommen können oder auch wieder entfernt werden. Wie sie operationalisiert werden, also welche beobachtbaren Symptome eine bestimmte Störung definieren, ist eine Frage der Konvention. Vergleicht man klinische Klassifikationssysteme wie die ICD-10 und das DSM-5, lassen sich vielfach Diskrepanzen feststellen: Eine Störung in einem System ist nicht exakt identisch mit der in einem anderen. Diese Argumente sprechen dafür, dass psychische Störungen zumindest partiell „menschengemacht“ sind. Aus pragmatischer Sicht stellt sich oftmals die Frage, ob ein diagnostisches Verfahren (Interview, Fragebogen, Beobachtungsverfahren) geeignet ist, eine bestimmte psychische Störung valide zu messen. Dazu wird man die Übereinstimmung mit anderen diagnostischen Verfahren überprüfen – und in vielen Einzelfällen wird es so sein, dass jedes Verfahren eine andere Antwort gibt, ob die Störung vorliegt oder nicht bzw. wie ausgeprägt sie ist. Auch wenn man die Frage, ob psychisch Störungen Konstrukte sind, weiter kontrovers diskutieren kann, ist es für die Evaluierung diagnostischer Verfahren (einschließlich strukturierter klinischer Interviews) nützlich, von Konstrukten auszugehen.

Konstruktvalidität eignungsdiagnostischer Interviews hängt vom Interviewtyp ab

Bei der Beurteilung von *eignungsdiagnostischen Interviews* ist die *Konstruktvalidität* sehr wichtig. Misst ein Interview beispielsweise wie gewollt die Anforderungen an einen bestimmten Beruf oder zu einem bedeutsamen Teil auch die Fähigkeit, sich gut darzustellen, Intelligenz oder beispielsweise andere Anforderungen bzw. Eigenschaften? Hier interessiert also auch die diskriminante Validität. Salgado und Moscoso (2002) haben in einer Metaanalyse die vorliegenden Untersuchungen danach eingeteilt, ob es sich eher um ein konventionelles Interview handelte oder eher um ein verhaltensbezogenes. Beide Interviewtypen weisen Gemeinsamkeiten auf; die wesentlichen Unterschiede liegen darin, dass in den als „verhaltensbezogen“ definierten Interviews überwiegend nach früheren Aktivitäten und Erfahrungen sowie nach konkretem Verhalten in früheren oder auch in fiktiven Situationen gefragt wurde. Diese beiden Fragetypen werden auch als biografische und situative Fragen bezeichnet. Als konventionell wurden Interviews eingestuft, in denen besonders nach biografischen Fakten (z. B. Noten, Größe des zu verwalteten Etats) und Selbstbewertungen („Was sind Ihre Stärken und Schwächen?“, „Wo wollen Sie in fünf Jahren stehen?“) gefragt wurde. Die Beurteilungen in konventionellen Interviews wiesen moderate Zusammenhänge mit Intelligenz, emotionaler Stabilität und sozialen Fertigkeiten auf. Die Beurteilungen in verhaltensbezogenen Interviews korrelierten dagegen niedriger mit Intelligenz und emotionaler Stabilität, aber deutlich höher mit sozialen Fertigkeiten (Tab. 3.30) und einigen anderen Variablen (insbesondere mit Berufserfahrung: $r = .71$).

■ Tab. 3.30 Ausgewählte Ergebnisse einer Metaanalyse zur Konstruktvalidität eignungsdiagnostischer Interviews

Art des Interviews	Validitätskriterium		
	Intelligenz	Emotionale Stabilität	Soziale Fertigkeiten
Konventionell	.41	.38	.46
Verhaltensbezogen	.28	.08	.65

Quelle: Metaanalyse von Salgado und Moscoso (2002); für Varianzeinschränkung im Prädiktor sowie für Reliabilität von Prädiktor und Kriterium korrigierte Korrelation

Drei Schlussfolgerungen können aus dieser Metaanalyse gezogen werden. Erstens unterscheiden sich die beiden Interviewtypen in dem, was sie (mit)erfassen. Zweitens hängt das Eignungsurteil auf der Basis des Interviews relativ eng mit den sozialen Fertigkeiten der interviewten Personen zusammen. Je stärker diese Fertigkeiten ausgeprägt sind, desto positiver werden die Personen beurteilt. Leider fehlen in der Publikation nähere Angaben dazu, wie soziale Fertigkeiten in den Untersuchungen operationalisiert wurden. Es lässt sich nicht feststellen, ob das Interviewergebnis durch die sozialen Fertigkeiten der Bewerberinnen und Bewerber beeinflusst oder sogar verfälscht wird. In manchen Berufen gehören soziale Fertigkeiten zum Anforderungsprofil, stellen also ein Eignungsmerkmal dar. Die Funktion der sozialen Fertigkeiten und ihr Beitrag zur Vorhersage von Berufs- und Ausbildungserfolg muss daher durch weitere Untersuchungen geklärt werden. Drittens ist der Zusammenhang mit Intelligenz nicht hoch. Dieses Ergebnis ist für die Berufseignungsdiagnostik sehr erfreulich. Intelligenz ist ein sehr guter Prädiktor für Ausbildungs- und Berufserfolg. Damit besteht die Chance, dass Interviews eine inkrementelle Validität über die Intelligenz hinaus haben. Würden die Interviewergebnisse hoch mit Intelligenz korrelieren, könnte man die Interviews durch einen Intelligenztest ersetzen. Man kann die Ergebnisse zum Vergleich beider Interviewtypen so interpretieren, dass verhaltensbezogene Interviews mit sozialen Fertigkeiten eher das messen, was sie vermutlich messen sollen und deren Ergebnisse weniger von der Intelligenz und der emotionalen Stabilität der interviewten Personen abhängen.

Die *Kriteriumsvalidität* ist bei *eignungsdiagnostischen Interviews* von zentraler Bedeutung. Sie werden zumindest in vielen Fällen dazu eingesetzt, Kandidatinnen und Kandidaten zu „entdecken“, die eine Ausbildung oder ein Studium erfolgreich abschließen oder in ihrem Beruf erfolgreich sein werden. Damit liegen auch passende Kriterien vor (Noten als Maß für Ausbildungs- oder Studienerfolg, Beurteilung durch Vorgesetzte beispielsweise als Maß für Berufserfolg). Zur Kriteriumsvalidität von Eignunginterviews liegen sehr viele Studien vor, die auch mehrfach metaanalytisch zusammengefasst wurden. Die Studien befassen sich mit der Vorhersage von Ausbildungs- und Berufserfolg. Die Validitätskoeffizienten insbesondere für standardisierte Interviews liegen im hohen Bereich (► Abschn. 6.2).

Ein spezielles Thema ist die Eignung für einen Studiengang. Eine Metaanalyse von Hell et al. (2007) ergab, dass Interviews nur eine sehr niedrige Validität aufweisen, also nur schwach mit späteren Studienleistungen korrelieren. Zur Auswahl von Medizinstudierenden wurden in Deutschland zeitweise auch Interviews eingesetzt. Die Ergebnisse waren enttäuschend: Die per Interview ausgewählten Studierenden wiesen deutlich schlechtere Leistungen in der Zwischenprüfung auf als diejenigen, die nach einem Test oder einer Kombination von Test und Abiturnote ausgewählt worden waren (Nauels und Klieme 1994). Bei einem Großeinsatz von standardisierten Auswahlgesprächen ist zu bedenken, dass die Fragen schnell bekannt werden und damit eine gezielte Vorbereitung auf das Interview stattfinden kann. Letztlich wird dann nur noch gemessen, wie gut sich jemand informiert und vorbereitet hat.

Verhaltensbezogene Interviews erfassen eher soziale Fertigkeiten

Hohe Validität von Interviews zur Vorhersage von Ausbildungs- und Berufserfolg

Interviews zur Vorhersage von Studienerfolg wenig valide

Erscheinungsbild, Impression Management, verbales und nonverbales Verhalten

Einschätzung im Interview anfällig für Urteilsfehler

► Beispiel

An einer großen deutschen Universität wurde ein standardisiertes Interview als Bestandteil des Auswahlverfahrens für Studierende im Fach Psychologie eingeführt. Ein Onlinenachrichtenmagazin berichtete darüber und gab die Fragen preis. Auch Ratschläge, wie man am besten darauf antwortet, waren zu lesen. Das Interview wurde daraufhin nicht mehr eingesetzt. ◀

Beurteilungsfehler Die Beurteilungen, die Menschen Interviews erfahren, sind für Beurteilungsfehler anfällig. Es sind die gleichen Fehler, die auch bei der Verhaltensbeurteilung zum Tragen kommen (► Abschn. 3.6.4). Die interviewten Personen tragen auch aktiv dazu bei, dass Beurteilungsfehler zu ihren Gunsten auftreten. Es ist bekannt, dass sich die Bewerberinnen und Bewerber auf unterschiedliche Weise vorteilhaft darzustellen versuchen; Levashina und Campion (2007) zufolge tun dies über 90 %. Die Forschung hat sich mit 3 potenziellen Einflussfaktoren so intensiv befasst, dass inzwischen hinreichend viele empirische Befunde dazu vorliegen: das äußere Erscheinungsbild, Impression Management sowie verbales und nonverbales Verhalten. Zum äußeren Erscheinungsbild gehören die körperliche Attraktivität (gutes Aussehen), aber auch angemessene Kleidung und ein gepflegtes Äußeres (Friseur etc.; □ Abb. 3.38). Unter Impression Management versteht man strategisches Verhalten, das dazu dient, die Stelle zu bekommen. Das Verhalten kann darauf abzielen, sich selbst positiv darzustellen und die Interviewerin oder den Interviewer durch nettes, liebenswürdiges Auftreten oder Konformität der Meinung, aber auch durch vorgetäuschte Hilflosigkeit für sich einzunehmen. Vorteilhaftes verbales Verhalten kann etwa eine klare Aussprache sein; günstige nonverbale Verhaltensweisen sind etwa Lächeln, Blickkontakt oder eine zugewandte Körperhaltung.

Eine Metaanalyse von Barrick et al. (2009) demonstriert eindrucksvoll, wie stark sich das äußere Erscheinungsbild und Selbstdarstellungsstrategien auf die Bewertung der interviewten Personen und damit auch negativ auf die Validität von Eignungsinterviews auswirken. Sie belegt, dass mit dem Grad der Standardisierung die Verzerrung durch Beurteilungsfehler stark abnimmt – ein starkes Argument für die Verwendung standardisierter Interviews.



□ Abb. 3.38 Ein gepflegtes Äußeres und gutes Aussehen führen zu einer besseren Beurteilung im Einstellungsgespräch. (© goodluz/stock.adobe.com)

Diagnostische Verfahren

Die in □ Tab. 3.31 aufgeführten Ergebnisse zeigen Folgendes: Erstens besteht ein deutlicher Zusammenhang zwischen dem äußeren Erscheinungsbild, dem Auftreten von Impression Management und verbalem und nonverbalem Verhalten der Bewerberin oder des Bewerbers mit der Beurteilung, die sie oder er durch die Interviewerin oder den Interviewer erfährt. Am stärksten ist der Einfluss des äußeren Erscheinungsbildes, am schwächsten der des verbalen und nonverbalen Verhaltens. Zweitens nimmt der Einfluss dieser Faktoren mit dem Grad der Strukturiertheit des Interviews stark ab; in hoch strukturierten Interviews ist die Beurteilung insbesondere für das äußere Erscheinungsbild und für Impression Management wenig anfällig. Drittens korrelieren diese Einflussfaktoren auch schwach mit dem Berufserfolg. Dieser wird häufig über Vorgesetztenbeurteilungen erfasst; möglicherweise spiegeln die Korrelationen den Einfluss der Selbstdarstellung auf die Urteile der Vorgesetzten wider. Vielleicht gehen diese Variablen aber auch mit realem Erfolg im Beruf einher – welche Interpretation richtig ist, lässt sich anhand der berichteten Ergebnisse nicht feststellen. Jedenfalls korrelieren die genannten Einflussfaktoren wesentlich höher mit der Beurteilung im Interview als mit dem Berufserfolg. Dies lässt den Schluss zu, dass Beurteilungsfehler die Validität von wenig standardisierten eignungsdiagnostischen Interviews massiv negativ beeinflussen.

Für die eignungsdiagnostische Praxis ergibt sich aus diesen Forschungsergebnissen, dass Eignungsinterviews so stark wie möglich strukturiert und standardisiert werden sollten. Das betrifft nicht nur die Fragen, sondern auch die Auswertung der Antworten. Weiterhin bietet es sich an, die Interviewerinnen und Interviewer gut zu trainieren und sie im Rahmen des Trainings auch für die genannten Präsentationsstrategien zu sensibilisieren.

Vorbereitung auf ein Eignungsinterview Den meisten Bewerberinnen und Bewerbern wird bewusst sein, dass ihr Abschneiden im „Auswahlgespräch“ einen großen Einfluss darauf hat, ob sie die Stelle bekommen oder nicht. Da

Präsentationsstrategien mindern die Validität wenig standardisierter Interviews

Standardisierung der Interviews und Interviewertraining wichtig

Training für die interviewten Personen

□ **Tab. 3.31** Metaanalyse zur Beziehung zwischen Präsentationsstrategien im Eignungsinterview und Bewertung im Interview sowie Berufserfolg

Einflussfaktor	r_c insgesamt	r_c für Interviewstruktur			r_c Berufserfolg
		Niedrig	Mittel	Hoch	
Äußeres Erscheinungsbild insgesamt	.53	.88	.52	.18	.14
– Professionelles Erscheinungsbild	.48				
– Körperliche Attraktivität	.54				
Impression Management insgesamt	.47	.46	.34	.21	.15
– Impression Management direkt	.55				
– Werbung für sich selbst	.32				
– Interviewer/in für sich einnehmen	.26				
Verbales und nonverbales Verhalten	.40	.69	.47	.37	.23
– Nur verbales Verhalten	.34				
– Nur nonverbales Verhalten	.40				

Quelle: Barrick et al. (2009, Tab. 1–4). Korrelationen nach Stichprobengröße gewichtet und für Reliabilität des Kriteriums korrigiert

liegt es nahe, sich möglichst gut vorzubereiten. Hilfe versprechen einschlägige Ratgeber wie *Erfolgreich im Bewerbungsgespräch für Dummies* (Yeung 2016). Man findet aber auch Online- und Präsenztrainingsangebote. Schließlich ist es möglich, dass man in Internetforen detaillierte Informationen über das Interview der Organisation entdeckt, bei der man sich beworben hat. Es liegt nahe, dass sich eine gute Vorbereitung positiv auf die Bewertung im Interview auswirken kann. In einer experimentelle Studie von Tross und Maurer (2008) wurden Psychologiestudierende in den USA unterschiedlich umfassend auf ein simuliertes Einstellungsgespräch vorbereitet. Alle wurden in einer etwa 20-minütigen Sitzung mit Vortrag und Diskussion über die Untersuchung, Methoden der Personalauswahl und über die Gründe für den Einsatz von Auswahlgesprächen informiert. Das war die niedrigste Stufe, die einer Kontrollgruppe entspricht. In der 2. Bedingung wurden die Teilnehmerinnen und Teilnehmer für die Anforderungsmerkmale „Leistungsmotivation“ und „Gewissenhaftigkeit“ sensibilisiert, die in dem späteren Interview erfasst werden sollten. Auch Beispielfragen kamen vor. Die 3. Gruppe erhielt zusätzlich ein Interviewtraining mit Übungscharakter. Die zusätzlichen Module nahmen 40 bzw. weitere 50 min in Anspruch. Das anschließend durchgeführte Interview dauerte 20–30 min. Es wurde mit einer Videokamera aufgezeichnet und von eigens trainierten Personen unter Verwendung von Ratingskalen (von 1 bis 7) ausgewertet.

Training ist effektiv

Bei der Leistungsmotivation erreichte die Kontrollgruppe nur einen Mittelwert von 3,6, während die voll trainierte Gruppe hier einen Mittelwert von 4,7 aufwies. Dies entspricht einer Effektstärke (Glass Delta) von 0,78. Bei Gewissenhaftigkeit zeigte sich ein ähnliches Bild ($M=3,7$ vs. 4,8; Glass Delta = 0,88). Damit war ein auch praktisch bedeutsamer Effekt des 1½-stündigen Training zu verzeichnen. Da anzunehmen ist, dass sich das Training nicht auf die Merkmale selbst ausgewirkt hat, sondern nur auf die Präsentation im Interview, wird sich die Validität von eignungsdiagnostischen Interviews durch eine gute Vorbereitung der interviewten Personen vermindern. Das gilt selbstverständlich nur, wenn die Bewerberinnen und Bewerber entweder nur zum Teil ein Training erfahren oder unterschiedlich gut trainiert sind.

Informationen über das Interview geben

Ein ganz anderes Trainingskonzept besteht darin, alle Bewerberinnen und Bewerber so auf das Interview vorzubereiten, dass sie sich angemessen darstellen können. Dahinter steht die Idee, dass mangelnde Erfahrung mit Interviews oder Angst das Interviewergebnis verzerren könnten. In einer Feldstudie von Maurer et al. (2008) wurde dieses Konzept erprobt. In einer großen Stadt in den USA fand eine Kampagne zur Rekrutierung von Mitarbeiterinnen und Mitarbeitern bei der Polizei und der Feuerwehr statt. Die 146 Bewerberinnen und Bewerber, die nach einer Vorauswahl übrig geblieben waren, sollten mit einem strukturierten Eignungsinterview untersucht werden. Ihnen wurde angeboten, etwa 3 Wochen vor dem Interview an einem rund 2-stündigen Vorbereitungskurs teilzunehmen. Die Zuordnung erfolgte also nicht nach Zufall, sondern durch Selbstzuweisung. Etwa die Hälfte machte davon Gebrauch. Diese Bewerberinnen und Bewerber erhielten u. a. *Informationen* über den Ablauf, über strukturierte Interviews allgemein, aber auch über das bevorstehende. In einem Rollenspiel vor anderen Mitgliedern der Trainingsgruppe hatten sie Gelegenheit, sowohl selbst ähnliche Fragen wie im späteren Interview zu stellen als auch solche Fragen selbst zu beantworten. Die trainierte Gruppe erhielt in dem späteren Interview geringfügig bessere Bewertungen als die untrainierte Gruppe. Von besonderem Interesse ist die Validität des Interviews. Der Berufserfolg konnte nach der Einstellung durch Vorgesetztenbeurteilung ermittelt werden. Während in der untrainierten Gruppe kein Zusammenhang zwischen dem Interviewergebnis und dem Berufserfolg

feststellbar war ($r=-.01$), war die Validität in der trainierten Gruppe mit $r=.24$ signifikant höher. Möglicherweise wird also die Validität von strukturierten Eignungsinterviews durch eine angemessene Information über das Interview erhöht.

3.7.2 Interviews selbst konstruieren

Bei den meisten Fragestellungen können Diagnostikerinnen und Diagnostiker nicht auf ausgearbeitete Interviews zurückgreifen. So wurden im Bereich der Berufseignungsdiagnostik zwar zahlreiche standardisierte Interviews entwickelt, diese sind jedoch meist für Anwenderinnen und Anwender aus anderen Betrieben nicht frei verfügbar. Und wenn doch ein solches Interview weitergegeben würde, wäre es wahrscheinlich nicht brauchbar, weil die Anforderungsmerkmale nicht auf die zu besetzende Stelle passen. Will man sich nicht mit einem unstandardisierten Gespräch begnügen, das bekanntlich viele Nachteile aufweist, muss man selbst ein Interview konstruieren. Dabei gilt es, einige Empfehlungen zu beachten. Die wichtigste und zugleich grundlegendste Empfehlung ist die, einen *Leitfaden* für das Interview auszuarbeiten. Hinweise zum Aufbau finden sich etwa bei Kici und Westhoff (2000) sowie ausführlicher bei Westhoff und Kluck (2008). Der Leitfaden gibt dem Interview eine feste Struktur und standardisiert die Durchführung des Interviews. Bei Bedarf kann auch die Auswertung standardisiert werden. Mithilfe eines Leitfadens ist es also auch möglich, ein standardisiertes strukturiertes Interview (vgl. ▶ Abschn. 3.7.1) zu entwickeln.

Interviewleitfaden erstellen

Die Erstellung eines Leitfadens gliedert sich in 3 Schritte:

- Grobaufbau des Leitfadens festlegen
- Fragen finden
- Ausarbeitungen im Detail vornehmen (Feinaufbau des Leitfadens)

3.7.2.1 Grobaufbau

Im Grobaufbau eines Leitfadens werden die Themenblöcke festgelegt und in eine angemessene Reihenfolge gebracht. Unabhängig vom Thema und Anwendungsbereich empfehlen wir als übergeordnete Grobstruktur, stets eine Einleitung, eine Informationserhebungsphase und einen Abschluss vorzusehen. Für die Einleitung und den Abschluss führen wir (in Anlehnung an Westhoff und Kluck 2008) Themen auf, die häufig vorkommen. Der Einleitungs- oder Eröffnungsphase geht, wenn die interviewte Person eigens zum Interview erscheint, eine Begrüßung und eventuell ein kurzer Small Talk zur Aufwärmung voraus.

Themen für Einleitungsphase

- Vorstellung weiterer beteiligter Personen mit Namen und Funktion
- Nennen der Ziele und der Fragestellung
- Angaben zum Ablauf (Themenbereiche, Dauer, ggf. auf Möglichkeit für eigene Fragen hinweisen)
- Aufklärung der interviewten Person über ihre Rechte (z. B. Antwortverweigerung möglich)
- Information der interviewten Person, wer erfährt, was sie im Interview sagt
- Hinweis, dass Interviewerin oder Interviewer der Schweigepflicht unterliegt
- Bei Ton- oder Videoaufnahmen Einholen des Einverständnisses
- Zur Überleitung auf die Informationserhebungsphase interviewte Person um kurze Schilderung des Problems aus deren Sicht bitten

Themen für Abschlussphase

- Zusammenfassung der wichtigsten aus dem Interview gewonnenen Informationen
- Nachfrage, ob etwas Wichtiges fehlt
- Klärung eventuell wichtiger Ergänzungen oder offener Fragen
- Ansprechen eines neutralen Themas, sollte die interviewte Person emotional sehr erregt sein
- Vorstellung des weiteren Vorgehens

Multimodales Einstellungsinterview

Multimodales Einstellungsinterview Schuler (1992) hat mit dem sog. „multimodalen Einstellungsinterview“ einen wichtigen Beitrag zur Gestaltung von Einstellungsgesprächen geleistet. Das Attribut „multimodal“ weist darauf hin, dass dem Interview unterschiedliche Methoden zugrunde liegen. Im Einzelnen sieht der Aufbau aus, wie unten dargelegt. Wie ersichtlich, finden sich als „Herzstücke“ biografiebezogene und situative Fragen, erstere in den Schritten (2) und (5), letztere in (7). Das multimodale Interview stellt lediglich ein Konstruktionsprinzip dar, also eine Leitlinie zum Aufbau eines Einstellungsgesprächs. Die inhaltliche Ausgestaltung variiert in Abhängigkeit von der Stelle, die zu besetzen ist. Für die konkrete Ausgestaltung der Abschnitte (5) und (7) ist eine detaillierte Anforderungsanalyse erforderlich.

Multimodales Einstellungsinterview

1. Gesprächsbeginn:
 - Kurze informelle Unterhaltung
 - Hauptfunktion: Aufbau des Verfahrensablaufs
2. Selbstvorstellung der Bewerberin oder des Bewerbers:
 - Berichtet in freier Form über persönlichen und beruflichen Hintergrund (ggf. Schwerpunkt definieren)
3. Berufsinteressen und Berufswahl:
 - 4 standardisierte Fragen zu Berufswahl, Berufsinteressen, Organisations- bzw. Institutionsauswahl und Bewerbung
4. Freies Gespräch:
 - Funktion: Auflockerung
 - Offene Fragen in Anknüpfung an Selbstvorstellung und Bewerbungsunterlagen
5. Biografiebezogene Fragen:
 - Biografische oder „Erfahrungsfragen“ (aus Anforderungsanalysen abgeleitet)
6. Realistische Tätigkeitsinformation:
 - Vermittlung positiver Seiten oder Erwartungen sowie auch von Problemen der Institution und des Arbeitsalltags
7. Situative Fragen:
 - Knappe Schilderung von mehreren erfolgskritischen Situationen (auf Critical-Incident-Basis entwickelt); interviewte Person schildert, wie sie sich in dieser Situation verhalten würde
8. Gesprächsabschluss:
 - Interviewte Person erhält Gelegenheit, Fragen zu stellen und verbliebene Unklarheiten zu erörtern
 - Zusammenfassung, weitere Vereinbarungen

Anmerkung: Nur in den Abschnitten 2, 3, 5 und 7 erfolgt eine Bewertung anhand von Ratingskalen.

3.7.2.2 Fragen finden

Welche Fragen im Interview zu stellen sind, richtet sich in erster Linie nach der Fragestellung für die diagnostische Untersuchung und den daraus abgeleiteten psychologischen Fragen. Nicht jede psychologische Frage (z. B. „Wie belastbar ist die Klientin oder der Klient?“) wird mithilfe des Interviews abgeklärt. Manchmal sind andere Methoden vorzuziehen, weil sie valider oder ökonomischer sind. Bei wichtigen psychologischen Fragen ist aber ein multi-methodales Vorgehen angebracht: Zur Abklärung eines Sachverhalt oder zur Erfassung einer Persönlichkeitseigenschaft versucht man, aus mehreren Quellen zu schöpfen. Informationen aus den Akten und aus einer Verhaltensbeobachtung werden dann eventuell durch das Interview ergänzt oder überprüft.

Soll die Eignung für einen bestimmten Beruf oder für ein bestimmtes Studium festgestellt werden? Soll das Vorliegen einer bestimmten Störung überprüft werden? Soll die Ursache für ein Schulversagen eruiert werden? Oder soll festgestellt werden, ob ein früherer Trunkenheitsfahrer künftig nüchtern am Steuer sitzen wird? Um die richtigen Fragen zu finden, ist Wissen über den Messgegenstand erforderlich. Wenn die Eignung für einen Beruf oder ein Studium ermittelt werden soll, bedarf es eingehender Kenntnisse über die Anforderungen des Berufs bzw. des Studiums. Liegt noch keine Anforderungsanalyse vor, muss diese durchgeführt werden (s. dazu Höft et al. 2018). In den anderen Fällen finden sich auch in der einschlägigen Fachliteratur nützliche Informationen. Das kann im pädagogischen Bereich etwa die Forschung zu den Gründen für Schulversagen sein. In der Verkehrseignungsdiagnostik sind bei dem Verdacht auf Alkoholmissbrauch die Kriterien für das Vorliegen von Alkoholmissbrauch relevant (► Abschn. 9.3). Wenn im klinischen Bereich abzuklären ist, ob eine Klientin oder ein Klient eine bestimmte psychische Störung hat, stellen die diagnostischen Kriterien nach dem DSM-5 oder der ICD-10 die Grundlage für Interviewfragen dar.

Grundsätzlich sollte man bei der Ausarbeitung des Interviews auch Informationen aus anderen Quellen nutzen. Bei Einstellungsgesprächen liefern die Bewerbungsunterlagen mit Vita, Arbeitszeugnissen, Qualifikationsnachweisen etc. wichtige Erkenntnisse. Zudem wäre es peinlich, in einem Einstellungsgespräch etwa nach dem beruflichen Werdegang zu fragen, wenn dieser in den Bewerbungsunterlagen gut dokumentiert ist. Im klinischen, forensischen, verkehrspsychologischen und pädagogischen Bereich liegen oft Vorgutachten und andere Akteninformationen vor, die es zu nutzen gilt.

Für die Informationserhebungsphase ist es meist sinnvoll, Themenblöcke zu bilden. Wie wir bereits am Beispiel des multimodalen Einstellungsinterviews (► Abschn. 3.7.2.1) gesehen haben, kann eine Untergliederung auch anhand von Fragetypen erfolgen. Themenblöcke können beispielsweise die Vorgeschichte, die Klassifikation der psychischen Störung und die Auswirkung auf Beruf und Alltag sein. Mit anderen Worten: Mit dem Grobaufbau erfährt das Interview eine Struktur. Genau das ist das wesentliche Merkmal „strukturierter“ Interviews.

3.7.2.3 Feinaufbau: Formulierung von Fragen

Für den Feinaufbau des Leitfadens werden die Fragen innerhalb eines jeden Blocks mehr oder weniger präzise ausformuliert. Die Fragen können durch Stichworte fixiert (z. B. „Streit mit Mitschülern?“) oder differenziert ausformuliert und mit Nachfragen versehen werden. Ein hoch standardisiertes diagnostisches Interview ist manchmal sehr angemessen, stellt manchmal aber auch eine Fiktion dar. Eine möglichst hohe Standardisierung ist immer dann angebracht, wenn mehrere Personen mit der gleichen Fragestellung untersucht werden. Typischerweise findet man solche Umstände in der Klinischen Psychologie bei der Klassifikation von psychischen Störungen oder von Persönlichkeitsstörungen und bei Einstellungsgesprächen. Das Prinzip lautet dann: Allen Personen werden die exakt gleichen Fragen gestellt. Und das

Welche psychologischen Fragen werden überhaupt mit dem Interview abgeklärt?

Anforderungsanalysen und Fachliteratur relevant für Fragen

Vorinformationen über die zu interviewende Person nutzen

Themenblöcke festlegen

Ausformulierte Fragen tragen zur Standardisierung bei

gelingt nur mit ausformulierten Fragen. Die Konstanz der Interviewdurchführung betrifft die Interviewerinnen und Interviewer sowie die interviewten Personen. Gerade bei sich wiederholenden Fragestellungen muss gewährleistet sein, dass eine andere Interviewerin oder ein anderer Interviewer das Interview auf die gleiche Weise durchführen kann. Dafür gibt es so triviale Anlässe wie Urlaub, Krankheit, andere wichtige Termine oder einfach auch Arbeitsentlastung. Es wäre ethisch unverantwortlich, wenn eine Diagnose oder ein Eignungsurteil davon abhängt, wer gerade Zeit hat, das Interview zu führen. Genauso muss gewährleistet sein, dass auch ein und dieselbe Interviewerin oder ein und derselbe Interviewer alle interviewten Personen auf die gleiche Weise befragt. Halbstandardisierte Fragen würden dazu führen, dass die ausformulierten Fragen (leicht) unterschiedlich ausfallen und damit auch (leicht) unterschiedliche Antworten provozieren.

Ein weiterer Vorteil von ausformulierten Fragen ist zu erwähnen. Ein Interview zu führen und dabei noch Notizen anzufertigen und die interviewte Person zu beobachten, kann gerade für ungeübte Interviewerinnen und Interviewer sehr anstrengend sein. Ein Leitfaden mit ausformulierten Fragen kann eine große Entlastung im Gespräch darstellen. Man braucht sich nicht um die „richtigen“ Worte zu bemühen – ein Blick in den Leitfaden genügt. Auch kann man bereits beantwortete Fragen abhaken und behält so einen Überblick über das, was noch erfragt werden muss. Manchmal geben die interviewten Personen nebenbei schon eine Antwort auf eine Frage, die erst später vorgesehen ist. In diesem Fall wäre es unangemessen, diese Frage noch einmal zu stellen. Das wäre nicht nur unökonomisch, sondern könnte bei der interviewten Person auch den Eindruck vermitteln, man habe ihr nicht richtig zugehört.

Die Ausarbeitung eines standardisierten Interviews kostet sehr viel Zeit. Die Kosten amortisieren sich jedoch nach vielen Anwendungen. Bei einmaligen Untersuchungsanlässen liegt es nahe, den Aufwand zu verringern. Die Festlegung einer Struktur für das Interview ist unverzichtbar. Aber die einzelnen Fragen können auch stichwortartig fixiert werden. Ein solches Interview könnte man als strukturiertes halbstandardisiertes Interview bezeichnen. Wenn die Interviewerin oder der Interviewer auch den Leitfaden selbst ausgearbeitet hat, weiß sie oder er, was mit den Stichworten gemeint ist. So ist weitgehend sichergestellt, dass alle Themen auch in angemessener Form angesprochen werden. Dass das Interview bei einer Wiederholung etwas anders ablaufen würde, ist unerheblich – es wird ja nie wiederholt. Wie eine Frage dann ausformuliert wird und welche Nachfragen gestellt werden, ergibt sich im Gespräch.

► Beispiel

In einem Interview soll festgestellt werden, wie die interviewte Person mit einer beruflich bedingten Trennung von der Familie umgehen wird.

In einem Einstellungsgespräch wird die Frage ausformuliert: „Die neue Tätigkeit als Monteur bringt es mit sich, dass Sie oft 2 oder 3 Tage in der Woche so weit entfernt von zu Hause arbeiten, dass Sie in einem Hotel übernachten müssen. Welche Auswirkungen hätte das auf Ihr Familienleben und auf Ihre Freizeitaktivitäten? Wie würden sie damit umgehen?“

Das gleiche Thema kommt auch in einem klinischen Interview mit einem Klienten vor, der wegen Beziehungsproblemen mit seiner Partnerin Hilfe sucht. In dem Interview soll u. a. erkundet werden, welche situativen Faktoren zu den Beziehungsproblemen beitragen. Im Leitfaden steht: „Auswirkung des Berufs auf das Privatleben“. Im Interview wird die Frage dann so gestellt: „Wie wirkt sich Ihr Beruf auf das Privatleben aus?“ [Klient berichtet, dass er aus finanziellen Gründen ein Stellenangebot als Monteur annehmen will und dann häufiger berufsbedingt 2 oder 3 Tage auswärts übernachten muss]. „Was bedeutet die Abwesenheit von zu Hause für Sie?“ [Klient sagt, dass er es bedauere, weniger Zeit für seine Kinder zu haben]. „Und was ist mit Ihrer Frau?“ ◀

Ausformulierte Fragen entlasten

Stichworte statt ausformulierte Fragen

Was bei der Formulierung von Fragen zu beachten ist, wird von Westhoff und Kluck (2008) ausführlich behandelt. Der Autor und die Autorin geben zahlreiche praxisnahe Empfehlungen für das Formulieren von „günstigen“ Fragen und weisen auf „ungünstige“ Fragen hin, die man möglichst vermeiden sollte.

Hinweise zur Formulierung von Fragen im Interview nach Westhoff und Kluck (2008)

- Offene Frage als Einstieg in einen Abschnitt
- Formulierung möglichst kurzer und verständlicher Sätze
- Erfragen von konkretem Verhalten
- Einbeziehen des Kontextes als Gedächtnissstütze
- Vermeidung von Fachbegriffen und Fremdwörtern
- Keine Suggestivfragen
- Keine Fragen, die das erfragte Verhalten bewerten

Keine Regel ohne Ausnahme: In begründeten Fällen mag ein Abweichen von solchen allgemein formulierten Empfehlungen durchaus sinnvoll sein. Entscheidend ist, dass die Absicht erkannt und umgesetzt wird, die hinter solchen Empfehlungen steht. Im Regelfall sind Fachbegriffe und Fremdwörter zu vermeiden, damit die befragte Person den Sinn der Fragen richtig versteht. Handelt es sich um eine Person mit abgeschlossenem Hochschulstudium, sind Fremdwörter und auch manche Fachbegriffe nicht nur erlaubt, sondern sogar angemessen. Andernfalls entsteht vielleicht der Eindruck, ihr Bildungsniveau würde unterschätzt.

Abweichen von der „Regel“ erlaubt

Suggestivfragen sind zu vermeiden, um die interviewte Person nicht bei ihrer Antwort zu beeinflussen. Unter Umständen wird sich eine erfahrene Interviewerin oder ein erfahrener Interviewer dennoch auch einmal absichtlich für eine Suggestivfrage entscheiden, um die eigentlich relevanten Fragen zum Thema anbringen zu können. Man stelle sich einen Schüler vor, der nach Angaben der Lehrerin mehrmals dem Unterricht ferngeblieben ist und nun wegen „Schulschwierigkeiten“ diagnostisch untersucht wird. Um Details über das Fernbleiben vom Unterricht zu erfahren, könnte man im Anschluss an die Frage, ob er gerne in die Schule gehe (die dieser vorsichtig verneint), direkt sagen: „Wenn man nicht so gerne in die Schule geht, kommt man doch schnell auf den Gedanken, einfach ab und zu blau zu machen. Wie ist das denn bei dir?“ Diese Frage „bricht das Eis“ und führt direkt zu dem sensiblen Thema. Ob der Schüler tatsächlich absichtlich vom Unterricht ferngeblieben ist, wird durch Nachfragen geklärt: Welche Unterrichtsstunden sind betroffen, was hat er in dieser Zeit getan, was hat er als Begründung für das Fehlen gegenüber den Lehrkräften geäußert?

3.7.2.4 Auswertung und Dokumentation

Bewertung der Antworten in standardisierten Interviews Die Antworten auf die Interviewfragen müssen in irgendeiner Form dokumentiert und ausgewertet werden. Bei (halb-)standardisierten Interviews werden die Antworten auf eine Frage in der Regel unmittelbar durch Ankreuzen von Antwortkategorien oder durch ein Rating ausgewertet. Bei klinischen Interviews zur Diagnosestellung gilt es, festzustellen, ob ein diagnostisches Kriterium vorliegt oder nicht. Dazu genügt es, das diagnostische Kriterium, auf das eine Frage abzielt, ausformuliert oder auch stichwortartig im Leitfaden aufzuführen. Die Antwortkategorien können lauten „ja“ und „nein“. Oft ist es aber nicht so einfach zu beurteilen, ob ein Kriterium klar vorliegt. Deshalb ist es sinnvoll, eine Kategorie „?“ oder 2 zusätzliche Kategorien „eher ja“ und „eher nein“ vorzusehen.

Bewertung der Antworten im klinischen Interview

Bewertung der Antworten im eignungsdiagnostischen Interview

Bei eignungsdiagnostischen Interviews möchte man wissen, ob aus der Antwort folgt, dass ein Anforderungsmerkmal vorliegt oder nicht. Auch hier sollte das Merkmal, also die Anforderung, im Leitfaden stichwortartig oder ausformuliert genannt werden. Erkennt man an, dass jemand eine Anforderung mehr oder weniger gut erfüllen kann, ist die Bewertung auf einer Dimension angebracht. Wie viel Skalenstufen adäquat sind, richtet sich primär nach dem Differenzierungsvermögen der Interviewerinnen und Interviewer oder auch der zusätzlichen Beobachterinnen und Beobachter. Stellt man bei einer Zielgruppe von Bewerberinnen und Bewerbern fest, dass die Antworten wenig variieren, muss die Skala auf weniger Stufen reduziert werden. Zusätzlich sollte auch eine Verankerung an Verhaltensweisen, die für eine hohe, mittlere oder niedrige Ausprägung sprechen, vorgenommen werden.

Skalen für die Ausprägung eines Merkmals						
1	2	3	4	5		
sehr niedrig	niedrig	durchschnittlich		hoch		sehr hoch

Oder:

1	2	3	4	5	6	7
weit unter-durch-schnittlich	unter-durch-schnittlich	leicht unter-durch-schnittlich	durch-schnittlich	leicht über-durch-schnittlich	über-durch-schnittlich	weit über-durch-schnittlich

Oder kriteriumsorientiert:

0	1	2	3	4
nicht vorhanden	ansatzweise vorhanden	teilweise vorhanden	weitgehend vorhanden	vollständig vorhanden

Standardisierung der Auswertung

Auswertung eines standardisierten Interviews Zu einem standardisierten Interview gehört auch die Standardisierung der Auswertung. Konkret bedeutet dies, dass die Verrechnung der Antworten vorab festgelegt wird. In Abb. 3.39 wird das Vorgehen für ein eignungsdiagnostisches Interview demonstriert. In dem Interview wurden die Anforderungsmerkmale „Kommunikationsfähigkeit“, „Durchsetzungsfähigkeit“ und „analytisches Denken“ durch geeignete Fragen erfasst und die Antworten auf einer Ratingskala von 1 („sehr niedrig“) bis 5 („sehr hoch“) bewertet. Schon vor der Durchführung des ersten Interviews wurde festgelegt, dass die Merkmale nach ihrer Relevanz für die zu besetzende Stelle gewichtet werden und dass Mindestanforderungen bestehen. Der akzeptable Bereich ist in Abb. 3.39 durch blaue Rahmen gekennzeichnet. Kommunikationsfähigkeit etwa wird mit dem Faktor 3 gewichtet, und die Mindestanforderung liegt bei der Ratingskalastufe 4. Bei keinem der 3 Merkmale gibt es eine Begrenzung im oberen Bereich. Grundsätzlich wäre es beispielsweise möglich gewesen, bei der Durchsetzungsfähigkeit den akzeptablen Bereich auf die Skalenstufen 3 und 4 festzulegen.

Anforderungsmerkmal	Gewicht 1–5	Ausprägung				
		1	2	3	4	5
Kommunikationsfähigkeit (K)	3	1	2	3	X	5
Durchsetzungsfähigkeit (D)	2	1	2	3	4	X
Analytisches Denken (A)	2	1	2	3	X	5

Abb. 3.39 Standardisierte Auswertung eines eignungsdiagnostischen Interviews. Die Kreuze zeigen, wie eine konkrete Person im Interview eingestuft worden ist. Regel: Eignung = wenn $K > 3$, $D > 2$, $A > 3$, dann $(3K + 2D + 2A) / 7$. Hier: alle Mindestwerte erreicht; dann $(3 \times 4) + (2 \times 5) + (2 \times 4) = 30/7 = 4,3$

Die Auswertung des Interviews einer Bewerberin oder eines Bewerbers erfolgt in 2 Schritten. In Schritt 1 wird geprüft, ob die Mindestanforderungen erfüllt sind. Im Beispiel ist das zu bejahen; alle Kreuzchen liegen im akzeptablen Bereich (innerhalb der blauen Rahmen). Nur dann erfolgt Schritt 2: Die Einstufungen werden mit den jeweiligen Gewichten multipliziert und anschließend durch die Summe der Gewichte (7) dividiert. So erhält man das gewichtete arithmetische Mittel aus den 3 Skalen. Die Umrechnung hat den Vorteil, dass das Gesamtnoteil (4,3) auf der Skala von 1 bis 5 abgebildet werden kann. Mithilfe einer Excel-Tabelle kann die Berechnung für alle Bewerberinnen und Bewerber auch automatisch durchgeführt werden.

Gewichtung von Antworten

Qualitative Auswertung von Interviews Dient ein Interview dazu, das familiäre Umfeld, eine Beziehung, die Arbeitsbedingungen, ein Drogenproblem, die Vorgeschiede einer Erkrankung etc. zu erkunden, sind quantitative Einstufungen der Antworten mit anschließender Verrechnung in der Regel nicht hilfreich. Bei solchen Interviews werden in der Regel die Antworten protokolliert. Im Interviewleitfaden lässt man dazu bei jeder Frage genügend Platz. Aus Zeitgründen kann eine Antwort während des Interviews stichwortartig notiert werden. Eventuell ergänzt man die Notizen zeitnah im Anschluss an das Interview oder überführt die Stichworte in Sätze. Diese Nachbearbeitung wird wesentlich erleichtert, wenn der Leitfaden als Textdatei auf einem Laptop zu lesen ist und die Stichworte dort in Antwortfelder eingetragen werden.

Antworten protokollieren

Zur Erinnerung sei erwähnt, dass sich die konkreten Fragen im Interview aus psychologischen Fragen ableiten. Bei der Auswertung sollen die Antworten im Interview dazu dienen, die einzelnen psychologischen Fragen zu beantworten. Dazu müssen die Antworten zunächst den psychologischen Fragen zugeordnet werden. Oft sind die Antworten im Leitfaden schon entsprechend geordnet. Es kommt aber auch vor, dass die Antwort einer Klientin oder eines Klienten zusätzlich für eine andere psychologische Frage relevant ist. In diesem Fall wird die Antwort doppelt verwendet oder „aufgeteilt“.

Zuordnung der Antworten zu psychologischen Fragen

Pro und Kontra abwägen

Liegt das Interview mit Fragen und Antworten als Ausdruck vor, empfiehlt es sich, den Text für jede psychologische Frage durchzugehen; Antworten, die für eine psychologische Frage relevant sind, werden am besten farbig markiert. Einfacher geht es, die Kopie einer Textdatei zu verwenden und die Antworten den einzelnen psychologischen Fragen zuzuordnen. Die eigentliche Auswertung besteht darin, unter Abwägung von Pro- und Kontraargumenten jede psychologische Frage zu beantworten. Am besten listet man auf, welche Antworten als pro und welche als kontra zu bewerten sind. Wenn eine psychologische Frage nicht eindeutig zu beantworten ist, wird dies vermerkt, und die Formulierung fällt entsprechend vorsichtig aus. Wenn eine psychologische Frage nicht beantwortet werden kann, wird dies ebenfalls vermerkt, eventuell mit Angabe der Gründe. Das Ergebnis liegt in Form eines Berichts vor, der nach den psychologischen Fragen gegliedert ist. Weitere Informationen zur quantitativen Auswertung von Interviews finden sich bei Krumm et al. (2015).

Von Notizen bis zur elektronischen Aufzeichnung

Dokumentation Ein wichtiges Prinzip ist, dass die Ergebnisse nachvollziehbar dokumentiert werden. Die Anforderungen an die Nachvollziehbarkeit können in der Praxis sehr unterschiedlich sein. Die Spanne beginnt mit der Aufbewahrung des Leitfadens mit stichwortartigen Notizen zu den Antworten. Falls die Fragen und Antworten zwecks Auswertung des Interviews wörtlich transkribiert wurden, wird man den Text einschließlich der eigenen Notizen und Markierungen zu den eigenen Akten nehmen. Wird das Interview im Rahmen eines Gutachtens verwendet, kann das transkribierte Interview auch als Anlage zum Gutachten beigefügt werden. Manchmal stellt ein Interview in einem Gerichtsverfahren ein wichtiges Beweismittel dar. So kommt es vor, dass Klientinnen oder Klienten gerichtlich gegen ein negatives verkehrspsychologisches Gutachten vorgehen (► Abschn. 9.3). Bei der sehr häufig vorkommenden Fragestellung, ob Alkoholmissbrauch vorliegt, hat das Interview eine herausragende Bedeutung. Dann sagt eine Mandantin vielleicht gegenüber dem eigenen Anwalt: „Was hier im Gutachten steht, stimmt nicht. Das habe ich so nie gesagt.“ Für Anwältinnen und Anwälte ist es sehr hilfreich, nun ein exakt dokumentiertes Interview zur Verfügung zu haben. So kann eventuell schon im Vorfeld die Aussichtslosigkeit einer Klage erkannt werden. Kommt es zum Gerichtsverfahren, stellt das Interview ein zentrales Beweismittel dar. Deshalb ist es aus anwaltlicher und richterlicher Sicht sehr wünschenswert, bei Bedarf (!) eine Ton- oder Videoaufzeichnung anfordern zu können. Mit einer Videoaufzeichnung ist die höchste Stufe der Nachvollziehbarkeit durch Dokumentation erreicht. Allerdings kann die Auswertung in einem Gerichtsprozess sehr aufwendig sein, wenn es um mehr als nur die Überprüfung einer einzelnen Aussage geht.

Ton- und Videoaufzeichnungen liefern deutlich exaktere Ergebnisse als nahezu wörtliche schriftliche Aufzeichnungen durch eine beim Interview anwesende weitere Person, wie eine vergleichende Studie zu verschiedenen Dokumentationstechniken für Interviews mit Kindern bei Verdacht auf Kindesmissbrauch ergab (Berliner und Lieb 2001). Diese Studie zeigt auch, wie sich eine elektronische Aufzeichnung auf das Interview selbst auswirkt. Aufzeichnungen hatten nahezu keinen Einfluss auf das Verhalten der Kinder. Die Interviewerinnen und Interviewer stellten fast 5× so viele Fragen wie bei der nahezu wörtlichen Protokollierung. Dennoch kamen bei ihnen signifikant weniger suggestive Fragen bezüglich eines möglichen Missbrauchs vor. Die

Verhaltensbeeinflussung durch elektronische Aufzeichnung

Tatsache, dass ihr Verhalten aufgezeichnet wurde, hatte möglicherweise eine „erzieherische“ Funktion, indem sie sich um Fehlervermeidung bemühten.

Werden die Antworten in einem Interview von mehreren Personen beurteilt, was bei standardisierten Interviews zur Personalauswahl üblich ist, so sollte auch die Beurteilendenübereinstimmung (Interrater-Reliabilität) ermittelt und dokumentiert werden.

Weiterführende Literatur

Für eine differenzierte Gegenüberstellung der Vor- und Nachteile der Dokumentationsmethoden handschriftliche Aufzeichnungen, Mitschrift am Computer, Tonaufnahme und Videoaufnahme sei auf Okulicz-Kozaryn et al. (2015) verwiesen.

3.7.3 Techniken der Gesprächsführung

Interviewerinnen und Interviewer können verschiedene Verhaltensweisen oder Techniken einsetzen, um die für die Fragestellung relevanten Informationen zu gewinnen. Es reicht nicht aus, in einem Leitfaden die „richtigen“ Fragen aufzuschreiben. Ein und dieselbe Information kann man mit sehr unterschiedlich formulierten Fragen gewinnen. Manchmal ist die eine Art von Fragen hilfreich, mal eine andere. Und auch zwischen den Fragen geschieht in einem Interview viel: Interviewerinnen und Interviewer beeinflussen mit ihrem Verhalten die Bereitschaft ihrer Gesprächspartnerinnen und Gesprächspartner, sich intensiver mit einem Thema auseinanderzusetzen, offen darüber zu sprechen oder endlich „zur Sache zu kommen“. Für den Erfolg eines Interviews ist der flexible Einsatz von Techniken der Gesprächsführung entscheidend. „Flexibel“ bedeutet in diesem Kontext, die richtige Technik zur richtigen Zeit einzusetzen. Im Folgenden werden einige Techniken vorgestellt, die generell hilfreich für die Gesprächsführung sind. Im Anschluss daran werden Probleme benannt, die in einem Interview auftreten können, und Lösungsvorschläge unterbreitet.

Gesprächstechnik wichtig

3.7.3.1 Offene und geschlossene Fragen

Mit offenen Fragen lädt man seine Gesprächspartnerinnen und Gesprächspartner ein, von sich aus etwas zu erzählen. Meist wird man dabei auch das Thema vorgeben. Im Alltag begegnet man offenen Fragen häufig, wenn man sich an jemanden mit der Bitte um Hilfe oder Rat wendet. Eine Ärztin fragt vielleicht: „Was führt Sie zu mir?“ (und nicht: „Sind Sie erkältet?“ – was eine geschlossene Frage wäre). In einer Bankfiliale eröffnet die Bankangestellte das Gespräch mit den Worten: „Was kann ich für Sie tun?“ (und nicht mit „Wollen Sie einen Kredit?“). Gerade in der Eröffnungsphase eines Interviews sind offene Fragen zumeist sehr angemessen. Sie signalisieren Interesse am Gegenüber und die Bereitschaft, zuzuhören. Ein erfahrener forensischer Gutachter beschreibt seine Erfahrungen so:

Bereitschaft zuzuhören signalisieren

- » Ich wundere mich oft darüber, wie schnell Menschen sich öffnen und auch über intime Sachverhalte sprechen, wenn man ein geeignetes Gesprächsklima schafft und sie nicht zu früh mit Detailfragen bedrängt.[...] Ich habe mir über die Herstellung dieser besonderen Gesprächssituation bei Gutachten häufig Gedanken gemacht. Mein Gegenüber soll spüren, dass ich es nicht vorverurteile. Dass ich jemand bin, der sich neutral und sachlich anhört, was die Person zu sagen hat. Dazu gehört auch, dass ich das Gespräch zu Beginn in keine Richtung lenke. Ich lasse die Leute erst einmal reden (Steller 2015, S. 72 f.).

Die genaue Wortwahl ist wichtig

► Beispiel

Steller (2015) beschreibt eine solche Gesprächssituation: Eine junge Frau, er nennt sie Susanne Fuchs, hatte im Alter von 14 Jahren ihr Elternhaus wegen Schul- und Erziehungsproblemen verlassen. Sie kam mehrfach mit dem Gesetz in Konflikt, und bei einer Vernehmung wegen eines Schmuckdiebstahls im Wert von 20 € bezichtigte sie ihren Vater, sie sexuell missbraucht zu haben. Die Staatsanwaltschaft eröffnete daraufhin ein Strafverfahren gegen den Vater. Prof. Steller wurde mit der Erstellung eines Glaubhaftigkeitsgutachtens beauftragt. Die inzwischen 20-Jährige erschien in einem Punkeroutfit. „Provozierend lief sie in meinem Arbeitszimmer herum und ließ dabei mehrmals rosafarbene Kaugummiblasen platzen.“ Der Gutachter begann das eigentliche Gespräch mit der Aufforderung „Erzählen Sie mal, worum es geht?“ und kommentiert diese Wortwahl so: „Es ist wichtig, eine unverfälschte Formulierung zu wählen. Eben nicht zu fragen: Erzählen Sie mal von der Sache mit Ihrem Vater.“ Auf die Antwort „Ja, aber das wissen Sie doch schon alles“ reagierte der Gutachter mit einer Begründung, die von der Zeugin offensichtlich akzeptiert wurde und sie ermutigte, bereitwillig ihre Erinnerungen zu schildern: „Leider ist es in unserem Strafprozessrecht nötig, dass eine Zeugin auch in einer späteren Gerichtsverhandlung noch einmal den gesamten Verlauf schildert. Die Besprechung hier dient der Vorbereitung so einer Gerichtsverhandlung. Das Gericht hat mich nach meiner Meinung gefragt, und wenn ich eine Meinung äußern soll, ist es notwendig, dass Sie mir den Sachverhalt noch einmal schildern.“ ◀

Raum für eigene Ausführungen geben

Offene Fragen Auch im Laufe eines Interviews lassen offene Fragen der interviewten Person viel Raum für ihre Ausführungen. Für die Interviewerin oder den Interviewer kann es diagnostisch aufschlussreich sein, *wie* die interviewte Person ein Ereignis, ihr Problem oder etwa ihre Stärken und Schwächen beschreibt.

Formulierung offener Fragen

Offene Fragen beginnen oft so:

- Was bewegt Sie ...?
- Erzählen Sie doch mal ...
- Schildern Sie mir bitte einmal, ...
- Wie muss ich mir das vorstellen?

An den Beispielen wird deutlich, dass hierbei nicht unbedingt eine Frage formuliert werden muss. Es kann sich auch um eine Aufforderung handeln, etwas zu einem Thema zu erzählen.

Offene Frage lassen sich mühelos auf ein bestimmtes Thema lenken: „Erzählen Sie mir doch einmal etwas über Ihre Eltern“, „Was sagen die Lehrer Ihres Kindes dazu, dass seine Leistungen so stark abgefallen sind?“, „Wo würden Sie gerne in fünf Jahren stehen?“ – dies sind Beispiele für die Fokussierung auf ein Thema mithilfe einer offenen Frage.

Effizient zur Erhebung konkreter Informationen

Geschlossene Fragen Das Gegenstück zu offenen Fragen sind geschlossene. Bei geschlossenen Fragen wird eine bestimmte, konkrete Antwort erwartet. Sie sind angemessen, wenn es um die Erhebung von konkreten Informationen geht. Sie werden oft im Nachgang zu offenen Fragen gestellt, um Details zu erfragen oder um Zusammenhänge besser zu verstehen. Wenn eine Klientin oder ein Klient ihre bzw. seine Partnerschaftsprobleme geschildert hat und

man nun wissen möchte, wie lange beide schon getrennt leben, wäre die offene Frage „Wie ist ihre Beziehung auseinandergegangen?“ nicht zielführend. Die Frage „Wie lange wohnen Sie nicht mehr zusammen?“ liefert bei minimalem Zeitaufwand die gewünschte Information.

Folgen zu viele geschlossene Fragen aufeinander, kann das von der befragten Person als Ausfragen erlebt werden. Sie verliert die Kontrolle über das, was sie sagen möchte. Deshalb ist ein Wechsel von offenen und geschlossenen Fragen oftmals eine kluge Wahl.

Reines Abfragen vermeiden

3.7.3.2 Aktives Zuhören

Beim aktiven Zuhören handelt es sich um ein ganzes Bündel von Einzelmaßnahmen, die alle dazu dienen, die interviewte Person zu ermuntern, weiter die gewünschten Informationen zu geben (s. Wittmann und Holling 2001). Durch verschiedene verbale und nonverbale Signale und die Art, wie man auf Äußerungen eingeht, kann man zum Ausdruck bringen, dass man aufmerksam und entspannt zuhört. Im Einzelnen lassen sich diese Techniken in folgende Kategorien einteilen:

- Nonverbale Verstärker
- Verbale Verstärker
- Reflexionen
- Nachfragen
- Zusammenfassen

Interesse signalisieren

Verschiedene nonverbale und verbale Signale werden auch als Verstärker bezeichnet, weil sie von Interviewenden als erwünscht angesehene Verhaltensweisen verstärken sollen. Erwünschte Verhaltensweisen können etwa sein: das Thema vertiefen, auf das Wesentliche kommen oder ein Thema nun abzuschließen. Mit nonverbalen und verbalen Verstärkern lässt sich auch steuern, wie die interviewte Person an ein Thema herangeht. Beispielsweise können selbstkritische Äußerungen, eine Betrachtung aus der Perspektive einer anderen Person oder eine emotionale Distanzierung verstärkt werden. Mit dem Einsatz verbaler und nonverbaler Verstärker kann man immer nur Verhaltensweisen steuern, die bereits aufgetreten sind; die Botschaft lautet „weiter so, mehr davon“. Zugleich lässt sich durch ein Aussetzen dieser Verstärker unerwünschtes Verhalten wie etwa ein Abschweifen vom Thema oder die Verneidlichung eines Problems abstellen – zumindest kann man das versuchen.

Erwünschte Verhaltensweisen verstärken

Auswahl nonverbaler und verbaler Verstärker

- Nonverbale Verstärker:
 - Nicken
 - Blickkontakt halten oder aufnehmen
 - Augenbrauen anheben (Interesse signalisieren)
- Verbale Verstärker:
 - „Ah“
 - „Ja“
 - „Bitte beschreiben Sie Ihre ... (z. B. Partnerbeziehung) so, dass ich mir ein gutes Bild davon machen kann.“

Reflexionen Dieses Konzept spielt in der motivierenden Gesprächsführung (Miller und Rollnick 2015) eine bedeutende Rolle und wird hier von uns auf rein diagnostische Interviews übertragen, die nicht den Anspruch haben, die Selbsterkenntnis der interviewten Person zu fördern. Im Wesentlichen geht es

Paraphrasieren oder vorsichtig interpretieren

darum, eine Äußerung aufzugreifen und mit anderen Worten zu wiederholen. Reflektiertes Zuhören im Sinne der motivierenden Gesprächsführung kann auf unterschiedlichen Ebenen stattfinden:

- *Einfache Reflexion*: Eine Äußerung der interviewten Person wird mit leicht veränderter Wortwahl wiederholt. Dies wird in der Literatur auch als „Paraphrasieren“ bezeichnet. Sie dient hauptsächlich dazu, der interviewten Person zu signalisieren, dass man ihre Aussage verstanden hat.
- *Komplexe Reflexion*: Die Äußerung wird in eine bestimmte Richtung hin interpretiert. „Eine komplexe Reflexion fügt dem, was der Klient gesagt hat, andere Bedeutungsaspekte hinzu oder setzt andere Schwerpunkte, indem sie eine Vermutung anstellt über das, was unausgesprochen mitschwingt oder als Nächstes kommt (also den Gedanken fortführt)“ (Miller und Rollnick, 2015, S. 79).

Gezielte Unter- und Übertreibung

Die Deutung der Interviewerin oder des Interviewers soll nicht länger sein als die Äußerung, auf die sie sich bezieht. Selbstverständlich darf die interviewte Person damit nicht in suggestiver Weise zu einer neuen Sicht ihrer Situation, ihrer Gefühle etc. „gebracht“ werden. Die Interviewerin oder der Interviewer will damit herausfinden, was die interviewte Person wirklich meint. Sie oder er achtet auf deren Reaktion: Zustimmung, empörte Zurückweisung, Korrektur? Als zusätzliches Mittel kann dabei die Reflexion eher abgeschwächt formuliert sein („Unterbieten“) oder übertrieben werden (von Miller und Rollnick „Überbieten“ genannt). In beiden Fällen ist es aufschlussreich, ob die interviewte Person die Interpretation akzeptiert oder sie korrigiert.

► Beispiel

Die interviewte Person spricht über das Arbeitsklima, das sie offensichtlich belastet. Ihr Vorgesetzter kritisiert sie oft und verlangt dennoch viel von ihr. „Wenn ich montags wieder an die Arbeit gehe, spüre ich einen inneren Widerstand. Ich fühle mich einfach nicht gut, habe so ein flausches Gefühl im Magen. Manchmal würde ich am liebsten zu Hause bleiben.“

Einfache Reflexion: „Sie gehen nach dem Wochenende nur sehr ungern wieder zur Arbeit“. Diese Form der Gesprächsführung ist weitgehend identisch mit dem Paraphrasieren. Die Interviewerin oder der Interviewer wiederholt eine Aussage mit eigenen Worten.

Komplexe Reflexion: „Der Gedanke, gleich wieder an der Arbeit zu sein, macht Ihnen richtig Angst?“ (Vermutung; Überbieten). Interviewte Person: „Angst, hm. Vielleicht schon. Ich denke manchmal, dass mein Job in Gefahr ist.“ Oder: „Das Arbeitsklima ist Ihrer Meinung nach nicht optimal“ (Vermutung; Unterbieten). Interviewte Person: „Nicht optimal – das ist ja wohl eine leichte Untertreibung. Es belastet mich, macht mir Angst.“ ◀

Nachfragen Bleiben die Ausführungen der interviewten Person zu vage, hilft es manchmal, nachzufragen. Nachfragen können eher unspezifisch sein (z. B. „Erzählen Sie mir doch mehr dazu“) oder auch fokussiert auf einen Teilaspekt der Ausführungen (z. B. „Wie haben Sie auf dieses Angebot reagiert?“). Nachfragen dienen der Informationsgewinnung und haben den positiven Nebeneffekt, Interesse an einem Thema und damit auch an der interviewten Person zu signalisieren.

Zusammenfassen Eine Interviewerin oder ein Interviewer kann die Ausführungen der interviewten Person nur treffend zusammenfassen, wenn sie oder er zuvor gut zugehört hat. Eine kurze Zusammenfassung der Ausführungen der interviewten Person signalisiert also: „Ich habe Ihnen aufmerksam zugehört und will mich noch einmal vergewissern, dass ich Sie richtig verstanden habe.“ Zusammenfassen ist dem Paraphrasieren (s. o.) ähnlich. Im Unterschied zum Paraphrasieren bezieht sich eine Zusammenfassung nicht auf eine einzelne Aussage, sondern auf längere Ausführungen. Der Unterschied ist aber nur graduell.

Nachfragen und Zusammenfassen signalisieren aufmerksames Zuhören

3.7.3.3 Umgang mit Problemen während eines Interviews

Während eines Interviews können verschiedene Probleme auftreten, die zunächst aufgelistet und dann näher mit Lösungsvorschlägen erläutert werden. Das Problem des optimalen Umgangs mit Kindern im Interview wird hier ausgeklammert; Empfehlungen dazu finden sich bei Querido et al. (2001).

Mögliche Probleme beim Interview

- Die interviewte Person erweist sich insgesamt als wenig gesprächig, und das Interview droht unergiebig zu werden. Wie kann man sie zum Reden motivieren?
- Sie ist zwar meist auskunftsbereit, scheint aber bei einem Thema abzublocken; sie gibt vor, sich nicht mehr richtig erinnern zu können, oder liefert vage Antworten. Was kann man unternehmen, um „das Eis zu brechen“?
- Sie schweift immer wieder vom Thema ab. Wie kann die Interviewerin oder der Interviewer sie sanft und doch wirkungsvoll dazu bringen, die Fragen ohne Umschweife zu beantworten?
- Die Interviewerin oder der Interviewer muss ein sehr heikles Thema angehen, das auch ihr oder ihm peinlich ist. Wie spricht man ein heikles Thema am besten an?

Zum Reden motivieren Die 1. Maßnahme zur Motivierung der interviewten Person findet bereits in der Eröffnungsphase des Interviews statt. Die interviewte Person wird über die Ziele und die Fragestellung sowie den Ablauf des Gesprächs informiert. Sie weiß nun, zu welchem Zweck welche Fragen gestellt werden. Sofern es auch in ihrem Interesse liegt, dass die genannten Ziele erreicht werden, wird sie gerne ihren Beitrag dazu leisten.

Vorab über Ziele und Fragestellung informieren

Die 2. Maßnahme zur Gewinnung der gewünschten Informationen sind die Fragen selbst. Im alltäglichen Umgang mit anderen Menschen gibt man auf eine Frage normalerweise eine passende Antwort. Dieses Verhaltensmuster kommt auch im Interview zum Tragen. Die Fragen müssen allerdings als angemessen für den Zweck erlebt werden und sollten nach bewährten Prinzipien formuliert sein (► Abschn. 3.7.2), um zu den „richtigen“ Antworten zu führen. Durch Überleitungen von einem Thema zum anderen und angemessene offene Fragen zu Beginn eines Themenblocks wird vermieden, dass sich die interviewte Person ausgefragt fühlt. Die Überleitung gibt ihr die nötige Orientierung über das Thema, und offene Fragen erlauben es ihr, ihre eigene Sicht darzulegen.

Angemessene Fragen

Aktives Zuhören

Während die beiden eben erläuterten Maßnahmen bereits bei der Abfassung des Leitfadens eingeplant werden und somit vorbereitet sind, muss eine 3. Maßnahme während des Gesprächs spontan und flexibel umgesetzt werden. Es handelt sich um das in ▶ Abschn. 3.7.3.2 bereits erläuterte aktive Zuhören.

Antwortalternativen vorgeben

Falls es der interviewten Person schwerfällt, die richtige Antwort zu finden, kann die Vorgabe von Antwortalternativen hilfreich sein. Wenn etwa auf die Frage, wann die Beschwerden erstmals aufgetreten sind, keine klare Antwort folgt, bieten sich die Vorgaben „in den letzten vier Wochen, eher vor ein paar Monaten oder vor mehr als einem Jahr“ an. Hat sich die interviewte Person für eine der Antwortalternativen entschieden, lässt sich eventuell durch Nachfragen der Zeitpunkt noch präzisieren.

Gefühle benennen

Im klinischen Kontext sind oft die Gefühle der Klientin oder des Klienten diagnostisch relevant. Um die interviewte Person anzuregen, mehr über ihre Gefühle zu berichten, kann die Interviewerin oder der Interviewer eigene Vermutungen über das emotionale Befinden der interviewten Person äußern. Anknüpfungspunkte ergeben sich durch den nonverbalen Ausdruck und vor allem durch Schilderungen von Situationen und Ereignissen durch die Klientin oder den Klienten. So kann die Interviewerin oder der Interviewer etwa sagen: „Ich habe den Eindruck, dass bei Ihnen Wut und Ärger hochkommen, wenn Sie über dieses Thema sprechen.“ Die Gefühle sollten dabei, wie im Beispiel, möglichst präzise benannt werden.

Eigene Gefühle nicht mit denen der interviewten Person verwechseln

Grundsätzlich besteht die Gefahr, dass Interviewerinnen oder Interviewer ihre eigenen Gefühle mit denen der interviewten Person verwechseln. Eine Interviewerin würde sich in der vom Klienten geschilderten Situation vielleicht traurig fühlen, der Klient hat aber Scham empfunden und nicht Traurigkeit. Daher sollten Interviewerinnen oder Interviewer bereit sein, auch andere Gefühle zu akzeptieren. Es kann hilfreich sein, Nachfragen zu Gefühlen eher als Vermutungen und Interpretationen kenntlich zu machen und nicht etwa als ein Expertinnen- bzw. Expertenurteil erscheinen zu lassen. Damit öffnet man der interviewten Person die Tür für Korrekturen. Über das Explorieren von Gefühlen und den Umgang mit Gefühlen informieren Morrison (2014, S. 62 ff.) sowie Wittmann und Holling (2001, S. 45 ff.).

Anzeichen für Widerstand erkennen

Umgang mit Widerstand Unter Widerstand wird jeder bewusste oder unbewusste Versuch der Klientin oder des Klienten verstanden, ein Gesprächsthema zu vermeiden (Morrison 2014, S. 198 ff.). Die folgenden Ausführungen zum Thema Widerstand sind an Morrison (2014) angelehnt. Manchmal wird der Widerstand von der interviewten Person offen thematisiert mit Aussagen wie „Darüber möchte ich nicht sprechen“, „Das Thema ist mir unangenehm, ich würde lieber nicht darüber reden“. Alle anderen Anzeichen von Widerstand sind mehrdeutig und zudem oft schwer zu erkennen. Mehrdeutig sind sie, weil nicht sicher ist, ob sie anzeigen, dass die Person nicht über das Thema sprechen will, oder ob ihr Verhalten andere Ursachen hat. So kann Schweigen auf angestrengtes und (noch) erfolgloses Nachdenken zurückzuführen sein, und „sich nicht richtig erinnern können“ kann tatsächlich auf mangelndem Erinnerungsvermögen beruhen. Deshalb sollte man sich bewusst sein, dass die in der Übersicht (s. u.) genannten Anzeichen von

Widerstand keine sicheren Anzeichen sind. Sie sollen vielmehr als Hinweis darauf verstanden werden, dass hier Widerstand vorliegen könnte. Die Interviewerin oder der Interviewer wird sich die Frage stellen, welche Gründe die Klientin oder der Klient dafür haben könnte, bei diesem Thema Widerstand zu zeigen.

Die *Gründe für Widerstand* können unterschiedlich sein. Einige Themen sind für die meisten Menschen heikel. Dazu zählen illegale Aktivitäten, Sexualität und Geld. Darüber zu sprechen kann Angst oder Verlegenheit auslösen. Die Befürchtung, die Antwort könnte für einen selbst oder eine nahestehende Person negative Konsequenzen haben, kann ebenfalls zu Widerstand führen. Eine Klientin fürchtet sich vielleicht vor einer bestimmten Diagnose, oder ein Bewerber vermutet, dass seine Einstellungschancen sinken, wenn er sich offen auf ein bestimmtes Thema einlässt. Die Ursache für Widerstand kann sogar bei der Interviewerin oder dem Interviewer liegen, wenn es ihnen nicht gelingt, eine vertrauensvolle Beziehung herzustellen, oder wenn sie negativ wertend auf Antworten reagieren.

Gründe für Widerstand

Beispiele für indirekte Anzeichen von Widerstand

- Verspätetes Erscheinen zum Interview
- Nonverbale Anzeichen dafür, dass das Thema unangenehm ist:
 - Gähnen
 - Erröten
 - Auf die Uhr schauen
 - Meiden des Blickkontakts
 - Schweigen
- Verbales Vermeidungsverhalten:
 - „Sich nicht erinnern können“
 - Themenwechsel
 - Unpräzise Angaben, Auslassungen

Der Umgang mit Widerstand richtet sich nach den vermuteten Ursachen. Es gilt, die Ursachen abzustellen, sofern dies möglich ist. Die Interviewerin oder der Interviewer kann Verständnis dafür zeigen, dass es der befragten Person schwerfällt, über das Thema zu sprechen, kann sie überzeugen, dass ihre Ängste unbegründet oder dass ihre negativen Gefühle bei diesem Thema „normal“ und angemessen sind.

Verständnis zeigen

Gelingt es nicht, die Gründe für den vermuteten Widerstand zu erkennen oder zu beseitigen, stehen immer noch einige *allgemeine Strategien* zur Wahl. Schweigt die befragte Person, kann die Interviewerin oder der Interviewer signalisieren, dass sie bzw. er bereit ist, zu warten. Alternativ oder als nächster Schritt kann sie oder er die Frage, eventuell leicht umformuliert, wiederholen. Unter Umständen ist es hilfreich, die Gefühle der interviewten Person zu verbalisieren, indem man beispielsweise sagt: „Ich sehe, dass es Ihnen schwerfällt, darüber zu sprechen, dass es Ihnen peinlich ist.“ Auch eine nachgereichte Begründung für die Frage (z. B. „Um Ihnen helfen zu können, muss ich auch von Ihnen wissen, …“) wird in manchen Fällen helfen, den Widerstand zu beseitigen.

Abwarten, Fragen umformulieren, Gefühle ansprechen

3

Konfrontation

Unter Umständen kann auch eine Konfrontation angemessen sein. Ist etwa bekannt, dass der Klient Drogen konsumiert hat, kann Widerstand beim Thema Drogen damit beseitigt werden, indem man konfrontativ vorgeht: „Sie haben doch Erfahrung mit Drogen. Ich möchte jetzt gerne mehr darüber wissen.“ Schließlich bleibt noch die Möglichkeit, das Thema auf einen späteren Zeitpunkt zu verschieben.

Ziele erläutern

Kontrolle über die Gesprächsführung Die Gefahr, dass eine Klientin oder ein Klient über irrelevante Themen spricht, unwichtige Details zu sehr ausbreitet oder auf Themen zurückkommt, die bereits hinreichend geklärt sind, besteht besonders dann, wenn sie oder er nicht weiß, was in diesem Interview wichtig und was unwichtig ist. Deshalb wird in der Eröffnungsphase die Zielsetzung erläutert. Bei der Überleitung zu einem neuen Themenblock ist es hilfreich, erneut eine Orientierung zu geben und die Notwendigkeit der Fragen herauszustellen.

Klare, verständliche Fragen

Im Allgemeinen sind klare, verständliche Fragen günstig, da sie keinen Spielraum für Missverständnisse bieten. Auch wenn diese Vorkehrungen getroffen werden, kann es passieren, dass die Klientin oder der Klient abschweift. Für diesen Fall stehen folgende Interventionsmöglichkeiten zur Verfügung.

Interventionsmöglichkeiten beim Abschweifen vom Thema

- Auf die Ausgangsfrage zurückkommen
- Äußerungen für Überleitung auf die eigentliche Frage nutzen
- Vermehrtes Stellen geschlossener Fragen
- Paraphrasieren, dabei das Wichtige aufgreifen
- Nonverbale Signale, dass die Äußerungen nicht wichtig sind
- Mit dem Anfertigen von Notizen aufhören
- Verstärker (Nicken, „hm“ etc.) aussetzen
- Nonverbale Verstärker bei angemessenen Antworten

Notwendigkeit der Fragen begründen

Peinliche Fragen stellen Einige Themen können für die interviewte Person peinlich sein, und darüber zu sprechen kann bei ihr zu Schamgefühlen und Verlegenheit führen. Deshalb ist es wichtig, dass die Interviewerin oder der Interviewer solche Fragen angemessen einleitet und sie dann mit den richtigen Worten stellt. Dies geschieht, indem man das Thema nennt und kurz begründet, warum man dazu Fragen stellt. Eventuell entschuldigt man sich dafür, dass man nun diese Fragen stellt: „Es tut mir leid, dass ich Sie nun über ... befragen muss.“ Die kritischen Fragen selbst sollten einfach und direkt formuliert sein.

Nicht um das Thema herumreden

Wenn die Interviewerin oder der Interviewer um das Thema herumredet, sich nur indirekt oder umständlich ausdrückt, kann dies die Peinlichkeit noch erhöhen. Die interviewte Person merkt, dass es der Interviewerin oder dem Interviewer selbst peinlich ist, über das Thema zu sprechen. Bei Morrison (2014, S. 96 ff.) finden sich Formulierungsvorschläge für Fragen zu peinlichen Themen, die im klinischen Kontext relevant sein können: Suizid, Gewalttätigkeit, Drogenmissbrauch, Sexualleben und sexuelle Probleme einschließlich Perversionen sowie sexueller Missbrauch.

Nachfolgend sollen die vorgeschlagenen verbalen Interventionsmaßnahmen an einem Beispiel illustriert werden.

► Beispiel

Das diagnostische Interview dient der Feststellung der Eignung und Interessen eines 15-jährigen Schülers zwecks Beratung bei der Berufswahl. Die Interviewerin fragt nach den früheren Leistungen in einzelnen Schulfächern. Der Proband fängt nun an, über die seiner Meinung nach ungerechte Bewertung seiner Leistungen durch den Klassenlehrer zu sprechen. Mit der Bemerkung „Lass uns zuerst einmal über deine Noten sprechen“ kann die Interviewerin auf die Ausgangsfrage zurückkommen. Sie kann seine Äußerung aufgreifen, indem sie fragt: „Welche Noten hat dir denn der Klassenlehrer gegeben?“ Sie kann zu geschlossenen Fragen wechseln: „Lass uns jetzt einmal deine Schulleistungen Fach für Fach betrachten. Welche Noten hast du in Deutsch, ...?“ Das Paraphrasieren könnte mit folgenden Worten geschehen: „Du fühlst dich durch den Lehrer ungerecht beurteilt. Welche Noten hat er dir denn gegeben?“ ◀

Abschließend muss betont werden, dass die Beherrschung von Gesprächsführungstechniken keinen Ersatz für einen sorgfältig ausgearbeiteten Leitfaden darstellt. Diese Techniken kommen idealerweise bei einem leitfadenbasierten Interview zum Einsatz.

Weiterführende Literatur und Internetressourcen

Über klinische Interviews informieren die folgenden beiden Bücher sehr gut, die bereits in der 4. bzw. 6. Auflage erschienen sind und inzwischen als Klassiker gelten können: *The first interview* von Morrison (2014) sowie *Clinical interviewing* von Sommers-Flanagan und Sommers-Flanagan (2017). Für die Planung, Durchführung und Auswertung strukturierter Einstellungsgespräche sind das Buch von Jetter (2008) sowie ein Beitrag von Strobel et al. (2018), der die Umsetzung der DIN 33430 zur Berufseignungsdiagnostik zeigt, zu empfehlen. Beide Werke orientieren sich an wissenschaftlichen Standards und sind zugleich sehr praxisbezogen. Ebenfalls sehr praxisnah und wissenschaftlich fundiert ist ein Beitrag von Krumm et al. (2015), der sich allgemein mit der Konstruktion, Durchführung und Auswertung von Interviews befasst.

Der Arbeitskreis Assessment Center e. V. (seit 2019 „Forum Assessment e. V.“) hat 2008 die Handreichung „Interview-Standards“ veröffentlicht (s. ▶ https://www.forum-assessment.de/images/standards/149_akac-interview-standards.pdf). Sie enthält nützliche Anregungen zur Gestaltung, Auswertung und Dokumentation von eignungsdiagnostischen Interviews.

?

Übungsfragen

— Abschn. 3.7:

- Was versteht man unter Anamnese und unter Exploration?
- Was bedeutet eine standardisierte Auswertung bei einem diagnostischen Interview?
- Wann bietet sich der Einsatz standardisierter Interviews besonders an?
- Was bedeutet Standardisierung konkret beim Strukturierten Klinischen Interview für DSM-5-Störungen – Klinische Version (SKID-5-CV)?
- Warum ist es problematisch, Kappa-Koeffizienten über .70 für die Übereinstimmung zweier Interviewer/-innen als gute Übereinstimmung zu interpretieren?
- Was sind die wesentlichen Merkmale des sog. „multimodalen Einstellungsinterviews“?
- Welches sind die wichtigsten Ergebnisse aus Metaanalysen zur Kriteriumsvalidität von Eignungsinterviews?

- Von welchen Faktoren hängt die Beurteilung im Eignungsinterview ab?
Welche Rolle spielt die Art des Interviews dabei?
- Welche 3 Phasen unterscheiden Westhoff und Kluck (2008) bei der Grobstruktur eines Interviews?
- Welche Vorteile hat man, wenn man einen Leitfaden ausformuliert?
- Nennen Sie Maßnahmen, um die interviewte Person zum Reden zu motivieren!
- Woran kann man erkennen, dass der Klient/die Klientin bei einem Thema oder dem ganzen Interview Widerstand leistet?
- Nennen Sie Strategien zum adäquaten Umgang mit Widerstand!
- Wie kann man intervenieren, wenn der Klient/die Klientin vom Thema abschweift?

3.8 Zusammenfassung

Die Psychologische Diagnostik verfügt über ein großes Repertoire an Messverfahren. Diese lassen sich nach dem allgemeinen Messgegenstand (Leistungen, Persönlichkeit, psychische Störungen, Verhalten) sowie nach dem methodischen Vorgehen (Testen, Befragen, Beobachten) unterteilen. Daraus ergeben sich Kombinationen wie „Leistungstests“ (Messung von Leistungen mittels Tests) oder Persönlichkeitsfragebögen (Messung von Persönlichkeit mittels Fragebogen). Der Messgegenstand wird zudem durch Konstrukte (Intelligenz, Konzentrationsfähigkeit, diverse Persönlichkeitsmerkmale etc.) inhaltlich näher spezifiziert.

Ein bestimmter Messgegenstand kann in der Regel mit unterschiedlichen Methoden erfasst werden. Wenn davon Gebrauch gemacht wird, spricht man von einem multimethodalen Vorgehen. Beispielsweise lassen sich zur Konzentrationsfähigkeit mithilfe von Tests, Verhaltensbeobachtung und diagnostischem Interview diagnostisch relevante Informationen gewinnen. Für Persönlichkeitsmerkmale kommen Fragebögen, objektive Persönlichkeitstests, projektive Verfahren und sogar der Einsatz künstlicher Intelligenz infrage: Ferner können eine systematische Verhaltensbeobachtung und ein diagnostisches Interview Erkenntnisse über die Persönlichkeit liefern.

Von wenigen Ausnahmen abgesehen zeichnen sich alle Verfahren durch eine hohe Objektivität aus. Es liegen Informationen zu ihrer Reliabilität und Validität vor, über die in diesem Kapitel informiert wurde. Die Informationen über die Gütekriterien sind wichtig für die Auswahl von diagnostischen Verfahren zwecks Beantwortung einer bestimmten Fragestellung. Die meisten in diesem Kapitel vorgestellten Verfahren liegen in standardisierter Form vor und können über Testverlage bezogen werden. Vorgestellt wurden Leistungstests (zur Konzentrationsfähigkeit, Aufmerksamkeit, Intelligenz sowie zu speziellen Fähigkeiten, dem Entwicklungsstand und zu schulischen Fertigkeiten wie Lesen, Schreiben und Rechnen), Fragebögen zur Persönlichkeit, zu einzelnen Persönlichkeitsmerkmalen sowie Motiven und Interessen und auch zum aktuellen Befinden. Mit den objektiven Persönlichkeitstests und projektiven Verfahren wurden alternative Messmethoden vorgestellt, die ohne Befragung der Testpersonen auskommen. Auch wenig standardisierte (und damit weniger objektive) Verfahren haben oftmals ihre Berechtigung. Besonders zu erwähnen sind hier freie die Verhaltensbeobachtung und halbstandardisierte Interviews.

Mit der Verhaltensbeobachtung (einschließlich Verhaltensbeurteilung) und dem diagnostischen Interview wurden 2 Verfahrensgruppen vorgestellt, die in der diagnostischen Praxis eine herausragende Bedeutung haben, wie Umfragen zeigen. Zu beiden Ansätzen liegen nur wenige käufliche Verfahren vor. Für viele Anwendungszwecke müssen sie eigens konstruiert werden. Deshalb wurden hier praktische Hinweise gegeben, wie dabei vorzugehen ist. Speziell für die Durchführung eines Interviews ist zudem die Beherrschung übergreifender Techniken der Gesprächsführung unerlässlich, über die deshalb ausführlich informiert wurde.

Literatur

- Abels, D. (1974). *KVT: Konzentrations-Verlaufs-Test*. Göttingen: Hogrefe.
- Abrell-Vogel, C., & Gerstenberg, F. (2014). TBS-TK Rezension: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren: Modul zur Selbstbeschreibung (BIP-6F). *Psychologische Rundschau* 65, 195–197.
- Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General* 131, 567–589.
- Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin* 121, 219–245.
- Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: Erlbaum.
- Albers, F., & Höft, S. (2009). „Do it again and again. And again?“ Übungseffekte bei einem computergestützten Test zum räumlichen Vorstellungsvermögen. *Diagnostica* 55, 71–83.
- Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs* 47, whole no 211.
- Amthauer, R. (1953). *I-S-T: Intelligenz-Struktur-Test*. Göttingen: Hogrefe.
- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000: Intelligenz-Struktur-Test 2000*. Göttingen: Hogrefe.
- Andresen, B. (2006). *IKP: Inventar Klinischer Persönlichkeitsakzentuierungen: Dimensionale Diagnostik nach DSM-IV und ICD-10*. Göttingen: Hogrefe.
- Andresen, B., & Beauducel, A. (2008). TBS-TK Rezension: "NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R)". *Report Psychologie* 33, 543–544.
- Arnold, W. (1975). *Der Pauli-Test. Anweisung zur sachgemäßen Durchführung, Auswertung und Anwendung des Kraepelinschen Arbeitsversuchs* (5. Aufl.). Berlin, Heidelberg: Springer.
- Asendorpf, J. (2020). Persönlichkeit. In M. A. Wirtz (Hrsg.), Dorsch – Lexikon der Psychologie. ► <https://portal.hogrefe.com/dorsch/persoenlichkeit/>. Zugegriffen: 20. März 2020.
- Ashton, M., & Lee, K. (2004). HEXACO Personality Inventory-Revised (HEXACO-PI-R). APA PsycTests. ► <https://doi.org/10.1037/t03138-000>.
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review* 11, 150–166.
- Ashton, M. C., Jackson, D. N., Helmes, E., & Paunonen, S. V. (1998). Joint factor analysis of the Personality Research Form and the Jackson Personality Inventory: Comparisons with the Big Five. *Journal of Research in Personality* 32, 243–250.
- Atkinson, J. W. (1953). The achievement motive and recall of interrupted and completed tasks. *Journal of Experimental Psychology* 46, 381–390.
- Bales, R. F. (1975). Die Interaktionsprozessanalyse: Ein Beobachtungsverfahren zur Untersuchung kleiner Gruppen. In R. König (Hrsg.), *Beobachtung und Experiment in der Sozialforschung* (8. Aufl., S. 148–167). Köln: Kiepenheuer & Witsch.
- Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology* 94, 1394–1411.
- Bartenwerfer, H. (1964). Allgemeine Leistungstests. In R. Heiss (Hrsg.), *Handbuch der Psychologie: Band VI. Psychologische Diagnostik* (S. 385–410). Göttingen: Hogrefe.
- Bayley, N. (2014). *BAYLEY-III: Bayley Scales of Infant and Toddler Development – Third Edition. Deutsche Bearbeitung: G. Reuner, J. Rosenkranz*. Frankfurt am Main: Pearson Clinical & Talent Assessment Deutschland.

- Becker, P. (2003). *TIPI: Trierer Integriertes Persönlichkeitsinventar*. Göttingen: Hogrefe.
- Beckmann, D., Brähler, E., & Richter, H.-E. (2012). *GT-II: Der Gießen-Test – II*. Göttingen: Hogrefe.
- Beesdo-Baum, K., Zaudig, M., & Wittchen, H.-U. (Hrsg.). (2019). *SCID-5-CV: Strukturiertes Klinisches Interview für DSM-5®-Störungen – Klinische Version. Deutsche Bearbeitung des Structured Clinical Interview for DSM-5® Disorders – Clinician Version von Michael B. First, Janet B. W. Williams, Rhonda S. Karg, Robert L. Spitzer*. Göttingen: Hogrefe.
- Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2 Restructured Form: Manual for administration, scoring and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Ben-Porath, Y. S., & Tellegen, A. (2016). *Part 5: MMPI-2-RF Development*. Minneapolis, MN: University of Minnesota Press. ► <https://conservancy.umn.edu/handle/11299/187605>. Zugriffen: 04. Mai 2020.
- Benton-Sivan, A., & Spreen, O. (2009). *Der Benton Test* (8. Aufl.). Bern: Huber.
- Bents, R., & Blank, R. (2003). *Der M.B.T.I.: Die 16 Grundmuster unseres Verhaltens nach C. G. Jung* (4. Aufl.). München: Claudio.
- Berg, K. C., Peterson, C. B., Frazier, P., & Crow, S. J. (2011). Convergence of scores on the interview and questionnaire versions of the Eating Disorder Examination: A meta-analytic review. *Psychological Assessment* 23, 714–724.
- Bergmann, C., & Eder, F. (2018). *AIST-3: Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-3) – Version 3*. Göttingen: Hogrefe.
- Berliner, L., & Lieb, R. (2001). *Child sexual abuse investigations: Testing documentation methods*. Olympia, WA: Washington State Institute for Public Policy.
- Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review* 23, 190–203.
- Bleidorn, W., Hopwood, C. J., & Lucas, R. E. (2018). Life events and personality trait change. *Journal of Personality* 86, 83–96.
- Blotenberg, I., & Schmidt-Atzert, L. (2019a). On the characteristics of sustained attention test performance: The role of the preview benefit. *European Journal of Psychological Assessment*. doi: ► <https://doi.org/10.1027/1015-5759/a000543>.
- Blotenberg, I., & Schmidt-Atzert, L. (2019b). On the locus of the practice effect in sustained attention tests. *Journal of Intelligence* 7: 12. doi: ► <https://doi.org/10.3390/intelligence7020012>
- Blotenberg, I., & Schmidt-Atzert, L. (2019c). Towards a process model of sustained attention tests. *Journal of Intelligence* 7: 3. doi: ► <https://doi.org/10.3390/intelligence7010003>.
- Bohm, E. (2004). *Lehrbuch der Rorschach-Psychodiagnostik* (8. Aufl.). Bern: Huber.
- Bölte, S., Adam-Schwebe, S., Englert, E., Schmeck, K., & Poustka, F. (2000). Zur Praxis der psychologischen Testdiagnostik in der deutschen Kinder- und Jugendpsychiatrie: Ergebnisse einer Umfrage. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie* 28, 151–161.
- Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI)* nach Costa & McCrae. Göttingen: Hogrefe.
- Borkenau, P., Friedel, A., & Wolfradt, U. (2011). Standardisierte Persönlichkeitssfragebögen. In L. Hornke, M. Amelang, & M. Kersting (Hrsg.), *Persönlichkeitssdiagnostik* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 4, S. 1–72). Göttingen: Hogrefe.
- Bös, K. (Hrsg.). (2017). *Handbuch Motorische Tests: Sportmotorische Tests, Motorische Funktionstests, Fragebögen zur körperlich-sportlichen Aktivität und sportpsychologische Diagnoseverfahren* (3. Aufl.). Göttingen: Hogrefe.
- Brähler, E., Holling, H., Leutner, D., & Petermann, F. (Hrsg.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (3. Aufl.). Göttingen: Hogrefe.
- Brandstätter, V. (2005). Der objektive Leistungsmotivations-Test OLMT von L. Schmidt-Atzert. Rezension. *Zeitschrift für Personalpsychologie* 4, 132–137.
- Brandt, I., & Sticker, E. J. (2001). *GES: Griffiths-Entwicklungsskalen zur Beurteilung der Entwicklung in den ersten beiden Jahren* (2. Aufl.). Göttingen: Beltz Test GmbH.
- Brem-Gräser, L. (2001). *Familie in Tieren: Die Familiensituation im Spiegel der Kinderzeichnung. Entwicklung eines Testverfahrens* (8. Aufl.). München: Reinhardt.
- Brickenkamp, R. (1962). *Test d2: Aufmerksamkeits-Belastungs-Test*. Göttingen: Hogrefe.
- Brickenkamp, R. (1994). *Test d2: Aufmerksamkeits-Belastungs-Test* (8. Aufl.). Göttingen: Hogrefe.
- Brickenkamp, R., Merten, T., & Hänsgen, K.-D. (1996). *d2-C: Aufmerksamkeits-Belastungs-Test*. Göttingen: Hogrefe.
- Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *d2-R: Test d2 – Revision. Aufmerksamkeits- und Konzentrationstest*. Göttingen: Hogrefe.
- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Diagnostische Verfahren

- Bühner, M., Schmidt-Atzert, L., Grieshaber, E., & Lux, A. (2001). Faktorenstruktur verschiedener neuropsychologischer Tests: Ergebnisse einer retrospektiven Studie mit hirngeschädigten Patienten. *Zeitschrift für Neuropsychologie* 12, 181–187.
- Bundesamt für Sicherheit in der Informationstechnik. (2020). E-Mails verschlüsseln, unerwünschte Mitleser ausschließen. ► https://www.bsi-fuer-buerger.de/BSIFB/DE/Service/Aktuell/Informationen/Artikel/EMails_verschluesseln.html. Zugegriffen: 08. Mai 2020.
- Butcher, J., Dahlstrom, W., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Manual for the administration and scoring of the MMPI-2*. Minneapolis, MN: University of Minnesota Press.
- Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC-V in four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools* 55, 741–769.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. In I. Dennis, & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 1–17). Mahwah, N.J.: Lawrence Erlbaum.
- Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology* 31, 161–179.
- Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests: A theoretical introduction and practical compendium*. Champaign, IL: University of Illinois Press.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology* 82, 300–310.
- Christiansen, E. (1983). *Die Arbeitskurve nach Emil Kraepelin und Richard Pauli: Mainzer Version*. Weinheim: Beltz.
- Clause, C. S., Delbridge, K., Schmitt, N., Chan, D., & Jennings, D. (2001). Test preparation activities and employment test performance. *Human Performance* 14, 149–167.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.
- Cohen, R. A. (1993). The neuropsychology of attention. New York: Plenum Press.
- Collins, C. J., Hanges, P. J., & Locke, E. A. (2004). The relationship of achievement motivation to entrepreneurial behavior: A meta-analysis. *Human Performance* 17, 95–117.
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment* 15, 110–117.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease* 175, 526–536.
- Czopp, T. S., & Zeligman, R. (2016). The Rorschach Comprehensive System (CS) psychometric validity of individual variables. *Journal of Personality Assessment* 98, 335–342.
- Daniel, M., & Wahlstrom, D. (2018). Raw-score equivalence of computer-assisted and paper versions of WISC-V. *Psychological Services* 16, 213–220.
- Daseking, M., & Putz, D. (2015). TBS-TK Rezension "Test d2-R: Aufmerksamkeits- und Konzentrationstest. *Report Psychologie* 40, 323–324.
- Deimann, P., Kastner-Koller, U., Esser, G., & Hänsch, S. (2010). TBS-TK Rezension: FRANKIS Fragebogen zur fröhdkindlichen Sprachentwicklung. FRANKIS (Standardform) und FRANKIS-K (Kurzform). *Psychologische Rundschau* 61, 169–171.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* 41, 417–440.
- Döpfner, M., Dietmair, I., Mersmann, H., Simon, K., & Trost-Brinkhues, G. (2005). *S-ENS: Screening des Entwicklungsstandes bei Einschulungsuntersuchungen*. Göttingen: Hogrefe.
- Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology* 22, 90–104.
- Düker, H., & Lienert, G. A. (1965). *Konzentrations-Leistungs-Test KLT*. Göttingen: Hogrefe.
- Düker, H., Lienert, G. A., Lukesch, H., & Mayrhofer, S. (2001). *KLT-R: Konzentrations-Leistungs-Test, revidierte Fassung*. Göttingen: Hogrefe.
- Eggert, D. (1974). *LOS KF 18: Lincoln-Oseretzký-Skala Kurzform. Deutsche Bearbeitung von D. Eggert* (2. Aufl.). Weinheim: Beltz-Test.
- Eimer, M., Nattkemper, D., Schröger, E., & Prinz, W. (1996). Involuntary attention. In O. Neumann, & A. F. Sanders (Eds.), *Handbook of perception and action: Attention* (Vol. 3, pp. 155–184). San Diego, CA: Academic Press.
- Endlich, D., Berger, N., Küspert, P., Lenhard, W., Marx, P., Weber, J., Schneider, W. (2017). *WVT-Würzburger Vorschultest: Erfassung schriftsprachlicher und mathematischer (Vorläufer-) Fertigkeiten und sprachlicher Kompetenzen im letzten Kindergartenjahr*. Göttingen: Hogrefe.

- Engel, R. R. (2003). Stellungnahme zur Testrezension des MMPI-2 durch Hank und Schwenkmezger (2003). *Report Psychologie* 28, 304–306.
- Engel, R. R. (2019). *MMPI-2-RF: Minnesota Multiphasic Personality Inventory – 2 Restructured Form: Deutschsprachige Adaptation des Minnesota Multiphasic Personality Inventory – 2 – Restructured Form® von Yossef Ben-Porath und Auke Tellegen*. Göttingen: Hogrefe.
- Entwistle, D. R. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin* 77, 377–391.
- Erdmann, G., & Janke, W. (2008). *SVF: Stressverarbeitungsfragebogen. Stress, Stressverarbeitung und ihre Erfassung durch ein mehrdimensionales Testsystem* (4. Aufl.). Göttingen: Hogrefe.
- Evers, A., Muniz, J., Bartram, D., Boben, D., Egeland, J., Fernandez-Hermida, J. R., Frans, Ö., et al. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist* 17, 300–319.
- Exner, J. E. (1974). *The Rorschach: A comprehensive system. Basic Foundations* (vol. 1). New York, NY: John Wiley & Sons Inc.
- Exner, J. E. (2003). *The Rorschach: A comprehensive system* (4th ed.). New York, NY: John Wiley & Sons Inc.
- Exner, J. E. (2010). *Rorschach-Arbeitsbuch für das Comprehensive System. Deutschsprachige Fassung des „A Rorschach Workbook for the Comprehensive System – Fifth Edition“*. Bern: Huber.
- Fahrenberg, J., Hampel, R., & Selg, H. (2010). *FPI-R: Freiburger Persönlichkeitsinventar* (8. Aufl.). Göttingen: Hogrefe.
- Fahrenberg, J., Hampel, R., & Selg, H. (2020). *FPI-R: Freiburger Persönlichkeitsinventar* (9. Aufl.). Göttingen: Hogrefe.
- Faßnacht, G. (1995). *Systematische Verhaltensbeobachtung: Eine Einführung in die Methodologie und Praxis* (2. Aufl.). München: Reinhardt.
- Fay, E. (2003). Bochumer Matrizentest (BOMAT – advanced – short version). In E. Fay (Hrsg.), *Tests unter der Lupe 4: Aktuelle psychologische Testverfahren – kritisch betrachtet* (S. 24–35). Göttingen: Vandenhoeck & Ruprecht.
- Fineman, S. (1977). The achievement motive construct and its measurement: Where are we now? *British Journal of Psychology* 68, 1–22.
- Fischer, S. (2010). US-Urteile über deutsche Politiker. Blamierte Regierung überspielt Deutschen-Affront. Der Spiegel. Artikel vom 29. November 2010. ► <https://www.spiegel.de/politik/deutschland/us-urteile-ueber-deutsche-politiker-blamierte-regierung-ueberspielt-deutschen-affront-a-731779.html>. Zugegriffen: 04. Mai 2020.
- Frintrup, A., & Schuler, H. (2007). *SMT: Sportbezogener Leistungsmotivationstest*. Göttingen: Hogrefe.
- Funsch, K., & Arias Martin, B. (2017). *DKT-K: Differentieller Konzentrationstest für Kinder*. Göttingen: Hogrefe.
- Garb, H. N., Wood, J. M., Lilienfeld, S. O., & Nezworski, M. T. (2005). Roots of the Rorschach controversy. *Clinical Psychology Review* 25, 97–118.
- Gardner, H. (2002). *Intelligenzen: Die Vielfalt des menschlichen Geistes*. Stuttgart: Klett Cotta.
- Gasteiger-Klicpera, B., & Sticker, E. (2011). TBS-TK Rezension: „Deutscher Rechtschreibtest für das erste und zweite/dritte und vierte Schuljahr, DERET 1-2+/3-4+“. *Psychologische Rundschau* 63, 75–77.
- Gawronski, B. (2006). Die Technik des Impliziten Assoziationstests als Grundlage für Objektive Persönlichkeitstests. In T. M. Ortner, R. T. Proyer, & K. D. Kubinger (Hrsg.), *Theorie und Praxis Objektiver Persönlichkeitstests* (S. 53–69). Bern: Huber.
- Gay, F. (2004). *Das persolog Persönlichkeits-Profil: Persönliche Stärke ist kein Zufall. Mit Fragebogen zur Selbstauswertung* (40. Aufl.). Offenbach, Remchingen: Gabal Verlag und persolog GmbH.
- Gignac, G. E. (2006). The WAIS-III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences* 27, 73–86.
- Goeters, K. M. (1981). Zur Taxonomie fehlerhaften Arbeitens in psychologischen Leistungstests. *Zeitschrift für Differentielle und Diagnostische Psychologie* 2, 237–251.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality* 40, 84–96.
- Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.
- Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology* 82, 379–398.
- Gosling, S. D., Kwan, V. S., & John, O. P. (2003). A dog's got personality: A cross-species comparative approach to personality judgments in dogs and humans. *Journal of Personality and Social Psychology* 85, 1161–1169.

Diagnostische Verfahren

- Green, E., Stroud, L., Bloomfield, S., Cronje, J., Foxcroft, C., Hurter, K., et al. (2015). *Griffiths III: Griffiths Scales of Child Development*. Oxford, UK: Hogrefe.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97, 17–41.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Boston: Pearson.
- Grieder, S., & Grob, A. (2019). Exploratory factor analyses of the Intelligence and Development Scales – 2: Implications for theory and practice. *Assessment*. doi: ▶ <https://doi.org/10.1177/1073191119845051>.
- Griffiths, R. (1954). *The abilities of babies: A study in mental measurement*. New York: McGraw-Hill.
- Grob, A., & Hagmann-von Arx, P. (2011). Replik auf die Rezension von Koch, H., Kastner-Koller, U., & Deimann, P. (2011). Testbesprechung „Grob, A., Meyer, C. S., & Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren. Bern: Huber.“. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 43, 246–249.
- Grob, A., & Hagmann-von Arx, P. (2018). *IDS-2: Intelligence and Development Scales – 2. Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche*. Göttingen: Hogrefe.
- Grob, A., Meyer, C. S., & Hagmann-von Arx, P. (2009). *IDS: Intelligence and Development Scales. Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren* (2. Aufl.). Bern: Huber.
- Grube, A., Schroer, J., Hentzschel, C., & Hertel, G. (2008). The event reconstruction method: An efficient measure of experience-based job satisfaction. *Journal of Occupational and Organizational Psychology* 81, 669–689.
- Gruber, N., & Tausch, A. (2016). TBS-TK-Rezension: CFT 20-R mit WS/ZF-R. Grundintelligenztest Skala 2 – Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest – Revision (WS/ZF-R). *Psychologische Rundschau* 67, 77–79.
- Grund, M., Leonhart, R., & Naumann, C. L. (2017). *DRT 5: Diagnostischer Rechtschreibtest für 5. Klassen* (3. Aufl.). Göttingen: Hogrefe.
- Guttmann, G., & Bauer, H. (2004). *RISIKO – Risikowahlverhalten*. Mödling: Schuhfried.
- Häcker, H., Schmidt, L. R., Schwenkmezger, P., & Utz, H. (1975). *Objektive Testbatterie OA-TB 75*. Weinheim: Beltz.
- Hahlweg, K. (2016). *FPD: Fragebogen zur Partnerschaftsdiagnostik* (2. Aufl.). Göttingen: Hogrefe.
- Hank, P., & Schwenkmezger, P. (2003). Das Minnesota Personality Inventory-2 (MMPI): Testbesprechung im Auftrag des Testkuratoriums. *Report Psychologie* 28, 294–306.
- Hasselhorn, M., & Margraf-Stiksrud, J. (2015). TBS-TK Rezension: "Entwicklungs test für Kinder von sechs Monaten bis sechs Jahren – Revision (ET 6-6 R)". *Psychologische Rundschau* 66, 208–210.
- Hathaway, S. R., & McKinley, J. C. (1942). *Manual for the Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.
- Hathaway, S. R., McKinley, J. C., & Engel, R. R. (2000). *MMPI-2: Minnesota Multiphasic Personality Inventory-2*. Bern: Huber.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology* 92, 373–385.
- Heckhausen, H. (1963). *Hoffnung und Furcht in der Leistungsmotivation*. Meisenheim am Glan: Anton Hain.
- Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik* 21, 251–270.
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research* 58, 47–77.
- Hergovich, A., Bognar, B., Arendasy, M., & Sommer, M. (2005). *WRBTW: Wiener Risikobereitschaftstest Verkehr* [Test & Manual]. Mödling: Schuhfried.
- Hesse, J., & Schrader, H.-C. (2015). *Testtraining 2000plus: Einstellungs- und Eignungstests erfolgreich bestehen*. Hallbergmoos: Stark.
- Hessler, R. M., Downing, J., Beltz, C., Pelliccio, A., Powell, M., & Vale, W. (2003). Qualitative research on adolescent risk using e-mail: A methodological assessment. *Qualitative Sociology* 26, 111–124.
- Heubrock, D., & Petermann, F. (2001). *Aufmerksamkeitsdiagnostik*. Göttingen: Hogrefe.
- Heyde, G. (2000). *INKA: Inventar komplexer Aufmerksamkeit*. Frankfurt am Main: Swets.
- Heyde, G. (2004). INKA – Inventar Komplexer Aufmerksamkeit. In G. Büttner, & L. Schmidt-Atzert (Hrsg.), *Diagnostik von Konzentration und Aufmerksamkeit* (S. 133–142). Göttingen: Hogrefe.

- Hibbard, S. (2003). A Critique of Lilienfeld et al.'s (2000) „The scientific status of projective techniques“. *Journal of Personality Assessment* 80, 260–271.
- Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment* 11, 278–296.
- Hof, J.-T. (2012). *Research Real Life: Aufmerksamkeit in Tests und im wahren Leben. [Unveröffentlichte Diplomarbeit]*. Marburg: Philipps-Universität.
- Hoffmeyer, M. (2017). Bewerbungsgespräch bei einem Computer. *Süddeutsche Zeitung*. Artikel vom 04. November 2017. ► <https://www.sueddeutsche.de/karriere/einstellungstest-bewerbungsgespraech-bei-einem-computer-1.3725219>. Zugegriffen: 05. Mai 2020.
- Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin* 31, 1369–1385.
- Höft, S., & Kersting, M. (2018). Anforderungsprofil, Verhaltensbeobachtung und Verhaltensbeurteilung. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430*(S. 27–63). Berlin, Heidelberg: Springer.
- Höft, S., Püttner, I., & Kersting, M. (2018). Anforderungsanalyse, Verfahren der Eignungsbeurteilung sowie rechtliche Rahmenbedingungen. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430*(S. 95–153). Berlin, Heidelberg: Springer.
- Holden, R. R., Wood, L. L., & Tomashevski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology* 81, 160–169.
- Holland, J. L. (1970). *The self-directed search for educational and vocational planning*. Palo Alto, Calif.: Consulting Psychologists Press.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Odessa, FL: Psychological Assessment Resources.
- Holling, H., Preckel, F., & Vock, M. (2005). *Intelligenzdiagnostik*. Göttingen: Hogrefe.
- Holmes, D. S. (1974). The conscious control of thematic projection. *Journal of Consulting and Clinical Psychology* 42, 323–329.
- Horn, W. (1983). *LPS: Leistungsprüfssystem* (2. Aufl.). Göttingen: Hogrefe.
- Horn, J. L., & Cattell, R. B. (1966). Refinement and test of theory of fluid and crystallized intelligence. *Journal of Educational Psychology* 57, 253–270.
- Hossiep, R., & Hasella, M. (2010). *BOMAT – Standard: Bochumer Matrizentest Standard*. Göttingen: Hogrefe.
- Hossiep, R., & Krüger, C. (2012). *BIP-6F: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren: Modul zur Selbstbeschreibung*. Göttingen: Hogrefe.
- Hossiep, R., & Mühlhaus, O. (2005). *Personalauswahl und -entwicklung mit Persönlichkeitstests*. Göttingen: Hogrefe.
- Hossiep, R., & Paschen, M. (2003). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* (2. Aufl.). Göttingen: Hogrefe.
- Hossiep, R., & Paschen, M. (2019). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* (3. Aufl.). Göttingen: Hogrefe.
- Hossiep, R., & Weiß, S. (2018). *BIP-Normenband 2018. Ergänzung zum Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung von Rüdiger Hossiep und Michael Paschen*. Göttingen: Hogrefe.
- Hossiep, R., & Weiß, S. (2020). *BIP-AM: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – Anforderungsmodul*. Göttingen: Hogrefe.
- Hossiep, R., Turck, D., & Hasella, M. (1999). *BOMAT – advanced: Bochumer Matrizentest*. Göttingen: Hogrefe.
- Hossiep, R., Turck, D., & Hasella, M. (2001). *BOMAT – advanced – short version: Bochumer Matrizentest*. Göttingen: Hogrefe.
- Hossiep, R., Schecke, J., & Weiß, S. (2015). Zum Einsatz von persönlichkeitsorientierten Fragebogen: Eine Erhebung unter den 580 größten deutschen Unternehmen. *Psychologische Rundschau* 66, 127–129.
- Hoyer, J., Margraf, J., & Schneider, S. (2009). Fragebogen, Ratingskalen und Tagebücher für die verhaltenstherapeutische Praxis. In J. Margraf, & S. Schneider (Hrsg.), *Lehrbuch der Verhaltenstherapie. Grundlagen, Diagnostik, Verfahren, Rahmenbedingungen* (Bd. 1, 3. Aufl., S. 409–433). Berlin, Heidelberg: Springer.
- Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment* 12, 262–273.

Diagnostische Verfahren

- Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment* 21, 264–276.
- Hüppé, M., Schmidt-Atzert, L., Elbert, I., Klasen, B., & Schmucker, P. (2000). Stimmung und Gedächtnis bei Personen mit altersassoziierten Gedächtnisstörungen. *Zeitschrift für Gerontopsychologie und Gerontopsychiatrie* 13, 61–77.
- Jacobs, C., & Petermann, F. (2012). *Diagnostik von Rechenstörungen* (2. Aufl.). Göttingen: Hogrefe.
- Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *BIS-Test: Berliner Intelligenzstruktur-Test, Form 4*. Göttingen: Hogrefe.
- Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., & Beauducel, A. (2006). *BIS-HB: Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdagnostik*. Göttingen: Hogrefe.
- James, N., & Busher, H. (2012). Internet interviewing. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd. ed., pp. 177–188). Thousand Oaks CA: Sage.
- Janke, B., & Janke, W. (2005). Untersuchungen zur Erfassung des Befindens von Kindern: Entwicklung einer Selbstbeurteilungsmethode (EWL40-KJ). *Diagnostica* 51, 29–39.
- Janke, W., & Debus, G. (1978). *Die Eigenschaftswörterliste EWL*. Göttingen: Hogrefe.
- Jetter, W. (2008). *Effiziente Personalauswahl: Durch strukturierte Einstellungsgespräche die richtigen Mitarbeiter finden* (3. Aufl.). Stuttgart: Schäffer-Poeschel.
- Joerin Fux, S., & Stoll, F. (2006). *EXPLOJOB – Das Werkzeug zur Beschreibung von Berufsanforderungen und -tätigkeiten. Deutschsprachige Adaptation und Weiterentwicklung des Position Classification Inventory (PCI) nach Garry D. Gottfredson und John L. Holland*. Bern: Huber.
- Joerin Fux, S., Stoll, F., Bergmann, C., & Eder, F. (2012). *EXPLORIX® – das Werkzeug zur Berufswahl und Laufbahnplanung. Deutschsprachige Adaptation und Weiterentwicklung des Self-Directed Search® (SDS) nach John Holland. Testset Ausgabe Deutschland*. Bern: Huber.
- Jungo, D., & Toggweiler, S. (2019). *Foto-Interessen-Test F-I-T Serie 2020* (6. Aufl.). Bern: Schweizerisches Dienstleistungszentrum Berufsbildung.
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 1776–1780.
- Kanning, U. P. (2003). Sieben Anmerkungen zum Problem der Selbstdarstellung in der Personalauswahl. *Zeitschrift für Personalpsychologie* 2, 193–195.
- Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin* 134, 404–426.
- Kastner-Koller, U., & Deimann, P. (2012). *WET: Wiener Entwicklungstest: Ein Verfahren zur Erfassung des allgemeinen Entwicklungsstandes bei Kindern von 3 bis 6 Jahren* (3. Aufl.). Göttingen: Hogrefe.
- Kaufman, A. S., & Kaufman, N. L. (2015). *KABC-II: Kaufman Assessment Battery for Children – Second Edition. Deutsche Bearbeitung von P. Melchers und M. Melchers*. Frankfurt am Main: Pearson.
- Kaufman, A. S., Kaufman, N. L., Melchers, P., & Preuß, U. (2001). *Kaufman Assessment Battery for Children, deutsche Version* (6. Aufl.). Göttingen: Hogrefe.
- Kersting, M., Althoff, K., & Jäger, A. O. (2008). *WIT-2: Wilde-Intelligenz-Test 2*. Göttingen: Hogrefe.
- Kici, G., & Westhoff, K. (2000). Anforderungen an psychologisch-diagnostische Interviews in der Praxis. *Report Psychologie* 25, 428–436.
- Kiphard, E. J., & Schilling, F. (2017). *KTK: Körperkoordinationstest für Kinder* (3. Aufl.). Göttingen: Hogrefe.
- Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data*. Washington, DC: National Highway Traffic Safety Administration.
- Klauer, S. G., Guo, F., Sudweeks, J., & Dingus, T. A. (2010). *An analysis of driver inattention using a case-crossover approach on 100-car data: Final report (Report No.DOT HS 811 334)*. Washington, DC: National Highway Traffic Safety Administration.
- Klinck, D. (2002). *Computergestützte Diagnostik: Beeinflusst das Medium der Testverarbeitung die Testcharakteristika, die Testfairness oder das Erleben der Testsituation?* Göttingen: Hogrefe.
- Koch, H., Kastner-Koller, U., & Deimann, P. (2011). Testbesprechung Grob, A., Meyer, C.S., & Hagmann-von Arx, P. (2009). Intelligence and Development Scales (IDS). Intelligenz und Entwicklungsskalen für Kinder von 5–10 Jahren. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 43, 108–113.

- Koch, T., Ortner, T. M., Eid, M., Caspers, J., & Schmitt, M. (2014). Evaluating the construct validity of objective personality tests using a multitrait-multimethod-multioccasion-(MT-MM-MO)-approach. *European Journal of Psychological Assessment* 30, 208–230.
- Köller, M., & Zettler, I. (2017). TBS-TK Rezension: „EXPLORIX®—das Werkzeug zur Berufswahl und Laufbahnanplanung“. *Psychologische Rundschau* 68, 98–100.
- Köllner, M. G., & Schultheiss, O. C. (2014). Meta-analytic evidence of low convergence between implicit and explicit measures of the needs for achievement, affiliation, and power. *Frontiers in Psychology* 5: 826.
- Komar, S., Komar, J. A., Robie, C., & Taggar, S. (2010). Speeding personality measures to reduce faking: A self-regulatory model. *Journal of Personnel Psychology* 9, 126–137.
- König, C. J., & Marcus, B. (2013). TBS-TK Rezension: „Persolog Persönlichkeits-Profil“. *Psychologische Rundschau* 64, 189–191.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 5802–5805.
- Krammer, G., Sommer, M., & Arendasy, M. E. (2017). The psychometric costs of applicants' faking: Examining measurement invariance and retest correlations across response conditions. *Journal of Personality Assessment* 99, 510–523.
- Kreuzpointner, L., Lukesch, H., & Horn, W. (2013). *LPS-2: Leistungsprüfsystem 2*. Göttingen: Hogrefe.
- Kroll, E., & Ziegler, M. (2016). Discrimination due to ethnicity and gender: How susceptible are video-based job interviews? *International Journal of Selection and Assessment* 24, 161–171.
- Krumm, S., & Schmidt-Atzert, L. (2009). *Leistungstests im Personalmanagement*. Göttingen: Hogrefe.
- Krumm, S., Schmidt-Atzert, L., & Eschert, S. (2008). Investigating the structure of attention: How do test characteristics of paper-pencil sustained attention tests influence their relationship with other attention tests? *European Journal of Psychological Assessment* 24, 108–116.
- Krumm, S., Stenzel, N. M., & Pauls, C. A. (2015). Diagnostische Interviews. In G. Stemmler, & J. Margraf-Stiksrud (Hrsg.), *Lehrbuch Psychologische Diagnostik* (S. 77–155). Bern: Huber.
- Krumm, S., Schäpers, P., & Göbel, A. (2016). Motive arousal without pictures? An experimental validation of a hybrid implicit motive test. *Journal of Personality Assessment* 98, 514–522.
- Kubinger, K. D. (2006). Ein Update der Definition von Objektiven Persönlichkeitstests: Experimentalpsychologische Verhaltensdiagnostik. In T. M. Ortner, R. T. Proyer, & K. D. Kubinger (Hrsg.), *Objektive Tests in der Persönlichkeitsforschung* (S. 38–52). Bern: Huber.
- Kubinger, K. D., & Ebenhöh, J. (1996). *Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität. Test: Software und Manual*. Frankfurt am Main: Swets.
- Kubinger, K. D., & Hagenmüller, B. (2019). *AID-G: Gruppentest zur Erfassung der Intelligenz auf Basis des AID*. Göttingen: Hogrefe.
- Kubinger, K. D., & Holocher-Ertl, S. (2014). *AID 3: Adaptives Intelligenz Diagnostikum 3*. Göttingen: Beltz Test Gesellschaft.
- Kubinger, K. D., & Wurst, E. (1995). *Adaptives Intelligenz Diagnostikum (AID)*. Weinheim: Beltz.
- Kuhl, J., & Scheffer, D. (1999). *Der Operante Multi-Motive-Test (OMT): Manual*. Osnabrück: Universität Osnabrück.
- Kuhn, J.-T., Holling, H., & Freund, P. A. (2008). Begabungsdiagnostik mit dem Grundintelligenztest (CFT 20-R): Psychometrische Eigenschaften und Messäquivalenz. *Diagnostica* 54, 184–192.
- Kunz, B. (2015). Konzentration im Alltag: Validität von Konzentrationstests für die Bewältigung alltagsnaher Aufgaben. [Unveröffentlichte Diplomarbeit]. Marburg: Philipps-Universität.
- Kuschel, A., Kamp-Becker, I., & Ständer, D. (2017). TBS-TK Rezension: „Kaufman Assessment Battery for Children-2 (KABC-II)“. *Psychologische Rundschau* 68, 321–323.
- Kutcher, E. J., & Bragger, J. D. (2004). Selection interviews of overweight job applicants: Can structure reduce the bias? *Journal of Applied Social Psychology* 34, 1993–2022.
- Lang, J. W. B. (2014). A dynamic Thurstonian item response theory of motive expression in the picture story exercise: Solving the internal consistency paradox of the PSE. *Psychological Review* 121, 481–500.
- Lang, J. W. B., Kersting, M., Hülsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology* 63, 595–640.
- Langan-Fox, J., & Grant, S. (2006). The Thematic Apperception Test: Toward a standard measure of the big three motives. *Journal of Personality Assessment* 87, 277–291.

Diagnostische Verfahren

- Langfeldt, H.-P., & Tent, L. (1999). *Pädagogisch-psychologische Diagnostik* (Bd. 2). Göttingen: Hogrefe.
- Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. (1981). *STAI: Das State-Trait-Angstinventar*. Weinheim: Beltz Test.
- Lehrl, S. (1977). *MWT-B: Mehrfachwahl-Wortschatz-Tests Form B*. Balingen: Pitta.
- von Lenz, C. A. (2013). *Der Fürst des Nicola Machiavell, erste deutsche Übersetzung, 1692*. Sandendorf-Brehna: Renneritz.
- Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology* 92, 1638–1656.
- Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology* 67, 241–293.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R: Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.
- Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). *IST-Screening: Intelligenz-Struktur-Test – Screening*. Göttingen: Hogrefe.
- Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology* 92, 1672–1682.
- Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest* 1, 27–66.
- Lindzey, G. (1959). On the classification of projective techniques. *Psychological Bulletin* 56, 158–168.
- Lohbeck, A., & Petermann, F. (2019). *FLM 3–6 R: Fragebogen zur Leistungsmotivation für Schülerinnen und Schüler der 3. bis 6. Klasse – Revision*. Göttingen: Hogrefe.
- Lück, H. E., & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-MC). *Diagnostica* 15, 134–141.
- Luria, A. R. (1970). The functional organization of the brain. *Scientific American* 222, 66–78.
- McDaniel, M. A., Whetzel, D. L., Schmitt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology* 79, 599–616.
- Macha, T., & Petermann, F. (2015). Bayley Scales of Infant and Toddler Development, Third Edition – Deutsche Fassung. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie* 63, 1–5.
- Macha, T., & Petermann, F. (2017). *FREDI 0–3: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 49, 50–56.
- Mähler, C. (2017). Replik auf die Testbesprechung von Thorsten Macha und Franz Petermann zu „Frühkindliche Entwicklungsdiagnostik für Kinder von 0 bis 3 Jahren (FREDI 0–3)“. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 49, 50–55.
- Mähler, C., Cartschau, F., & Rohleder, K. (2016). *FREDI 0–3: Frühkindliches Entwicklungsdiagnostikum für Kinder von 0–3 Jahren*. Göttingen: Hogrefe.
- Marchese, M. C., & Muchinski, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment* 1, 18–26.
- Marcus, B. (2003). Das Wunder sozialer Erwünschtheit in der Personalauswahl. *Zeitschrift für Personalauswahl* 2, 129–132.
- Marcus, B. (2004). Rezension der 2. Auflage des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) von R. Hossiep und M. Paschen. *Zeitschrift für Arbeits- und Organisationspsychologie* 48, 79–86.
- Mariacher, H., & Neubauer, A. (2005). *PAI 30: Test zur Praktischen Alltagsintelligenz*. Göttingen: Hogrefe.
- Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin* 137, 856–879.
- Marschner, G. (1980). *Revisions-Test: Ein allgemeiner Leistungstest zur Untersuchung anhaltender Konzentration bei geistiger Tempoarbeit. Handanweisung (Teil II). Revisions-Test (Form S) – Parallelform zum Rev.T-A nach Dr. B. Stender*. Göttingen: Hogrefe.
- Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences* 32, 247–256.
- Martinez Méndez, R., Schneider, W., & Hasselhorn, M. (2015). *DERET 5-6+: Deutscher Rechtschreibtest für fünfte und sechste Klassen*. Göttingen: Hogrefe.
- Mattlar, C.-E. (2004). The Rorschach Comprehensive System is reliable, valid, and cost-effective. *Rorschachiana* 26, 158–186.

- Maurer, T. J., Solamon, J. M., & Lippstreu, M. (2008). How does coaching interviewees affect the validity of a structured interview? *Journal of Organizational Behavior* 29, 355–371.
- McCarthy, J. M., Van Iddekinge, C. H., Lievens, F., Kung, M.-C., Sinar, E. F., & Campion, M. A. (2013). Do candidate reactions relate to job performance or affect criterion-related validity? A multistudy investigation of relations among reactions, selection test scores, and job performance. *Journal of Applied Psychology* 98, 701–719.
- McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review* 96, 690–702.
- McCrae, R. R., Costa, P. T. J., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO Personality Inventory. *Journal of Personality Assessment* 84, 261–270.
- McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment* 26, 207–214.
- McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of cognitive abilities. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York: Guilford Press.
- Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science* 26, 184–190.
- Mehl, M. R., & Conner, T. S. (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.
- Merten, T. (2000). Die Computerversion d2 und die Frage der Transferäquivalenz. *Psychologische Beiträge* 42, 572–589.
- Meuwly, N., Schoebi, D., & Bierhoff, H.-W. (2018). TBS-TK Rezension: „Fragebogen zur Partnerschaftsdiagnostik (FPD; 2. Aufl.). *Psychologische Rundschau* 69, 391–393.
- Mickley, M., & Renner, G. (2019). Auswahl, Anwendung und Interpretation deutschsprachiger Intelligenztests für Kinder und Jugendliche auf Grundlage der CHC-Theorie: Update, Erweiterung und kritische Bewertung. *Praxis der Kinderpsychologie und Kinderpsychiatrie* 68, 323–343.
- Mihura, J. L., Meyer, G. J., Dumitrescu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin* 139, 548–605.
- Mihura, J. L., Meyer, G. J., Bombel, G., & Dumitrescu, N. (2015). Standards, accuracy, and questions of bias in Rorschach meta-analyses: Reply to Wood, Garb, Nezworski, Lilienfeld, and Duke (2015). *Psychological Bulletin* 141, 250–260.
- Mihura, J. L., Meyer, G. J., Dumitrescu, N., & Bombel, G. (2016). On conducting construct validity meta-analyses for the Rorschach: A reply to Tibon Czopp and Zeligman (2016). *Journal of Personality Assessment* 98, 343–350.
- Miller, W. R., & Rollnick, S. (2015). *Motivational Interviewing: 3. Auflage des Standardwerks in Deutsch (amerik. Orig. Motivational interviewing: Helping people change (3rd ed.)*. Freiburg im Breisgau: Lambertus-Verlag.
- Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology* 55, 1–22.
- Moosbrugger, H., & Goldhammer, F. (2006). Aufmerksamkeits- und Konzentrationsdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 83–102). Berlin, Heidelberg: Springer.
- Moosbrugger, H., & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test II* (2. Aufl.). Bern: Huber.
- Moosbrugger, H., & Oehlschlägel, J. (1996). *FAIR: Frankfurter Aufmerksamkeits-Inventar*. Bern: Huber.
- Moosbrugger, H., & Oehlschlägel, J. (2011). *FAIR-2: Frankfurter Aufmerksamkeits-Inventar 2* (2. Aufl.). Göttingen: Hogrefe.
- Morrison, J. (2014). *The first interview* (4th ed.). New York, NY: Guilford Press.
- Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars. *Diagnostica* 60, 86–97.
- Muck, P. M. (2004). Rezension des "NEO-Persönlichkeitssinventar nach Costa und McCrae (NEO-PI-R)" von F. Ostendorf und A. Angleitner. *Zeitschrift für Arbeits- und Organisationspsychologie* 48, 203–210.
- Mummendey, H. D., & Grau, I. (2014). *Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeit-, Einstellungs- und Selbstkonzeptforschung* (6. Aufl.). Göttingen: Hogrefe.
- Murray, H. (1936). *Thematic apperception test*. New York: Grune & Stratton.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Murray, H. A. (1991). *TAT: Thematic Apperception Test* (3. Aufl.). Göttingen: Hogrefe.

Diagnostische Verfahren

- Musewicz, J., Marczyk, G., Knauss, L., & York, D. (2009). Current assessment practice, personality measurement, and Rorschach usage by psychologists. *Journal of Personality Assessment* 91, 453–461.
- Nai, A., Martínez i Coma, F., & Maier, J. (2019). Donald Trump, populism, and the age of extremes: Comparing the personality traits and campaigning styles of Trump and other leaders worldwide. *Presidential Studies Quarterly* 49, 609–643.
- Nauels, H.-U., & Klieme, E. (1994). Wie hat sich das „besondere Auswahlverfahren“ bewährt? Prüfungsleistungen und Erfolgsraten von Medizinstudenten, die nach verschiedenen Kriterien zugelassen worden sind. In G. Trost (Hrsg.), *Tests für Medizinische Studiengänge (TMS): Studien zur Evaluation (18. Arbeitsbericht)* (S. 138–152). Bonn: Institut für Test- und Begabungsforschung.
- Nell, V., Bretz, J., & Sniehotta, F. F. (2004). *KT 3–4 R: Konzentrationstest für 3. und 4. Klassen (revidierte Fassung)*. Göttingen: Hogrefe.
- Nimax, C. (2012). Aufmerksamkeit im Alltag: Validierung von Videoarbeitsproben an Aufmerksamkeitsleitungen im Feld. [Unveröffentlichte Diplomarbeit]. Marburg: Philipps-Universität.
- O'Boyle, E. H., Forsyth, D. R., Banks, G. C., & McDaniel, M. A. (2012). A meta-analysis of the Dark Triad and work behavior: A social exchange perspective. *Journal of Applied Psychology* 97, 557–579.
- Oden, M. H. (1968). The fulfillment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs* 77, 3–93.
- Okulicz-Kozaryn, M., Banse, R., Kluck, M.-L., & Schubert, W. (2015). Dokumentation im Explorationsgespräch, Tonband- und Videomitschnitt. In A. Patermann, W. Schubert, & W. Gräw (Hrsg.), *Handbuch des Fahreignungsrechts* (S. 378–400). Bonn: Kirschbaum.
- Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology* 60, 995–1027.
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology* 81, 660–679.
- Ortner, T. M., & Sokolowski, K. (2008). TBS-TK. Rezension: Objektiver Leistungsmotivations-Test (OLMT). *Report Psychologie* 6, 305–306.
- Ortner, T. M., & Proyer, R. T. (2015). Objective personality tests. In T. M. Ortner, & F. J. van de Vijver (Eds.), *Psychological Assessment – Science and Practice: Behavior-based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains* (Vol. 1, pp. 133–149). Göttingen: Hogrefe.
- Ortner, T. M., Proyer, R. T., & Kubinger, K. D. (2006). *Theorie und Praxis Objektiver Persönlichkeitstests*. Bern: Huber.
- Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R., & Litzenberger, M. (2007). BAcO-D: Belastbarkeits-Assessment: Computerisierte Objektive Persönlichkeits-Testbatterie. Frankfurt: Harcourt.
- Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung*. Göttingen: Hogrefe.
- Oswald, W. D., & Hagen, B. (1997). Test d2: Aufmerksamkeits-Belastungs-Test (Rezension). *Zeitschrift für Differentielle und Diagnostische Psychologie* 18, 87–89.
- Oswald, W. D., & Roth, E. (1987). *ZVT: Der Zahlen-Verbindungs-Test. Ein sprachfreier Intelligenz-Test zur Messung der „kognitiven Leistungsgeschwindigkeit“* (2. Aufl.). Göttingen: Hogrefe.
- Pang, J. S. (2010). Content coding methods in implicit motive assessment: Standards of measurement and best practices for the picture story exercise. In O. C. Schultheiss, & J. C. Brunstein (Eds.), *Implicit motives* (vol. 1, pp. 119–151). New York, NY: Oxford University Press.
- Pargent, F., Hilbert, S., Eichhorn, K., & Büchner, M. (2018). Can't make it better nor worse: An empirical study about the effectiveness of general rules of item construction on psychometric properties. *European Journal of Psychological Assessment* 35, 891–899.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology* 46, 598–609.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. R. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
- Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality* 36, 556–563.
- Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences* 37, 1137–1151.
- Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the big five? Plenty! *Journal of Personality* 68, 821–835.

- Pawlak, K. (2006a). Objektive Tests in der Persönlichkeitsforschung. In T. M. Ortner, R. T. Proyer, & K. D. Kubinger (Hrsg.), *Theorie und Praxis Objektiver Persönlichkeitstests* (S. 16–23). Bern: Huber.
- Pawlak, K. (Hrsg.). (2006b). *Handbuch Psychologie: Wissenschaft – Anwendung – Berufsfelder*. Berlin, Heidelberg: Springer.
- Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology* 9, 743.
- Pearson Clinical Assessment Deutschland. (2019). *Raven's Progressive Matrices 2, Clinical Edition (Raven's 2)*. Deutsche Fassung. Frankfurt am Main: Pearson Clinical & Talent Assessment Deutschland.
- Petermann, F., & Daseking, M. (2015). *Diagnostische Erhebungsverfahren*. Göttingen: Hogrefe.
- Petermann, F., & Macha, T. (2015). *ET 6-6 R: Entwicklungstest für Kinder von 6 Monaten bis 6 Jahren – Revision* (2. Aufl.). Frankfurt am Main: Pearson.
- Petermann, F., Metz, D., & Fröhlich, L. P. (2010). *SET 5–10: Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren*. Göttingen: Hogrefe.
- Petermann, F., Melzer, J., & Rißling, J.-K. (2016). *Sprachdiagnostik im Kindesalter*. Göttingen: Hogrefe.
- Petermann, F., & Winkel, S. (2015). *FLM 7–13: Fragebogen zur Leistungsmotivation für Schüler der 7. bis 13 Klasse* (2. Aufl.). Göttingen: Hogrefe.
- Post, A., Gilljam, H., Bremberg, S., & Galanti, M. R. (2008). Maternal smoking during pregnancy: A comparison between concurrent and retrospective self-reports. *Paediatric and Perinatal Epidemiology* 22, 155–161.
- Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin* 100, 67–77.
- PRECIRE. (2016). *JobFit: Manual zum Testverfahren. Informationen zu theoretischen Hintergründen, zur Testkonstruktion, den Gütekriterien, Durchführung, Auswertung und Interpretation*. Aachen: PRECIE Technologies.
- Poyer, R. T., & Häusler, J. (2008). *MOI: Multimethodische Objektive Interessentestbatterie*. Mödling: Schuhfried.
- Querido, J., Eyberg, S., Kanfer, R., & Krahn, G. (2001). The process of the clinical child assessment interview. In C. E. Walker, & M. C. Roberts (Eds.), *Handbook of clinical child psychology* (3rd. ed., pp. 75–89). New York, NY: John Wiley & Sons Inc.
- Rammstedt, B., Kemper, C., Klein, M. C., Beierlein, C., & Kovaleva, A. (2013). Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: Big-Five-Inventory-10 (BFI-10). *Methoden, Daten, Analysen (mda)* 7, 233–249.
- Raven, J. C. (2009). *SPM: Raven's Standard Progressive Matrices. Deutsche Bearbeitung und Normierung* hrsg. von R. Horn (2. Aufl.). Frankfurt am Main: Pearson Clinical & Talent Assessment.
- Raven, J. C., Raven, J., & Court, J. H. (1998). *APM: Raven's Advanced Progressive Matrices. Deutsche Bearbeitung*: S. Bulheller, H. O. Häcker. Frankfurt am Main: Pearson Assessment.
- Raven, J. C., Raven, J., & Court, J. H. (2002). *CPM: Ravens Coloured Progressive Matrices. Deutsche Bearbeitung und Normierung von S. Bulheller und H. O. Häcker*. Frankfurt am Main: Pearson Assessment.
- Renner, G. (2011). Grob, A., Meyer, C.S., & Hagemann-von Arx, P.: Intelligence and Development Scales (IDS). Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren. *Praxis der Kinderpsychologie und Kinderpsychiatrie* 60, 481–494.
- Renner, G. (2019). Neuere Testverfahren: Grob, A., Hagemann-von Arx, P. (2018). IDS2. Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche. *Praxis der Kinderpsychologie und Kinderpsychiatrie* 68, 655–670.
- Renner, G., & Renner, K.-H. (2015). Kubinger, K. D & Holocher-Ertl, S. (2014). AID 3: Adaptives Intelligenz Diagnostikum 3. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 47, 173–176.
- Renner, G., & Schroeder, A. (2018). *Testinformation zur Wechsler Intelligence Scale–Fifth Edition (WISC-V)*. Dia-Inform Verfahrensinformationen 003–01. Ludwigsburg: Pädagogische Hochschule Ludwigsburg.
- Renziehausen, A. (2003). Wiener Entwicklungstest (WET) von Ursula Kastner-Koller und Pia Deimann. Ein Verfahren zur Erfassung des allgemeinen Entwicklungsstandes bei Kindern von 3 bis 6 Jahren (1. Auflage 1998; 2., überarbeitete und neu normierte Auflage 2002) [Testrezension]. *Diagnostika* 49, 140–146.
- Revers, W. J. (1958). *Der thematische Apperceptionstest*. Bern: Huber.
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin* 138, 353–387.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin* 132, 1–25.

Diagnostische Verfahren

- Roch, S. G., Woehr, D. J., Mishra, V., & Kiesczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology* 85, 370–395.
- Rohrmann, S., & Spinath, F. M. (2011). TBS-TK Rezension: "FPI-R. Freiburger Persönlichkeitssinventar". *Psychologische Rundschau* 62, 268–270.
- Roick, T., Göltz, D., & Hasselhorn, M. (2018). *DEMAT 3+: Deutscher Mathematiktest für dritte Klassen* (2. Aufl.). Göttingen: Hogrefe.
- Rorschach, H. (1949). *Psychodiagnostik. Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments* (4. Aufl.). Bern: Huber.
- Rosenthal, R., Hiller, J. B., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (2001). Meta-analytic methods, the Rorschach, and the MMPI. *Psychological Assessment* 13, 449–451.
- Roth, M., & Herzberg, P. Y. (2008). Psychodiagnostik in der Praxis: State of the Art? *Klinische Diagnostik und Evaluation* 1, 5–18.
- Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology* 11, 299–324.
- Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology* 23, 3–30.
- Schaarschmidt, U., & Fischer, A. W. (2008). *AVEM: Arbeitsbezogenes Verhaltens- und Erlebensmuster* (3. Aufl.). Frankfurt am Main: Pearson.
- Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence* 67, 44–66.
- Schellig, D., Drechsler, R., Heinemann, D., & Sturm, W. (Hrsg.). (2009). *Handbuch neuropsychologischer Testverfahren: Aufmerksamkeit, Gedächtnis, exekutive Funktionen* (Bd. 1). Göttingen: Hogrefe.
- Schellig, D., Heinemann, D., Schächtele, B., & Sturm, W. (Hrsg.). (2018a). *Handbuch neuropsychologischer Testverfahren* (Bd. 2). Göttingen: Hogrefe.
- Schellig, D., Heinemann, D., Schächtele, B., & Sturm, W. (Hrsg.). (2018b). *Handbuch neuropsychologischer Testverfahren* (Bd. 3). Göttingen: Hogrefe.
- Schermuly, C. C., Schröder, T., Nachtwei, J., & Scholl, W. (2010). Das Instrument zur Kodierung von Diskussionen (IKD). *Zeitschrift für Arbeits- und Organisationspsychologie* 54, 149–170.
- Schmalt, H. D., Sokolowski, K., & Langens, T. A. (2000). *MMG: Das Multi-Motiv-Gitter für Anschluss, Leistung und Macht*. Frankfurt am Main: Pearson Clinical & Talent Assessment.
- Schmidt, L. R. (2006). Objektive Persönlichkeitstests in der Tradition Cattells: Forschungslinien und Relativierungen. In T. M. Ortner, R. T. Proyer, & K. D. Kubinger (Hrsg.), *Theorie und Praxis Objektiver Persönlichkeitstests* (S. 24–37). Bern: Huber.
- Schmidt, J. U., & König, F. (1986). Untersuchungen zur Validität der revidierten Form des Freiburger Persönlichkeitssinventars (FPI-R). *Diagnostica* 3, 197–208.
- Schmidt-Atzert, L. (2001). Rezension des „Leistungsmotivationsinventar (LMI)“ von H. Schuler und M. Prochaska. *Zeitschrift für Arbeits- und Organisationspsychologie* 45, 142–145.
- Schmidt-Atzert, L. (2005). Prädiktion von Studienerfolg bei Psychologiestudenten. *Psychologische Rundschau* 56, 131–133.
- Schmidt-Atzert, L. (2006). Erwachsenendiagnostik. In K. Pawlik (Hrsg.), *Handbuch Psychologie: Wissenschaft, Anwendung, Berufsfelder* (S. 599–612). Berlin, Heidelberg: Springer.
- Schmidt-Atzert, L. (2007). *Objektiver Leistungsmotivations-Test OLMT (unter Mitarbeit von Markus Sommer, Markus Bühner und Astrid Jurecka). Software und Manual* (2. Aufl.). Mödling: Schuhfried.
- Schmidt-Atzert, L. (2009). Verbale Daten: Fragebogenverfahren. In J. H. Otto & V. Brandstätter-Morawietz (Hrsg.), *Handbuch der Allgemeinen Psychologie – Motivation und Emotion* (S. 532–539). Göttingen: Hogrefe.
- Schmidt-Atzert, L., & Brickenkamp, R. (2017). *d2-R: Elektronische Fassung des Aufmerksamkeits- und Konzentrationstests d2-R*. Göttingen: Hogrefe.
- Schmidt-Atzert, L., & Deter, B. (1993). Intelligenz und Ausbildungserfolg: Eine Untersuchung zur prognostischen Validität des I-S-T 70. *Zeitschrift für Arbeits- und Organisationspsychologie* 37, 52–63.
- Schmidt-Atzert, L., & Funsch, K. (2021). *KoKi – Konzentrationstest für Kinder*. Göttingen: Hogrefe. [Im Druck].
- Schmidt-Atzert, L., & Rauch, W. (2008). TBS-TK Rezension: "Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). 2., erweiterte und überarbeitete Auflage". *Report Psychologie* 33, 303–304.
- Schmidt-Atzert, L., Hommers, W., & Heß, M. (1995). Der I-S-T 70: Eine Analyse und Neubewertung. *Diagnostica* 41, 108–130.
- Schmidt-Atzert, L., Büttner, G., & Bühner, M. (2004a). Theoretische Aspekte von Aufmerksamkeits-/Konzentrationsdiagnostik. In G. Büttner & L. Schmidt-Atzert (Hrsg.), *Diagnostik von Konzentration und Aufmerksamkeit* (S. 3–22). Göttingen: Hogrefe.
- Schmidt-Atzert, L., Bühner, M., Rischen, S., & Warkentin, V. (2004b). Erkennen von Simulation und Dissimulation im Test d2. *Diagnostica* 50, 124–133.

- Schmidt-Atzert, L., Bühner, M., & Enders, P. (2006). Messen Konzentrationstests Konzentration? Eine Analyse von Konzentrationstestleistungen. *Diagnostica* 52, 33–44.
- Schmidt-Atzert, L., Krumm, S., & Bühner, M. (2008). Aufmerksamkeitsdiagnostik: Ableitung eines Strukturmodells und systematische Einordnung von Tests. *Zeitschrift für Neuropsychologie* 19, 59–82.
- Schmidt-Atzert, L., Künecke, J., & Zimmermann, J. (2019). TBS-DTK Rezension: „PRECIRE JobFit“. *Psychologische Rundschau* 70, 299–301.
- Schmidt-Atzert, L., Stemmler, G., & Peper, M. (2014). *Emotionspsychologie: Ein Lehrbuch* (2. Aufl.). Stuttgart: Kohlhammer.
- Schmukle, S. C., & Egloff, B. (2011). Indirekte Verfahren zur Erfassung von Persönlichkeit („Objektive Persönlichkeitstests“). In L. Hornke, M. Amelang, & M. Kersting (Hrsg.), *Persönlichkeitssdiagnostik* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 4, S. 73–120). Göttingen: Hogrefe.
- Schneewind, K. A., & Graf, J. (1998). *16 PF-R: 16-Persönlichkeit-Faktoren-Test, revidierte Fassung*. Göttingen: Hogrefe.
- Schneibel, G. (2009). Psychologen: Wikipedia gefährdet die Wirkung des Rorschach-Tests. *Die Welt*. Artikel vom 12. August 2009. ► https://www.welt.de/welt_print/wissen/article4304160/Psychologen-Wikipedia-gefaehrdet-die-Wirkung-des-Rorschach-Tests.html. Zugegriffen: 07. Mai 2020.
- Schönbrodt, F. D., & Gerstenberg, F. X. R. (2012). An IRT analysis of motive questionnaires: The Unified Motive Scales. *Journal of Research in Personality*, 46(6), 725–742.
- Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. *Diagnostica* 41, 3–20.
- Schuhfried, G. (2020a). COG: Cognitron. Mödling: Schuhfried. ► <https://www.schuhfried.at/test/COG>. Zugegriffen: 26. März 2020.
- Schuhfried, G. (2020b). RT: Reaktionstest. Mödling: Schuhfried. ► <https://www.schuhfried.at/test/RT>. Zugegriffen: 26. März 2020.
- Schuler, H. (1992). Das Multimodale Einstellungsinterview. *Diagnostica* 38, 281–300.
- Schuler, H., & Prochaska, M. (2001). *LMI: Leistungsmotivationsinventar*. Göttingen: Hogrefe.
- Schüler, J., Brandstätter, V., Wegner, M., & Baumann, N. (2015). Testing the convergent and discriminant validity of three implicit motive measures: PSE, OMT, and MMG. *Motivation and Emotion* 39, 839–857.
- Schultheiss, O. C., & Brunstein, J. C. (2001). Assessment of implicit motives with a research version of the TAT: Picture profiles, gender differences, and relations to other personality measures. *Journal of Personality Assessment* 77, 71–86.
- Schultheiss, O. C., Liening, S. H., & Schad, D. (2008). The reliability of a Picture Story Exercise measure of implicit motives: Estimates of internal consistency, retest reliability, and ipsative stability. *Journal of Research in Personality* 42, 1560–1571.
- Schulz, R., & Weiß, S. (2018). Forschungsbericht: BIP-6F-Fi-R2, Gütekriterien: Objektivität – Reliabilität – Validität. Ruhr-Universität Bochum. ► https://www.testentwicklung.de/mam/forschungsbericht_bip-6f-fi-r2.pdf. Zugegriffen: 12. Juni 2020.
- Schwarz, N., & Sudman, S. (Hrsg.). (1994). *Autobiographical memory and the validity of retrospective reports*. New York, NY: Springer.
- Schwarzinger, D., & Schuler, H. (2016). *TOP: Dark triad of personality at work*. Göttingen: Hogrefe.
- Searls, D. (2017). *The inkblots: Hermann Rorschach, his iconic test, and the power of seeing*. Portland, OR: Broadway Books.
- Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research* 4, 27–41.
- Seitz, W., & Rausche, A. (2019). *PFK 9–14: Persönlichkeitsfragebogen für Kinder zwischen 9 und 14 Jahren* (5. Aufl.). Göttingen: Hogrefe.
- Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence* 53, 72–80.
- Sommer, M., Arendasy, M. E., & Schützhofer, B. (2017). Psychometric costs of retaking driving-related cognitive ability tests. *Transportation Research Part F: Traffic Psychology and Behaviour* 44, 105–119.
- Sommers-Flanagan, J., & Sommers-Flanagan, R. (2017). *Clinical interviewing* (6th ed.). Hoboken, NJ: John Wiley & Sons Inc.
- Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin* 112, 140–154.
- Spearman (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology* 15, 201–292.
- Spielberger, C. D. (1966). Theory and research on anxiety. In C. D. Spielberger (Ed.), *Anxiety and behavior* (pp. 3–22). New York, NY: Academic Press.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1968). *State-Trait Anxiety Inventory (STAII): Test Manual for Form X*. Palo Alto: Consulting Psychologists Press.
- Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2012). *SELLMO: Skalen zur Erfassung der Lern- und Leistungsmotivation* (2. Aufl.). Göttingen: Hogrefe.

Diagnostische Verfahren

- Spreen, O., Hathaway, S. R., McKinley, J. C., & Sundberg, N. D. (1963). *MMPI Saarbrücken: Handbuch zur deutschen Ausgabe des Minnesota Multiphasic Personality Inventory*. Bern: Huber.
- Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Büchner, M. (2019). Personality research and assessment in the era of machine learning. *PsyArXiv (psychology archive) Preprint*. ► <https://doi.org/10.31234/osf.io/efnj8>.
- Steck, P. (1996). Die Prüfung der Dauerkonzentration mit einer Apparateversion des Pauli-Tests. *Diagnostica* 42, 332–351.
- Steck, P. (1997). Psychologische Testverfahren in der Praxis: Ergebnisse einer Umfrage unter Test-anwendern. *Diagnostica* 43, 267–284.
- Steinmayr, R., & Amelang, M. (2006). Erste Untersuchungen zur Kriteriums-Validität des I-S-T 2000 R an Erwachsenen beiderlei Geschlechts. *Diagnostica* 52, 181–188.
- Steinmayr, R., Schütz, A., Herte, J., & Schröder-Abé, M. (2011). *MSCEIT: Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz. Deutschsprachige Adaptation des Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) von John D. Mayer, Peter Salovey & David R. Caruso*. Bern: Huber.
- Steller, M. (2015). *Nichts als die Wahrheit? Warum jeder unschuldig verurteilt werden kann*. München: Heyne.
- Stemmler, G., & Margraf-Stiksrud, J. (2015). Verhaltensbeobachtung. In G. Stemmler, & J. Margraf-Stiksrud (Hrsg.), *Lehrbuch Psychologische Diagnostik* (S. 13–76). Bern: Huber.
- Stemmler, M., & Kornhuber, J. (2018). *Demenzdiagnostik*. Göttingen: Hogrefe.
- Stemmler, G., Hagemann, D., Amelang, M., & Spinath, F. M. (2016). *Differentielle Psychologie und Persönlichkeitsforschung* (8. Aufl.). Stuttgart: Kohlhammer.
- Stock, C., & Schneider, W. (2008a). *DERET 1-2+: Deutscher Rechtschreibtest für das erste und zweite Schuljahr*. Göttingen: Hogrefe.
- Stock, C., & Schneider, W. (2008b). *DERET 3-4+: Deutscher Rechtschreibtest für das dritte und vierte Schuljahr*. Göttingen: Hogrefe.
- Strobel, A., Franke-Bartholdt, L., Püttner, I., & Kersting, M. (2018). Eignungsinterviews/direkte mündliche Befragung. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 65–93). Berlin, Heidelberg: Springer.
- Stumpf, H., Angleitner, A., Wiek, T., Jackson, D. N., & Beloch-Till, H. (Hrsg.). (1985). *Deutsche Personality Research Form (PRF)*. Göttingen: Hogrefe.
- Sturm, W. (2008). *WAF: Wahrnehmungs- und Aufmerksamkeitsfunktionen*. Mödling: Schuhfried.
- Sturm, W., Willmes, K., & Horn, W. (2015). *LPS 50+: Leistungsprüfsystem für 50- bis 90-Jährige* (2. Aufl.). Göttingen: Hogrefe.
- Styck, K. M., & Walsh, S. M. (2016). Evaluating the prevalence and impact of examiner errors on the Wechsler scales of intelligence: A meta-analysis. *Psychological Assessment* 28, 3–17.
- Süß, H. M., & Beauducel, A. (2011). Intelligenztests und ihre Bezüge zu Intelligenztheorien. In L. F. Hornke, M. Amelang, & M. Kersting (Hrsg.), *Leistungs-, Intelligenz- und Verhaltensdiagnostik* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 3, S. 97–234). Göttingen: Hogrefe.
- Sutherland, S. (1992). *Irrationality: Why we don't think straight!* New Brunswick, NJ: Rutgers University Press.
- Szagun, G., Stumper, B., & Schramm, S. A. (2009). *FRAKIS: Fragebogen zur frühkindlichen Sprachentwicklung*. Frankfurt: Pearson.
- Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology* 75, 277–294.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin* 133, 859–883.
- Tross, S. A., & Maurer, T. J. (2008). The effect of coaching interviewees on subsequent interview performance in structured experience-based interviews. *Journal of Occupational and Organizational Psychology* 81, 589–605.
- Trost, G. (1996). Interview. In K. Pawlik (Hrsg.), *Grundlagen und Methoden der Differentiellen Psychologie* (Enzyklopädie der Psychologie, Serie Differentielle Psychologie und Persönlichkeitsforschung, Bd. 1, S. 463–505). Göttingen: Hogrefe.
- Tröster, H., Flender, J., & Reineke, D. (2005). Dortmunder Entwicklungsscreening für den Kindergarten (DESK 3-6). *Kindheit und Entwicklung* 14, 140–149.
- Tröster, H., Flender, J., Reineke, D., & Wolf, S. M. (2016). *DESK 3-6 R: Dortmunder Entwicklungsscreening für den Kindergarten – Revision*. Göttingen: Hogrefe.
- Visser, B. A., Book, A. S., & Volk, A. A. (2017). Is Hillary dishonest and Donald narcissistic? A HEXACO analysis of the presidential candidates' public personas. *Personality and Individual Differences* 106, 281–286.
- Waller, N. G., & Waldman, I. D. (1990). A reexamination of the WAIS-R factor structure. *Psychological Assessment* 2, 139–144.

- Wechsler, D. (2012). *WAIS-IV: Wechsler Adult Intelligence Scale – Fourth Edition. Deutsche Fassung von F. Petermann*. Frankfurt am Main: Pearson.
- Wechsler, D. (2017). *WISC-V: Wechsler Intelligence Scale for Children – Fifth Edition. Deutsche Bearbeitung von F. Petermann*. Frankfurt am Main: Pearson.
- Wechsler, D. (2018). *Wechsler Preschool and Primary Scale of Intelligence – Forth Edition (WPPSI-IV)*. Deutsche Bearbeitung durch F. Peterman u. M. Daseking. Frankfurt am Main: Pearson.
- Weiner, I. B. (2018). Society for Personality Assessment/Journal of Personality Assessment: A history. *Journal of Personality Assessment* 100, 2–15.
- Weiβ, R. H. (1997). Replik zur Rezension des CFT 20. *Zeitschrift für Differentielle und Diagnostische Psychologie* 18, 56–61.
- Weiβ, R. H. (2006). *CFT 20-R mit WS/ZF-R: Grundintelligenztest Skala 2 – Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest – Revision (WS/ZF-R)*. Göttingen: Hogrefe.
- Weiβ, R. H. (2019). *CFT 20-R mit WS/ZF-R: Grundintelligenztest Skala 2 – Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest – Revision (WS/ZFR)* (2. Aufl.). Göttingen: Hogrefe.
- Weiβ, R. H., & Osterland, J. (2012). *CFT 1-R: Grundintelligenztest Skala 1 – Revision*. Göttingen: Hogrefe.
- Westhoff, K. (1995). Aufmerksamkeit und Konzentration. In M. Amelang (Hrsg.), *Verhaltens- und Leistungsunterschiede* (Enzyklopädie der Psychologie, Serie Differentielle Psychologie und Persönlichkeitsforschung, Bd. 2, S. 375–402). Göttingen: Hogrefe.
- Westhoff, K., & Dewald, D. (1990). Effekte der Übung in der Bearbeitung von Konzentrations- tests. *Diagnostica* 36, 1–15.
- Westhoff, K., & Kluck, M. L. (2008). *Psychologische Gutachten schreiben und beurteilen* (5. Aufl.). Berlin, Heidelberg: Springer.
- Whippman, R. (2017). Tell Me What You See: The Rorschach Test and Its Inventor. *The New York Times*. Artikel vom 14. März 2017. ► <https://www.nytimes.com/2017/03/14/books/review/the-inkblots-hermann-rorschach-biography-damion-searls.html>. Zugegriffen: 07. Mai 2020.
- Wildman, R. W., & Wildman, R. W. I. (1975). An investigation into the comparative validity of several diagnostic tests and test batteries. *Journal of Clinical Psychology* 31, 455–458.
- Wilhelm, S. (2005). Die Erfassung von Leistungsmotivation durch Selbst- und Fremdbeurteilung. [Unveröffentlichte Diplomarbeit]. Marburg: Philipps-Universität Marburg.
- Winter, D. G. (1994). *Manual for scoring motive imagery in running text (4th ed.)*, unpublished manuscript. Ann Arbor, MI: University of Michigan.
- Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.
- Wirtz, M. A., & Kutschmann, M. (2007). Analyse der Beurteilerübereinstimmung für kategoriale Daten mittels Cohens Kappa und alternativer Maße. *Die Rehabilitation* 46, 370–377.
- Wittmann, A. J., & Holling, H. (2001). *Hochbegabtenberatung in der Praxis*. Göttingen: Hogrefe.
- Wood, J. M., Lilienfeld, S. O., Nezworski, M. T., Garb, H. N., Allen, K. H., & Wildermuth, J. L. (2010). Validity of Rorschach Inkblot scores for discriminating psychopaths from nonpsychopaths in forensic populations: A meta-analysis. *Psychological Assessment* 22, 336–349.
- Wood, J. M., Garb, H. N., Nezworski, M. T., Lilienfeld, S. O., & Duke, M. C. (2015). A second look at the validity of widely used Rorschach indices: Comment on Mihura, Meyer, Dumitrascu, and Bombel (2013). *Psychological Bulletin* 141, 236–249.
- Yeung, R. (2016). *Erfolgreich im Bewerbungsgespräch für Dummies* (2. Aufl.). Weinheim: Wiley-VCH.
- Ziegler, M., & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement* 59, 197–210.
- Ziegler, M., & Reichert, A. (2017). TBS-TK Rezension: Adaptives Intelligenz Diagnostikum 3 (AID 3). *Psychologische Rundschau* 68, 237–239.
- Ziegler, M., Schmidt-Atzert, L., Bühner, M., & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: Questionnaire, semi-projective, and objective. *Psychology Science* 49, 291–307.
- Zimmermann, P. (2020). Fremde Situation oder Fremde Situations Test (FST). In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. ► <https://m.portal.hogrefe.com/dorsch/fremde-situation-oder-fremde-situations-test-fst/>. Zugegriffen: 20. März 2020.
- Zimmermann, P., & Fimm, B. (1993). *Testbatterie zur Erfassung von Aufmerksamkeitsstörungen – Version 1.02*. Freiburg: Psytest.
- Zimmermann, P., & Fimm, B. (2017). *TAP: Testbatterie zur Aufmerksamkeitsprüfung (TAP, Version 2.3.1)*. Herzogenrath: Psytest.

Durchführung einer diagnostischen Untersuchung und Gutachtenerstellung

Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang

Inhaltsverzeichnis

- 4.1 Persönliche Voraussetzungen und ethisch verantwortliches Vorgehen – 481**
- 4.2 Auftragsannahme und Fragestellung – 483**
- 4.3 Ableitung von psychologischen Fragen – 486**
 - 4.3.1 Psychologische Fragen finden – 487
 - 4.3.2 Darstellung der psychologischen Fragen – 489
- 4.4 Auswahl der Verfahren und Untersuchungsplanung – 489**
 - 4.4.1 Geeignete diagnostische Verfahren finden – 490
 - 4.4.2 Untersuchungsplanung – 494
- 4.5 Durchführung und Auswertung diagnostischer Verfahren – 497**
 - 4.5.1 Aufklärung – 497
 - 4.5.2 Gute Arbeitsbedingungen – 502
 - 4.5.3 Standardisierung der Untersuchungsbedingung – 503
 - 4.5.4 Testauswertung – 504
 - 4.5.5 Darstellung und Interpretation der Ergebnisse – 505
- 4.6 Das psychologische Gutachten – 512**
 - 4.6.1 Der Befund – 512
 - 4.6.2 Stellungnahme – 514
 - 4.6.3 Wenn der Begutachtungsprozess nicht erfolgreich verläuft – 515
 - 4.6.4 Formale Gestaltung des Gutachtens – 516
 - 4.6.5 Beurteilung der Qualität eines Gutachtens – 520
- 4.7 Zusammenfassung – 522**
- Literatur – 524**

■ Vorbemerkungen

Die Überschrift dieses Kapitels könnte einen Fehlschluss nahelegen: Eine Psychologin oder ein Psychologe führt zunächst eine diagnostische Untersuchung durch und erstellt dann darüber ein Gutachten. Warum ist diese Interpretation falsch? Der Begriff „Gutachten“ meint nicht, dass man nur einen Untersuchungsbericht mit daran anschließenden Interpretationen verfasst. Zwar könnte man ein solches Dokument mit der Überschrift „Gutachten“ versehen. Aber: Über eine unsachgemäß durchgeführte diagnostische Untersuchung kann man kein Gutachten schreiben, zumindest keines, das den Qualitätsanforderungen an ein Gutachten genügt! In den „Qualitätsstandards für psychologische Gutachten“ (Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen 2017, S. 8) wird das so formuliert:

- » Fehler auf der ersten Ebene [gemeint ist das gutachterlichen Handeln] können durch eine einwandfreie Darstellung auf der zweiten Ebene [gemeint ist das schriftliche Gutachten] nicht wettgemacht werden.

Die Begutachtung umfasst alle Teile des diagnostischen Prozesses

Daraus folgt, dass man sich bereits bei der Planung einer diagnostischen Untersuchung, ja sogar schon bei der Auftragsannahme, an den Qualitätsanforderungen an dem zu erstellenden Gutachten orientieren muss. Diesen Qualitätsanforderungen (sie werden in diesem Kapitel genannt und im Detail erläutert) zufolge wird jede Begutachtung als ein Prozess betrachtet, der von der Auftragsannahme über die Durchführung einer Untersuchung bis zum Bericht reicht. Die Untersuchungsdurchführung ist darin nur ein Element von vielen.

Dennoch lohnt es sich, die Planung und Durchführung einer diagnostischen Untersuchung separat zu betrachten, weil sie nicht zwangsläufig ein Gutachten nach sich zieht. In der Klinischen Psychologie können Untersuchungen beispielsweise durchgeführt werden, um eine Grundlage für eine Beratung oder Therapie zu schaffen oder etwa um eine abgeschlossene Therapie zu evaluieren. In der beruflichen Eignungsdiagnostik ist es unüblich, Gutachten zu verfassen. Allenfalls trägt einmal ein Feedback zum Assessment-Center oder einem Persönlichkeitsfragebogen (zu Unrecht) die Bezeichnung „psychologisches Gutachten“. Die Entscheidungsträger hätten keine Zeit, über viele Bewerberinnen und Bewerber erstellte Gutachten (die den Qualitätsstandards entsprechen) zu lesen. Auch würden sie die hohen Kosten und die zeitliche Verzögerung durch die Ausarbeitung von Gutachten scheuen.

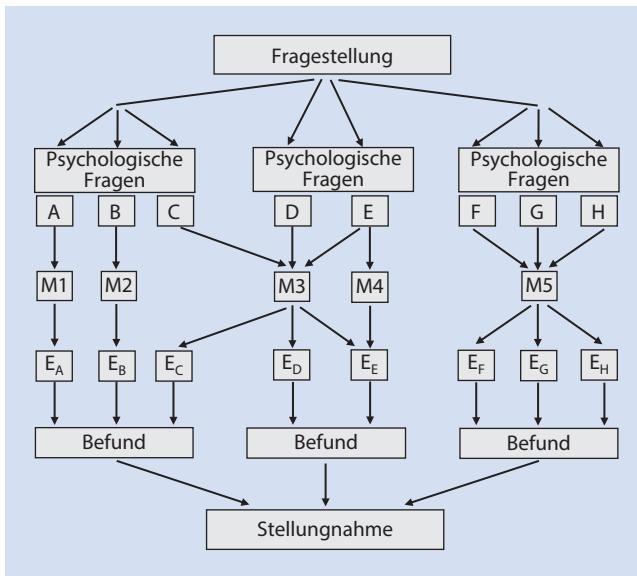
Wir gliedern dieses Kapitel nach dem in ▶ Abschn. 1.5 kurz skizzierten diagnostischen Prozess (s. u.) und gehen an geeigneter Stelle auch auf Anforderungen an ein psychologisches Gutachten ein. Das Thema „Abfassen von Gutachten“ wird dann in ▶ Abschn. 4.6.4 relativ kompakt behandelt werden.

Schritte des diagnostischen Prozesses

- Schritt 1: Auftragsannahme und Formulierung der globalen Fragestellung
- Schritt 2: Differenzierung der globalen Fragestellung in dafür infrage kommende Teilfragen (sog. „psychologische Fragen“)
- Schritt 3: Auswahl der zur Beantwortung der Teilfragen bestmöglichen diagnostischen Instrumente
- Schritt 4: Durchführung und Auswertung der diagnostischen Instrumente
- Schritt 5: Integration der Ergebnisse zur Beantwortung der Teilfragen und der globalen Fragestellung

Ablaufschema

Übersicht Für die Einordnung der notwendigen Schritte des diagnostischen Vorgehens, die in weiteren Abschnitten dieses Kapitels ausführlich erläutert werden, ist eine grafische Darstellung (Abb. 4.1) hilfreich – auch wenn sie auf den ersten Blick kompliziert aussehen mag (s. auch ▶ Abschn. 1.5).



■ Abb. 4.1 Schematische Darstellung des Begutachtungsprozesses (in Anlehnung an Schmidt-Atzert und Krumm 2007, Abb. 1, © Georg Thieme Verlag KG). Für ein beispielhaft aus gefülltes Schema s. ▶ Abschn. 1.5

Die Begutachtung beginnt mit der Auftragsannahme und damit der Fragestellung (▶ Abschn. 4.2). Dies ist die Fragestellung der Auftraggeberin oder des Auftraggebers. Um sie beantworten zu können, werden psychologische Fragen abgeleitet. Das sind Hypothesen oder Arbeitsaufträge. In der Grafik sind sie mit A bis H gekennzeichnet. Die Unterteilung in 3 Gruppen soll ausdrücken, dass die psychologischen Fragen thematisch geordnet sind (▶ Abschn. 4.3). Zur Beantwortung der psychologischen Fragen A bis H werden geeignete diagnostische Verfahren, d. h. Methoden (hier M1 bis M5) ausgewählt (▶ Abschn. 4.4). In dem Schema wird deutlich, dass nicht immer das Prinzip „eine Frage=eine Methode“ gilt. Es kommt vor, dass eine Methode zur Beantwortung von mehreren Fragen verwendet werden kann (in ■ Abb. 4.1 trifft dies auf M3 und M5 zu). Beispiele für eine solche Methode sind das diagnostische Interview und mehrdimensionale Fragebögen. Wenn wir nachfolgend von diagnostischen Verfahren sprechen, schließt dies also nicht nur Tests ein, sondern explizit auch Interviews, Aktenanalysen, Verhaltensbeobachtungen etc. Idealerweise versucht man, psychologische Fragen mit mehr als einer Methode zu beantworten. In ■ Abb. 4.1 ist dies bei Frage E der Fall, für die Methoden M3 und M4 vorgesehen sind. Nach Durchführung und Auswertung der Untersuchung (▶ Abschn. 4.5) liegt zu jeder Fragestellung ein Ergebnis (hier E_A bis E_H) vor, manchmal sind es auch mehrere Ergebnisse (zwei bei E_E). Die Ergebnisse werden nun integriert, d. h., sie werden so zusammengeführt, dass auch Übereinstimmungen und Widersprüche zwischen einzelnen Ergebnissen erkennbar werden. Dies ist der Befund, wobei auch hier wie bei den psychologischen Fragen eine geordnete Darstellung angestrebt wird. Die einzelnen Befunde werden anschließend dazu verwendet, die Fragestellung zu beantworten. Befund und Stellungnahme sind wesentliche Elemente eines Gutachtens und werden deshalb in ▶ Abschn. 4.6.1 und 4.6.2 behandelt.

Qualitätsstandards Auf internationaler Ebene und speziell in den USA wurde eine Reihe von Richtlinien und Standards zur Qualitätssicherung Psychologischer Diagnostik erarbeitet. Ein Standardwerk sind die *Standards for Educational and Psychological Testing*, die in der 5. Revision vorliegen (AERA et al. 2014). Sie wurden von mehreren amerikanischen Berufsverbänden gemeinsam ausgearbeitet und stellen (Stand: Juni 2020) die vorläufig aktuellste

Richtlinien und Standards

Version dar. Nur ein kleiner Teil befasst sind mit der Durchführung von Tests. Die 1. Auflage erschien bereits 1954 unter der Bezeichnung *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (s. Plake und Wise 2014; der Beitrag informiert über den Aufbau der „Standards“ und Veränderungen gegenüber der vorigen Auflage). Die „International Test Commission“ (kurz: ITC), eine Vereinigung von u. a. psychologischen Fachgesellschaften, Testkommissionen und Testverlagen, hat Richtlinien zur Testanwendung erstellt, die auch ins Deutsche übersetzt wurden. Diese liegen als „Internationale Richtlinien für die Testanwendung (Version 2000), deutsche Fassung“ vor (ITC 2001). Darüber hinaus gibt es weitere Standards, Richtlinien und Empfehlung (für eine Übersicht s. Leibniz-Zentrum für Psychologische Information und Dokumentation [ZPID] unter ► <https://www.psychlinker.de/category.php?cat=97>). Für das vorliegende Kapitel orientieren wir uns vor allem an den Qualitätsstandards für psychologische Gutachten (Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen 2017), die den ganzen Begutachtungsprozess betreffen, also nicht nur das Abfassen eines als „Gutachten“ bezeichneten Berichts.

Qualitätsstandards für psychologische Gutachten

Die „Qualitätsstandards für psychologische Gutachten“ (hier kurz: Qualitätsstandards) wurden am 18. Oktober 2017 vom Vorstand der Föderation Deutscher Psychologenvereinigungen verabschiedet. Sie ersetzen die „Richtlinien für die Erstellung psychologischer Gutachten“ der Föderation Deutscher Psychologenvereinigungen von 1988. Sie sind in einem langen Prozess entstanden. Die Deutsche Gesellschaft für Psychologie (DGPs) beauftragte im August 2009 eine Arbeitsgruppe, Qualitätsstandards für psychodiagnostische Gutachten auszuarbeiten und deren Konsequenzen für die Lehre an den Hochschulen aufzuzeigen. Die Arbeitsgruppe setzte sich aus je einer Fachvertreterin oder einem Fachvertreter der Fachgruppen zusammen, für die das Thema Begutachtung relevant ist: Arbeits-, Organisations- und Wirtschaftspsychologie, Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik, Klinische Psychologie und Psychotherapie, Pädagogische Psychologie, Rechtspsychologie sowie Verkehrspychologie. Die Kommission legte der Fachöffentlichkeit 1 Jahr später einen ersten Entwurf vor, der ausführlich diskutiert wurde. Die Arbeitsgruppe sprach auch weitere Expertinnen und Experten mit der Bitte an, den Text zu kommentieren und ggf. auch Änderungsvorschläge zu machen. Diese machten zahlreiche konstruktive Vorschläge, die von der Arbeitsgruppe diskutiert wurden. Ergebnis war eine Überarbeitung der ersten Version mit dem Ziel, das Gesamtanliegen noch klarer darzustellen, ohne jedoch in eine „Überregulierung“ abzugleiten. Diese Version „Qualitätsstands für psychologische-diagnostische Gutachten (Version 2.2)“ (Arbeitsgruppe „Qualitätsstandards für psychodiagnostische Gutachten“, 2011) wurde im Dezember 2011 vom Vorstand der DGPs verabschiedet.

Der Berufsverband Deutscher Psychologinnen und Psychologen (BDP), das 2. Mitglied der Föderation Deutschen Psychologenvereinigungen, legte 2012 einen von drei durch einschlägige Veröffentlichungen zum Thema Gutachten ausgewiesene Expertinnen und Experten verfassten Gegenentwurf „Standards psychologischer Gutachten“ vor. Das Diagnostik- und Testkuratorium wurde schließlich beauftragt, auf der Grundlage beider Vorschläge die nun vorliegenden „Qualitätsstandards für psychologische Gutachten“ zu erstellen.

Die Vorgeschichte lässt erkennen, dass sehr viel Fachkompetenz in die Qualitätsstandards eingeflossen ist. Die Qualitätsstandards sind als Mindestanforderungen an ein psychologisches Gutachten zu verstehen und sollen dem Schutz vor unsachgemäßen Gutachten dienen. Sie sind aber juristisch nicht verbindlich.

Wir werden diese Qualitätsstandards im Laufe dieses Kapitels auszugsweise vorstellen. Neben diesen allgemeinen Qualitätsstandards liegen von Fachgesellschaften und Verbänden eigene Standards oder Richtlinien vor, die aber nicht als Konkurrenz zu verstehen sind, sondern verfasst wurden, um auch den spezifischen Anforderungen in bestimmten Anwendungsbereichen gerecht zu werden. Zum Teil sind auch andere Berufsgruppen (Juristen/Juristinnen und Mediziner/Medizinerinnen) involviert. Zu nennen sind die Leitlinie „Neuropsychologische Begutachtung“ (Gesellschaft für Neuropsychologie et al. 2009), die „Mindestanforderungen an die Qualität von Sachverständigengutachten im Kindschaftsrecht“ (Arbeitsgruppe Familienrechtliche Gutachten 2019), die „Mindestanforderungen für Prognosegutachten“ für den forensischen Bereich (Boetticher et al. 2007) und die „Begutachtungsleitlinien zur Kraftfahreignung“ (Schubert et al. 2018).

Zusätzlich fachspezifische Leitlinien

4.1 Persönliche Voraussetzungen und ethisch verantwortliches Vorgehen

Fachliche Kompetenz Die Auswahl, Durchführung und Auswertung eines diagnostischen Verfahrens und die Interpretation der Ergebnisse erfordert fachliche Kompetenz. Fehler können zu gravierenden Konsequenzen für die untersuchte Person führen. Die Personen, die diagnostische Verfahren auswählen, durchführen, auswerten und die Ergebnisse interpretieren, sollten deshalb über eine angemessene Ausbildung verfügen und sich kontinuierlich fortbilden. Die Forderung nach fachlicher Kompetenz findet sich auch in den berufsethischen Richtlinien der beiden großen deutschen Psychologenvereinigungen BDP und DGPs (Föderation Deutscher Psychologenvereinigungen 2016).

Berufsethische Richtlinien verlangen fachliche Kompetenz

» Berufsethische Richtlinien

1. Präambel

1.2 Ethische und fachliche Grundhaltungen

Psychologinnen und Psychologen:

[...]

(15) erbringen Dienstleistungen eigenständig nur in den Tätigkeitsfeldern, für die sie durch eine wissenschaftlich fundierte Ausbildung, fachliche Fortbildung und berufliches Handeln qualifiziert sind.

[...]

(© Föderation Deutscher Psychologenvereinigungen)

In den berufsethischen Richtlinien werden explizit 4 Prinzipien der „Berufsethischen Prinzipien der europäischen Psychologenvereinigung in Europa“ anerkannt, darunter (Föderation Deutscher Psychologenvereinigungen 2016, S. 9):

» Berufsethische Richtlinien

2.2 Kompetenz

Psychologinnen und Psychologen streben danach, einen hohen Kompetenzstandard in ihrer Arbeit sicherzustellen und zu erhalten. Sie wissen um die Grenzen ihrer spezifischen Kompetenzen und ihrer Fachkenntnis. Sie bieten nur solche Dienstleistungen an und verwenden nur diejenigen Methoden, für die sie durch Ausbildung, Fortbildung oder Erfahrung qualifiziert sind.

(© Föderation Deutscher Psychologenvereinigungen)

Teils gesetzliche Anforderungen an die berufliche Qualifikation

In einigen Anwendungsfeldern, z. B. in der Personalauswahl, wird Psychologische Diagnostik inzwischen häufig von Nichtpsychologinnen und -psychologen durchgeführt. Zumindest in zwei Anwendungsbereichen hat der Gesetzgeber hingegen für eine Tätigkeit als Sachverständige oder Sachverständiger, die Gutachten erstellen, Folgendes zur Auflage gemacht: Bei familiengerichtlichen Gutachten wird als Mindestvoraussetzung eine psychologische, psychotherapeutische, kinder- und jugendpsychiatrische, psychiatrische, ärztliche, pädagogische oder sozialpädagogische Berufsqualifikation verlangt; bei einer pädagogischen oder sozialpädagogischen Berufsqualifikation ist der Erwerb ausreichender diagnostischer und analytischer Kenntnisse durch eine anerkannte Zusatzqualifikation nachzuweisen (► Abschn. 9.2). Das war dringend nötig, weil gelegentlich auch psychologische Laien familienpsychologische Gutachten für ein Gericht erstellt hatten. In der Verkehrspychologie (► Abschn. 9.3) wird in vorbildlicher Weise durch den Gesetzgeber geregelt, dass die psychologische Begutachtung von verkehrsauftällig gewordenen Kraftfahrerinnen und Kraftfahrern nur von Psychologinnen und Psychologen mit Diplom- oder gleichwertigem Masterabschluss und mindestens 2-jähriger praktischer Berufstätigkeit durchgeführt werden darf. Zusätzlich wird eine mindestens 1-jährige Praxis in der Begutachtung der Eignung von Kraftfahrerinnen und Kraftfahrern in einer Begutachtungsstelle für Fahreignung verlangt. Sie sind zu jährlichen Fortbildungen verpflichtet. Zusätzlich wird ihre Arbeit durch eine Bundesbehörde kontrolliert (Fahrerlaubnisverordnung, Anlage 14 zu § 66 Absatz 2).

Anwendung üben

Grundsätzlich müssen Personen, die eine Untersuchung durchführen, mit den diagnostischen Verfahren vertraut sein. Einige Verfahren sollten nur nach eingehender Übung durchgeführt werden – dazu zählen Interviews, systematische Verhaltensbeobachtungen, Verhaltensbeurteilungen, einige Tests wie die Wechsler-Intelligenztests und bestimmte projektive Tests. Für die Auswertung des Rorschach-Tests ist sogar ein mehrwöchiges Training erforderlich. Auch Psychologinnen und Psychologen mit großer Erfahrung bei der Untersuchungsdurchführung und -auswertung müssen sich mit Tests, die für sie neu sind, vorher gründlich vertraut machen.

Da zumindest die Durchführung und Auswertung diagnostischer Verfahren häufig delegiert werden, ist es wichtig, auch Mitwirkende angemessen zu schulen und für eine Supervision zu sorgen.

Im Rahmen der berufsbezogenen Eignungsbeurteilung nach DIN 33430 (s. auch ► Abschn. 6.4) wurden *Personenlizenzen* eingeführt, zu deren Erwerb man erfolgreich an einer wissensbasierten schriftlichen Prüfung teilnehmen muss (s. ► <https://www.din33430portal.de/din33430/din33430>). Neben einer umfassenden „Lizenz E für Eignungsdiagnostiker(innen)“ ist es möglich, eine „Lizenz BE für Beobachter(innen), die an direkten mündlichen Befragungen beteiligt sind“ und eine „Lizenz BV für Beobachter(innen), die an Verhaltensbeobachtungen und -beurteilungen beteiligt sind“ zu erwerben. Es werden entsprechende Fortbildungen angeboten, die jedoch keine formale Voraussetzung für die Lizenzprüfung darstellen. Auch wird kein besonderer Abschluss (Studium oder Berufsausbildung) vorausgesetzt. Dies ist eine von mehreren Möglichkeiten, Mitwirkende an der Eignungsdiagnostik zu qualifizieren.

Das folgende Beispiel macht deutlich, dass eine diagnostische Strategie als unzulänglich kritisiert werden kann und der Einsatz von nicht hinreichend qualifizierten Mitarbeitenden problematisch ist.

Mitwirkende schulen

Lizenzerwerb für Mitwirkende

Unqualifizierter Auswahltest (Eyde et al. 2010, S. 31 f.)

Ein amerikanisches College vergibt Stipendien an begabte Studierende. In der Vergangenheit hatten 2 Stipendiaten das Studium aufgegeben, weil sie nicht hinreichend emotional stabil waren. Das College entschied daraufhin, emotionale Stabilität als Auswahlkriterium mit aufzunehmen. Aufgrund begrenzter Ressourcen wurde eine Psychologin beauftragt, 2 Krankenschwestern des College-Krankenhauses anzulernen, Persönlichkeitstests durchzuführen und zu interpretieren. Die Psychologin trainierte die Krankenschwestern in der Durchführung und Auswertung des Minnesota Multiphasic Personality Inventory-2 (MMPI-2; ► Abschn. 3.3.3.1) und gab ihnen ihre Telefonnummer für den Fall, dass irgendwelche Schwierigkeiten auftreten sollten. Lag der T-Wert der Skala 1 (Hypochondrie), 2 (Depression), 3 (Hysterie), 6 (Paranoia), 7 (Psychastenie) oder 8 (Schizophrenie) über 65, mussten Kandidatinnen und Kandidaten wegen emotionaler Labilität abgelehnt werden. Die Antwortbögen wurden sicher aufbewahrt, nur die Entscheidung gelangte in die Bewerbungsunterlagen. Die Eltern eines abgelehnten Bewerbers erfuhren Details des Auswahlverfahrens und brachten die Angelegenheit vor das Ethikkomitee des Berufsverbands. Dieses stellte fest, dass die Psychologin nicht alle Probleme bedacht hatte, die bei einer solchen Testung ohne Supervision auftreten können: Die Krankenschwestern mussten sich aufgrund der fehlenden Diagnostikausbildung starr an die Auswertungsregeln halten. Situative Faktoren, die sich auf die Testergebnisse auswirken könnten, blieben daher unberücksichtigt. Die Entscheidungsregel berücksichtigte zudem keine Informationen zur Gültigkeit der Ergebnisse; so wurden die Ergebnisse der Validitätsskalen nicht einbezogen.

Es wäre besser gewesen, sich bei der Ablehnung von Bewerberinnen und Bewerbern nicht alleine auf – im Extremfall – einen einzigen Testwert über 65 zu verlassen. Bewerberinnen und Bewerber mit einem auffälligen Testergebnis hätten in einer 2. Sitzung gründlicher untersucht werden müssen. Da mit einer Ablehnung gravierende finanzielle Nachteile verbunden sind, hätte das Auswahlverfahren bzw. die Entscheidungsregel einer Validitätsprüfung unterzogen werden müssen.

4.2 Auftragsannahme und Fragestellung

Jede psychologisch-diagnostische Untersuchung beginnt mit einem Auftrag. Dieser kann mehr oder weniger formell und in mündlicher oder in schriftlicher Form von einer anderen Person oder einer Institution stammen, er kann aber auch selbst gesetzt werden.

Ein Beispiel für einen eher informellen mündlichen Auftrag wäre, wenn Eltern mit ihrem Kind zu einer schulpsychologischen Beratungsstelle kommen und um Abklärung der Gründe für die „Schulprobleme“ ihres Kindes bitten. In der Regel wird diese Bitte mit dem Wunsch nach konkreterer Hilfe oder Beratung verbunden sein. Die Psychologin oder der Psychologe sieht diese Bitte als Auftrag an, wird aber abklären, was die Auftraggebenden genau erwarten. Dazu gehört, ob der Auftrag Empfehlungen einschließt oder „nur“ eine diagnostische Abklärung gewünscht wird. Wollen die Eltern wissen, ob ihr Kind einen Lehrer unbewusst ablehnt (sie haben den Verdacht, konnten sich aber im Gespräch mit ihrem Kind keine Klarheit verschaffen), wird zumindest dieser Teil des Auftrags vermutlich gleich beim Erstkontakt abgelehnt. Die Psychologin oder der Psychologe wird den Eltern erklären, dass es sinnvoll ist, verschiedene Gründe für die Schulprobleme in Erwägung zu ziehen. Dieses Beispiel macht deutlich, dass eventuell der Umfang eines Auftrags abgeklärt werden muss und dass ein Auftrag auch modifiziert oder

Informeller Auftrag

abgelehnt werden kann, falls die Eltern auf der unangemessen engen Fragestellung beharren.

Formeller Auftrag

4

Vielfalt an möglichen Aufträgen

Ein formeller schriftlicher Auftrag kann beispielsweise von einem Familiengericht kommen und die Klärung der Frage zum Gegenstand haben, ob im konkreten Fall das Kindeswohl gefährdet ist. Es bestehe der Verdacht, dass der Vater die eigene Tochter sexuell missbraucht hat. In diesem Fall liegt es nahe, dass auch eine aussagenpsychologische Untersuchung (s. dazu ▶ Abschn. 9.2) nötig sein wird. Die angefragte Psychologin lehnt diesen Auftrag möglicherweise ab, weil Aussagenpsychologie nicht zu ihrem Kompetenzbereich gehört.

Im Rahmen einer Psychotherapie möchte der behandelnde Psychologe wissen, ob und wie stark sich die depressive Symptomatik seines Klienten im Laufe der Behandlung verändert. Er gibt sich selbst einen entsprechenden Auftrag.

Die Vielfalt möglicher Aufträge wird ersichtlich, wenn man die folgende Auflistung von Fragestellungen aus verschiedenen Arbeitsfeldern betrachtet, die keinen Anspruch auf Vollständigkeit hat.

Beispiele für Fragestellungen

- Schule: Schulfähigkeit, Lernfähigkeit; Lern-/Leistungsstörungen; Verhaltensauffälligkeiten; Schullaufbahnberatung
- Hochschule: Zulassung zum Studium, z. B. Härtefälle; Beratung bei Studienwahl, Wechsel des Studienfachs
- Versicherungsträger: Berufsunfähigkeit, Begründung psychotherapeutischer Interventionen, Rehabilitationsmaßnahmen mit beruflichen Einsatzmöglichkeiten
- Gesundheitswesen: Notwendigkeit klinisch-psychologischer Interventionen, psychologische Vorbereitung, Begleitung und Nachsorge bei medizinischen Interventionen (z. B. entstellende Operationen, Sterilisation, Geschlechtsumwandlung)
- Öffentliche Verwaltung: Namensänderung
- Bundesagentur für Arbeit: Studieneignung, Berufseignung, Berufslaufbahnberatung
- Verkehrsbehörden: insbesondere Kraftfahreignung
- Betriebe und Organisationen: Personalentscheidungen, Arbeitsplatzgestaltung, betrieblichen Organisation
- Betreuungsgericht: Pflegeschaft, Vormundschaft, Adoption, Vernachlässigung oder Misshandlung des Kindes
- Familiengericht: Sorgerechtsentscheidung, Umgangsregelung
- Nachlassgericht: Testierfähigkeit
- Jugendgericht: Beurteilung von Reife, Schuldfähigkeit, Feststellung schädlicher Neigungen
- Strafgerichte: Schuldfähigkeit, Glaubhaftigkeit von Zeugenaussagen, Aussetzung einer Strafe auf Bewährung
- Strafvollzug: Haftfähigkeitsüberprüfung, Vollzugslockerung, Prognose zur bedingten Entlassung
- Zivilgericht: Prozessfähigkeit, zivilrechtliche Delikthaftung, Schmerzensgeldforderungen
- Arbeitsgericht: Kündigungen, arbeitsgerichtliche Auseinandersetzungen
- Sozialgericht: Berufs- und Arbeitsfähigkeit
- Verwaltungsgericht (Schullaufbahnen, Fahreignungsuntersuchungen)

(Modifizierte und erweiterte Auflistung nach Föderation Deutscher Psychologenvereinigungen 1994).

Durchführung einer diagnostischen Untersuchung ...

Die Aufträge können also sehr unterschiedlich sein. Manchmal bedürfen sie einer Abklärung, die auch zu einer Ablehnung führen kann (► Abschn. 1.5). Der Auftragsannahme folgt der eigentliche Begutachtungsprozess. Mit der adäquaten Formulierung der globalen Fragestellung wird „der Ball zurechtgelegt“ für die weitere Arbeit. Eventuell muss der genaue Wortlaut des Auftrags und damit der globalen Fragestellung mit der Auftraggeberin oder dem Auftraggeber in einem Beratungsprozess einvernehmlich neu oder umformuliert werden. Dies geschieht in beiderseitigem Interesse. In diesem komplexen Umfeld sind Empfehlungen hilfreich, wie sie sich in den *Qualitätsstandards für psychologische Gutachten* (Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen 2017, S. 3 f.) finden. Dort wird zum Thema „Auftragsklärung und Auftragsannahme“ ausgeführt:

Formulierung des Auftrags wichtig
für die weitere Arbeit

» 3.1 Auftragsklärung und Auftragsannahme

Die Gutachterin/der Gutachter muss vor Auftragsannahme prüfen, ob

1. der Auftrag ethisch verantwortbar ist und er die rechtlichen Vorgaben erfüllt,
2. bei ihr/ihm die nötige Sachkunde zur Beantwortung der Frage vorliegt inklusive der Kenntnisse relevanter rechtlicher Regelungen,
3. im Allgemeinen genügend wissenschaftliche Erkenntnisse und geeignete Methoden zur fundierten Beantwortung der Frage verfügbar sind und
4. weitere fachfremde Gutachten zur Beantwortung der Fragestellung notwendig sind.

Für den Fall, dass die Gesamtfragestellung oder eine Teilfragestellung von der Gutachterin/vom Gutachter nicht beantwortet werden kann, muss die Gutachterin/der Gutachter die Begutachtung ablehnen bzw. – soweit dies rechtlich möglich und vertretbar ist – die (Teil-) Fragestellung mit der Auftraggeberin/dem Auftraggeber so abwandeln, dass sie beantwortet werden kann.

Sofern eine Auftragsklärung erforderlich ist, muss die Gutachterin/der Gutachter der Auftraggeberin/dem Auftraggeber verständlich vermitteln, wie sie/er den Auftrag verstanden hat.

Sofern erkennbar ist, dass die Auftraggeberin/der Auftraggeber Erwartungen an die Durchführung oder das Ergebnis des Gutachtens hat, die nicht mit dieser Richtlinie vereinbar sind, soll die Gutachterin/der Gutachter ggf. vorab informieren, dass:

- der Aussagekraft des geplanten psychologischen Gutachtens zeitliche und inhaltliche Grenzen gesetzt sind,
- stets höchstmögliche Objektivität angestrebt wird,
- der Begutachtungsprozess ergebnisoffen ist.

Ob das Gutachten schriftlich und/oder mündlich erstattet wird, ist – soweit möglich – im Rahmen der Auftragsklärung mit der Auftraggeberin/dem Auftraggeber zu vereinbaren.

(© Föderation Deutscher Psychologenvereinigungen)

In einem Manuskript, das den 2017 veröffentlichten Qualitätsstandards für psychologische Gutachten mit zugrunde lag, den „Qualitätsstandards für psychodiagnostische Gutachten (Version 2.2)“ (Arbeitsgruppe „Qualitätsstandards für psychodiagnostische Gutachten“ 2011) wurde das erforderliche Vorgehen konkreter beschrieben, begründet und mit Umsetzungsempfehlungen versehen.

Konkrete Handlungsempfehlungen

Empfehlungen aus den Qualitätsstandards für psychodiagnostische Gutachten (Version 2.2)

- „Auftragsannahme erfolgt erst nach positiver Prüfung
- der eigenen Sachkunde,
 - des zu erwartenden Erkenntnisgewinns für den Auftraggeber,
 - ob der Auftrag neutral (ergebnisoffen) bearbeitet werden kann,
 - ob der Auftrag mit den gesetzlichen Vorschriften sowie
 - dem eigenen Gewissen vereinbar ist.“

Eine Umsetzungsempfehlung Punkt a, c und d betreffend lautet: „Bei fehlender eigener Sachkunde, bei Befangenheit (fehlender Neutralität) oder bei einem möglichen Verstoß gegen gesetzliche Vorschriften ist der Auftrag abzulehnen“.

Zu Punkt b wird empfohlen: „Wird der Erkenntnisgewinn – gemessen an der Beanspruchung des Probanden und den Kosten – voraussichtlich unverhältnismäßig klein sein, sollte der Auftraggeber darüber informiert werden. Eine einvernehmliche Zurücknahme des Auftrages, ggf. Ablehnung, ist dann anzustreben“.

(Arbeitsgruppe „Qualitätsstandards für psychodiagnostische Gutachten“, 2011, S. 8, © Deutsche Gesellschaft für Psychologie)

Es ist zu klären, welche Dienstleistung der Auftraggeber oder die Auftraggeberin genau wünscht. Nicht jeder Auftrag soll oder kann angenommen werden. Manchmal wird ein Auftrag einvernehmlich modifiziert oder spezifiziert.

! Zu beachten sind folgende Punkte:

- Ethische und rechtliche Aspekte
- Eigene Sachkunde
- Eigene Unvoreingenommenheit
- Erwarteter Nutzen für die Auftraggeberin oder den Auftraggeber in Relation zum Aufwand

4.3 Ableitung von psychologischen Fragen

Ein Auftrag ist allenfalls in Ausnahmefällen so beschaffen, dass er ein routinemäßiges Vorgehen auslöst. Etwas zugespielt lässt sich das Problem mit Bezug auf die oben aufgelisteten Fragestellungen so benennen: Es gibt bei der Studienzulassung keinen „Härtefalltest“ oder für das Sozialgericht keinen „Arbeitsfähigkeitstest“. Damit die globale Fragestellung beantwortet werden kann, also beispielsweise ob bei Frau M. ein Härtefall vorliegt oder ob Herr K. arbeitsfähig ist, muss eine ausgeklügelte Strategie entwickelt werden. Dazu wird die „große“ Ausgangsfrage in Teilfragen zerlegt. Deren Beantwortung führt dazu, dass am Ende auch auf die Ausgangsfrage (in Abb. 4.1 „Fragestellung“) eine angemessene Antwort (in Abb. 4.1 „Stellungnahme“) möglich ist.

Die Teilfragen werden auch „psychologische Fragen“ genannt. Diese Formulierung ist für Laien fast selbsterklärend. Um die im Auftrag fixierte globale Frage zu beantworten, muss die Psychologin oder der Psychologe zunächst bestimmte Fachfragen stellen und abklären. Der entsprechende Abschnitt im schriftlichen Gutachten wird deshalb auch „psychologische Fragen“ oder alternativ z. B. auch „Präzisierung der Fragestellung“ genannt.

Definition

Psychologische Fragen sind ein wesentliches Element des Begutachtungsprozesses und auch des schriftlichen Gutachtens. Sie sind „Hypothesen“ oder selbst gesetzte Arbeitsaufträge, die zur Beantwortung der globalen Fragestellung benötigt werden.

Mit den psychologischen Fragen wird die globale Fragestellung in konkrete Unterfragen „übersetzt“. Sie dienen der Strukturierung des weiteren diagnostischen Prozesses. Damit sie diese Funktion gut erfüllen können, müssen sie bestimmte Eigenschaften aufweisen:

- Sie müssen nachvollziehbar aus der globalen Fragestellung hergeleitet werden.
- Dazu ist es erforderlich, für jede psychologische Frage zu begründen, warum sie zur Beantwortung der globalen Fragestellung erforderlich ist. Argumente können wissenschaftlich gesicherte Gesetzmäßigkeiten oder Erkenntnisse, gesicherte allgemeine oder eigene Erfahrungen und/oder Anknüpfungstatsachen aus den vorliegenden Informationen sein.
- Sie müssen so formuliert werden, dass sie grundsätzlich mit den zur Verfügung stehenden diagnostischen Verfahren beantwortet werden können. „Grundsätzlich“ impliziert, dass dies im konkreten Fall auch einmal nicht gelingen kann, etwa weil die Klientin oder der Klient nicht kooperativ ist. Besteht jedoch bei einer psychologischen Frage von vorneherein keine Aussicht auf eine brauchbare Antwort, ist sie wertlos.
- Auf keinen Fall soll ein Vorgriff auf eine bestimmte Methode erfolgen (falsch wäre also eine Formulierung wie diese: „Mit einem Intelligenztest soll deshalb überprüft werden, ob die Intelligenz im sehr niedrigen Bereich liegt“). Die Auswahl geeigneter diagnostischer Verfahren erfolgt erst im nächsten Schritt.

Anforderungen an die psychologischen Fragen

In der Definition werden die psychologischen Fragen als „Hypothesen oder selbst gesetzte Arbeitsaufträge“ bezeichnet. Es erscheint uns wichtig, zu erwähnen, dass nicht nur Hypothesen zulässig sind. Diese sind wünschenswert, weil sie eine gerichtete Fragestellung darstellen und damit eine einseitige Prüfung der anfallenden Ergebnisse erlauben. Das Konfidenzintervall wird kleiner und die Chance, eine eindeutige Antwort zu erhalten, steigt. Viele psychologische Fragen haben jedoch einen explorativen Charakter. Sie werden dann in Form eines Arbeitsauftrages formuliert: „Deshalb ist zu prüfen, ob ...“ Auch wenn eine Hypothese formuliert werden kann, ist manchmal eine neutrale Formulierung gegenüber der zu untersuchenden Person respektvoller. Beispielsweise kann aufgrund der bekannten Vorgeschichte die Hypothese aufgestellt werden, dass die Person einen Intelligenzquotienten (IQ) unter 70 hat und damit als intelligenzgemindert gelten kann. Respektvoller ist Formulierung als Arbeitsauftrag: „Deshalb ist zu prüfen, ob die Intelligenz im Vergleich zu etwa gleichaltrigen Personen unter einem IQ von 70 liegt.“ Eine Formulierung als Hypothese kann von Laien leicht als eine Unterstellung aufgefasst werden.

Hypothese, explorative Fragestellung oder Arbeitsauftrag

4.3.1 Psychologische Fragen finden

Das Finden und die Begründung psychologischer Fragen stellt besonders für Psychologinnen und Psychologen mit wenig einschlägiger Erfahrung anfangs eine große Herausforderung dar. Es ist sehr zu empfehlen, systematisch nach möglichen Erklärungen zu suchen.

Inhaltsbereiche

Erstens kann es hilfreich sein, bewährte Inhaltsbereiche näher zu betrachten. Dies sind der kognitive Bereich (Intelligenz, Wissen, Merkfähigkeit, Aufmerksamkeit und Konzentration), Persönlichkeitsmerkmale (die „Big Five“ sowie spezifische Merkmale wie Belastbarkeit oder Ordnungsliebe), Interessen, Motivation und nicht zuletzt die Lebensbedingungen (berufliches und soziales Umfeld, finanzielle Situation, Wohnsituation etc.).

Zweitens kann eine verhaltensanalytische Betrachtung weiterführen. Dabei bietet sich die Verhaltensgleichung nach Westhoff und Kluck (2008, S. 24 f.) an. Diese Formel fasst relevante Variablen zusammen, die zur Erklärung, Vorhersage und Beeinflussung individuellen Verhaltens bedeutsam sind:

$$V = f(U, O, K, E, M, S)$$

Danach ist Verhalten eine Funktion folgender Variablengruppen:

- Umgebungsvariablen (U): äußere Lebensbedingungen, z. B. Wohnsituation, finanzielle Situation
- Organismusvariablen (O): körperliche Bedingungen, z. B. Krankheiten, Behinderungen
- Kognitive Variablen (K): Leistungsfähigkeit und Inhalte des Wahrnehmens, Lernens und Denkens, z. B. Allgemeine Intelligenz, Intelligenzstruktur, Konzentration
- Emotionale Variablen (E): z. B. emotionale Belastbarkeit, Umgang mit Gefühlen und Belastungen
- Motivationale Variablen (M): z. B. Leistungs- und Machtmotiv, Interessen, Werte
- Soziale Variablen (S): soziale Intelligenz, Normen, Einflüsse von „bedeutenden Anderen“ und deren Wechselwirkungen

Verhaltensgleichung

Anforderungsmerkmale

Drittens ist es manchmal hilfreich, sich an Anforderungen zu orientieren, die erfüllt sein müssen. Bei eignungsdiagnostischen Fragestellungen wird explizit der Begriff „Anforderungen“ bzw. „Anforderungsprofil“ für die Gesamtheit aller Anforderungen verwendet. Beispielsweise stellt man fest, dass Pilotinnen und Piloten über englische Sprachkenntnisse, eine gute Aufmerksamkeitsleistung, eine hohe Intelligenz, ein gutes räumliches Vorstellungsvermögen etc. verfügen müssen. Deshalb beziehen sich bei eignungsdiagnostischen Gutachten die psychologischen Fragen auf solche Anforderungsmerkmale. Das Konzept der Anforderungen lässt sich aber auch auf viele andere Bereiche übertragen, ohne dass dieser Begriff dabei unbedingt verwendet wird. So sind psychische Störungen laut ICD-10 bzw. DSM-5 durch bestimmte Symptome definiert; um festzustellen, ob jemand eine bestimmte Störung hat, überprüft man, ob diese Anforderungen (die Symptome) vorliegen. Damit eine Straftäterin bzw. ein Straftäter nach der Entlassung nicht mehr rückfällig wird, muss er/sie bestimmte Anforderungen, also Bedingungen erfüllen, die für eine gute Prognose sprechen. Grundsätzlich lassen sich kompensierbare von nichtkompenzierbaren, stabile von instabilen und veränderbare von unveränderbaren Anforderungsmerkmalen unterscheiden (vgl. Westhoff und Kluck 2008, S. 18 f.). Nichtkompenzierbare Anforderungen (s. auch ► Abschn. 5.1.3) müssen unbedingt erfüllt sein; hier ist es sinnvoll, Mindestwerte zu fordern. Beispielsweise wird zur Eignung zur Pilotin bzw. zum Piloten unabdingbar ein gutes Sehvermögen gehören. Anforderungen können sich im Laufe der Zeit ändern, weil Berufsausbildungen, Schulsysteme, technische Hilfsmittel oder etwa die Software am Arbeitsplatz einem Wandel unterliegen. Anforderungsmerkmale, die auch in Zukunft von Bedeutung sein werden, sind wichtiger als solche, die nur noch vorübergehend relevant sind (vorausgesetzt, dass eine Position nicht nur vorübergehend besetzt werden soll). Veränderbare Merkmale der Person lassen sich durch Training oder Weiterbildung nachträglich

so modifizieren, dass sie den Anforderungen entsprechen. Deshalb kann es angemessen sein, hier nur ein „soll“ anstelle eines „muss erfüllt sein“ zu fordern. Generell müssen die Anforderungen so definiert sein, dass sie messbar sind.

4.3.2 Darstellung der psychologischen Fragen

Von einem Gutachten wird verlangt, dass es transparent und nachvollziehbar ist. Die psychologischen Fragen an sich tragen dazu wesentlich bei. Sie erklären, warum was untersucht werden soll. Sie müssen aber auch selbst diesen Prinzipien folgen. Die Herleitung einer psychologischen Frage aus der Fragestellung muss nachvollziehbar sein. Die Begründung, warum beispielsweise die emotionale Belastbarkeit der Klientin oder des Klienten zur Beantwortung der Fragestellung untersucht werden muss, soll für die Empfängerin oder den Empfänger des Gutachtens verständlich formuliert und inhaltlich überzeugend sein. Eine geordnete Darstellung, beispielsweise nach Inhaltsbereichen, erleichtert das Verständnis. Wichtiges sollte immer zuerst genannt werden.

Nachvollziehbare Herleitung aus der Fragestellung

- !** Mit den psychologischen Fragen wird „der Ball zurechtgelegt“ für den weiteren diagnostischen Prozess. Ausgehend von der Fragestellung der Auftraggeberin oder des Auftraggebers wird nachvollziehbar dargelegt, warum bestimmte Hypothesen oder explorative Fragen naheliegen und zu klären sind. Um die „richtigen“ psychologischen Fragen zu finden, kann man sich an oft relevanten Inhaltbereichen orientieren, ein Verhalten mithilfe der Verhaltensgleichung analysieren und/oder Anforderungen festlegen. Psychologische Fragen müssen so gewählt werden, dass sie grundsätzlich mit den zur Verfügung stehenden Methoden zu beantworten sind.

Wichtig bei psychologischen Fragen

4.4 Auswahl der Verfahren und Untersuchungsplanung

Zur Beantwortung der psychologischen Fragen werden geeignete diagnostische Verfahren benötigt. Ziel ist es, zuverlässige und valide diagnostische Informationen zu gewinnen. Der mit einem Verfahren verbundene Aufwand spielt bei der Auswahl eine Rolle. Zentrale und wichtige Fragen rechtfertigen den Einsatz aufwendiger Verfahren. Auch der Einsatz mehrerer Verfahren, beispielsweise eines Tests, eines diagnostischen Interviews und einer Verhaltensbeobachtung, ist zu erwägen.

Wichtigkeit einer psychologischen Frage entscheidend für Aufwand

Es existieren enorm viele diagnostische Verfahren (► Kap. 3), sodass schon die richtige Auswahl entscheidend dafür ist, ob man brauchbare Erkenntnisse gewinnt oder nicht. Der Aufwand für die Auswahl des am besten geeigneten Verfahrens kann sehr groß sein. Wir stellen im Folgenden nützliche Strategien vor.

Richtige Auswahl und Anwendung entscheidend

! Kompetente Testauswahl bedeutet laut International Test Commission (ITC 2001):

- Beurteilung der Brauchbarkeit des Tests für die geplante diagnostische Untersuchung
- Auswahl technisch einwandfreier und für die diagnostische Situation angemessener Tests
- Beachtung der Fairness des Tests bei der geplanten Testanwendung

Kompetente Testauswahl

Die hier eher allgemein gehaltenen Forderungen der International Test Commission werden in ► Abschn. 5.2.1 im Detail beschrieben.

4.4.1 Geeignete diagnostische Verfahren finden

- » Fachkompetente Testanwender legen eine begründete Rechtfertigung für die Anwendung eines Tests vor (ITC 2001, S. 15).

Auch in Gutachten soll zwecks Nachvollziehbarkeit des methodischen Vorgehens aufgeführt werden, warum ein bestimmtes Verfahren ausgewählt hat. Nun gibt es natürlich Hunderte oder sogar Tausende von diagnostischen Verfahren. Wir führen zunächst die Auswahlkriterien in Form von Fragen auf und gehen anschließend auf pragmatische Aspekte der Auswahl ein.

■ Frage 1: Was soll das Verfahren messen?

Messanspruch muss passen und durch Validität belegt sein

Wir gehen stets von einer psychologischen Frage aus, die wir mithilfe von einem oder auch mehreren diagnostischen Verfahren beantworten wollen. Ein Verfahren wird danach ausgewählt, dass es das misst, was man zur Klärung einer Frage benötigt. Beispielsweise möchte man wissen, wie intelligent eine Person ist. Intelligenz kann sehr gut mit Tests erfasst werden, jedenfalls besser als mittels Interview oder Analyse von Schulzeugnissen. Also suchen wir nach einem „Intelligenztest“, der nachweislich auch Intelligenz misst. Bei der Auswahl eines Tests sind grundsätzlich nicht nur dessen Name und Messanspruch zu beachten, sondern auch seine psychometrischen Gütekriterien (► Abschn. 2.6) – und hier zunächst die Validität. Ein „Intelligenztest“ misst nicht zwangsläufig Intelligenz. Ohne überzeugende Belege zur Validität ist der „Intelligenztest“ nicht zur Messung der Intelligenz geeignet.

■ Frage 2: Ist das Verfahren für diese Person angemessen?

Angemessene Normen vorhanden?

Für viele Fragestellungen ist es erforderlich, die Ausprägung des Merkmals anhand von Normtabellen zu interpretieren. Dann kommen *biografische Merkmale* ins Spiel. In diesen Fällen kann es unabdingbar sein, dass aktuelle, repräsentative Normen für einen bestimmten Altersbereich und/oder eine bestimmte Bildungsstufe vorliegen. Beispielsweise handelt es sich bei der untersuchten Person um ein 10-jähriges Kind mit einer vermutlich eher niedrigen Intelligenz, wie die Vorinformationen vermuten lassen. In diesem Fall suchen wir einen Intelligenztest, der auch im unteren Bereich gut differenziert sowie für die Altersgruppe von 10 Jahren eine hinreichend große Eichstichprobe ($N > 200$) vorweisen kann.

Akzeptanz

Der Bildungsstand und der berufliche Hintergrund spielen noch aus einem anderen Grund eine Rolle: Die Akzeptanz eines Verfahrens kann darunter leiden, dass es Aufgabengruppen oder Items enthält, die nicht zu der untersuchten Person passen. Einer Person mit Hochschulabschluss einen Test vorzugeben, der Aufgaben wie „Fritz hatte sieben Buntstifte, zwei hat er verloren ...“, wäre unangemessen.

Körperliche und kognitive Einschränkungen beachten

Die Fragestellung ist stets in einen bestimmten Anwendungskontext eingebettet. Im Berufseignungskontext wäre es akzeptanzmindernd, wenn ein Persönlichkeitsfragebogen viele Items enthält, die sich auf die Freizeit und die Familie beziehen. Ein Fragebogen, der für den klinischen Bereich entwickelt wurde, wird ebenso inadäquat sein, weil er nach Symptomen psychischer Störungen fragt („Was geht das meinen künftigen Arbeitgeber an?“).

Es gilt, auch Besonderheiten der untersuchten Person zu beachten. Grundsätzlich ist zu beachten, dass das Verfahren die Testperson nicht systematisch benachteiligt, also fair ist (► Abschn. 2.6.5.5). Relevante *körperliche Faktoren* können sein: eingeschränktes Sehvermögen (einige Aufmerksamkeits- und Konzentrationstests enthalten relativ kleine Zeichen als Items), motorische Behinderung oder Beeinträchtigung (bei Tests, die unter großem Zeitdruck bearbeitet werden müssen, kann die Markierung von Antwortalternativen dann zu viel Testzeit kosten). *Kognitive Faktoren*, die gegen die

Verwendung eines bestimmten Verfahrens sprechen können, sind eine niedrige Intelligenz und eine niedrige Sprachkompetenz (beide Faktoren können dazu führen, dass das Instruktionsverständnis nicht sichergestellt ist oder Fragebogenitems nicht richtig verstanden werden) sowie mangelndes Wissen, etwa aufgrund eines anderen kulturellen Hintergrunds (Intelligenztests können Items enthalten, die ein für uns eher selbstverständliches Vorwissen voraussetzen).

Wenn die Person einen in die Auswahl genommenen Leistungstest in den letzten Jahren schon einmal durchgeführt hat, wird die mit dem Test gemessene Fähigkeit überschätzt (z. B. durch Übungseffekte; ► Abschn. 3.2.1). Besonders wenn der Test in der letzten Zeit mehrfach bearbeitet worden ist, sollte man möglichst auf ein anderes Verfahren ausweichen. Das gilt für den Einzelfall genauso wie bei der Untersuchung vieler Menschen mit gleicher Fragestellung. Wenn bekannt ist, dass viele Unternehmen in der Region einen bestimmten Leistungstest zur Personalauswahl einsetzen, bietet es sich an, einen anderen Test zu verwenden.

Testerfahrung, Übung

■ Frage 3: Ist das Verfahren für diese Fragestellung geeignet?

Bei vielen Fragestellungen spielen Prognosen eine Rolle. Es sollen Ausbildungs-, Schul-, Berufs- oder Studienerfolge vorhergesagt werden oder etwa das Rückfallrisiko bei einer forensischen Fragestellung. Immer dann ist die *Kriteriumsvalidität* des Verfahrens von Bedeutung. Bei Prognosen ist es wichtig, dass sich das Merkmal, wie es mit dem Test erfasst wird, über den Prognosezeitraum wenig ändert. Deshalb sollte das Verfahren eine hohe *Retest-Reliabilität* aufweisen.

Validitätsbefunde mit Bezug zur Fragestellung

Bei vielen Fragestellungen ist grundsätzlich damit zu rechnen, dass Probandinnen und Probanden stark motiviert sind, ein „gutes“ oder auch ein „schlechtes“ Ergebnis zu erzielen. In diesem Fall ist mit *Verfälschung* zu rechnen. Folgende (Gegen-)Maßnahmen bieten sich an:

Verfälschung verhindern oder erkennen

- Bei Leistungstests in Gruppenuntersuchungen kann das Abschreiben vom Nachbarn durch große räumliche Abstände zwischen den Testpersonen verhindert werden, oder man setzt Paralleltests bzw. Pseudoparalleltests (andere Abfolge der Items) ein.
- Bei Persönlichkeitsfragebögen werden Verfahren eingesetzt, die schwer verfälschbar sind (Beispiel MMPI-2-RF; ► Abschn. 3.3.3.2) oder Kontrollskalen haben, die sozial erwünschtes Antwortverhalten anzeigen.
- Man kann den Einsatz von Fragebögen mit Forced-Choice-Format (► Abschn. 2.4.2.6 und 3.3.2) erwägen.
- Mit speziellen Verfahren lässt sich eine Simulation erkennen (► Abschn. 9.1).

Der Einsatz eines Verfahrens kann mit juristischen Problemen verbunden sein. In der Berufseignungsdiagnostik ist es wichtig, dass der Anforderungsbezug gegeben ist. Verfahren, die über die beruflichen Anforderungen hinausgehen, sind unzulässig (Höft et al. 2018). In der verkehrspychologischen Diagnostik dürfen zur Messung der „geistigen Anforderungen“ nur noch Verfahren verwendet werden, deren Eignung von einer Prüfstelle nach einer gründlichen Begutachtung festgestellt worden ist (► Abschn. 9.3).

Ist das Verfahren rechtlich zulässig?

4.4.1.1 Vorgehen bei der Auswahl

Die oben genannten Auswahlkriterien sind in keinem Katalog und in keiner Datei aufgelistet. Die große Suche nach einem passenden Verfahren, die „ganz von vorne“ anfängt, also beispielsweise zunächst alle Intelligenztests oder alle mehrdimensionalen Persönlichkeitsfragebögen in Betracht zieht, wäre extrem aufwendig. Dennoch lässt sie sich manchmal nicht vermeiden.

Auswahl unter den verfügbaren Verfahren

Einsatz suboptimaler Verfahren ethisch nicht vertretbar

Effizient suchen

Diagnostikerinnen und Diagnostiker, die immer nur Untersuchungen zu einer Fragestellung (z. B. die Auswahl von Auszubildenden für kaufmännische Berufe oder die Diagnostik von Essstörungen) durchführen, kommen in der Regel mit einer kleinen Auswahl von diagnostischen Verfahren zurecht. Wenn man „seine“ Verfahren gut kennt, fällt die Auswahl nach den obigen Kriterien leicht.

Man sollte sich jedoch stets die Frage stellen, ob das gerade verfügbare Verfahren wirklich optimal ist und ob es vielleicht inzwischen bessere Alternativen gibt. Da sich aus den Ergebnissen einer diagnostischen Untersuchung manchmal erhebliche Konsequenzen wie Schulwechsel, Ablehnung einer Bewerberin oder eines Bewerbers oder die (Nicht-)Finanzierung einer Therapie ergeben, ist es ethisch nicht vertretbar, wider besseres Wissen suboptimale Verfahren einzusetzen. Der eigene Testbestand bedarf einer kontinuierlichen Pflege.

Wir setzen uns nun mit dem Fall auseinander, dass ein Verfahren zur Messung eines bestimmten Merkmals benötigt wird, dieses aber nicht am eigenen Arbeitsplatz verfügbar ist. Nun beginnt die oben angesprochene „große Suche“. Manchmal kann ein Lehrbuch wie das vorliegende helfen, innerhalb einer Kategorie von Verfahren wie etwa „mehrdimensionale Persönlichkeitsfragebögen“ zumindest eine gute Vorauswahl zu treffen. Grundsätzlich benötigt man für die Suche nach einem geeigneten Verfahren eine effiziente Strategie und gute Informationsquellen.

Als Vorbild für die *Suchstrategie* bietet sich das sequenzielle Vorgehen an, das in ▶ Abschn. 5.1.4 für die Selektion von Personen beschrieben wird. Wir beginnen damit, die Anforderungen an das benötigte Verfahren in unbedingt erforderliche und nachrangige einzuteilen. Nachrangig kann (in manchen Kontexten) etwa die Akzeptanz sein. Notfalls entschuldigt man sich bei einer Testperson dafür, dass manche Fragen oder Items scheinbar unangemessen sind: „Sie erfassen schon das richtige Merkmal, stammen aber aus einem Kontext, den Sie bei der vorliegenden Fragestellung zurecht als unpassend erleben.“ Die nächsten Schritte sind in folgender Übersicht genannt.

Sequentielle Suchstrategie für ein Verfahren

1. Anforderungen an das benötigte Verfahren (s. Text) aufschreiben.
2. Anforderungen, die unbedingt erfüllt sein müssen, markieren.
3. Unter den unbedingt erforderlichen Anforderungen eine auswählen, die am leichtesten überprüfbar ist.
4. Verfahren suchen, die diese Anforderung erfüllen.
5. Für diese Verfahren die nun am leichtesten überprüfbare Anforderung auswählen.
6. Schritt 4 und 5 so lange wiederholen, bis alle unbedingt erforderlichen Anforderungen abgearbeitet sind.

4.4.1.2 Informationsquellen

In der Datenbank *PSYNDEX Tests* finden sich 7979 Testnachweise (Stand: Juli 2019). „Test“ ist dabei umfassend zu verstehen und schließt nicht nur Fragebögen, sondern auch Interview- und Beobachtungsverfahren ein. Der Zugriff ist von vielen Hochschulen für Berechtigte kostenfrei möglich. Alle Verfahren werden im deutschen Sprachraum verwendet. Für rund die Hälfte liegen ausführliche Verfahrensbeschreibungen („PSYNDEX Tests Review“) vor. Das Abstract ist wie folgt gegliedert: diagnostische Zielsetzung, Aufbau, Grundlagen und Konstruktion, empirische Prüfung und Gütekriterien.

Durchführung einer diagnostischen Untersuchung ...

Es schließen sich weiter untergliederte Ausführungen zu Testkonzept, Durchführung, Testkonstruktion, Gütekriterien und kurze Angaben zu Anwendungsmöglichkeiten sowie eine Bewertung an. Liegen Testrezensionen vor, wird mit genauen Quellenangaben darauf verwiesen.

Wir haben die Gliederung aufgegriffen, weil so erkennbar ist, dass eine gezielte Suche nach bestimmten Informationen leicht möglich ist. Beispielsweise sucht man bestimmte Angaben zur Validität, die dann unter Gütekriterien/Validität zu finden sind. Oder man möchte wissen, für welchen Altersbereich das Verfahren normiert ist und wie lange die Durchführung dauert; unter Durchführung/Altersbereiche bzw./Durchführungszeit wird man fündig. Der Aufbau der Testbeschreibung ist für alle Verfahren identisch, was die Suche zusätzlich erleichtert. Die Suche kann mit Schlagworten (z. B. Intelligenztest, Kinder) eingeengt werden.

Übersichtlich gegliederte Testinformationen

Freie Literatursuche

Ein freier Zugriff auf Testinformationen ist über die Suchmaschine PubPsych möglich, die man über das ZPID erreicht (► <https://pubpsych.zpid.de/pubpsych>). Auf ein Stichwort hin findet man aber nicht nur Tests, sondern auch viel Fachliteratur zu dem Thema. Die Suche kann auf Tests eingeengt werden, indem bei „Publikationstyp“ die Kategorie „Tests/Questionnaires“ angeklickt wird. Bei den aufgeführten Verfahren kommt man durch Anklicken von „PSYNDEX Tests Zusatzinformationen“ (soweit vorhanden) zu den oben beschriebenen Testinformationen.

Testverlage als Informationsquelle

Die diagnostischen Verfahren lassen sich danach unterteilen, ob sie primär für die Anwendung in der Praxis entwickelt wurden oder erst einmal nur für Forschungszwecke. Verfahren der 1. Kategorie werden in der Regel von *Testverlagen* angeboten. Deshalb kann eine Suche auch über die Seiten der Testverlage, das sind vor allem Hogrefe, Schuhfried (nur Computertests) und Pearson Clinical & Talent Assessment durchführen. Der Hogrefe-Verlag vertreibt über seine „Testzentrale“ auch Tests von kleineren anderen Verlagen. Über diese Quellen findet man ebenfalls diagnostische Verfahren zu einem bestimmten Thema sowie zu den einzelnen Verfahren zumeist kurze Übersichten über Messgegenstand, Gütekriterien, Erscheinungsjahr etc. Es ist zu empfehlen, sich über infrage kommende Verfahren bei PSYNDEX Tests weitere unabhängige Informationen einzuholen oder gezielt nach Testrezensionen zu suchen, insbesondere nach solchen, die nach den Standards des Diagnostik- und Testkuratoriums (s. ► <https://www.psyndex.de/index.php?wahl=Testkuratorium>) verfasst wurden.

Elektronisches Testarchiv

Bei der 2. Kategorie von Verfahren handelt es sich überwiegend um Fragebögen, die im Rahmen von Forschungsprojekten entwickelt wurden. Anstelle eines Testmanuals liegt oft eine Publikation in einer Fachzeitschrift vor. Diese Verfahren kommen vor allem für Forschungszwecke in Betracht, weil zumeist keine Normen verfügbar sind. Das ZPID (► <https://www.zpid.de/>) bietet mit seinem „Elektronischen Testarchiv“ (► <https://www.testarchiv.eu/>) den Zugriff auf (Stand: Juni 2020) 201 Forschungsinstrumente, die aber nur zu nicht kommerziellen Zwecken verwendet werden dürfen. Zu jedem Test findet man eine Beschreibung aus der Datenbank PSYNDEX Tests (s. o.) sowie das Testmaterial zum Download.

4.4.1.3 Zugriff auf diagnostische Verfahren

Testing on Demand

Leider beschränken sich Testanwenderinnen und Testanwender oft auf die wenigen Verfahren, die sie einmal angeschafft haben. Um Kosten zu sparen, werden in manchen Einrichtungen Tests verwendet, von denen längst überarbeitete und neu normierte Nachfolger auf dem Markt sind. Besonders bei Intelligenztests besteht dabei die Gefahr, dass mit den alten Normen die Intelligenz überschätzt wird. Die Anschaffung eines gerade benötigten Tests kann manchmal daran scheitern, dass der Etat für das laufende Jahr ausgeschöpft

ist. Wenn es nicht möglich ist, sich das Verfahren aus einer anderen Einrichtung auszuleihen, kann „Testing on Demand“ eine Lösung sein. Testverlage bieten zu manchen Verfahren eine internetbasierte Testung mit Auswertung an. Man bezahlt also nur für die einzelne Testanwendung. Für Forschungszwecke findet man eventuell in dem oben genannten „Elektronischen Testarchiv“ ein geeignetes Verfahren.

4

4.4.2 Untersuchungsplanung

Bei der Planung der Untersuchung sind aus ökonomischen Gründen einige grundsätzliche Entscheidungen zu treffen:

- Sollen bestimmte Informationen vorab zu Hause erhoben werden?
- Sollen einige oder alle Verfahren in einer Gruppenuntersuchung durchgeführt werden?
- Sollen Verfahren als Papier-und-Bleistift-Version oder computerbasiert eingesetzt werden?

4.4.2.1 Durchführung zu Hause vs. unter Anleitung in Untersuchungsräumen

Bewerber-Screenings häufig über Internet

Nur die Anleitung und Aufsicht durch eine Psychologin oder einen Psychologen oder eine von ihnen instruierte und supervidierte Hilfskraft bieten die Gewähr, dass eine diagnostische Untersuchung standardisiert durchgeführt wird. Die wünschenswerte Standardisierung wird jedoch mit der Arbeitszeit der Untersuchungsleiterin oder des Untersuchungsleiters und mit den Kosten für die Anreise der zu untersuchenden Personen bezahlt; oftmals müssen sich diese zudem für die Untersuchung einen halben oder ganzen Tag Urlaub nehmen. Aus ökonomischen, ökologischen und aus Komfortgründen ist es wünschenswert, einen Teil der Untersuchung zu Hause durchführen zu können. Die klar erkennbaren ökonomischen Vorteile führen dazu, dass insbesondere Unternehmen das Internet nutzen, um Psychologische Diagnostik zu den Probandinnen oder Probanden zu bringen. In den USA nutzen einem von Gnambs et al. (2011) zitierten Bericht zufolge etwa 60 % der US-amerikanischen Unternehmen das Internet oder das Telefon, um Bewerberinnen oder Bewerber diagnostisch zu untersuchen (s. auch ▶ Abschn. 3.7). Im deutschen Sprachraum ist internetbasierte Diagnostik bei Fragebögen und Leistungstests im Rahmen von Studienberatungen weitverbreitet (s. dazu Online-Self-Assessments in ▶ Abschn. 7.1.3). Die Reliabilität internetbasierter Fragebögen zur Persönlichkeit und zu klinischen Störungen ist gegenüber Papier-und-Bleistift-Verfahren mindestens gleichwertig (Gnambs et al. 2011). Wenn internetbasierte Diagnostik zur Auswahl von Bewerberinnen und Bewerbern eingesetzt wird, geschieht dies in der Regel nur zur Vorauswahl.

Die International Test Commission hat Richtlinien zur computer- und internetbasierten Testung veröffentlicht (ITC 2006). Die Anforderungen werden aus 3 unterschiedlichen Perspektiven betrachtet: der von Menschen, die Tests entwickeln, die Tests anbieten und die einen Test durchführen. Zentrale Themen für internetbasiertes Testen sind:

- Hard- und Software-Voraussetzungen müssen sichergestellt sein.
- Die Internetverbindung muss stabil sein.
- Die Verfahren müssen bestimmte ergonomische Anforderungen (z. B. angemessene Schriftgröße bei verschiedenen Bildschirmen und die Möglichkeit, zur Instruktion zurückzugehen) erfüllen.
- Niemand soll ausgeschlossen werden (ggf. andere Testmöglichkeit anbieten).
- Verständliche Informationen und technische Unterstützung müssen vorhanden sein.

Richtlinien zur internetbasierten Testung

Durchführung einer diagnostischen Untersuchung ...

- Die Authentizität der Testperson und die Möglichkeit zum Schummeln sind zu beachten.
- Die Testergebnisse müssen sicher über das Internet übertragen werden.

4.4.2.2 Gruppen- oder Einzeltestung

Durch die gleichzeitige Testung mehrerer Personen verringert sich der Zeitaufwand pro Testperson und für die Testleiterin oder den Testleiter ganz erheblich. Letztere müssen nicht nur die Instruktion vortragen, sondern während der Bearbeitung für Fragen zur Verfügung stehen und zumeist auch die Einhaltung von Bearbeitungszeiten kontrollieren. Wenn eine Einzeltestung 2 h in Anspruch nimmt, reduziert sich der Aufwand pro Person bei einer Gruppenuntersuchung mit 20 Testpersonen auf 6 min (Abb. 4.2).

Gruppenuntersuchungen sind ökonomisch

Dennoch gibt es auch gute Gründe für Einzeltests: Einige Tests erlauben grundsätzlich keine Gruppenanwendungen, bei Einzeluntersuchungen sind genauere Verhaltensbeobachtungen möglich als bei der Untersuchung in Gruppen, und reine Powertests mit individuell unterschiedlich langer Bearbeitungszeit lassen sich nur schwer in eine Gruppensitzung integrieren. Darüber hinaus schafft die Gruppensituation selbst gewisse Probleme, die aber bewältigt werden können: Dadurch, dass die getesteten Personen häufig eng zusammensitzen, besteht die Möglichkeit, die Ergebnisse von Sitznachbarn bzw. -nachbarinnen zu übernehmen.

Gründe für Einzeltests

Ob die Einzel- oder die Gruppentestung aus diagnostischer Sicht günstiger sind, ist nicht grundsätzlich zu klären, sondern nur unter der Berücksichtigung sowohl individueller Erfordernisse als auch des konkreten Untersuchungsziels. Manchmal ist eine Aufteilung sinnvoll: Ein Teil der Tests wird in einer Gruppensitzung durchgeführt und der Rest in einer Einzelsitzung.

Aufteilung sinnvoll

4.4.2.3 Papier-und-Bleistift- oder Computertest

Viele Persönlichkeits- und Leistungstests können wahlweise in der klassischen Papier-und-Bleistift-Version oder computerbasiert durchgeführt werden. Normalerweise ist heute nicht mehr mit Akzeptanzproblemen zu rechnen, wenn Testaufgaben oder Fragen am Computer zu bearbeiten sind. Dennoch kann es Menschen geben, für die eine Testbearbeitung am Computer sehr ungewöhnlich wäre; in diesem Fall ist es angemessen, die Papier-und-Bleistift-Version zu verwenden. Ein wichtiger Gesichtspunkt ist die Zuverlässigkeit der Testauswertung. Die Auswertung von Papier-und-Bleistift-Tests ist fehleranfällig und sollte deshalb nur geschultem Personal anvertraut werden.

Akzeptanz und Fehleranfälligkeit der Auswertung



Abb. 4.2 Gruppenuntersuchungen sind ökonomisch, aber nicht immer sinnvoll. (© Robert Kneschke/stock.adobe.com)

Kostenüberlegungen

Außerdem kann sie zeitintensiv sein, was sich erschwerend auf eine geplante Ergebnisrückmeldung auswirken kann. Die Auswertung bei Computertests erfolgt sehr schnell und zuverlässig.

Meist sind ökonomische Überlegungen dafür ausschlaggebend, welche Testvariante angeschafft oder im Einzelfall ausgewählt wird. Bei Gruppenuntersuchungen stellt eventuell die Anzahl verfügbarer Computerarbeitsplätze eine Begrenzung dar. Bei Papier-und-Bleistift-Tests entstehen zumeist nur geringe Kosten für das Testmaterial – dafür schlagen die Personalkosten für die Durchführung und Auswertung in der Regel erheblich zu Buche. Bei der Verwendung von Computertests verursacht die Anschaffung der Testsoftware oft relativ hohe Kosten, während jede einzelne Anwendung nur wenig Geld kostet. Die Anschaffungskosten für Software entfallen, wenn man sich für „Testing on Demand“ entscheidet; die einzelne Anwendung ist dafür teurer als beim Computertest. Ein Test wird dann über das Internet beim Testverlag durchgeführt, und die Ergebnisse stehen direkt nach der Durchführung zur Verfügung.

! Die Auswahl der „richtigen“ Verfahren ist mitentscheidend dafür, ob eine Fragestellung überhaupt angemessen beantwortet werden kann. Wird ein Gutachten (► Abschn. 4.5) erstellt, gehört die Begründung für die Auswahl jedes einzelnen Verfahrens in das Gutachten. Dies dient der Nachvollziehbarkeit des methodischen Vorgehens.

4.4.2.4 Abfolge der Verfahren festlegen

Steht die Auswahl der Verfahren fest, wird deren Abfolge festgelegt. Dabei ist die Belastbarkeit der untersuchten Personen ebenso zu beachten wie eventuelle Anforderungen an eine bestimmte Testdurchführung. Bei manchen Leistungstests ist im Abschnitt „Durchführung“ des Manuals zu lesen, dass die Testpersonen ausgeruht sein sollen. Generell ist es günstig, Leistungstests zu Beginn einer Untersuchung oder auch nach einer angemessenen Pause durchzuführen – auch wenn zumindest studentische Testpersonen auch lange Testsungen relativ gut bewältigen können.

Leistungsabfall nach lang dauernden Leistungstests?

Kein Zusammenhang zwischen Testdauer und Testleistung

Sind die Ergebnisse am Ende der Sitzung vielleicht nicht mehr aussagekräftig, weil die Testpersonen erschöpft sind? Ackerman und Kanfer (2009) ließen Studierende einen kognitiven Leistungstest, der einschließlich kurzer Pausen 3,5, 4,5 oder 5,5 h dauerte, bearbeiten. Erwartungsgemäß fühlten sich die Testpersonen am Ende umso erschöpfter, je länger die Untersuchung gedauert hatte; auch innerhalb der Testsitzungen nahm die Erschöpfung zu. Eine Leistungsabnahme mit der Untersuchungsdauer war jedoch nicht festzustellen: Die Leistung in den letzten 50 min der Sitzung, in denen immer die gleichen Tests vorkamen, war unter allen 3 Bedingungen praktisch identisch. Die Autoren betonen, dass dieses Ergebnis im Einklang mit den Ergebnissen anderer Studien steht. Die Sorge, dass die Ergebnisse in einem Leistungstest nicht mehr typisch für die Testperson sind, wenn sie zuvor andere Leistungstests bearbeitet hat und daher erschöpft ist, scheint also unbegründet zu sein. Entscheidend ist vielmehr, ob sich jemand hinreichend anstrengt und damit ein Nachlassen der Leistungsfähigkeit durch Ermüdung kompensiert.

Belastbarkeit der Testperson beachten

Wenn sich Testpersonen aufgrund einer zu langen Testdauer erschöpft fühlen, kann ihre Compliance und Motivation abnehmen. Deshalb und auch aus ethischen Gründen sollten Testpersonen nicht zu stark belastet werden.

Durchführung einer diagnostischen Untersuchung ...

Welche Belastung zumutbar ist, muss im Einzelfall entschieden werden. Wie belastend eine Testsitzung ist, hängt vom Alter der untersuchten Personen, deren Testfahrung und eventuellen Einschränkungen ab. Bei depressiven Patientinnen und Patienten etwa kann die Grenze der Belastbarkeit schnell erreicht sein. Jedenfalls gilt es, die Untersuchung so zu planen, dass die Testpersonen sie gut bewältigen können. Empfehlenswerte Maßnahmen sind: anstrengende Tests gegen Anfang oder nach Pausen, abwechslungsreicher Ablauf, Einplanen von Pausen, ggf. Verteilung der Untersuchung auf mehrere Termine.

- !** Bei der Auswahl geeigneter Verfahren zur Beantwortung der psychologischen Fragen ist zu beachten, wie wichtig eine psychologische Frage ist, was genau man messen will und ob ein Verfahren für diese Person und für diese konkrete Fragestellung passend ist. Neben den klassischen Testgütekriterien können daher auch Fairness, Akzeptanz, Verfälschbarkeit und Ökonomie/Aufwand relevant sein. Im Gutachten ist zu begründen, warum man sich für ein bestimmtes Verfahren unter den vielen zur Auswahl stehenden entschieden hat. Die Abfolge der Verfahren einschließlich Pausen wird vorab festgelegt, wobei die Belastbarkeit der Untersuchungsperson zu beachten ist. Ökonomische Gründe spielen eine Rolle bei der Entscheidung, ob eine Einzel- oder Gruppenuntersuchung vorgesehen und ob ein Test computer- bzw. internetbasiert (zu Hause oder in den Untersuchungsräumen) durchgeführt wird.

4.5 Durchführung und Auswertung diagnostischer Verfahren

4.5.1 Aufklärung

Für viele Teilnehmerinnen und Teilnehmer an diagnostischen Untersuchungen ist die Untersuchungssituation sehr ungewohnt. Deshalb ist es angebracht, sie zu Beginn über wichtige Aspekte der Untersuchung aufzuklären. Dafür gibt es ethische und auch pragmatische Gründe.

Definition

Ein wichtiges ethisches Prinzip ist die **informierte Einwilligung** (engl. informed consent). Die Teilnahme an einer diagnostischen Untersuchung ist grundsätzlich freiwillig. Die Teilnehmerinnen und Teilnehmer sollen daher vor Beginn der Untersuchung über wichtige Details informiert werden. Sie können dann entscheiden, ob sie sich der Untersuchung unterziehen oder nicht bzw. ob sie die Teilnahme an einem bestimmten Verfahren verweigern.

Gegenstand der Aufklärung können, je nach Untersuchungsanlass und Randbedingungen, folgende Aspekte sein:

- Zweck der Untersuchung; ggf. mit Begründung der Notwendigkeit der diagnostischen Untersuchung
- Wer führt die Untersuchung durch bzw. wer ist daran beteiligt?
- Welche Verfahren kommen zum Einsatz?
- Wie lange dauert die Untersuchung, wann gibt es Pausen?
- Wer erfährt die Ergebnisse?
- Schweigepflicht der beteiligten Personen
- Freiwilligkeit der Untersuchungsteilnahme, ggf. aber auch Konsequenzen einer Nichtteilnahme

Aufklärungsaspekte

Bei Interviews und Persönlichkeitsfragebögen ist Vertraulichkeit wichtig

Mit diagnostischen Interviews und Fragebögen sollen häufig sehr persönliche Informationen erhoben werden. Spitznagel (1982) erklärt den Stellenwert der *Vertraulichkeit* mit der in Alltagssituationen erlernten Regel, sich Fremden gegenüber zurückzuhalten oder sich zumindest nicht zu einem frühen Zeitpunkt zu offenbaren. Dieser Regel zuwiderzuhandeln, bedarf offensichtlich erheblicher Überwindung. Deshalb ist es wichtig, in der Einführungsphase eines Interviews darauf hinzuweisen, dass die Angaben vertraulich behandelt werden (sofern diese Aussage gerechtfertigt ist) bzw. wer von den Gesprächsinhalten erfährt.

Für den eignungsdiagnostischen Bereich wird in der DIN 33430 (DIN 2016 S. 18) festgestellt:

- » Spätestens zu Beginn der Untersuchung – soweit möglich bereits im Rahmen der Einladung – sind Kandidaten über die infrage stehende Tätigkeit [d. h. über die Stelle; Anm. der Autoren], Ziele und Funktion der Eignungsuntersuchung, ihren Ablauf und ihre Dauer sowie über mitwirkende Personen und deren Funktion zu informieren. Ebenso sind die Kandidaten darüber aufzuklären, wie die Verfahrensergebnisse verwendet werden, in welcher Form und wie lange sie aufbewahrt werden und wer von ihnen Kenntnis erlangt.[...] Es muss die Einwilligung der Kandidaten in die Eignungsuntersuchung vor dem Hintergrund dieser Informationen sowie die Zustimmung zur Weitergabe der Verfahrensergebnisse eingeholt werden.

Über wesentliche Aspekte der Untersuchung informieren Datenschutz

Anzumerken ist, dass die DIN 33430 nicht rechtlich verbindlich ist, sondern den Charakter einer Leitlinie hat.

Bei einer diagnostischen Untersuchung werden zwangsläufig personenbezogene Daten erhoben, verarbeitet und gespeichert. Deshalb stellt sich die Frage nach dem *Datenschutz*. Relevant ist die „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr ...“ (kurz: Datenschutz-Grundverordnung oder DSGVO). Das Bundesdatenschutzgesetz (neu) 2018 (kurz: BDSG; ebenfalls anwendbar seit dem 25. Mai 2018) ist dem europäischen Recht nachgeordnet; die DSGVO enthält aber einige sog. „Öffnungsklauseln“, die es den einzelnen europäischen Ländern gestatten, hier spezielle Regelungen zu treffen.

Weiterführende Internetressourcen

Einsicht in die DSGVO können Sie unter ► <https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:32016R0679> nehmen. Das Bundesdatenschutzgesetz ist unter folgendem Link zu finden: ► <https://dsgvo-gesetz.de/bdsg/>.

Bedeutung der DSGVO für die Psychologische Diagnostik

Die DSGVO und das BDSG thematisieren diagnostische Untersuchungen nicht. Wir versuchen, die Bedeutung der DSGVO für die Psychologische Diagnostik anhand von selbst formulierten Fragen ohne Anspruch auf Vollständigkeit abzuschätzen und in Form von eigenen Kommentaren nach bestem Wissen und Gewissen auch zu erläutern. Auf Fragen zum Datenschutz in der „wissenschaftlichen Forschung“ können wir hier nicht eingehen, weil es dazu sehr viele Aussagen gibt.

■ Was sind „personenbezogene Daten“ laut DSGVO?

» Artikel 4 Begriffsbestimmungen

Im Sinne dieser Verordnung bezeichnet der Ausdruck:

1. „personenbezogene Daten“ alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder

Durchführung einer diagnostischen Untersuchung ...

indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, [...] identifiziert werden kann [...].

Kommentar: Praktisch alle in der Psychologischen Diagnostik erhobene Daten (biografische Daten, Befragungsergebnisse, Verhaltensbeobachtungen, Testergebnisse etc.) stellen personenbezogene Daten dar, wenn die untersuchte Person über den Namen, eine Kennnummer, ein Foto etc. identifiziert werden kann (was in der Regel der Fall ist).

■ Gibt es Daten, die als besonders schutzwürdig gelten?

» Artikel 9 Verarbeitung besonderer Kategorien personenbezogener Daten

(1) Die Verarbeitung personenbezogener Daten, aus denen die [...] ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie die Verarbeitung von genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person ist untersagt.

Allerdings wird in Artikel 9 Absatz 2 eine Reihe von Ausnahmen genannt:

» Artikel 9 Verarbeitung besonderer Kategorien personenbezogener Daten

(2) Absatz 1 gilt nicht in folgenden Fällen:

- a. Die betroffene Person hat in die Verarbeitung der genannten personenbezogenen Daten für einen oder mehrere festgelegte Zwecke ausdrücklich eingewilligt, es sei denn, nach Unionsrecht oder dem Recht der Mitgliedstaaten kann das Verbot nach Absatz 1 durch die Einwilligung der betroffenen Person nicht aufgehoben werden,
- b. [...]
- c. die Verarbeitung ist für Zwecke der Gesundheitsvorsorge oder der Arbeitsmedizin, für die Beurteilung der Arbeitsfähigkeit des Beschäftigten, für die medizinische Diagnostik, die Versorgung oder Behandlung im Gesundheits- oder Sozialbereich oder für die Verwaltung von Systemen und Diensten im Gesundheits- oder Sozialbereich auf der Grundlage des Unionsrechts oder des Rechts eines Mitgliedstaats oder aufgrund eines Vertrags mit einem Angehörigen eines Gesundheitsberufs und vorbehaltlich der in Absatz 3 genannten Bedingungen und Garantien erforderlich,
- d. [...]

In Absatz 3 wird dann erläutert, dass die in Absatz 1 genannten Daten für die in Absatz 2, Buchstabe h genannten Zwecke verwendet werden dürfen, wenn diese Daten von Fachpersonal oder unter dessen Verantwortung verarbeitet werden:

» Artikel 9 Verarbeitung besonderer Kategorien personenbezogener Daten

(3) Die in Absatz 1 genannten personenbezogenen Daten dürfen zu den in Absatz 2 Buchstabe h genannten Zwecken verarbeitet werden, wenn diese Daten von Fachpersonal oder unter dessen Verantwortung verarbeitet werden und dieses Fachpersonal nach dem Unionsrecht oder dem Recht eines Mitgliedstaats oder den Vorschriften nationaler zuständiger Stellen dem Berufsgeheimnis unterliegt, oder wenn die Verarbeitung durch eine andere Person erfolgt, die ebenfalls nach dem Unionsrecht oder dem Recht eines Mitgliedstaats oder den Vorschriften nationaler zuständiger Stellen einer Geheimhaltungspflicht unterliegt.

Kommentar: Die in Artikel 9 Absatz 1 genannten Daten sind besonders schutzwürdig und dürfen nur unter ganz bestimmten Zwecken verarbeitet werden. Berufspsychologen und -psychologinnen mit staatlich anerkannter wissenschaftlicher Abschlussprüfung gehören im Strafrecht nach § 203 Strafgesetzbuch (StGB) zu den Berufsgeheimnisträgern/-trägerinnen. Möglicherweise öffnet sich eine Tür zur rechtmäßigen Verarbeitung von personenbezogenen Daten zum Sexualleben oder der sexuellen Orientierung im Rahmen der Versorgung oder Behandlung im Gesundheits- oder Sozialbereich – etwa zwecks Diagnostik für eine Eheberatung, Paartherapie oder bestimmte forensische Fragestellungen. Die anderen in Absatz 1 genannten Daten spielen in der Psychologischen Diagnostik eher keine Rolle.

■ **Welche Rechte haben die betroffenen Personen?**

» **Artikel 12 Transparente Information, Kommunikation und Modalitäten für die Ausübung der Rechte der betroffenen Person**

(1) Der Verantwortliche trifft geeignete Maßnahmen, um der betroffenen Person alle Informationen gemäß den Artikeln 13 und 14 und alle Mitteilungen gemäß den Artikeln 15 bis 22 und Artikel 34, die sich auf die Verarbeitung beziehen, in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln; dies gilt insbesondere für Informationen, die sich speziell an Kinder richten.[...] Die Übermittlung der Informationen erfolgt schriftlich oder in anderer Form, gegebenenfalls auch elektronisch.[...] Falls von der betroffenen Person verlangt, kann die Information mündlich erteilt werden, sofern die Identität der betroffenen Person in anderer Form nachgewiesen wurde.

Die untersuchte Person hat u. a. ein Auskunftsrecht (Artikel 16), ein Recht auf Berichtigung ihrer personenbezogenen Daten (Artikel 16) sowie auf deren Löschung (Artikel 17).

Kommentar: Die untersuchten Personen haben ein Recht, ausführlich, präzise und in verständlicher Form darüber informiert zu werden, wie ihre Daten verarbeitet werden. Wie an anderer Stelle ausgeführt wird, gehört dazu u. a., zu welchem Zweck die Daten verarbeitet und wie lange sie gespeichert werden (Artikel 13 Absatz 1). Das gilt auch, wenn die personenbezogenen Daten nicht bei der untersuchten Person erhoben werden (Artikel 14); also beispielsweise eine Fremdanamnese durchgeführt wird.

■ **In welcher Form muss die betroffene Person der Verarbeitung ihrer Daten zustimmen?**

In Artikel 7 steht:

» **Artikel 7 Bedingungen für die Einwilligung**

(1) Beruht die Verarbeitung auf einer Einwilligung, muss der Verantwortliche nachweisen können, dass die betroffene Person in die Verarbeitung ihrer personenbezogenen Daten eingewilligt hat.

(2) Erfolgt die Einwilligung der betroffenen Person durch eine schriftliche Erklärung, die noch andere Sachverhalte betrifft, so muss das Ersuchen um Einwilligung in verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache so erfolgen, dass es von den anderen Sachverhalten klar zu unterscheiden ist. Teile der Erklärung sind dann nicht verbindlich, wenn sie einen Verstoß gegen diese Verordnung darstellen.

(3) Die betroffene Person hat das Recht, ihre Einwilligung jederzeit zu widerrufen. Durch den Widerruf der Einwilligung wird die Rechtmäßigkeit der aufgrund

Durchführung einer diagnostischen Untersuchung ...

der Einwilligung bis zum Widerruf erfolgten Verarbeitung nicht berührt. Die betroffene Person wird vor Abgabe der Einwilligung hiervon in Kenntnis gesetzt. Der Widerruf der Einwilligung muss so einfach wie die Erteilung der Einwilligung sein.

Kommentar: In Verbindung mit Artikel 12 Absatz 1 (s. o.) ergibt sich hier das vertraute Prinzip der informierten Einwilligung. Weder die Information noch die Einwilligung ist an die schriftliche Form gebunden. Allerdings muss die Psychologin oder der Psychologe nachweisen können, dass eine Einwilligung vorliegt. Am einfachsten ist dieser Nachweis durch eine schriftliche (unterschriebene) Erklärung zu führen.

■ Was ist noch besonders zu beachten?

In Artikel 5 sind einige Grundsätze formuliert:

» Artikel 5 Grundsätze für die Verarbeitung personenbezogener Daten

(1) Personenbezogene Daten müssen

- a. auf rechtmäßige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden („Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz“);
- b. für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; eine Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke gilt gemäß Artikel 89 Absatz 1 nicht als unvereinbar mit den ursprünglichen Zwecken („Zweckbindung“);
- c. dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein („Datenminimierung“);
- d. [...]
- e. in einer Form gespeichert werden, die die Identifizierung der betroffenen Personen nur so lange ermöglicht, wie es für die Zwecke, für die sie verarbeitet werden, erforderlich ist; ... („Speicherbegrenzung“);
- f. in einer Weise verarbeitet werden, die eine angemessene Sicherheit der personenbezogenen Daten gewährleistet, einschließlich Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung durch geeignete technische und organisatorische Maßnahmen („Integrität und Vertraulichkeit“);

(2) Der Verantwortliche ist für die Einhaltung des Absatzes 1 verantwortlich und muss dessen Einhaltung nachweisen können („Rechenschaftspflicht“).

Kommentar: Die verantwortliche Person (die Psychologin oder der Psychologe) muss dafür sorgen, dass personenbezogene Daten rechtmäßig, nur für einen legitimen Zweck und auf das notwendige Maß beschränkt verarbeitet werden. Die Sicherheit der Daten muss durch Begrenzung der Speicherdauer und technische und organisatorische Maßnahmen gewährleistet werden. Wichtig ist, dass die verantwortliche Person nicht nur entsprechende Maßnahmen ergreift, sondern deren Einhaltung ggf. auch nachweisen muss (Rechenschaftspflicht).

Eine Aufklärung über Zweck und Gegenstand der diagnostischen Untersuchung erfüllt nicht nur juristische Vorgaben, sondern hat auch positive Nebeneffekte:

Positive Effekte von Aufklärung

- Eine Untersuchung wird eher als fair wahrgenommen, wenn die Teilnehmerinnen und Teilnehmer über wesentliche Aspekte informiert werden (► Abschn. 4.5.1).
- Die Testangst (► Abschn. 4.5.2) der Teilnehmerinnen und Teilnehmer kann damit reduziert werden.
- Bei Gruppenuntersuchungen sind Zwischenfragen während der Untersuchung störend; deshalb sollten möglichst alle grundsätzlichen Fragen zu Beginn geklärt werden.

Aufklärung bei der Personalauswahl

Wie sich Informationen über die Untersuchung im Kontext der Personalauswahl auswirken, wurde in einer Metaanalyse genauer betrachtet (Truxillo et al. 2009).

Eine Aufklärung zu Beginn der Untersuchung geht damit einher, dass die Untersuchung als fairer wahrgenommen ($r=.12$) und das Unternehmen leicht positiver beurteilt wird ($r=.06$), die Motivation der Teilnehmenden, gute Ergebnisse zu erzielen (test-taking motivation), höher ist ($r=.21$), und in kognitiven Leistungstests die Ergebnisse leicht besser ausfallen ($r=.09$). (Die Angaben in Klammern geben den durchschnittlichen Zusammenhang zwischen der Bedingung „Aufklärung ja – nein“ und der abhängigen Variablen an.)

Truxillo et al. (2009) konnten im Rahmen eines Strukturgleichungsmodells zudem zeigen, dass sich die Aufklärung über eine Erhöhung der Testmotivation positiv auf die Testleistung auswirkt.

4.5.2 Gute Arbeitsbedingungen

Optimale Arbeitsbedingungen

Beim Einsatz von Leistungstests ist es wichtig, den Testpersonen optimale Arbeitsbedingungen zu bieten; schließlich will man ihnen Gelegenheit zur maximalen Entfaltung ihrer Fähigkeiten und Fertigkeiten geben. Wichtige Merkmale der Arbeitsbedingungen sind: genügend Platz, gute Lichtverhältnisse, keine Störungen, angenehme Temperatur, ausreichend Frischluft. Diese und eventuell auch weitere Arbeitsbedingungen werden meist in den Testmanualen im Abschnitt „Durchführung“ genannt.

Störungen während der Untersuchung sind ein besonderes Problem, sie lassen sich aber weitgehend durch einfache Maßnahmen verhindern:

Vermeidung von Störungen

- Schild an der Tür „Untersuchung – bitte nicht stören!“
- Eigenes Telefon und Handy ausschalten bzw. stumme Rufumleitung einrichten
- Teilnehmerinnen und Teilnehmer bitten, ihre Handys auszuschalten
- Pausenzeiten bekannt geben, damit in Gruppenuntersuchungen einzelne Teilnehmerinnen bzw. Teilnehmer nicht den Raum für Toilettenbesuche etc. verlassen müssen

Test- oder Prüfungsangst

Ein Hindernis für die Realisierung maximaler Leistungen ist die Test- oder Prüfungsangst. Sie ist in selbstwertrelevanten Situationen wie einer Testuntersuchung oftmals besonders stark ausgeprägt. Die mangelnde Vertrautheit mit der Test- oder der Prüfungssituation verstärkt den hemmenden Einfluss der Testangst.

Zur Beziehung zwischen Testangst und Testleistung wurden sehr viele Studien durchgeführt (s. auch ▶ Abschn. 3.2.1).

Einfluss von Testangst auf die Testleistung

Hembree fand bereits 1988 über 500 einschlägige Studien, die er metaanalytisch auswertete. Die Testangst wurde mittels Fragebögen gemessen; als Kriterien lagen Noten oder Leistungstestergebnisse vor. Die Korrelation zwischen Testangst und IQ betrug im Durchschnitt $r = -.23$ (für Schulnoten: $r = -.29$). Natürlich sagen Korrelationen nichts über die Kausalität aus: Aus diesen Studien lässt sich daher nicht zwingend schließen, dass sich Testangst negativ auf die Testleistung auswirkt; es wäre auch denkbar, dass die Testangst eine Folge niedriger Fähigkeiten ist.

Hembree (1988) trug deshalb auch Studien zusammen, in denen versucht wurde, die Testangst durch geeignete Interventionen zu reduzieren. Durch systematische Desensitivierung, Entspannungstraining oder Erzeugung von Vertrautheit mit Tests können nicht nur die Testangst reduziert ($d = 0,40, 0,68$ bzw. $0,55$), sondern auch die Testleistungen verbessert werden ($d = 0,32, 0,13$ bzw. $0,26$). Die Ergebnisse sprechen also insgesamt dafür, dass Testangst einen negativen Einfluss auf die Testleistung hat – und dass man diesen Effekt zumindest reduzieren kann.

Bei hoher Testangst niedrigerer IQ

Abbau der Testangst hilft

Bei der Gestaltung der Testsituation sind deshalb Maßnahmen sinnvoll, die sich mildernd auf die Testangst auswirken. In der Regel ist vor der Durchführung von Leistungstests eine *Aufwärmphase* nützlich, die mit Testsituation und -verfahren vertraut macht. Die meisten Tests bieten Übungsaufgaben oder sog. „Eisbrecheraufgaben“ an, durch die die Testpersonen mit den Testaufgaben vertraut gemacht werden, deren Ergebnis aber nicht gewertet wird.

Aufwärmphase

4.5.3 Standardisierung der Untersuchungsbedingung

In den Testmanualen finden sich Anleitungen zur Durchführung des Tests, die unbedingt zu beachten sind. Abweichungen von den Vorgaben führen dazu, dass die Ergebnisse nicht mit denen der Eichstichprobe vergleichbar sind.

Anweisungen in Testmanualen beachten

Auch wenn keine Normdaten verwendet werden, ist es wichtig, die Durchführungsbedingungen für alle Teilnehmerinnen und Teilnehmer *konstant* zu halten. Es wäre ausgesprochen unfair, etwa Bewerberinnen und Bewerber unter unterschiedlichen Bedingungen zu untersuchen. Durch genaue Anweisungen und Schulung muss sichergestellt werden, dass die Untersuchung auch von unterschiedlichen Personen immer auf die gleiche Weise durchgeführt wird. So wird verhindert, dass beispielsweise einmal Nachfragen beantwortet werden und ein andermal nicht oder dass Testanweisungen einmal durch zusätzliche Beispiele erläutert werden und ein andermal nicht.

Durchführungsbedingungen konstant halten

Verantwortung für korrekte Testdurchführung bedeutet

- Treffen der notwendigen Vorbereitungen für die Testdurchführung
- Fachlich kompetente Vorgabe des Tests

Dies sind zentrale Forderungen in den „Internationalen Richtlinien für die Testanwendung“ (ITC 2001), die in den Richtlinien weiter spezifiziert werden.

Sorgfalt ist angebracht

Testauswerteprogramme

4.5.4 Testauswertung

Die manuelle Auswertung von Tests und Fragebögen stellt eine potenzielle Fehlerquelle dar. Es ist daher wichtig, die damit betrauten Personen gründlich zu schulen und ihre Arbeit stichprobenartig zu kontrollieren (Abb. 4.3). Besonders gravierend sind systematische Fehler wie die Verwendung einer falschen Normtabelle oder von Schablonen der Parallelform des Tests oder die instruktionswidrige Verrechnung fehlender Antworten.

Die höchste Sicherheit bietet ein computergestützter Test, bei dem die Antworten der Testperson direkt erfasst und verrechnet werden. Aus verschiedenen Gründen können nicht immer Computertests eingesetzt werden, sodass Alternativen zu erwägen sind. Die Auswertungsobjektivität kann durch Verwendung eines *Testauswerteprogramms* erhöht werden. Die Antworten der Testperson (in der Regel die Nummer der angekreuzten Antwortalternative) werden in den Computer eingegeben, woraufhin das Programm die Rohwerte für jede Skala bzw. den Gesamttest sowie die dazugehörigen Normwerte ermittelt.

Allerdings kann selbst die automatisierte Auswertung durch Einscannen von Testbögen fehleranfällig sein, wie das folgende Beispiel zeigt.

Teure Auswertungsfehler

In den USA werden Studierfähigkeitstests wie der Scholastic Assessment Test (SAT) zur Auswahl von Studierenden eingesetzt. So bearbeiteten etwa 495.000 Schüler den College Board SAT I. Nachdem ihnen die Ergebnisse zugestellt wurden, verlangte eine Reihe von Bewerberinnen und Bewerbern eine manuelle Kontrolle, um die Richtigkeit ihrer Testwerte zu überprüfen. Es zeigte sich, dass tatsächlich Fehler aufgetreten waren, und es bestand der Verdacht, dass ein systematischer Fehler in der Auswertung vorlag. Daraufhin wurden alle Antwortbögen erneut eingescannt. Bei etwa 4000 Testteilnehmenden mussten die Testergebnisse korrigiert werden. Die korrigierten Ergebnisse wurden den Bewerberinnen und Bewerbern sowie den Universitäten und Colleges mitgeteilt. Die nun korrekten Ergebnisse führten dazu, dass sich in vielen Fällen das Eignungsurteil änderte. Die Korrekturen erfolgten zum Glück überwiegend kurz vor der Zulassung, sodass ein größerer Schaden verhindert wurde. Als Ursache des Fehlers konnte ein Scanner in einem Scanner-Zentrum in Texas ausgemacht werden, das von einem Subunternehmen betrieben wurde (Eyde et al. 2010, S. 104; dort wird auch über 2 weitere Pannen bei Zulassungstests berichtet).



Abb. 4.3 Bei der manuellen Auswertung von Tests ist größte Sorgfalt angebracht

4.5.5 Darstellung und Interpretation der Ergebnisse

4.5.5.1 Uneinheitliche Empfehlungen für Wertebereiche

Im Rahmen der Auswertung eines Tests erfolgt eine Einordnung der Testrohwerte anhand einer (repräsentativen) Vergleichsgruppe (vgl. ▶ Abschn. 2.6.4). Erfolgt diese Einordnung in Form von Prozenträngen, so könnte die Erkenntnis entstehen, dass die getestete Person besser oder gleich gut ist wie 65 % der Vergleichsgruppe. Eine solche erste Einordnung bedarf jedoch weiterer Interpretation. (Es ist deshalb auch ein Irrtum, einem Test allein deshalb Interpretationsobjektivität zu bescheinigen, weil Normtabellen vorliegen.) Ist der Wert beispielsweise als durchschnittlich zu beschreiben oder bereits als überdurchschnittlich?

Die Antwort auf diese Frage ist: Es kommt auf die Konvention an. Häufig wird der Bereich von +/– einer Standardabweichung um den Mittelwert, bei Standardwerten also der Bereich von 90 bis 110, als „durchschnittlich“ bezeichnet. Allerdings ist dieser Bereich sehr groß; 68 % der Menschen werden damit „durchschnittliche“ Testleistungen bescheinigt. Daher machen manche Testautorinnen und Testautoren andere Vorschläge. In der Testbatterie Wahrnehmungs- und Aufmerksamkeitsfunktionen (WAF) von Sturm (2008) beispielsweise wird der Durchschnittsbereich durch die mittleren 50 %, also von Prozentrang 25 bis 75 definiert, was einem Standardwertbereich von 93,3 bis 106,7 gleichkommt. Im Wiener Entwicklungstest (WET) von Kastner-Koller und Deimann (2012) werden C-Werte von 4 bis 6 (die Standardwerten von 95 und 105 und damit einer Häufigkeit von 38,3 % entsprechen) als „Normalbereich – altersgemäße Entwicklung“ definiert.

Fehlende Konventionen

Leider reicht auch eine Konvention zur Verbalisierung von Normwerten (in beispielsweise durchschnittlich oder überdurchschnittlich) nicht aus, um ein einheitliches Verständnis herbeizuführen. Die Aussage „Frau Müller beschrieb sich in dem Fragebogen als überdurchschnittlich gewissenhaft“ kann sehr unterschiedlich aufgefasst werden. Was bedeutet hier „überdurchschnittlich“? Das folgende Beispiel zeigt, wie falsch selbst Expertinnen und Experten eine Aussage (hier) zur Häufigkeit interpretieren können.

Fragen Sie Ihren Arzt oder Apotheker

Medikamente können bestimmte Nebenwirkungen haben. Im Beipackzettel ist darauf hinzuweisen. In wie viel Prozent der Fälle tritt eine Nebenwirkung auf, die laut Beipackzettel „häufig“ vorkommt? In einer Umfrage (Ziegler et al. 2013) wurden Ärztinnen und Ärzte sowie Apothekerinnen und Apotheker gebeten, die prozentuale Häufigkeit von Nebenwirkungen anzugeben, wenn diese dem Beipackzettel zufolge „häufig“ vorkommt. Die Schätzungen lagen im Durchschnitt beider Berufsgruppen bei 60 bzw. 50 %. Das Bundesinstitut für Arzneimittelsicherheit und Medizinprodukte (BfArM) hat den Begriff „häufig“ in diesem Kontext definiert als 1 bis <10 %. Ähnlich falsch wurden auch die Begriffe „gelegentlich“ und „selten“ verstanden.

Auch standardisierte Formulierungen werden falsch verstanden

4.5.5.2 Vorschläge für Messwertbereiche

Im Folgenden unterbreiten wir Vorschläge zur Lösung der beiden genannten Probleme, nämlich der Festlegung von Bereichen in der Normalverteilung und der verständlichen „Übersetzung“ der Begriffe (z. B. „durchschnittlich“), mit denen die Bereiche benannt werden.

In der Diagnostikausbildung am Fachbereich Psychologie der Universität Marburg hat sich ein Modell bewährt, das sich durch folgende Eigenschaften auszeichnet:

- Zur Klassifikation der Ausprägung eines Merkmals werden 5 Bereiche unterschieden.
- Der mittlere Bereich („durchschnittlich“) umfasst den Mittelwert $+/-$ eine halbe (!) Standardabweichung, bei Standardwerten also den Bereich von 95 bis 105. Damit werden etwa 38 % der Personen als durchschnittlich klassifiziert. Die weiteren Bereiche umfassen jeweils eine ganze Standardabweichung (z. B. hoch = >105 bis 115). Natürlich kann das Modell auch auf andere Normwerte angewandt werden, denen eine Normalverteilung zugrunde liegt (► Tab. 4.1).
- Die Merkmalsausprägung wird zudem bei Bedarf auf einer 5-stufigen Skala visualisiert, was insbesondere bei mehrdimensionalen Verfahren angebracht ist (► Abb. 4.4).
- Die Konfidenzintervalle (► Abschn. 2.6.2.2) werden dann ebenfalls visualisiert (► Abb. 4.4).
- Bei der Verbalisierung wird immer Bezug auf die herangezogene Referenzgruppe genommen.

„Marburger Modell“

In ► Tab. 4.1 werden die nach dem „Marburger Modell“ vorgeschlagenen Messwertbereiche aufgeführt.

„Standardmodell“

Das Modell mit der Festlegung des Durchschnittsbereichs auf $+/-$ einer Standardabweichung um den Mittelwert ist, wie bereits oben erwähnt, sehr verbreitet. Wir stellen es deshalb hier auch vor (► Tab. 4.2).

Stanine-Werte verbalisieren

Die vorgestellten Verbalisierungsvorschläge können leider nicht ohne Weiteres auf Stanine-Werte übertragen werden. Beim Zusammenlegen von Kategorien entstehen Intervalle, die weder mit denen des „Marburger Modells“ noch denen des „Standardmodells“ übereinstimmen. Für Stanine-Werte schlagen wir vor, den Durchschnittsbereich wie im FPI-R (► Abschn. 3.3.3.3) durch die Stanine-Werte 4 bis 6 zu definieren. Damit gelten die Testwerte von 54 % aller Personen als durchschnittlich. Wir schlagen vor, die angrenzenden

► Tab. 4.1 Vorschlag zur Festlegung der Messwertbereiche nach dem „Marburger Modell“

Ausprägung	z	Zugehöriger Normwertebereich				Häufigkeit
		SW	IQ	T	Sten	
„sehr hoch“	$>1,5$	>115	$>122,5$	>65	1–2	6,7 %
„hoch“	0,51 bis 1,5	105,01–115	107,51–125,5	55,01–65	3–4	24,2 %
„durchschnittlich“	−0,5 bis 0,5	95–105	92,5–107,5	45–55	5–6	38,3 %
„niedrig“	−1,5 bis −0,51	85–94,99	77,5–92,49	35–44,99	7–8	24,2 %
„sehr niedrig“	$<-1,5$	<85	$<77,5$	<35	9–10	6,7 %

Die Häufigkeiten (rechte Spalte) addieren sich rundungsbedingt nicht exakt auf 100 %. Das Wesentliche an diesem Modell ist die Festlegung der Breite der 5 Ausprägungsbereiche. Zu deren Benennung könnte man auch andere Begriffe, beispielsweise die aus ► Tab. 4.2, verwenden. Wichtig ist es, die gewählten Begriffe im Gutachten einheitlich zu verwenden und nicht etwa aus stilistischen Gründen zu verändern (z. B. kein Wechsel zwischen „hoch“ und „überdurchschnittlich“).

Tab. 4.2 Vorschlag zur Festlegung der Messwertbereiche nach dem „Standardmodell“

Ausprägung	z	Zugehöriger Normwertebereich				Häufigkeit
		SW	IQ	T	Sten	
„weit überdurchschnittlich“	> 2	> 120	> 130	> 70	1	2,3 %
„überdurchschnittlich“	> 1 bis 2	110,01–120	115,01–130	60,1–70	2–3	13,6 %
„durchschnittlich“	-1 bis 1	90–110	85–115	40–60	4–7	68,3 %
„unterdurchschnittlich“	< -1 bis -2	80–89,99	70–84,99	30–39,99	8–9	13,6 %
„weit unterdurchschnittlich“	< 2	< 80	< 70	< 30	10	2,3 %

Siehe Fußzeile in **Tab. 4.1**.**Tab. 4.3** Vorschlag zur Festlegung der Messwertbereiche von Stanine-Werten

	„sehr niedrig“	„niedrig“	„durchschnittlich“	„hoch“	„sehr hoch“
Stanine	1	2–3	4–6	7–8	9
Häufigkeit	4 %	19 %	54 %	19 %	4 %
z	< -1,75	-0,7501 bis -1,75	-0,75 bis 0,75	0,7501 bis 1,75	> 1,75

Für die Verbalisierung der Messwertbereiche gelten die Ausführungen in der Fußzeile in **Tab. 4.1**.

Bereiche über Stanine 2 und 3 und anschließend 1 bzw. 7 und 8 und anschließend 9 zu definieren (**Tab. 4.3**). Zur Bestimmung der Konfidenzintervalle sei auf ▶ Abschn. 2.6.2.2 verwiesen.

4.5.5.3 Vorschlag zur Erläuterung der Messwertbereiche

Wie geht man am besten mit dem Problem um, dass auch per Konvention in ihrer Bedeutung festgelegte Begriffe oft falsch verstanden werden? Was bedeutet beispielsweise „durchschnittlich“? Der Versuch, Testergebnisse für Laien (!) in einer Anmerkung etwa so zu erläutern: „durchschnittlich bedeutet, dass das Ergebnis im Bereich von +/– einer Standardabweichung um den Mittelwert liegt“, trägt kaum zur Klärung bei. Auf Beipackzetteln von Medikamenten findet man manchmal eine Tabelle, in der die Begriffe „häufig“, „gelegentlich“, „selten“ etc. durch konkrete Häufigkeitsangaben erläutert werden. Diese Lösung lässt sich auf die Messwertbereiche für die Merkmalsausprägung übertragen (**Tab. 4.4** und **4.5**). Für Stanine-Werte lässt sich unter Verwendung der Angaben in **Tab. 4.3** eine entsprechende Interpretationshilfe erstellen.

Die Mitteilung eines Testergebnisses sollte 4 Kernelemente enthalten:

- Das Merkmal (z. B. „Lebenszufriedenheit“) wird benannt.
- Die Ausprägung des Merkmals (des beobachteten Wertes) wird auf einheitliche Weise sprachlich eingeordnet (z. B. „hoch“). Dazu kann man sich am „Marburger Modell“ (**Tab. 4.1**) oder dem „Standardmodell“ (**Tab. 4.2**) oder auch dem Vorschlag für Stanine-Werte (**Tab. 4.3**) orientieren. Wenn gewünscht, können die verwendeten Begriffe zur Merkmalsausprägung („sehr hoch“ etc.) zusätzlich erläutert werden; bei häufiger Verwendung der Begriffe geschieht dies am besten in tabellarischer Form (**Tab. 4.4** und **4.5**).
- Die Referenzgruppe (z. B. „etwa gleichaltrige Frauen“) wird erwähnt.
- Das Konfidenzintervall wird mitgeteilt (z. B. „kann unter Berücksichtigung der Messgenauigkeit auch sehr hoch sein“).

4 Kernelemente der mitgeteilten Ergebnisse

Tab. 4.4 Vorschlag zur Erläuterung der Messwertbereiche im „Marburger Modell“

Formulierung	Testwert	Häufigkeit	Von 100 Personen, die diesen Test bearbeiten, erreichen ...
„sehr hoch“	> 115	7 %	etwa 7 % ebenfalls sehr hohe Werte; etwa 93 % erreichen niedrigere Werte
„hoch“	105,1–115	24 %	etwa 24 % ebenfalls hohe Werte; etwa 7 % erreichen noch höhere und etwa 69 % niedrigere Werte
„durchschnittlich“	95–105	38 %	etwa 38 % ebenfalls durchschnittliche Werte; etwa 31 % erreichen niedrigere und etwa 31 % höhere Werte
„niedrig“	85–94,9	24 %	etwa 24 % ebenfalls niedrige Werte; etwa 69 % erreichen höhere und etwa 7 % noch niedrigere Werte
„sehr niedrig“	< 85	7 %	etwa 7 % ebenfalls sehr niedrige Werte; etwa 93 % erreichen höhere Werte

Erläuterungsbeispiel: Als Vergleichsgruppe dienen hier für die Gesamtbevölkerung weitgehend repräsentative Personen im Alter von 20 bis 40 Jahren. Die Testwerte sind Standardwerte. (Bei Bedarf hier andere Normwerte mit den zugehörigen Normwertbereichen aufzuführen).

Tab. 4.5 Vorschlag zur Erläuterung der Messwertbereiche im „Standardmodell“

Formulierung	Testwert	Häufigkeit	Von 100 Personen, die diesen Test bearbeiten, erreichen ...
„weit überdurchschnittlich“	> 120	2 %	etwa 2 % ebenfalls sehr hohe Werte; etwa 98 % erreichen niedrigere Werte
„überdurchschnittlich“	110,1–120	14 %	etwa 14 % ebenfalls hohe Werte; etwa 2 % erreichen noch höhere und etwa 84 % niedrigere Werte
„durchschnittlich“	90–110	68 %	etwa 68 % ebenfalls durchschnittliche Werte; etwa 16 % erreichen niedrigere und etwa 16 % höhere Werte
„unterdurchschnittlich“	80–89,9	14 %	etwa 14 % ebenfalls niedrige Werte; etwa 84 % erreichen höhere und etwa 2 % noch niedrigere Werte
„weit unterdurchschnittlich“	< 80	2 %	etwa 2 % ebenfalls sehr niedrige Werte; etwa 98 % erreichen höhere Werte

Erläuterungsbeispiel: Als Vergleichsgruppe dienen hier für die Gesamtbevölkerung weitgehend repräsentative Personen im Alter von 20 bis 40 Jahren. Die Testwerte sind Standardwerte. (Bei Bedarf sind hier andere Normwerte mit den zugehörigen Normwertbereichen aufzuführen).

Für Laien Fließtext ohne Fachbegriffe

Konfidenzintervalle wurden bereits in ► Abschn. 4.3 und in ► Abschn. 2.6.2.2 eingehend thematisiert. Für die Gutachtenerstellung ist Folgendes zu beachten: Begriffe wie „Konfidenzintervall“ oder „Vertrauensintervall“, „Standardwert“, „Reliabilität“ oder „Standardabweichung“ haben für Laien keine klare Bedeutung und sollten vermieden werden. Auch alle Versuche, den Sachverhalt zu erläutern, führen meist eher zur Verwirrung. Der negativ assoziierte Begriff „Messfehler“ wird nicht verwendet – mit „Messgenauigkeit“ lässt sich der Sachverhalt genauso gut beschreiben. Da die Ergebnisdarstellung für Expertinnen und Experten nachvollziehbar sein muss, sollten diese Begriffe bzw. Abkürzungen in Klammern gesetzt werden. Im folgenden Beispiel ist das geschehen.

► Beispiel

Formulierungsvorschlag mit 4 Kernelementen

„In diesem Fragebogen wurden die Angaben der Probandin mit denen etwa gleichaltriger Frauen verglichen. Dem Ergebnis zufolge (Stanine 8) berichtete Frau Müller eine hohe Lebenszufriedenheit, die unter Berücksichtigung der Messgenauigkeit ($\alpha = .76$) auch sehr hoch oder durchschnittlich sein kann (Konfidenzintervall von 6 bis 9; zweiseitige Fragestellung, Urteilsicherheit: 90 %).“ ◀

Durchführung einer diagnostischen Untersuchung ...

Expertinnen und Experten sollte man mitteilen, wann die Begriffe „hoch“ etc. verwendet werden. Mit einer Information am Anfang eines längeren Ergebnisberichts stellt man klar, dass die Konvention im ganzen Dokument einheitlich verwendet wird. Liegen nur zu einem Verfahren Ergebnisse vor, erfüllt eine Fußnote die gleiche Funktion. Die Fußnote kann z. B. lauten: „Die Angaben zur Merkmalsausprägung beziehen sich auf die Position der Person in einer Normalverteilung (z-Werte); durchschnittlich = Mittelwert $\pm 0,5$ Standardabweichungen (SD), niedrig = $-0,5$ bis $-1,5$ SD, sehr niedrig = $\leq -0,5$ SD, hoch = $0,5$ bis $1,5$ SD, sehr hoch = $\geq 1,5$ SD vom Mittelwert.“ Für die Abfassung des Ergebnisberichts können in □ Tab. 4.1 für unterschiedliche Normwerte die passenden Verbalisierungen nach dem „Marburger Modell“ und in □ Tab. 4.2 nach dem „Standardmodell“ sowie in □ Tab. 4.3 für Stanine-Werte nachgeschlagen werden.

Häufig enthält ein Ergebnisbericht viele Normwerte, die zu erläutern sind. Bei einer Aneinanderreihung vieler ähnlicher Formulierungen zu einzelnen Ergebnissen wird der Text schnell schwer lesbar. Deshalb empfehlen wir bei umfangreichen Ergebnissen eine tabellarische Darstellung wie in □ Abb. 4.4. Wir beschränken uns hier aus Platzgründen auf 3 Skalen und erläutern das Vorgehen im Detail. Im Anschluss zeigen wir, wie eine Verbalisierung der gleichen Ergebnisse ohne Tabelle aussehen kann.

Übersetzung von Normwerten mit z-Werten erläutern

Merkmal	Testergebnis	Ausprägung	Berechnungsgrundlagen ¹		
			Reliabilität	Konfidenzintervall von ... bis	
Schlussfolgern-des Denken	IQ = 122	1 2 3 4 5	,95	115,5	126,3
Konzentrations-fähigkeit	SW = 102	1 2 3 4 5	,98	99,7	104,3
Gewissenhaftig-keit	T = 38	1 2 3 4 5	,84	33,9	46,0

□ Abb. 4.4 Tabellarische Ergebnisdarstellung nach dem „Marburger Modell“. Die Testwerte wurden im Vergleich zu etwa gleichaltrigen Menschen in 5 Bereiche eingeordnet: 1=sehr niedrig ($z < -1,5$), 2=niedrig ($z = -0,51$ bis $-1,5$), 3=durchschnittlich ($z = -0,5$ bis $0,5$), 4=hoch ($z = 0,51$ bis $1,5$), 5=sehr hoch ($z > 1,5$). Das Testergebnis fällt in das schwarze Kästchen. Unter Berücksichtigung der Messgenauigkeit könnte die Merkmalsausprägung auch in dem Bereich der blauen Kästchen liegen. ¹ Reliabilität: Cronbachs α für die Altersgruppe. Konfidenzintervall: Angabe der unteren und oberen Grenze; Urteilsicherheit: 90 %, zweiseitige Fragestellung; Berechnung unter Berücksichtigung der Regression zur Mitte. IQ=IQ-Wert, SW=Standardwert, T=T-Wert

Der Aufbau soll kurz erklärt werden, damit Leserinnen und Leser bei Bedarf selbst eine solche Tabelle erstellen können:

1. Tragen Sie in den beiden ersten Spalten die Merkmale (meist sind das die Skalenbezeichnungen) mit den zugehörigen Testergebnissen ein. Wenn alle Ergebnisse in einer Metrik (z. B. Standardwerte) vorliegen, kann diese auch im Kopf der Spalte in Klammern benannt werden (Testergebnis [SW]).
2. Die „Berechnungsgrundlagen“ sind bewusst in einer kleineren Schrifttype dargestellt. Sie dienen der Nachvollziehbarkeit für Expertinnen und Experten. Tragen Sie dort zunächst die Reliabilitäten der Skalen ein.

3. Bestimmen Sie unter Verwendung der Testwerte und der Reliabilitäten die Konfidenzintervalle. Mithilfe des Normwertrechners (► <https://www.psychometrika.de/normwertrechner.html>) geht das sehr einfach. Tragen Sie die Werte ein; bei einseitiger Fragestellung wird nur ein Wert eingetragen (z. B. „bis ...“).
4. Nennen Sie die Details zur Berechnung der Konfidenzintervalle in den Anmerkungen: Welche Reliabilitätsschätzung wurde verwendet? Welche Urteilssicherheit (im Normwertrechner = „Koeffizient“) wurde gewählt? Handelt es sich um eine ein- oder zweiseitige Fragestellung? Welche Methode zur Schätzung der Konfidenzintervalle (der Normwertrechner verwendet die Regressionsmodell) wurde verwendet?
5. Jetzt werden die Merkmalsausprägungen grafisch veranschaulicht. Jeder Bereich wird durch ein Kästchen symbolisiert. Stellen Sie je nach Modell (Tab. 4.1, Tab. 4.2 oder bei Stanine-Werten Tab. 4.3) fest, in welchen Bereich der Testwert fällt und markieren Sie das entsprechende Feld z. B. schwarz. Nach dem „Marburger Modell“ gilt ein IQ von 122 als „hoch“, deshalb ist in Abb. 4.4 das entsprechende Feld schwarz.
6. Analog verfahren Sie mit den oberen und unteren Grenzen der Konfidenzintervalle. Stellen Sie nach dem gewählten Modell jeweils fest, in welchen Bereich das obere und untere Konfidenzintervall fällt, und markieren Sie die entsprechenden Felder z. B. blau. Wenn das Feld bereits für den beobachteten Testwert markiert worden ist, entfällt dieser Schritt. Nach dem „Marburger Modell“ fällt ein IQ von 115,5 ebenfalls in den Bereich „hoch“. Deshalb ist in Abb. 4.4 keine zusätzliche Markierung nötig. Das obere Ende des Konfidenzintervalls liegt in unserem Beispiel mit einem IQ von 126,3 im „sehr hohen“ Bereich. Deshalb wurde hier der Bereich 5 blau markiert.
7. In den Anmerkungen machen Sie noch Angaben zur gewählten Referenzgruppe (im Beispiel „etwa gleichaltrige Menschen“) und zur Definition der Merkmalsbereiche. In Abb. 4.4 wurden die Bereiche in z-Werten angegeben, weil die Testwerte in unterschiedlichen Metriken angegeben wurden.

Erläuterungen zur tabellarischen Darstellung

Tabellarisch dargestellte Ergebnisse im Fließtext kurz erläutern

Inhaltlich ist zu den exemplarisch dargestellten Ergebnissen anzumerken, dass beim schlussfolgernden Denken der Testwert am oberen Ende des „hohen“ Bereichs liegt und das Konfidenzintervall daher (bei $\alpha=.95$) nur den nach oben angrenzenden „sehr hohen“ Bereich tangiert. Bei der Konzentrationsfähigkeit liegen die Grenzen des Konfidenzintervalls aufgrund der hohen Reliabilität ($\alpha=.98$) und weil der Testwert nicht am Rand des Durchschnittsbereichs liegt, im gleichen Bereich wie der Testwert. Bei der Gewissenhaftigkeit „rächt“ sich die Verwendung einer Skala mit einer relativ niedrigen Reliabilität; das Konfidenzintervall tangiert beide angrenzenden Bereiche.

Die hier beschriebene Art der Darstellung erlaubt es, die Ergebnisse eines mehrdimensionalen Tests, aber auch die von verschiedenen Tests mit unterschiedlichen Normwerten in einer Tabelle aufzuführen. Laien finden in Spalte 1 und 3 alle für sie relevanten Informationen. Bei einer tabellarischen Ergebnisdarstellung wird empfohlen, im Fließtext globale Aussagen zu machen und die tabellarische Darstellung zu erläutern. Beispielsweise könnte die Erläuterung lauten: „Die Ergebnisse sind in Tab. X zusammenfassend dargestellt. Man sieht, in welchem Bereich (z. B. „hoch“) ein Testergebnis liegt (schwarze Felder) und in welchem Bereich es auch liegen könnte (blaue Felder). Die Ergebnisse sprechen dafür, dass Herr Klug im Vergleich zu anderen etwa

Durchführung einer diagnostischen Untersuchung ...

gleichaltrigen Menschen eine hohe Fähigkeit zum schlussfolgernden Denken, eine durchschnittliche Konzentrationsfähigkeit und eine niedrige Gewissenhaftigkeit hat. Bei Berücksichtigung der Messgenauigkeit der Tests könnten das schlussfolgende Denken auch ‚sehr hoch‘ und die Gewissenhaftigkeit sowohl ‚sehr niedrig‘ als auch ‚durchschnittlich‘ sein (Sicherheitswahrscheinlichkeit von 90 %).“

Würde man die gleichen Ergebnisse ohne Hilfe einer Tabelle beschreiben, könnte sich das wie folgt lesen: „Herr Klug erreichte beim schlussfolgernden Denken einen hohen Wert ($IQ=122$), der unter Berücksichtigung der Messgenauigkeit des Tests auch ‚sehr hoch‘ sein könnte (Konfidenzintervall = 115,5 bis 126,3; alle Konfidenzintervalle nach dem Regressionsmodell unter Verwendung von Cronbachs α bei zweiseitiger Fragestellung mit einer Sicherheitswahrscheinlichkeit von 90 % berechnet). Seine Konzentrationsfähigkeit liegt dem Test zufolge – auch bei Berücksichtigung der Messgenauigkeit – im durchschnittlichen Bereich (Standardwert = 102, Konfidenzintervall = 99,7 bis 104,3). Sein Ergebnis bei der Gewissenhaftigkeit spricht für eine niedrige Ausprägung ($T\text{-Wert}=38$), die unter Berücksichtigung der Messgenauigkeit auch sehr niedrig oder durchschnittlich (Konfidenzintervall = 33,9 bis 46,0) sein kann. Zum Vergleich wurden stets die Ergebnisse von etwa gleichaltrigen Menschen herangezogen“. Durch die notwendigen Angaben für Expertinnen und Experten in Klammern ist der Text schon bei 3 Merkmalen nicht leicht zu verstehen. Es fehlt hier zudem die Erläuterung der Begriffe zur Ausprägung („durchschnittlich“ etc.), für die man eine Fußnote anlegen müsste. Bei vielen Einzelergebnissen fällt es den Leserinnen und Lesern oft schwer, das Wesentliche zu erfassen.

! Bei der Darstellung und der Interpretation der Ergebnisse ist es wichtig, eine für Laien verständliche Form zu finden und den Bericht zugleich für Expertinnen und Experten nachvollziehbar zu gestalten. Hilfreich sind dabei

- Klare und exakte Angabe der Testergebnisse
- Angemessene Interpretation der Testergebnisse
- Angemessene und einheitliche Benennung der Merkmale
- Interpretation der Ausprägung der Merkmale nach einem festen Schema
- Berücksichtigung der Konfidenzintervalle
- Einheitliche Benennung und verständliche Erläuterung der Wertebereiche
- Tabellarische Darstellung bei vielen Einzeltestergebnissen
- Verständliches Abfassen von Text und Tabellen für Laien
- Für Expertinnen und Experten Angabe aller nötigen Informationen in Klammern und Tabellenerläuterungen

Die beiden ersten Punkte sind zentrale Forderungen in den „Internationalen Richtlinien für die Testanwendung“ (ITC 2001), die in den Richtlinien weiter spezifiziert werden.

Weiterführende Literatur und Internetressourcen

Eine Übersicht über die zahlreichen Standards und Richtlinien zur Testanwendung wird vom ZPID ([► www.zpid.de/redact/category.php?cat=88](http://www.zpid.de/redact/category.php?cat=88)) gepflegt. Von den dort aufgeführten Quellen sind vielleicht die von mehreren amerikanischen Berufsverbänden herausgegebenen *Standards for Educational and Psychological Testing* (AERA et al. 2014) und die auch in deutscher Sprache vorliegenden International Guidelines for Test Use (ITC 2001) besonders hervorzuheben. Das ZPID ([► https://leibniz-psychology.org/](https://leibniz-psychology.org/)) bietet auch einen Zugriff auf die Suchmaschine PubPsych, mit deren Hilfe verschiedene Datenbanken nach Informationen über diagnostische Verfahren durchsucht werden können. In dem Buch von Eyde et al. (2010) finden sich zahlreiche Beispiele für Fehler im Umgang mit Test, die jeweils analysiert werden.

4.6 Das psychologische Gutachten

Definition

„Ein **psychologisches Gutachten** dokumentiert ein wissenschaftlich fundiertes Vorgehen und beantwortet eine von einer Auftraggeberin/einem Auftraggeber vorgegebene Fragestellung (oder mehrere Teilfragestellungen). Die Fragestellung betrifft bestimmte Aspekte des Erlebens und Verhaltens von einer Person oder mehreren Personen. Die Fragestellung wird im Rahmen des nachfolgend beschriebenen diagnostischen Prozesses beantwortet. Im Gutachten werden dieser Prozess und die Beantwortung der Fragestellung nachvollziehbar dargestellt. Die im Rahmen der Begutachtung eingesetzten Methoden werden so beschrieben, dass sie nach wissenschaftlich akzeptierten Gütekriterien beurteilt werden können“ (Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen 2017, S. 2).

Begutachtungsprozess bis auf Befund und Stellungnahme bereits behandelt

In dieser Definition wird auf den diagnostischen Prozess verwiesen, der von der Fragestellung bis zu deren Beantwortung reicht. Dieser Prozess wurde zu Beginn dieses Kapitels grafisch dargestellt (► Abb. 4.1). Eine Begutachtung ist identisch mit dem in ► Abb. 4.1 beschriebenen diagnostischen Prozess. In einem Gutachten wird dieser Prozess transparent und nachvollziehbar dargestellt; es ist der Bericht an die Auftraggeberin oder den Auftraggeber. In ► Abb. 4.1 wird die Beantwortung der Fragestellung (wie auch in einem Gutachten) „Stellungnahme“ genannt. In ▶ Abschn. 4.2 bis 4.5 wurde bereits der größte Teil des Begutachtungsprozesses behandelt – und zwar von der Auftragsannahme bis zur Interpretation der einzelnen Ergebnisse. Diese Etappen der Begutachtung sind Gegenstand des Gutachtens. In der schematischen Übersicht über den Begutachtungsprozess in ► Abb. 4.1 sind wir nun beim Befund und der Stellungnahme angelangt. Auch diese Teile gehören unbedingt zu einem Gutachten.

Wir werden im Folgenden zunächst 3 inhaltliche Themen des Begutachtungsprozesses behandeln: Die beiden ersten inhaltlichen Themen sind der Befund (► Abschn. 4.6.1) und die Stellungnahme (► Abschn. 4.6.2). Manchmal führt der Begutachtungsprozess nicht zum gewünschten Erfolg, etwa weil zwischen einzelnen Ergebnissen unauflösbare Widersprüche vorliegen. Wir stellen deshalb die „Metafrage“: Was ist zu tun, wenn der Begutachtungsprozess nicht erfolgreich verlaufen ist (► Abschn. 4.6.3)? Danach gehen wir auf die formale Gestaltung eines Gutachtens ein (► Abschn. 4.6.4). Abschließend stellen wir die Frage, wie man die Qualität eines Gutachtens beurteilen kann (► Abschn. 4.6.5).

4.6.1 Der Befund

Interpretation der Ergebnisse zu einer psychologischen Frage

In einem Gutachten werden Fakten (Ergebnisse) und deren Bewertung bzw. Interpretation für die Leserin oder den Leser des Gutachtens erkennbar getrennt. Die Interpretation der Ergebnisse zwecks Beantwortung der psychologischen Fragen erfolgt im Befund (den man auch Interpretation der Ergebnisse nennen kann). Man könnte hier einwenden, dass bereits bei der Ergebnisdarstellung empfohlen wurde, die Ergebnisse durch Bezugnahme auf Normwerte zu interpretieren. An einem Beispiel soll der Unterschied zwischen der Angabe und Erläuterung von Normwerten und einer Ergebnisinterpretation zur Beantwortung einer psychologischen Frage verdeutlicht werden.

► Beispiel

Die Fragestellung in einem Gutachten lautet, ob bei Marco S. (Schüler, 8 Jahre) eine Intelligenzminderung vorliegt. Daraus wurde u. a. die psychologische Frage abgeleitet, ob Marcos IQ unter 70 liegt (ein etabliertes Kriterium einer Intelligenzminderung). In dem eingesetzten Intelligenztest erreichte Marco einen IQ von 64. Dieser wurde im Ergebnisbericht so interpretiert: „Der IQ-Wert von 64 spricht dafür, dass Marcos Intelligenz im Vergleich zu gleichaltrigen Kindern als sehr niedrig zu beurteilen ist. Auch unter Berücksichtigung der Messgenauigkeit des Tests ($\alpha = .97$, Sicherheitswahrscheinlichkeit = 95 %, einseitige Fragestellung) kann sein IQ-Wert auch bis zu 69 betragen, womit er weiter im sehr niedrigen Bereich liegt.“ Im Befund wird dieses Ergebnis unter Berücksichtigung weiterer Ergebnisse dahingehend interpretiert, dass mit dem Testergebnis seine Intelligenz unterschätzt wird. Das Ergebnis spricht zwar für eine sehr niedrige Intelligenz mit einem IQ unter 70, ist aber nicht als Beleg für eine Intelligenzminderung zu werten. Dafür sprechen die Verhaltensbeobachtung („wirkte unmotiviert“) und das diagnostische Interview („fühlte sich mit dem Test sehr an schulische Aufgaben erinnert und hatte deshalb ‚keine Lust‘; „fasste Fragen gut auf und beantwortete sie für sein Alter angemessen differenziert“). ◀

Ziel ist es, die psychologischen Fragen zu beantworten. Dazu werden alle verfügbaren Informationen herangezogen, also nicht nur die selbst gewonnenen Ergebnisse, sondern bei Bedarf auch Informationen aus vorliegenden Quellen (der Vorgeschichte). Es gilt, die Informationen aus mehreren Quellen zu integrieren, um eine Antwort auf eine psychologische Frage zu finden. So können beispielsweise Angaben über intellektuelle Fähigkeiten aus Zeugnissen (Vorgeschichte), einem Intelligenztest und der Verhaltensbeobachtung vorliegen, die dann zusammengefasst und aufeinander bezogen werden müssen. Übereinstimmungen werden ebenso erwähnt wie widersprüchliche Ergebnisse. Die Gründe für beobachtete Widersprüche können in den methodischen Besonderheiten der Verfahren (z. B. Selbst- vs. Fremdbeurteilung, eingeschränkter Geltungsbereich von Testergebnissen) und in den Durchführungsbedingungen (z. B. Ermüdung der Testperson) liegen und verdienen unbedingt eine Erörterung.

Damit alle psychologischen Fragen unter Verwendung aller dafür relevanter Informationen beantwortet werden, empfehlen wir zur Vorbereitung dieses Abschnitts die Erstellung einer als „Befundbogen“ bezeichneten Übersicht (Tab. 4.6). Der Befundbogen liefert einen vollständigen Überblick über die vorliegenden Ergebnisse und hilft sicherzustellen, dass keine Informationen unberücksichtigt bleiben und dass Übereinstimmungen und Widersprüche erkannt werden.

In der linken Spalte werden alle vorhandenen Datenquellen genannt. Die Nummerierung dient dazu, Verweise anzubringen (z. B. bei Übereinstimmungen „1, 4, 5“). Auf der rechten Seite stehen im Tabellenkopf die psychologischen Fragen, am besten in Form von Stichworten. Anstelle von „Bereich 1“ in Tab. 4.6 könnte hier „kognitiver Bereich“ stehen; Frage A könnte „Konzentration“ lauten, Frage B „Intelligenz“ etc. In die Zellen trägt man zunächst die Informationen stichpunktartig ein. Viele Zellen werden dabei leer bleiben, weil jede Datenquelle (Methode) in der Regel nur zu einer oder zu wenigen psychologischen Fragen Ergebnisse liefert. Das im Beispiel oben genannte Ergebnis zum IQ etwa würde in die Zeile 3. Intelligenztest (Spalte Intelligenz) als „IQ 64 (obere Konfidenzintervallgrenze von 69)“ eingetragen und in Zelle Interview/Intelligenz würde „keine Lust, guter sprachl. Ausdruck“ stehen. Auch die Kommentare zu Übereinstimmungen und Widersprüchen und das Fazit werden stichwortartig eingetragen.

Die psychologischen Fragen beantworten

Befundbogen als Hilfsmittel

Tab. 4.6 Aufbau eines Befundbogens

Datenquelle	Psychologische Fragen				
	Bereich 1			Bereich 2	
	Frage A	Frage B	Frage C	Frage D	Frage E
1. Akten					
2. Interview					
3. Intelligenztest					
etc.					
Übereinstimmungen					
Widersprüche					
Fazit					

Fakten und deren Bewertung trennen

Individuelle Gültigkeit der Ergebnisse beachten

Liegen auch andere Interpretationen nahe?

Antwort auf die Fragestellung

Beim Abfassen des Befunds bietet es sich an, die Struktur der psychologischen Fragen beizubehalten, also beispielsweise eine Unterteilung in „kognitiver Bereich“, „Persönlichkeit“, „soziale Bedingungen“ vorzunehmen. Um den Befund transparent und nachvollziehbar zu gestalten, soll deutlich werden, welche psychologische Frage gerade beantwortet wird und auf welche Ergebnisse Bezug genommen wird. Das Fazit (die Antwort auf eine psychologische Frage) kann am Ende eines Absatzes stehen. Eine gute Alternative besteht darin, jeden Abschnitt mit einer Antwort auf die psychologische Frage zu eröffnen. Beispielsweise kann unter der Überschrift „allgemeine intellektuelle Leistungsfähigkeit“ stehen: „Insgesamt sprechen die Ergebnisse dafür, dass Herr C. über eine hohe Allgemeine Intelligenz verfügt.“ Im Anschluss daran werden die Belege genannt, die dieses Fazit unterstützen. Dem Fazit (scheinbar) widersprechende Ergebnisse (beispielsweise schlechte Schulleistungen) werden genannt, und es wird dargelegt, warum diese keinen echten Widerspruch zu der zuvor getroffenen Aussage darstellen. Beispielsweise könnten die schlechten Schulleistungen auf häufiges krankheitsbedingtes Fehlen im Unterricht zurückzuführen sein.

Zu einer angemessenen Interpretation der Ergebnisse gehört es, deren individuelle Gültigkeit zu beachten. Ist die individuelle Gültigkeit vielleicht dadurch eingeschränkt, dass die Testperson die Instruktion oder Testitems nicht richtig verstanden hat? Gab es Störungen, die sich vielleicht auf das Ergebnis ausgewirkt haben? Gibt es Hinweise, dass sich die Testperson bei der Beantwortung eines Fragebogens oder im Interview absichtlich falsch präsentiert hat? Hat sie sich vielleicht bei einem Leistungstest auffällig wenig angestrengt oder wirkte unmotiviert? Gibt es Hinweise dafür, dass sie absichtlich auf ein schlechtes Ergebnis hingearbeitet hat?

Manchmal bieten sich mehrere Interpretationen an. Man sollte prüfen, ob die vorgenommene Interpretation die einzige mögliche ist oder ob die Ergebnisse auch mit anderen Interpretationen vereinbar sind. Sind auch andere Interpretationen plausibel, sollte die Bevorzugung einer bestimmten Interpretation begründet werden.

Schließlich ist es möglich, dass eine Frage trotz aller Bemühungen nicht geklärt werden konnte. Nicht Aufklärbares sollte unter der Angabe der Gründe benannt werden.

4.6.2 Stellungnahme

In der Stellungnahme findet die Leserin oder der Leser des Gutachtens schließlich eine klare und vollständige Antwort auf die Fragestellung.

Unentscheidbares wird als solches kenntlich gemacht. Wenn auch eine andere Antwort naheliegend ist, soll dies erwähnt werden; die Gründe, warum die Gutachterin oder der Gutachter diese Schlussfolgerung nicht zieht, sind darzulegen. Die Stellungnahme soll auch für sich alleine verständlich sein. Dies wird erreicht, indem die Erkenntnisse, auf die sich eine Schlussfolgerung stützt, genannt werden. Dies dient der Transparenz und Nachvollziehbarkeit: Die Leserin oder der Leser kann nachvollziehen, wie die Stellungnahme zu stande gekommen ist. Ausführungen, die über die Fragestellung hinausgehen, sind nicht mit dem Auftrag vereinbar. Das gilt insbesondere für unverlangte Empfehlungen.

4.6.3 Wenn der Begutachtungsprozess nicht erfolgreich verläuft

Bei der Formulierung der psychologischen Fragen knüpft die Gutachterin oder der Gutachter an Vorinformationen an, die im Gutachten auch genannt werden (s. u.). Man sollte nicht der Hybris verfallen, dass man immer von Anfang an schon die „richtigen“ psychologischen Fragen gefunden hat. Im Laufe der Untersuchung können neue Erkenntnisse anfallen, die ein Umdenken erfordern und andere Erklärungen als bisher bedacht nahelegen.

► Beispiel

Die Akten lieferten keinen Hinweis darauf, dass die *Zeugin* den Beschuldigten kennt. Sie lebt in einer weit entfernten Stadt und hielt sich nur für ein paar Tage wegen eines Bekanntenbesuchs vor Ort auf. Sie war zufällig an dem Ort, an dem ein Tötungsdelikt geschah. Ihre Aussage entlastet den Beschuldigten; die Tötung könnte aus Notwehr geschehen sein. Die Rekonstruktion des Tathergangs und die gerichtsmedizinischen Befunde passen nicht zu dieser Erklärung. Das Gericht fordert ein Gutachten zur Glaubhaftigkeit der Zeugenaussage an. Im diagnostischen Interview nennt die Zeugin den Beschuldigten mit dem Vornamen. Auf geschickte Nachfragen hin räumt sie ein, dass sie den Beschuldigten vor Jahren im Urlaub kennengelernt hat und mit ihm freundschaftlich verbunden ist. Nun muss die Möglichkeit einer absichtlichen Falschaussage in Erwägung gezogen werden. Deshalb wird eine entsprechende psychologische Frage mit der Begründung eingeführt, dass sich im Interview Hinweise auf eine persönliche Bekanntschaft mit dem Beschuldigten ergeben haben. ◀

Ebenso kommt es vor, dass die Ergebnisse zu einer psychologischen Frage nur den Schluss zulassen, dass die favorisierte Hypothese in keiner Weise unterstützt wird. Deshalb muss nach anderen Erklärungen gesucht werden. Obwohl die Untersuchung schon abgeschlossen war, muss man zumindest für einen Teilbereich noch einmal mit neuen psychologischen Fragen von vorne anfangen. Oder es liegen zu einer wichtigen psychologischen Frage widersprüchliche Ergebnisse vor, und es gelingt nicht, eine plausible Lösung zu finden. In diesem Fall braucht man vermutlich keine neue psychologische Frage, sondern man wird sich dazu entscheiden, eine andere Untersuchungsmethode zur Klärung der Frage einzusetzen.

Am Prinzip des diagnostischen Vorgehens bei der Begutachtung ändert sich nichts. Man geht lediglich einen oder mehrere Schritte zurück und fängt im Ablaufmodell (Abb. 4.1) wieder weiter oben an. Die Gründe dafür werden kurz im Gutachten dargelegt. Solche Rückwärtsschleifen können bei Bedarf auch mehrfach erfolgen.

[Im Ablaufmodell zurückgehen](#)

4.6.4 Formale Gestaltung des Gutachtens

Formale Angaben

Ein psychologisches Gutachten soll den *Qualitätsstandards für psychologische Gutachten* zufolge folgende formale Angaben enthalten (Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen 2017, S. 7):

» 4. Formale Gestaltung

4

Über die [...] genannten Aspekte hinaus umfasst ein schriftliches Gutachten die folgenden Elemente

- Name, akademischer Titel und Adresse der Gutachterin/des Gutachters,
- Name und Adresse der Auftraggeberin/des Auftraggebers,
- die Fragestellung der Auftraggeberin/des Auftraggebers,
- Name(n) und Geburtsdatum/Geburtsdaten der untersuchten Person(en),
- ggf. herangezogene zusätzliche Informationsquellen (z. B. Akten, Epikrisen),
- ggf. beauftragte Zusatzgutachten,
- das jeweilige Datum der Untersuchung(en),
- das Datum der schriftlichen Abfassung des Gutachtens,
- die rechtsverbindliche Unterschrift des Gutachters,
- ein Nachweis der im Gutachten verwendeten Fachliteratur inklusive Quellen-nachweise der eingesetzten Verfahren.

(© Föderation Deutscher Psychologenvereinigungen)

Die einzelnen Schritte des Begutachtungsprozesses müssen dokumentiert werden; außerdem gehören die oben genannten formalen Angaben ins Gutachten. Deshalb schlagen wir folgende Gliederung vor:

Gliederung eines Gutachtens

- Titelseite mit folgenden Angaben:
 - Absenderadresse (im Briefkopf)
 - Adressatin oder Adressat
 - Auftraggeberin oder Auftraggeber
 - Aktenzeichen (optional)
 - Überschrift (z. B. „Psychologisches Gutachten“)
 - Begutachtete Person (Name, Adresse)
 - Datum
 - Name der Gutachterin oder des Gutachters
- Inhaltsverzeichnis (optional bei langen Gutachten)
- Zusammenfassung (optional bei langen Gutachten)
- Untersuchungsanlass
- Fragestellung
- Vorgesichte (optional)
- Psychologische Fragen
- Untersuchungsmethoden
- Untersuchungsergebnisse
- Interpretation der Ergebnisse
- Stellungnahme
- Empfehlungen (optional)
- Unterschrift mit Ort und Datum
- Literatur
- Anhang (optional)

Titelseite Die Titelseite informiert darüber, wer in wessen Auftrag von wem wann begutachtet worden ist. Die Überschrift „Psychologisches Gutachten“ kann im Untertitel oder Nachsatz Informationen über den Gegenstand des Gutachtens enthalten (z. B. „zur Feststellung der Kraftfahreignung“). Die untersuchte Person soll so genau spezifiziert werden, damit keine Verwechslungen möglich sind. Neben dem vollen Namen können das Geburtsdatum und der Geburtsort genannt werden. Als Absender ist im Briefkopf oft eine Institution (z. B. eine Klinik bzw. Abteilung oder eine Praxisgemeinschaft) aufgeführt. Deshalb ist es sinnvoll, auf der Titelseite auch explizit die Person zu nennen, die das Gutachten verfasst hat. Für eventuelle Rückfragen ist dies sehr hilfreich.

Inhalte der Titelseite

Inhaltsverzeichnis und Zusammenfassung Bei sehr umfangreichen Gutachten wird dem eigentlichen Gutachten manchmal eine Zusammenfassung und/oder eine Gliederung mit Seitenzahlen vorangestellt. Das Inhaltsverzeichnis hilft der Leserin oder dem Leser, wenn sie oder er gezielt bestimmte Informationen sucht. Bei langen Gutachten ist auch zu bedenken, dass bestimmte Personen, die mit dem Gutachten befasst sind, auch nur die Zusammenfassung und vielleicht die Stellungnahme lesen wollen.

Untersuchungsanlass Viele Fragestellungen werden nur verständlich, wenn der Untersuchungsanlass bekannt ist. Unter diesem Gliederungspunkt wird der Hintergrund beschrieben, vor dem sich die Begutachtung ergeben hat. Daraus wird meist auch der Zweck der Begutachtung ersichtlich.

Hintergrund der Begutachtung

Fragestellung Für die Fragestellung gilt, dass deren Formulierung exakt mit den diesbezüglich getroffenen Vereinbarungen zwischen Gutachterin oder Gutachter und Auftraggeberin oder Auftraggeber übereinstimmen muss. Wurde die Fragestellung nach Rücksprache modifiziert, wird die zuletzt vereinbarte Version aufgeführt.

Vorgeschichte Unter der Überschrift „Vorgeschichte“ oder „Vorliegende Informationen“ werden alle für die Beantwortung der Fragestellung relevanten Informationen unter Nennung der Quellen (Vorgutachten, Gerichtsakten etc.) aufgeführt, die nicht von der Gutachterin oder dem Gutachter selbst erhoben worden sind.

Untersuchungsmethoden und -ergebnisse Der Untersuchungsbericht umfasst die Methoden und die Ergebnisse. Beide Teile können separate Gliederungspunkte darstellen oder auch zusammen unter dem Abschnitt „Untersuchungsmethoden und -ergebnisse“ behandelt werden. Bei der zweiten Variante sind besonders für Leserinnen und Leser, die mit diagnostischen Verfahren vertraut sind, die Ergebnisse oftmals leichter zu verstehen. Auf die Beschreibung der Methode folgen jeweils die angefallenen Ergebnisse. Die eingesetzten Untersuchungsmethoden (einzelne Tests, Fragebögen, diagnostisches Interview etc.) werden kurz und verständlich beschrieben. Wenn die Adressatin oder der Adressat des Gutachtens die Verfahren kennt, kann man auf deren Beschreibung verzichten.

Verfahren präzise benennen

Ein wesentlicher Bestandteil der Methoden ist die Angabe, was mit diesem Verfahren erfasst werden soll. Eine Formulierung kann beispielsweise lauten: „Zur Erfassung der Allgemeinen Intelligenz wurde [Nennung des Tests] eingesetzt.“ Tests werden präzise benannt: Der volle Testname, ggf. mit Angabe der Auflage (später kann eine hier eingeführte Abkürzung verwendet werden) gehört genauso dazu wie Autor(en) und das Jahr der Publikation. Manchmal ist es entscheidend, ob die alte oder neue Auflage eines Tests verwendet

4

Auswahl der Verfahren begründen

worden ist; die Normen können sich etwa geändert haben – daher ist eine präzise Angabe wichtig.

In der Regel stehen zur Beantwortung einer Frage mehrere Methoden zur Auswahl; deshalb sollte begründet werden, warum dieses und nicht ein anderes Verfahren gewählt wurde. Argumente für die Auswahl eines bestimmten Verfahrens können beispielsweise sein: „bewährtes Verfahren zur Erfassung von ...“, „aktuelle Normen“, „einschlägige Validitätsbelege“, „hohe Messgenauigkeit“, „kann auch von Probanden mit geringen Deutschkenntnissen bearbeitet werden“, „kurze Bearbeitungszeit kommt der geringen Belastbarkeit des Probanden entgegen“ (► Abschn. 4.4).

Durchführungsbedingungen beschreiben

Die Durchführungsbedingungen mit Angaben zu Ort und Zeit der Untersuchung, zur Abfolge der Verfahren, zur Durchführung in Einzel- oder Gruppenpensitzung sowie Angaben zur Person, die die Untersuchung durchgeführt hat (z. B. „die Gutachterin“), und eventuell die Erwähnung besonderer Vorkommnisse (z. B. Störungen) sind ein fester Bestandteil der Untersuchungsmethoden.

Hier keine Interpretationen

Zur Verbalisierung von Untersuchungsergebnissen wurden in ► Abschn. 4.5.5 bereits Vorschläge gemacht. Wichtig ist, dass an dieser Stelle noch keine Interpretation im Hinblick auf die Beantwortung der Fragen vorgenommen wird. Eine Relativierung von Rohwerten durch Vergleich mit den Werten einer Normstichprobe ist hier nicht als Interpretation zu werten. Zu den Untersuchungsergebnissen gehört ggf. auch eine Beschreibung des Verhaltens der untersuchten Person (Erscheinungsbild, Testverhalten, sprachlicher Ausdruck etc.).

Alle relevanten Ergebnisse berichten

Grundsätzlich werden alle Ergebnisse berichtet, die zur Beantwortung der psychologischen Fragen erhoben wurden. Das bedeutet, dass auch „unauffällige“ oder im Nachhinein als unergiebig eingestufte Ergebnisse aufgeführt werden. Weggelassen werden hingegen Ergebnisse, die zwar miterhoben, aber für die Beantwortung der psychologischen Fragen nicht benötigt werden. Manchmal verwendet man mehrdimensionale Fragebögen, von denen nur einige wenige Skalen relevant sind. Es wäre unangemessen, die Ergebnisse der nicht relevanten Skalen mit aufzuführen.

Aussagen über die Ergebnisse werden in Vergangenheitsform abgefasst. Ein Testergebnis besagt nicht, welche Eigenschaften etc. jemand hat (das ist bei der Interpretation angemessen), sondern was jemand vor einiger Zeit in einer bestimmten (Test-)Situation angegeben oder geleistet hatte.

Psychologische Fragen beantworten

Interpretation der Ergebnisse Dieser Teil des Gutachtens wird traditionell auch „Befund“ genannt. Da im medizinischen Bereich das Wort „Befund“ meist mit Untersuchungsergebnissen gleichgesetzt wird, sollte dieser Begriff zumindest in der Kommunikation mit Ärztinnen und Ärzten sowie mit Personen, die häufig auch ärztliche Gutachten lesen, nicht verwendet werden. In Bezug auf den Aufbau richtet sich die Interpretation der Ergebnisse nach den psychologischen Fragen. Die Themen der psychologischen Fragen können dabei als Zwischenüberschriften verwendet werden (z. B. „allgemeine intellektuelle Leistungsfähigkeit“). Ansonsten wird auf die Ausführungen in ► Abschn. 4.6.1 verwiesen.

Antwort auf die eigenen Fragen

Stellungnahme Für die meisten Leserinnen und Leser des Gutachtens ist dies der wichtigste Abschnitt des Gutachtens, weil er ihre Fragen beantwortet. Deshalb sollte bei der Stellungnahme auch besonderer Wert auf Verständlichkeit und Nachvollziehbarkeit gelegt werden. In ► Abschn. 4.6.2 zur Stellungnahme finden sich weitere Erläuterungen. Die persönliche Prädikation, z. B.

Durchführung einer diagnostischen Untersuchung ...

„Frau X ist überdurchschnittlich erregbar“, ist angemessen. Die Stellungnahme wird wie die Interpretation im Präsens verfasst.

Empfehlungen Empfehlungen zu therapeutischen und anderen Maßnahmen sollten nur gegeben werden, wenn dies vorher ausdrücklich vereinbart worden ist. Dieser Gliederungspunkt wird also in vielen Gutachten fehlen.

Unterschrift, Literaturverzeichnis und Anhang Die Gutachterin oder der Gutachter setzt am Ende ihre bzw. seine Unterschrift (mit Ort und Datum) unter das Werk. Im Literaturverzeichnis stehen exakte Angaben zu den verwendeten Verfahren (mit Angabe der Auflage) sowie zu eventuell zitierten Werken. Der Unsitte, auch weitere einschlägige Publikationen aufzuführen, sollte man nicht folgen. Der Anhang kann Materialien enthalten, die für das Gutachten nur auszugsweise Verwendung gefunden haben, beispielsweise ein Interview im Wortlaut. Er kann dazu genutzt werden, den Text im Gutachten durch Verweise (z. B. „s. Anlage 1.3“) von relativ unwichtigen Details zu befreien, um den „roten Faden“ nicht abreißen zu lassen. Testmaterialien (verwendete Fragebögen, Testbögen) gehören nicht in den Anhang.

Was gehört in den Anhang?

Für die Erstellung von psychologischen Gutachten sind auch die *berufsethischen Richtlinien der Psychologenverbände* (Föderation Deutscher Psychologenvereinigungen 2016, S. 27) relevant. Bezuglich der Inhalte fordern sie eine formgerechte Gestaltung und inhaltliche Nachvollziehbarkeit; die Regeln gehen aber auch darüber hinaus und betreffen u. a. die fristgerechte Erstellung und die Gewährung von Einsichtnahme.

Berufsethische Richtlinien

» Berufsethische Richtlinien

8.2 Gutachten und Untersuchungsberichte

Psychologinnen und Psychologen, die gutachterlich tätig sind:

- (1) pflegen eine größtmögliche sachliche und wissenschaftliche Fundiertheit, Sorgfalt und Gewissenhaftigkeit bei der Erstellung und Verwendung von Gutachten und Untersuchungsberichten;
- (2) fertigen Gutachten und Untersuchungsberichte frist- und formgerecht unter Einhaltung der „Richtlinien für die Erstellung Psychologischer Gutachten“ von BDP und DGP in ihrer jeweiligen Fassung an;
- (3) fertigen Gutachten und Untersuchungsberichte so an, dass sie für die Adressaten inhaltlich nachvollziehbar sind;
- (4) gewähren begutachteten Personen auf deren Wunsch Einsichtnahme in Gutachten und Untersuchungsberichte und befürworten solche erwünschten Einsichtnahmen, sofern für die begutachtete Person kein gesundheitlicher Schaden zu befürchten ist, es sei denn, dass ein Auftraggeber dazu keine Einwilligung gibt, und informieren die begutachteten Personen, falls der Auftrag zu einem Gutachten eine Einsichtnahme von vornherein ausschließt;
- (5) enthalten sich der Erstellung von Gefälligkeitsgutachten und der Abgabe von Gutachten im eigenen Namen, die von Dritten ohne eigene Mitwirkung erstellt sind;
- (6) geben Stellungnahmen zu Gutachten von Kolleginnen unter Berücksichtigung der Aussagen dieser Berufsethischen Richtlinien zum Verhältnis zu Berufskolleginnen ab (vgl. 7 Abschn. 4.4).

Richtigkeit praktisch kaum überprüfbar

Qualität aus Sicht von Laien

4.6.5 Beurteilung der Qualität eines Gutachtens

Woran kann man die Qualität eines Gutachtens erkennen? Zuerst wird man hier an die Richtigkeit denken. Leider ist die Richtigkeit nicht zur Beurteilung der Qualität geeignet, da oftmals kein Kriterium verfügbar ist, an dem man überprüfen kann, ob sich Aussagen als zutreffend erweisen oder nicht. Selbst wenn die Empfängerin oder der Empfänger eines Gutachtens die Möglichkeit hätte, die Richtigkeit zu überprüfen, so käme diese Erkenntnis in der Regel zu spät, nämlich nachdem aufgrund des Gutachtens bereits wichtige Entscheidungen getroffen worden sind. So mag ein Kraftfahrer als wieder geeignet zum Führen eines Kraftfahrzeugs beurteilt worden sein, und die Behörde hat ihm daraufhin wieder die Fahrerlaubnis erteilt; 1 Jahr später stellt sich vielleicht heraus, dass der Betreffende wieder mit seinem alten Problem Trunkenheit am Steuer auffällig geworden ist.

Die Qualität eines Gutachtens kann aus der Perspektive der Adressaten und aus einer fachlichen Perspektive beurteilt werden. Für die Empfängerin oder den Empfänger eines Gutachtens müssen folgende Fragen mit „Ja“ beantwortet werden:

Qualität für Adressaten/Adressatin

1. *Habe ich eine Antwort auf meine Frage(n) bekommen?* Diese Frage ist zu bejahen, wenn in der Stellungnahme die Fragestellung vollständig beantwortet wurde. Nicht Entscheidbares wurde ggf. als solches kenntlich gemacht.
2. *Kann ich inhaltlich nachvollziehen, wie die Antwort auf meine Frage(n) zustande gekommen ist?* Konkreter: Wurde transparent und nachvollziehbar dargestellt, wie die Gutachterin oder der Gutachter vorgegangen ist, wie sie oder er zu den Ergebnissen gelangt ist, was die Ergebnisse bedeuten und wie die Ergebnisse dazu verwendet wurden, „meine“ Fragestellung zu beantworten?
3. *Habe ich verstanden, was im Gutachten steht?* Hierzu ist anzumerken, dass ein Gutachten nicht auf einem barrierefreien Niveau abgefasst werden muss, wenn die Empfängerin oder der Empfänger ein sehr niedriges Sprachniveau hat. In diesem Fall gibt es verschiedene Möglichkeiten, ein Verstehen zu gewährleisten. Erstens kann das Gutachten zusätzlich mündlich erläutert werden. Zweitens kann es so abgefasst werden, dass es von einem zumindest halbwegs gebildeten Laien verstanden werden kann. Die Empfängerin oder der Empfänger kann sich Hilfe, z. B. im persönlichen Umfeld, holen.

Expertinnen und Experten können die Qualität eines Gutachtens anhand der Kriterien „Transparenz“ und „Nachvollziehbarkeit“, aber auch anhand weiterer fachlicher Kriterien bewerten. Die folgenden Fragen können auch als Checkliste für die Verfasserinnen und Verfasser eines Gutachtens dienen. Wir haben hier eine mittlere Auflösung gewählt; nicht alle Details sind erwähnt.

Wichtige Anforderungen an ein Gutachten

1. Enthält das Gutachten alle notwendigen *formalen Angaben* (► Abschn. 4.6.4)?
2. Sind die *psychologischen Fragen* gut begründet aus der globalen Fragestellung hergeleitet? Dazu gehören folgende Aspekte: Anknüpfung an relevante bereits vorliegende Informationen, Bezugnahme auf wissenschaftliche Erkenntnisse und/oder gesicherte allgemeine oder eigene Erfahrungen, ggf. Festlegung von Entscheidungskriterien wie etwa erforderliche Mindestwerte.
3. Sind die *psychologischen Fragen* grundsätzlich mit den verfügbaren diagnostischen Verfahren beantwortbar? Wurden also auch zu allen psychologischen Fragen Verfahren eingesetzt?
4. Wurden *geeignete Verfahren* zur Beantwortung der psychologischen Fragen ausgewählt? Ersichtlich ist dies aus den Begründungen der Gutachterin oder des Gutachters für ein Verfahren. Hinzu kommen ggf. generell wichtige Kriterien:
 - a) Ist das Verfahren präzise benannt (Name, Messgegenstand) und sind genaue Angaben im Literaturverzeichnis zu finden?
 - b) Ist das Verfahren für den Messgegenstand geeignet (Validitätsbelege)?
 - c) Sind die Normen angemessen (Aktualität, Größe und Repräsentativität der Normierungsstichprobe bezüglich der gewählten Vergleichsgruppe)?
 - d) Ist die Messgenauigkeit für die Fragestellung hoch genug? Sind ggf. bessere Alternativen vorhanden?
 - e) Ist das Verfahren für die Person geeignet? Sind eventuell Besonderheiten der Person zu beachten?
5. Wurde die *Durchführung* der Verfahren dokumentiert (Untersuchungssituation, beteiligte Personen, Abfolge der Verfahren, Pausen, Vorkommnisse)?
6. Wurden die *Ergebnisse* angemessen dargestellt und erläutert? Werden alle Ergebnisse berichtet (keine selektive Nutzung), ist die Quelle (Verfahren) genannt, sind Normwerte mit gewählter Referenzgruppe und ggf. Konfidenzintervall aufgeführt, wird die Ausprägung mit einheitlich verwendeten Begriffen verbalisiert?
7. Ist die *Interpretation* der Ergebnisse angemessen? Werden alle für die psychologischen Fragen relevanten Ergebnisse verwendet, die psychologischen Fragen beantwortet sowie ggf. Widersprüche und Übereinstimmungen genannt und erklärt? Wird die individuelle Gültigkeit der Ergebnisse thematisiert?
8. Wurde die *Fragestellung* vollständig und mit angemessen Begründungen beantwortet? Alle psychologischen Fragen samt den dazu angefallenen Ergebnissen sind dabei zu verwenden. Fakten (Ergebnisse) und deren Interpretation müssen unterscheidbar sein.
9. Ist die Argumentation transparent und nachvollziehbar?

Angesichts der Tatsache, dass allgemeine und auch auf bestimmte Anwendungsbereiche spezifizierte Richtlinien und Qualitätsstandards für Gutachten vorliegen sollte man erwarten, dass die Gutachten in der Praxis weitgehend den Anforderungen entsprechen. Das ist aber leider nicht der Fall.

► Beispiel

Ein besonders gravierender Fall, der auch große öffentliche Aufmerksamkeit erfahren hat, ist der von Gustl Mollath. Mollath verbrachte viele Jahre zu Unrecht im psychiatrischen Maßregelvollzug. Dazu haben psychiatrische Gutachten wesentlich beigetragen. Ein Diplom-Psychologe (Sponsel 2013) hat sich die Mühe gemacht, die insgesamt 4 psychiatrischen Gutachten anhand der „Mindestanforderungen für Prognosegutachten“ (Boetticher et al. 2007) zu bewerten. Er vergab für jedes Qualitätskriterium von –1 bis +1 Punkte; der Gesamtwert konnte daher von –35 bis +35 reichen. Die Ergebnisse lauten: –22 (Nürnberger Gutachter), –26 (Bayreuther Gutachter), –27,5 (Berliner Gutachter) und –13,5 (Ulmer Gutachter). Eines der Gutachten wurde von Mollaths Anwalt veröffentlicht und kann als Anschauungsmaterial eingesehen werden unter: ► <https://www.strate.net/de/dokumentation/Mollath-Gutachten-Kroeber-2008-06-27.pdf>. ◀

Nun könnte man meinen, nur bei psychiatrischen Gutachten gäbe es Qualitätsprobleme. In ► Abschn. 9.2 wird eine Studie von Salewski und Stürmer (2015) vorgestellt, der zufolge sehr viele der analysierten familienrechtspychologischen Gutachten, die ganz überwiegend von Psychologinnen und Psychologen verfasst wurden, gravierende Mängel aufweisen.

Weiterführende Literatur

Von den Büchern und Buchbeiträgen zur Erstellung von Gutachten sind folgende besonders zu empfehlen, da sie auch konkrete Handlungsanweisungen enthalten: Margraf-Stiksrud und Schmidt-Atzert (2015), Poyer und Ortner (2017) sowie Westhoff und Kluck (2020).

4.7 Zusammenfassung

In diesem Kapitel wurden mit der diagnostischen Untersuchung und der Gutachtenerstellung 2 große Themen behandelt. Diagnostische Untersuchungen werden oftmals auch ohne ein anschließendes Gutachten durchgeführt. Die Ergebnisse und deren Interpretation finden dann etwa für eine mündliche Rückmeldung, eine Aktennotiz oder einen Kurzbericht Verwendung. Wird ein psychologisches Gutachten erstellt, ist die Dokumentation der diagnostischen Untersuchung und ihrer Ergebnisse ein fester Bestandteil. Ein solches Gutachten ist jedoch weit mehr als ein Bericht über die Untersuchung mit ihren Ergebnissen.

Die Planung einer diagnostischen Untersuchung beginnt stets mit einer (mehr oder weniger globalen) Fragestellung. Bei einem Gutachten ist es die Fragestellung des Auftraggebers. Daraus werden Unterfragen abgeleitet, die auch „psychologische Fragen“ genannt werden. Darunter versteht man Hypothesen, explorative Fragen und an sich selbst gerichtete Arbeitsaufträge („um die Fragestellung beantworten zu können, muss festgestellt werden, ob ...“). Die richtigen Fragen zu stellen, ist nicht immer einfach, da sie nachvollziehbar aus der globalen Fragestellung abgeleitet werden müssen. Bei der Suche nach relevanten Fragen kann man sich an breiten Themengebieten wie kognitiver Bereich, Persönlichkeitsmerkmale, sozialer Bereich, Umweltbedingungen etc. orientieren. Zur Beantwortung der psychologischen Fragen wählt man geeignete diagnostische Verfahren aus. Dafür wurden Auswahlkriterien genannt.

Bei der Durchführung einer Untersuchung sind bestimmte Standards zu beachten. Dazu gehören die Aufklärung der Testperson, die Schaffung guter Arbeitsbedingungen (keine Störungen, genügend Platz etc.) und die Einhaltung von Vorgaben aus den Testmanualen zwecks Standardisierung. Letzteres gilt auch für die Auswertungen. Generell muss die Person, die die Untersuchung durchführt und die Auswertung vornimmt, über die notwendigen

Kompetenzen verfügen. Konkrete Hinweise, was zu beachten ist, finden sich u. a. in den „Internationalen Richtlinien für die Testanwendung“.

Bei der Kommunikation von Ergebnissen bestehen mehrere Herausforderungen: Das gemessene Merkmal ist korrekt zu benennen, bei Tests ist die Ausprägung des Merkmals (Normwerte) auf einheitliche Weise in Bereiche einzuordnen und zu verbalisieren. Schließlich ist der begrenzten Messgenauigkeit Rechnung zu tragen, indem für die Testwerte Konfidenzintervalle bestimmt und berichtet werden. Und all das soll für Laien nachvollziehbar verständlich geschehen und zugleich auch für Expertinnen und Experten (in Klammern) spezifiziert werden.

Wenn die Fragestellung komplex ist und nicht nur aus einer spezifischen Frage besteht, werden die Ergebnisse nun zwecks Beantwortung der einzelnen psychologischen Fragen interpretiert. Im Gutachten wird der entsprechende Abschnitt „Interpretation der Ergebnisse“ oder „Befund“ genannt. Da meist mehrere Ergebnisse zu einer psychologischen Frage vorliegen, empfiehlt sich eine übersichtliche Darstellung in einem „Befundbogen“, der nicht im Gutachten aufgeführt wird. Darin werden auch Übereinstimmungen und Widersprüche stichpunktartig benannt. In der Stellungnahme erfolgt schließlich eine Integration der Erkenntnisse zur Beantwortung der Fragestellung. Nur wenn dies ausdrücklich vereinbart wurde, gehören auch Empfehlungen in das Gutachten. Für den formalen Aufbau eines psychologischen Gutachtens wurden konkrete Empfehlungen gegeben.

Der geschilderte Ablauf, der sich an Ausführungen in ► Kap. 1 zu diagnostischen Prozess anschließt, wird in □ Abb. 4.1 zusammenfassend grafisch dargestellt. Für das diagnostische Vorgehen und speziell für die Gutachtererstellung liegen einschlägige Standards von Fachgesellschaften vor, auf die an geeigneter Stelle verwiesen wurde. Abschließend ist zu sagen, dass in der Praxis nicht alle Gutachten diesen Standards gerecht werden. Nicht die Richtigkeit (die in der Regel allenfalls nach langer Zeit überprüft werden kann), sondern die Einhaltung der genannten Standards entscheidet über die Qualität eines Gutachtens. Für eine Überprüfung der Qualität wurden 3 Fragen genannt, die von Laien als Adressatinnen/Adressaten positiv beantwortet werden sollten sowie 9 wichtige Anforderungen an ein Gutachten aus Expertensicht. Diese können von der Verfasserin oder dem Verfasser als Checkliste zur Überprüfung ihres Gutachtens genutzt werden.

?

Übungsfragen

— ► Abschn. 4.1–4.5:

- Nennen Sie die 5 zentralen Schritte des diagnostischen Prozesses!
- Nennen Sie 2 zentrale Anforderungen der International Test Commission (ITC) bezüglich der Darstellung und Interpretation der Ergebnisse!
- Die berufsethischen Richtlinien der deutschen Psychologenverbände verlangen „fachliche Kompetenz“ für Dienstleistungen, wozu auch diagnostische Untersuchungen und Begutachtungen gehören. Was wird konkret von den Psychologinnen und Psychologen erwartet?
- Nennen Sie je ein Beispiel für einen informellen und einen formellen Auftrag für eine diagnostische Untersuchung!
- Was kann man tun, um zu verhindern, dass die Testperson die Untersuchungsergebnisse verfälscht, und wie kann man eine Verfälschung eventuell erkennen?
- Bei der Auswahl eines diagnostischen Verfahrens ist zu prüfen, ob es auch für die zu untersuchende Person angemessen ist. Welche Aspekte sind dabei zu beachten?
- Wo findet man Richtlinien zur computer- und internetbasierten Testung?
- Welche Effekte kann eine für die Testperson zu lange Testdauer haben?

- Nennen Sie Vor- und Nachteile von Gruppenuntersuchungen (in Abgrenzung zu Einzeluntersuchungen)!
- Warum sollte man die zu untersuchenden Personen vor Beginn über wichtige Aspekte der Untersuchung aufklären?
- Welche Möglichkeiten der Testauswertung gibt es?
- Welche 4 Kernelemente sollte die Mitteilung eines Testergebnisses an Laien enthalten?
- ► Abschn. 4.6:
- Aus welchen Gründen sollte eine Diagnostikerin/ein Diagnostiker einen Begutachtungsauftrag ablehnen?
- Wie könnte eine leicht verständliche tabellarische Ergebnisdarstellung in einem Gutachten aussehen? Beschränken Sie sich auf die Elemente, die für Laien relevant sind.
- Welche Funktion hat der Befund im Gutachten? Wie kann man diesen Teil des Gutachtens auch nennen?
- Welche formalen Angaben sollten sich in einem Gutachten finden?
- Welche Funktion hat die Stellungnahme im Gutachten?

Literatur

- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied* 15, 163–181.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arbeitsgruppe „Qualitätsstandards für psychodiagnostische Gutachten“. (2011). *Qualitätsstandards für psychologische Gutachten (Version 2.2)*. Berlin: Deutsche Gesellschaft für Psychologie e. V. (DGPs).
- Arbeitsgruppe Familienrechtliche Gutachten. (2019). *Mindestanforderungen an die Qualität von Sachverständigengutachten im Kindschaftsrecht* (2. Aufl.). Berlin: Deutscher Psychologen Verlag.
- Boetticher, A., Kröber, H.-L., Müller-Isbner, R., Böhm, K. M., Müller-Metz, R., & Wolf, T. (2007). Mindestanforderungen für Prognosegutachten. *Forensische Psychiatrie, Psychologie, Kriminologie* 1, 90–100.
- Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen. (2017). Qualitätsstandards für psychologische Gutachten. ► https://www.psychologie.de/downloads/GA_Standards_DTK_10_Sep_2017_Final.pdf. Zugegriffen: 21. März 2020.
- Deutsches Institut für Normung e. V. (DIN). (2016). *DIN 33430:2016-07: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.
- Eyde, L. D., Robertson, G. J., & Krug, S. E. (2010). *Responsible test use: Case studies for assessing human behavior* (2nd ed.). Washington, DC: American Psychological Association.
- Föderation Deutscher Psychologenvereinigungen (Hrsg.). (1994). *Richtlinien für die Erstellung Psychologischer Gutachten*. Bonn: Deutscher Psychologen Verlag GmbH.
- Föderation Deutscher Psychologenvereinigungen. (2016). *Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V. und der Deutschen Gesellschaft für Psychologie e. V., zugleich Berufsvorschrift des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V.* Berlin: Föderation Deutscher Psychologenvereinigungen.
- Gesellschaft für Neuropsychologie, Neumann-Zielke, L., Riepe, J., Roschmann, R., Schötzau-Fürwentsches, P., & Wilhelm, H. (2009). Leitlinie „Neuropsychologische Begutachtung“. *Zeitschrift für Neuropsychologie* 20, 69–83.
- Gnambs, T., Bartinic, B., & Hertel, G. (2011). Internetbasierte psychologische Diagnostik. In L. F. Horneke, M. Amelang, & M. Kersting (Hrsg.), *Verfahren zur Leistungs-, Intelligenz- und Verhaltensdiagnostik* (Enzyklopädie der Psychologie, Psychologische Diagnostik, Bd. 3, S. 448–498). Göttingen: Hogrefe.
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research* 58, 47–77.
- Höft, S., Püttner, I., & Kersting, M. (2018). Anforderungsanalyse, Verfahren der Eignungsbeurteilung sowie rechtliche Rahmenbedingungen. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 95–153). Berlin, Heidelberg: Springer.

Durchführung einer diagnostischen Untersuchung ...

- International Test Commission (ITC). (2001). Internationale Richtlinien für die Testanwendung (Version 2000): Deutsche Fassung der International Guidelines for Test Use. ► https://www.intestcom.org/files/guideline_test_use_german.pdf. Zugriffen: 21. März 2020.
- International Test Commission (ITC). (2006). International Guidelines on Computer-Based and Internet-Delivered Testing. *International Journal of Testing* 6, 143–171.
- Kastner-Koller, U., & Deimann, P. (2012). *WET: Wiener Entwicklungstest. Ein Verfahren zur Erfassung des allgemeinen Entwicklungsstandes bei Kindern von 3 bis 6 Jahren* (3. Aufl.). Göttingen: Hogrefe.
- Margraf-Stiksrud, J., & Schmidt-Atzert, L. (2015). Das psychologische Gutachten. In G. Stemmler, & J. Margraf-Stiksrud (Hrsg.), *Lehrbuch Psychologische Diagnostik* (S. 321–378). Bern: Huber.
- Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice* 33, 4–12.
- Proyer, R. T., & Ortner, T. M. (2017). *Praxis der Psychologischen Gutachtenerstellung: Schritte vom Deckblatt bis zum Anhang* (2. Aufl.) Göttingen: Hogrefe.
- Salewski, C., & Stürmer, S. (2015). Qualität familienrechtspychologischer Gutachten. *Zeitschrift für Kindschaftsrecht und Jugendhilfe* 10, 4–9.
- Schmidt-Atzert, L., & Krumm, S. (2007). Diagnostische Urteilsbildung und Begutachtung. *Rehabilitation* 46, 9–15.
- Schubert, W., Huetten, M., Reimann, C., Graw, M., Schneider, W., & Stephan, E. (Hrsg.). (2018). *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl.). Bonn: Kirschbaum.
- Spitznagel, A. (1982). Die diagnostische Situation. In K. J. M. Groffmann (Hrsg.), *Grundlagen psychologischer Diagnostik* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 1, S. 248–294). Göttingen: Hogrefe.
- Sponsel, R. (2013). Mindestanforderungen für forensische Prognosegutachten und ihre Einhaltung bei Gustl F. Mollath durch den Nürnberger, Bayreuther, Berliner und Ulmer Gutachter. Erlangen: IP-GIPT. ► <https://www.sgipt.org/forpsy/NFPMRG/MS-Prog.htm>. Zugriffen: 21. März 2020.
- Sturm, W. (2008). *WAF: Wahrnehmungs- und Aufmerksamkeitsfunktionen*. Mödling: Schuhfried.
- Truxillo, D. M., Bodner, T. E., Bertolino, M., Bauer, T. N., & Yonce, C. A. (2009). Effects of explanations on applicant reactions: A meta-analytic review. *International Journal of Selection and Assessment* 17, 346–361.
- Westhoff, K., & Kluck, M. L. (2008). *Psychologische Gutachten schreiben und beurteilen* (5. Aufl.). Berlin, Heidelberg: Springer.
- Westhoff, K., & Kluck, M. L. (2020). *Psychologische Gutachten schreiben und beurteilen* (6. Aufl.). Berlin, Heidelberg: Springer.
- Ziegler, A., Hadlak, A., Mehlbeer, S., & König, I. R. (2013). Verständnis von Nebenwirkungsrisiken im Beipackzettel: Eine Umfrage unter Ärzten, Apothekern und Juristen. *Deutsches Ärzteblatt* 110, 669–673.



Diagnostische Strategien und Evaluation des Vorgehens

Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang

Inhaltsverzeichnis

- 5.1 Diagnostische Strategien – 528**
 - 5.1.1 Status- vs. Veränderungsdiagnostik – 528
 - 5.1.2 Selektion vs. Modifikation – 530
 - 5.1.3 Strategien der Integration von Daten zu einer diagnostischen Entscheidung – 532
 - 5.1.4 Einstufige vs. mehrstufige diagnostische Entscheidungen – 553
- 5.2 Evaluation des Vorgehens – 556**
 - 5.2.1 Prozessevaluation der Psychologischen Diagnostik – 556
 - 5.2.2 Ergebnisevaluation der Psychologischen Diagnostik – 558
 - 5.2.3 Schätzung des Nutzens Psychologischer Diagnostik – 559
- 5.3 Zusammenfassung – 563**
- Literatur – 564**

5.1 Diagnostische Strategien

Definition

Diagnostische Strategien lassen sich definieren als „auf diagnostischen Daten aufbauende Konzeptionen, mit deren Hilfe der Diagnostiker sein [bzw. die Diagnostikerin ihr] antizipiertes Ziel zu erreichen sucht“ (Jäger 1988, S. 117).

5

Unterscheidbare Prinzipien des Diagnostizierens

Wahl des bestmöglichen Vorgehens

Diagnostische Strategien beschreiben also unterscheidbare Prinzipien des Diagnostizierens. Die Wahl der Strategie richtet sich danach, wie das diagnostische Ziel unter den gegebenen Randbedingungen am besten zu erreichen ist. Besteht das Ziel beispielsweise in der Selektion von Bewerberinnen und Bewerbern unter der Randbedingung, dass sich sehr viele Personen bewerben, so könnte eine mehrstufige Entscheidungsstrategie gewählt werden. Es würden also nach einem ersten Screening bereits Bewerberinnen und Bewerber abgelehnt. Nur die verbleibenden Kandidatinnen und Kandidaten würden an einem Eignungsinterview teilnehmen.

Eine exhaustive Liste diagnostischer Strategien ist schwer zu erstellen. Die Vielzahl diagnostischer Fragestellung in unterschiedlichen Anwendungsbereichen der Psychologie und in der psychologischen Forschung sowie unterschiedliche Randbedingungen erfordern eine komplexe Entscheidung der Diagnostikerinnen bzw. des Diagnostikers, was im gegebenen Kontext das bestmögliche Vorgehen ist. Wenngleich der in ► Abschn. 1.5 dargestellte diagnostische Prozess das elementare Grundgerüst des diagnostischen Vorgehens ist – quasi die diagnostische Basisstrategie –, werden in der Umsetzung mannigfaltige Kombinationen weiterer Strategien nötig.

- ! Diagnostikerinnen und Diagnostiker begeben sich bei der Wahl der diagnostischen Strategien in eine ähnliche Rolle wie Forscherinnen und Forscher, die für eine vorliegende Forschungsfrage oder Hypothese die Datenerhebung so planen müssen, dass die bestmögliche Evidenz zur Beantwortung der Fragestellung entsteht (Westmeyer 2006).

Wir unterscheiden und diskutieren nachfolgend diese Strategien:

- Status- vs. Veränderungsdiagnostik
- Selektion vs. Modifikation
- Strategien der Integration von Daten zu einer diagnostischen Entscheidung
- Einstufige vs. mehrstufige diagnostische Entscheidungen

Abschließend wenden wir uns in der Praxis gebräuchlichen, aber aus diagnostischer Sicht nicht zu empfehlenden Strategien zu.

5.1.1 Status- vs. Veränderungsdiagnostik

Statusdiagnostik = Erfassung des aktuellen Stands

Diagnostik wird häufig durchgeführt, um zu prüfen, ob eine Intervention erforderlich ist. Bei der Intervention kann es sich beispielsweise um eine Personalentwicklungsmaßnahme (z. B. Training von Verhandlungstaktik), eine pädagogische (z. B. Einzelunterricht zur Verbesserung des Leseverständnisses) oder eine klinische Maßnahme (z. B. Verhaltenstherapie zum Abbau von Zwängen) handeln. Da die Diagnostik notwendigerweise vor der Intervention stattfindet, spricht man auch von einer Eingangsdagnostik. Auf jeden Fall wird der aktuelle Stand erfasst; deshalb handelt es sich um eine Statusdiagnostik.

Veränderungsdiagnostik dient dazu, einen Unterschied zwischen 2 oder mehreren Messungen zu identifizieren. Mit Messungen vor Beginn und nach Beendigung einer Intervention versucht man, festzustellen, ob die durch die Intervention angestrebte Veränderung erreicht wurde. Es können somit Fragen beantwortet werden wie: Hat sich das Leseverständnis des Kindes substanzial verbessert? Hat sich die Symptomatik einer Patientin oder eines Patienten bedeutsam verbessert? Häufig ist jedoch nicht nur interessant, ob sich eine Verbesserung im Vergleich zur Eingangsdiagnostik eingestellt hat, sondern auch, ob ein bestimmtes Niveau erreicht wurde. Die zuvor genannten Fragen könnten daher auch lauten: Hat sich das Leseverständnis des Kindes so weit verbessert, dass es wieder dem regulären Unterricht folgen kann? Hat sich die Symptomatik der Patientin oder des Patienten so weit gebessert, dass sie oder er im Alltags- und Berufsleben nicht weiter eingeschränkt ist? Im Übrigen sind solche Erfolgskontrollen im Sinne der Klientin oder des Klienten, denn sie dienen dazu, festzustellen, ob noch weitere Maßnahmen nötig sind oder nicht.

Bei einer Erfolgskontrolle durch Veränderungsmessung sind 2 Aspekte zu beachten: Erstens kann die eingetretene Verbesserung oder auch Verschlechterung gegenüber dem Ausgangszustand auch auf Faktoren zurückzuführen sein, die nichts mit der Intervention zu tun haben oder gar zufallsbedingt sind. So könnten die Eltern des Kindes erkannt haben, dass auch sie etwas zur Verbesserung des Leseverständnisses beitragen können und nun regelmäßig am Abend gemeinsam in einem Kinderbuch lesen.

! Veränderungsmessungen, die nicht im Rahmen kontrollierter experimenteller Untersuchungen vorgenommen werden, erlauben keine kausalen Schlüsse.

Zweitens ist bekannt, dass Messungen alleine durch ihre Wiederholung beeinflusst werden. Dies können Übungseffekte in Leistungstests (z. B. Hausknecht et al. 2007) oder Erinnerungseffekte in Fragebögen sein. Im Einzelfall sollte zumindest auf mögliche Einflussfaktoren geachtet werden und es sollten ggf. diagnostische Verfahren bevorzugt werden, für die solche Einflüsse eher gering sind.

Bei einer Verlaufs- oder Prozessdiagnostik werden die zu verändernden Merkmale kontinuierlich erfasst. So ist es möglich, die Intervention ggf. an die Veränderungen anzupassen. Wenn sich schnell Fortschritte zeigen, kann ein Training intensiviert werden; eine Unterforderung wird so vermieden. Umgekehrt ist auch eine Verlangsamung möglich, wenn eine Überforderung droht. Denkbar ist auch eine vorzeitige Beendigung der Maßnahme, wenn Zwischenziele nicht erreicht werden oder ein Erfolg am Ende nicht zu erwarten ist.

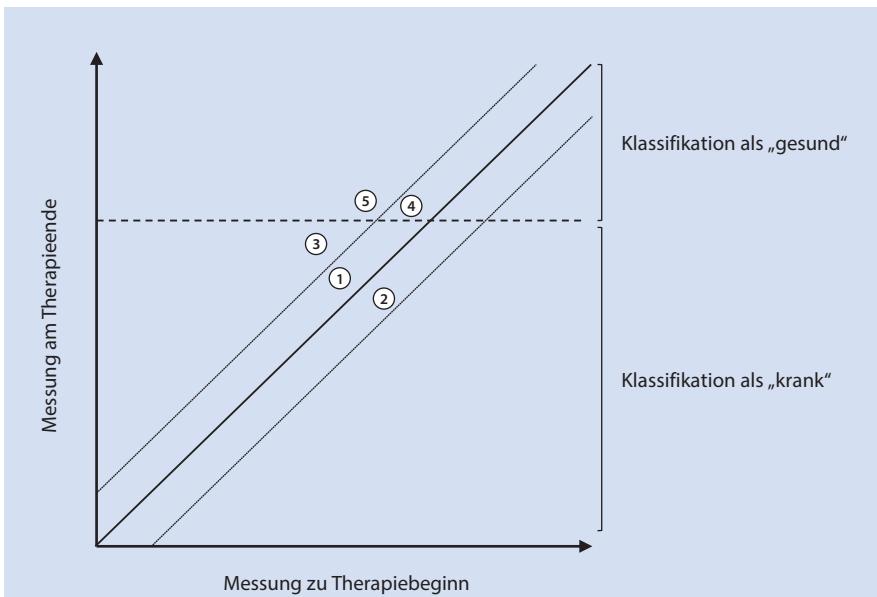
Veränderungsdiagnostik =
Unterschied zwischen mehreren
Messungen feststellen

Klinisch bedeutsame Veränderung

In ▶ Abschn. 2.6.2.2 haben wir bereits den Reliable Change Index (RCI) kennengelernt. Er dient zur Beurteilung, ob durch diagnostische Instrumente erfasste Veränderungen substanzial sind (also beispielsweise eine Verbesserung der Symptomatik darstellen) oder wahrscheinlich eher messfehlerbedingt sind. Allerdings kann man bei Veränderungsmessungen im Rahmen von Psychotherapien noch nicht alleine damit zufrieden sein, dass eine Veränderung größer ausfällt, als dass sie messfehlerbedingt sein könnte. Es ist zwar zu begrüßen, dass Patientinnen und Patienten substanzial verbesserte Symptomatiken aufweisen – noch schöner wäre es natürlich, wenn sie wieder als gesund bezeichnet werden könnten. Daher versteht man unter einer klinisch bedeutsamen Veränderung, dass Patientinnen und Patienten die in Messinstrumenten als kritisch definierten Cut-off-Werte überschreiten (oder unterschreiten, je nach Polung des Instruments) und somit als gesund klassifiziert werden können. Für Details zur Identifikation dieser Werte im Bereich der klinisch-psychologischen Diagnostik s. ▶ Abschn. 8.6.2.

Verlaufs- oder Prozessdiagnostik

Diagnostik klinisch bedeutsamer Veränderungen



Klinisch bedeutsame Veränderungen. (In Anlehnung an Jacobson und Truax 1991, S. 17, © American Psychological Association).

Die Abbildung visualisiert diese Überlegungen anhand von 2 Messungen (vor Therapiebeginn und nach Therapieende). Wir nehmen an, dass höhere Werte in diesen Messungen besser sind und dass Werte über der horizontalen, gestrichelten Linie dafür sprechen, dass Personen gesund sind. Die diagonale Linie kennzeichnet den Bereich, in dem Prä- und Postmessung exakt identisch sind. Punkte über der Diagonalen stellen also eine Verbesserung dar, Punkte darunter eine Verschlechterung. Ob diese Veränderungen als messfehlerbedingt gelten müssen, kennzeichnen die gestrichelten Linien parallel zur Diagonalen. Punkte innerhalb dieses Bereichs müssen als möglicherweise messfehlerbedingte Veränderungen angesehen werden.

Zur weiteren Erläuterung sind 5 hypothetische Personen eingezeichnet. Person 1 weist zwar eine Verbesserung auf, sie muss aber als vermutlich messfehlerbedingt bezeichnet werden. Person 2 zeigt eine Verschlechterung, aber auch diese ist wahrscheinlich messfehlerbedingt. Person 3 weist eine Verbesserung auf, die nicht mehr nur als messfehlerbedingt gelten kann. Allerdings erreicht Person 3 auch in der Postmessung keine so hohen Werte, dass sie als gesund gelten kann. Person 4 erreicht zwar in der Postmessung ein Niveau, dass sie als gesund klassifiziert, allerdings ist ihre Veränderung zu Prämessung so gering, dass dies als messfehlerbedingt gelten muss. Lediglich Person 5 zeigt eine Veränderung über Messfehlerniveau und erreicht gleichzeitig Werte in der Postmessung, die sie als gesund klassifizieren.

5.1.2 Selektion vs. Modifikation

Passung zwischen Bedingungen der Umwelt und Personenmerkmalen

Bei vielen diagnostischen Fragestellungen wird nach einer Passung zwischen Personen und Bedingungen gesucht. Ist eine Schülerin oder ein Schüler am besten auf einer Schule für Lernbehinderte aufgehoben? Ist eine Bewerberin oder ein Bewerber für die ausgeschriebene Stelle geeignet? Wird eine Patientin oder ein Patient von dem Therapieverfahren profitieren? Dies sind Beispiele, in denen Bedingungen (das Förderangebot einer Schule für Lernbehinderte, die mit der Stelle verbundenen Anforderungen, die Besonderheiten des Therapieverfahrens) zu den Merkmalen der Person (genauer Förderbedarf,

Ausprägung der Anforderungsmerkmale, personelle Voraussetzungen für das Therapieverfahren) in Beziehung zu setzen sind.

Bei einer *Selektion von Personen* steht zuvor eine Bedingung fest, und es werden Personen ausgewählt, die die größte Passung mit dieser Bedingung aufweisen. Ein typisches Beispiel ist die Auswahl von Bewerberinnen und Bewerbern für eine Stelle (s. auch ► Abschn. 6.2.1). Bei der *Selektion von Bedingungen* steht die Person im Vordergrund, und es werden die für sie passenden Bedingungen (z. B. interessante Berufe) ausgewählt. Manchmal wird keine gute Passung zwischen Person und Bedingung vorliegen.

Bei der *Modifikation* stellt sich die Frage: Welche Merkmale der Person oder der Bedingung sind zu ändern, damit eine Passung hergestellt wird? So können nur teilweise geeignete Bewerberinnen und Bewerber geschult werden, sodass sie danach die Anforderungen eines Ausbildungs- oder Arbeitsplatzes erfüllen. Andererseits kann auch die Bedingung, in diesem Beispiel der Arbeitsplatz, an die Person angepasst werden. Im Rahmen einer Psychotherapie dient Psychologische Diagnostik der Modifikation, indem Defizite und psychische Störungsbilder festgestellt werden, um danach zielgerichtete Maßnahmen zu deren Verbesserung umzusetzen.

Die diagnostischen Strategien zur Selektion vs. Modifikation unterscheiden sich deutlich. Steht eine Selektion im Vordergrund, sind alle infrage kommenden Entitäten (alle sich bewerbenden Personen, alle verfügbaren Therapieformen, alle infrage kommenden Berufe, alle verfügbaren Schulformen etc.) zu betrachten und in eine Rangreihe hinsichtlich ihrer Passung zu bringen. Die unter Beachtung verschiedener relevanter Kriterien beste Option wird empfohlen. Ob die zur Beurteilung der Passung herangezogenen Kriterien veränderbar sind oder nicht, spielt keine oder nur eine untergeordnete Rolle. Steht die Modifikation im Vordergrund, werden die Merkmale *einer* zu modifizierende Person oder *eines* zu modifizierenden Kontextes betrachtet. Ein bewertender Vergleich mit anderen Personen oder Kontexten ist nicht zentral. Vielmehr sind auslösende und aufrechterhaltende Bedingungen für eine problematische Gegebenheit sowie die tatsächlich modifizierbaren Anteile des „Problems“ zu identifizieren.

Selektionsdiagnostik = Auswahl von Personen, die zur Umwelt passen oder vice versa

Modifikationsdiagnostik = Veränderung von Person und/oder Umwelt, sodass Passung besser wird

Selektion vs. Modifikation implizieren unterschiedliche diagnostische Strategien

Verhaltensanalyse nach dem SORKC-Schema

Ein gängiges Vorgehen im Rahmen der Modifikationsdiagnostik im Vorfeld einer Psychotherapie ist die Verhaltensanalyse. Hier wird das Problemverhalten detailliert betrachtet und anhand der Komponenten „Stimulus“, „Organismus“, „Reaktion“, „Kontingenz“ und „Konsequenz“ bewertet.

S	Stimulus	Reiz, der auf die Person einwirkt; z. B. Halten eines Referats, alle schauen erwartungsvoll auf die referierende Person
O	Organismus	Körperliche und psychische Merkmale der Person; z. B. ängstlich-vermeidende Persönlichkeitszüge, wenig Schlaf vor dem Referat
R	Reaktion	Gedanken, Gefühle, körperliche Reaktionen, Verhalten der Person; z. B. „oh Gott, ich schaffe das nicht“, „alle werden denken, ich sei dumm“, Angst vor der Blamage, Schwitzen, Zittern, Herzrasen, Person entschuldigt sich und verlässt unter Vorwand den Raum
K	Konsequenz	Reaktionen auf das Problemverhalten; z. B. Wegfall der Anspannung, Selbstbild als minderwertig verfestigt sich, Referat muss noch einmal gehalten werden oder Seminar wird nicht bestanden
C	Kontingenz	Regelmäßigkeit mit der Konsequenzen eintreten; z. B. tritt immer bei Referaten auf

5.1.3 Strategien der Integration von Daten zu einer diagnostischen Entscheidung

5.1.3.1 Klinische oder mechanische Urteilsbildung?

Oftmals müssen zur Beantwortung einer Fragestellung Daten aus mehreren Quellen zu einem Gesamтурteil integriert werden. Die folgenden beiden Strategien der Aggregation von Daten zu einem Gesamтурteil werden unterschieden: die klinische und die mechanische Urteilsbildung. Der Begriff „klinische“ Urteilsbildung beschreibt die individuelle Integration von Daten durch die diagnostischen Expertinnen und Experten. Bei der „mechanischen“ Urteilsbildung werden die Daten nach einer Formel verrechnet, die zuvor festgelegt wurde. Oft wird die Formel durch die statistische Analyse der Daten vieler ähnlicher Fällen begründet; in diesem Fall spricht man auch von statistischer Urteilsbildung.

Ein bekanntes statistisches Urteilsmodell ist der Goldberg-Index (Goldberg, 1965); mithilfe des Goldberg-Index kann anhand von Testergebnissen des Minnesota Multiphasic Personality Inventory (MMPI-2) festgestellt werden, ob eine Patientin oder ein Patient psychotisch ist oder nicht. Fünf Skalenwerte (T-Werte) werden nach der Formel $L + Pa + Sc - Hy - Pt$ (das sind die Abkürzungen für Skalen im MMPI-2; ► Abschn. 3.3.3.1) verrechnet. Liegt der Index über 45, gilt eine Patientin bzw. ein Patient als psychotisch. Im Gegensatz dazu würden bei einer klinischen Urteilsbildung Expertinnen und Experten ohne feststehendes Schema (ohne eine Formel) sich die jeweiligen Informationen zu einer Patientin bzw. einem Patienten ansehen und urteilen, ob diese bzw. dieser psychotisch ist oder nicht; sie verlassen sich dabei auf ihre klinische Erfahrung.

- !** Es stellt sich die Frage, welche Strategie zu bevorzugen ist – klinische oder mechanische Urteilsbildung. Die Antwort auf diese Frage findet sich in mittlerweile 4 Metaanalysen, die einheitlich eine Überlegenheit der mechanischen Urteilsbildung zeigen.

Überlegenheit der mechanischen Urteilsbildung

Bereits 1954 wertete Meehl 22 Studien zum Vergleich klinischer und statistischer Urteile aus und fand eine Überlegenheit der statistischen Urteilsbildung. Meehl war nicht nur Psychologieprofessor, sondern auch praktizierender Psychoanalytiker. Als solcher hatte er auch viel Sympathie für klinische Urteile (Grove und Lloyd 2006). Er erkannte, dass die Alternative zum statistischen Urteil zumeist nicht das klassische klinische Urteil ist, sondern das Urteil eines Menschen, der die Formel kennt und entscheidet, ob er ihr folgt oder nicht.

In einer weiteren Metaanalyse haben Grove et al. (2000) die bis dato vorliegenden Untersuchungen zur mechanischen und klinischen Urteilsbildung einer vergleichenden Bewertung unterzogen. Sie nahmen 136 Untersuchungen in ihre Analyse auf, die sich mit der Genauigkeit von Urteilen aus dem psychologischen und medizinischen Bereich befassten. Insgesamt erwies sich die mechanische Vorhersage der klinischen als überlegen. Die mittlere Effektstärke erwies sich mit $d=0,089$ aber als klein. Die große Streuung der Effektstärken veranlasste die Autorin und die Autoren, nach Moderatorvariablen zu suchen. Dazu untersuchten und überprüften sie die verwendete Definition der Effektstärke, das Publikationsjahr, die Stichprobengröße, das vorhergesagte Kriterium, das Training oder die Erfahrung der Urteilenden, die Informationsmenge sowie die Informationsart. Lediglich bei 2 Variablen – dem vorhergesagten Kriterium und der Informationsart – entdeckten sie einen Effekt. Die mechanische Urteilsbildung scheint der klinischen besonders dann überlegen zu sein, wenn medizinische und forensische Kriterien vorherzusagen sind und wenn die Informationen in Form von Interviewdaten vorliegen.

Bekanntes statistisches Urteilsmodell: Goldberg-Index

In einer weiteren Metaanalyse werteten Ægisdóttir et al. (2006) insgesamt 69 Studien ausschließlich aus dem klinischen Bereich aus, in denen mechanische und klinische Urteile direkt miteinander verglichen wurden (die Studie von Goldberg, 1965, gehörte dazu, s. o.). Über alle Studien mit ihren 173 Effektstärken hinweg ermittelten die Autorinnen und Autoren eine Gesamteffektstärke von $d=0,16$ zugunsten der mechanischen Methode. Wurden nur die 41 Studien herangezogen, deren Ergebnisse keine Ausreißer darstellten und bei denen auch eine Kreuzvalidierung vorgenommen wurde, betrug die Effektstärke $d=0,12$. Obwohl diese Schätzung deutlich konservativer ist, belegt sie immer noch die Überlegenheit der mechanischen/statistischen Vorhersage. Innerhalb dieser Gruppe von Studien suchten Ægisdóttir et al. (2006) nach Moderatorvariablen. Die erste Frage war die nach demjenigen Merkmal, bei dessen Vorhersage sich die statistische und die klinische Vorhersage am stärksten unterscheiden. Die mit $d=0,17$ größte Effektstärke betraf die Vorhersage von Straftaten bzw. Gewalttätigkeit. Das Ergebnis deckt sich mit dem von Grove et al. (2000), die für den forensischen Bereich die größte Überlegenheit der mechanischen Urteilsbildung festgestellt hatten. Eine zweite Gruppe von Moderatorvariablen betraf die Art der mechanischen Vorhersage. Lineare statistische Modelle hatten mit $d=0,15$ die höchste durchschnittliche Effektstärke und waren zugleich das am häufigsten angewandte Urteilsmodell. Schlechte Ergebnisse wurden mit lediglich rational begründeten Urteilsmodellen (mechanisch, aber nicht statistisch) erreicht ($d=0,03$). Zwei weitere Befunde sind erstaunlich: Man sollte vermuten, dass sich die Güte der klinischen Urteilsbildung der mechanischen zumindest annähert, wenn den klinisch Urteilenden einerseits zusätzliche Informationen zu Verfügung stehen, die nicht in die Formel einfließen, und andererseits die zu integrierenden Informationen aus ihrem jeweiligen Arbeitsgebiet stammen – sie also mit dem Datenmaterial grundsätzlich vertraut sind. Es ist jedoch jeweils das Gegenteil der Fall: Standen den klinisch Urteilenden genau dieselben Informationen zur Verfügung, die mit dem mechanischen Urteil verwertet wurden, war der Unterschied beider Urteilsmodelle mit $d=0,06$ geringer, als wenn die klinisch Urteilenden zusätzliche Informationen nutzen konnten ($d=0,13$). Mehr Informationen scheinen also das klinische Urteil nicht zu verbessern. Erstaunlicherweise zeigte sich dieses Bild auch für die Vertrautheit mit dem Datenmaterial – klinisch Urteilende näherten sich eher der Qualität von mechanischen Urteilen an, wenn die zu integrierenden Daten nicht aus ihrem Arbeitsgebiet stammten. Schließlich zeigt die Metaanalyse von Ægisdóttir et al. (2006) auch, dass es nicht ratsam ist, klinisch Urteilenden die Formeln zu zeigen, die zur mechanischen Urteilsbildung verwendet werden – klinische Urteile sind näher am mechanischen Urteil, wenn Urteilende die Formeln nicht kennen.

Moderatorvariablen beachten

Die jüngste Metaanalyse, die sich mit dem Vergleich klinischer und mechanischer Urteile befasst, liegt von Kuncel et al. (2013) vor. Von der Autorin und den Autoren wurden nur Studien aus dem Bereich der Eignungsdiagnostik in ihre Analyse einbezogen. Dabei korrespondierten die „klinisch“ zu einem Eignungsurteil integrierten Ergebnisse im Mittel zu $r=.28$ mit beruflicher Leistung. Wurden Eignungsurteile auf Basis „mechanisch“ integrierter Ergebnisse ermittelt, lag deren Korrelation mit beruflicher Leistung im Mittel bei $r=.44$. Wenngleich diese Korrelationen nur anhand weniger Studien errechnet wurden, so ist die in den Studien inkludierte Stichprobe mit jeweils über 1000 Personen beachtlich. Die Ergebnisse der Metaanalysen von Grove et al. (2000); Ægisdóttir et al. (2006); Kuncel et al. (2013) sind in Abb. 5.1 zusammengefasst.

Mechanische Urteilsbildung auch in der Eignungsdiagnostik überlegen

Insgesamt besteht also kein Zweifel daran, dass die mechanische (und hier besonders die statistische) Vorhersage der klinischen Urteilsbildung überlegen

Klinische Urteile anfällig für Fehler

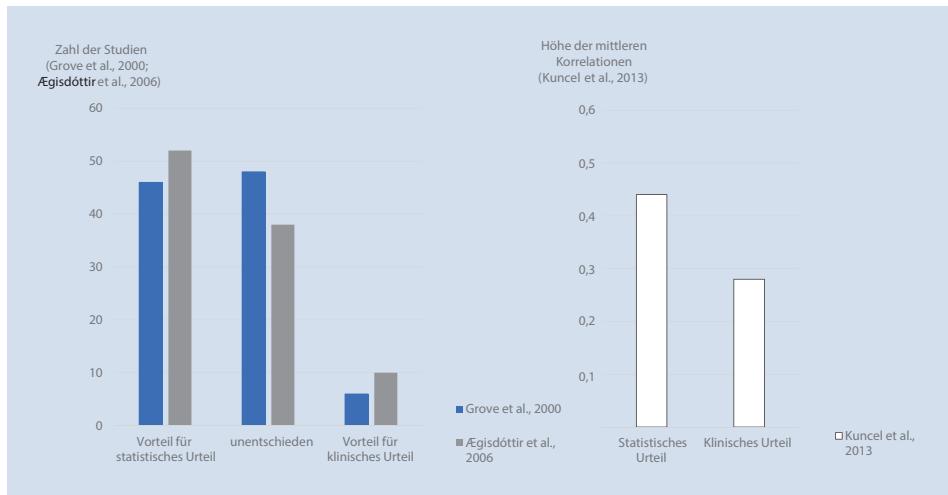


Abb. 5.1 Ergebnisse dreier Metaanalysen zum Vergleich klinischer vs. mechanischer Urteilsbildung

ist. Aber warum erreichen klinische Urteile nicht die Genauigkeit, die bei Anwendung von mechanischen Urteilsmodellen möglich ist? Grove et al. (2000) vermuten, dass die Anfälligkeit für bestimmte Urteilsfehler dafür verantwortlich ist. Verschiedene Untersuchungen belegen, dass Menschen oft die Basisrate ignorieren (also zu oft Diagnosen stellen, die statistisch selten und damit unwahrscheinlich sind), Informationen falsch gewichten, indem sie sich etwa hauptsächlich nach leicht verfügbaren Informationen richten („Availability-Heuristik“), oder die Regression zur Mitte vernachlässigen.

Trotz ihrer offensichtlichen Überlegenheit sollte die mechanische Vorhersage nicht kritiklos als universelle Lösung angesehen werden. Ein statistisches Urteilsmodell kann nur mit den Informationen konstruiert werden, die für viele Personen vorliegen. Zudem sind große Fallzahlen bei einer einheitlichen Fragestellung erforderlich. In vielen Fällen gibt es keine Alternative zum klinischen Urteil, da einschlägige Forschungsergebnisse, die eine Verrechnungsformel begründen könnten, schllichtweg fehlen. Aber auch statistische Urteile sollten nicht „blind“ angewandt werden. Meehl (1954) führt den fiktiven Fall des gebrochenen Beines als Argument an (sog. „broken leg cue“). Dazu nimmt er Folgendes an: Professorin M. geht am Dienstagabend ins Kino; das hat ein (in diesem Fall sehr einfaches) statistisches Urteilsmodell durch Analyse der Daten zu ihren bisherigen Kinobesuchen erkannt. Die Trefferquote lag bisher bei 90 %. Das Modell sagt auch für den nächsten Dienstag vorher, dass Professorin M. ins Kino gehen wird. Eingeweihte Urteilerinnen und Urteiler wissen aber, dass sich Professorin M. ein Bein gebrochen hat. Sie sagen vorher, dass der Kinobesuch am Dienstagabend ausfällt – obwohl sie auch die Prognose des statistischen Modells kennen.

„Broken leg cues“ beachten

! Es gilt also: Wenn ein seltenes Ereignis vorliegt, das von dem Prognosemodell nicht berücksichtigt wurde, aber für das zu treffende Urteil sehr relevant ist, sollte dieses Wissen – abweichend von einem mechanischen Vorgehen – genutzt werden.

Ganz allgemein gilt: In der diagnostischen Praxis sollten die Vorteile und Chancen beider Urteilsmodelle genutzt werden. Diagnostikerinnen und Diagnostiker sollten mechanische Vorhersagemodelle nutzen – ihnen aber nicht blind vertrauen. In begründeten Fällen sollte die mechanische Vorhersage korrigiert oder ganz durch eine klinische ersetzt werden (s. „broken leg cue“).

Vorhersagemodelle sollten zudem darauf geprüft werden, ob sie inhaltlich nachvollziehbar bzw. wissenschaftlich plausibel sind. Statistische Modelle beschreiben Zusammenhänge, indem sie viele, möglicherweise relevante Randbedingungen ignorieren. Eine Variable kann sich als guter Prädiktor herausstellen, ohne dass man den Wirkungsmechanismus versteht (was im Bereich der durch künstliche Intelligenz optimierten Prognosen besonders häufig vorkommen kann). In dieser Situation kann man auf die Weisheit des statistischen Modells vertrauen und sagen: „Es ist so, die empirischen Belege sind eindeutig.“ Man kann aber auch weiter forschen und nach moderierenden Faktoren suchen. Ein Beispiel ist die Rückfallprognose bei gewalttätigen Delinquentinnen und Delinquenten nach einem Psychiatrieaufenthalt. Rückfälle hängen von einer Reihe von Randbedingungen ab, darunter das soziale Umfeld der Patientin bzw. des Patienten nach der Entlassung: So erhöht kriminelle Nachbarschaft das Risiko erneuter Gewalt (Monahan 2003).

Schließlich sollten auch statistische Modelle kontinuierlich überprüft werden. Üblicherweise wird ein Modell an einer hinreichend großen Stichprobe entwickelt. Erst eine Kreuzvalidierung an unabhängigen Stichproben zeigt, ob das Modell noch Bestand hat. Es ist nicht zwangsläufig für alle Populationen und alle Zeiten gültig.

Eine interessante Frage ist, was praktisch tätige Psychologinnen und Psychologen über mechanische Urteilsmodelle denken und ob sie bereit sind, sie auch anzuwenden. Vrieze und Grove (2009) befragten dazu klinisch tätige Psychologinnen und Psychologen in den USA und wollten wissen, ob sie Informationen klinisch oder mechanisch integrieren. Es antworteten 180 Psychologinnen und Psychologen, die immerhin etwa 20 % ihrer Arbeitszeit mit Diagnostik verbrachten und damit über ausreichend Erfahrung verfügen sollten. Eine überwältigende Mehrheit von 98 % berichtet, klinische Urteilsbildung zu nutzen. Viele davon berichteten, normalerweise auch Informationen über statistische Modelle in ihr Urteil zu integrieren (56 %). Einige wandten nur die klinische Urteilsbildung an (47 %). Nur 31 % Befragten gaben an, mechanisch generierte Urteile (ohne weitere klinische Integration) zu nutzen.

Mechanischem Urteil nicht blind vertrauen

Wirkmechanismen und konfundierende Variablen beachten

Stetige Überprüfung von statistischen Modellen

Klinische Urteilsbildung überwiegt in der Praxis

Subjektive Gründe für den Verzicht auf mechanische Urteilsbildung

Diejenigen, die bei der Befragung von Vrieze und Grove (2009) angaben, keine mechanische Urteilsbildung zu nutzen, wurden um eine Begründung gebeten. Die am häufigsten genannten Gründe waren (in Klammern prozentuale Nennungshäufigkeit bei den Urteilenden, die keine mechanische Urteilsbildung anwenden; Mehrfachnennungen waren möglich):

- Mechanisches Urteilsmodell nicht verfügbar (40 %)
- Nicht gut genug mit der Methode vertraut, um sie bequem anzuwenden (36 %)
- Kann nicht alle Faktoren berücksichtigen, die für ein Urteil nötig sind (32 %)
- Nicht so genau wie andere Methoden (32 %)
- Zu teuer (27 %)
- Ineffizient (23 %)

Es gibt jedoch auch gute Gründe für die Anwendung eines mechanischen Urteilsmodells.

Gründe für die Anwendung der mechanischen Urteilsbildung

Folgende Gründe sprechen für die Anwendung der mechanischen Urteilsbildung:

- Bessere Urteile (s. die zuvor genannten Metaanalysen)
- Gleichbehandlung aller beurteilten Personen
- Keine Täuschung durch irrelevante Einflüsse (Sympathie etc.)
- Nachvollziehbares und prüfbares Vorgehen
- Übereinstimmung des Vorgehens mit Qualitätsstandards (wie z. B. der DIN 33430; ▶ Abschn. 6.4)

5.1.3.2 Verrechnungsregeln in Urteilsmodellen

Kompensatorische, konjunktive und disjunktive Entscheidungsstrategie

Neben der Frage, ob diagnostische Erkenntnisse intuitiv (klinisch) oder regelgeleitet (mechanisch) integriert werden, müssen weitere Annahmen über das spezielle Zusammenwirken der erhobenen Daten getroffen werden. Folgendes ist zusätzlich zu klären: Sollen sich mehrere Merkmale einer Person gegenseitig kompensieren können, d. h., kann eine geringe Ausprägung eines Merkmals durch eine hohe Ausprägung eines anderen Merkmals kompensiert werden? Oder sind Mindestausprägungen erforderlich, ohne dass andere Merkmale kompensatorisch wirken könnten, falls die geforderten Mindestausprägungen nicht gegeben sind? Oder reichen hohe Ausprägungen in einer bzw. wenigen Merkmalen aus, um zu einem diagnostischen Urteil zu kommen? Mit diesen 3 Fragen sind die kompensatorische, die konjunktive und die disjunktive Entscheidungsstrategie eingeführt.

Ein *kompensatorisches Entscheidungsmodell* bedeutet, dass sich die Prädiktoren gegenseitig ausgleichen (kompensieren) können. Mit anderen Worten: Niedrige Leistungen in einem Prädiktor können durch hohe in dem anderen wettgemacht werden. Für die diagnostische Entscheidung wird ein Gesamtwert berechnet, in den die Ausprägungen aller relevanten Merkmale gemeinsam einfließen. Die optimale Gewichtung kann man empirisch, u. a. mittels multipler Regression, ermitteln.

Kompensatorische Modelle liegen der diagnostischen Praxis sehr häufig zugrunde. So kann beispielsweise das Ziel der Versetzung in die nächste Schulkasse auch bei starken Defiziten in bestimmten Fächern erreicht werden, wenn diese durch besonders gute Leistungen in anderen Fächern ausgeglichen werden (d. h., eine Fünf in einem Nebenfach ist durch eine Zwei in einem anderen Fach kompensierbar). Bei der Auswahl von Studienbewerbern bzw. -bewerberinnen für die medizinischen Studiengänge wurden in Deutschland lange Zeit die Testergebnisse in einem Studieneignungstest mit der Abiturnote zu einem Gesamtwert verrechnet. Der Gesamtwert war entscheidend für die Zulassung.

Die kompensatorische Entscheidungsstrategie ist für 2 zu integrierende Informationen (Test A und Test B) in ▶ Abb. 5.2 dargestellt. Wir nehmen an,

Kompensatorische Entscheidungsstrategie kommt häufig vor

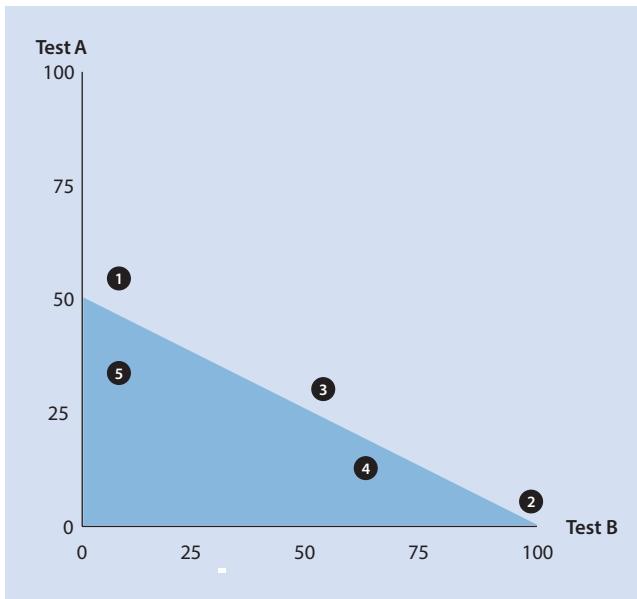


Abb. 5.2 Kompensatorisches Entscheidungsmodell. Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil

dass eine mechanische Urteilsbildung erfolgt und dass aufgrund früherer Daten ermittelt wurde, dass die optimale Vorhersage des Kriteriums (\hat{y}) mit folgender Formel zu erreichen ist:

$$\hat{y} = 2 \times A + 1 \times B$$

Zusätzlich soll angenommen werden, dass ein Gesamtwert von 100 nötig ist, um ein positives Urteil zu erhalten.

Für die Personen 1, 2 und 3 werden die Ergebnisse der beiden Tests zu einem positiven Gesamturteil verrechnet. Im Rahmen der Eignungsdiagnostik würden diese Personen beispielsweise ein Stellenangebot erhalten. Dabei schneiden sie ganz unterschiedlich in beiden Tests ab. Es liegt in der Natur des kompensatorischen Vorgehens, dass Person 1 sehr niedrige Werte in Test B durch das (doppelt gewichtete) Ergebnis in Test A kompensieren kann. Gleichsam kann Person 2 sehr niedrige Werte in Test A durch sehr hohe Werte in Test B ausgleichen. Auch ein mittelmäßiges Ergebnis in beiden Tests reicht aus (s. Person 3). Person 4 schneidet in Test B zwar besser ab als Person 3, erhält jedoch aufgrund ihres schlechteren Ergebnisses in Test A kein positives Urteil. Ebenso ergeht es Person 5 – sie erzielt zwar in Test B ein ebenso schlechtes Ergebnis wie Person 1, kann dies aber aufgrund des niedrigen Resultats in Test A nicht kompensieren.

Ein Testergebnis kann durch ein anderes kompensiert werden

Bei der *disjunktiven Entscheidungsstrategie*, auch „Oder-Strategie“ genannt, genügen entsprechend hohe Punktwerte in einem einzelnen der Prädiktoren (d. h., die Ergebnisse mehrerer diagnostischer Verfahren werden nicht zu einem Gesamtwert verrechnet). Eine solche Auswahlstrategie liegt dann nahe, wenn die durch das Kriterium geforderte Leistung entweder auf

Disjunktive Strategie: Eine deutlich ausgeprägte Variable kann ausreichen

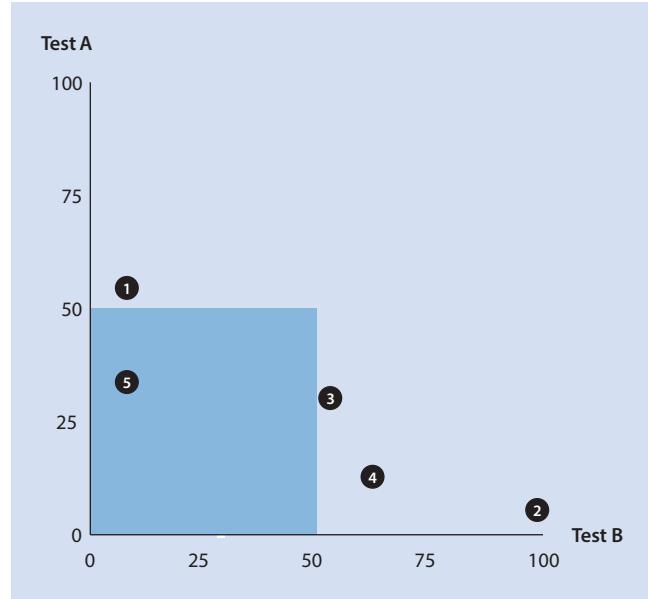
5

Konjunktive Entscheidungsstrategie fordert Mindestausprägungen

die eine oder auf die andere Weise erbracht werden kann. Eine gute Lehrerin bzw. ein guter Lehrer mag sich durch eine große Fähigkeit in Mathematik oder in Geschichte empfehlen – für das jeweils andere Fach wird sie bzw. er dann nicht eingesetzt. Natürlich kann auch ein niedrig ausgeprägter Prädiktor ausreichen, sofern eine niedrige Ausprägung gefordert war (beispielsweise geringe Neurotizismuswerte).

Wendet man die disjunktive Entscheidungsstrategie auf das Beispiel aus □ Abb. 5.2 an und geht davon aus, dass ein Wert von 50 in Test A oder in Test B erreicht werden muss, fallen die Urteile für die Personen 1 bis 5 teilweise anders aus (□ Abb. 5.3). Nun erhält neben den Personen 1, 2 und 3 auch Person 4 ein positives Urteil.

In vielen Fällen sind Mindestausprägungen in einzelnen Merkmalen erforderlich, um insgesamt als „positiv“ klassifiziert zu werden. In der Klinischen Psychologie gilt dies für verschiedene Symptombereiche, die in einer Mindestausprägung vorliegen müssen, um von dem Vorliegen einer psychischen Störung sprechen zu können. Auch in der Eignungsdiagnostik gibt es diese Fälle häufig. So kann eine Chirurgin bzw. ein Chirurg mangelnde feinmotorische Kompetenz nicht durch Intelligenz kompensieren. Dasselbe gilt für Pilotinnen und Piloten – sie können fehlende Sehtüchtigkeit nicht durch gute räumliche Orientierung ausgleichen. Hier ist eine *konjunktive Entscheidungsstrategie* angebracht, die auch als „Und-Strategie“ bezeichnet wird. Da für ein positives Urteil mehrere Cut-off-Werte überwunden bzw. mehrere „Hürden“ übersprungen werden müssen, sind auch die Bezeichnungen „Multiple-Cut-off-Modell“ und „Multiple-Hurdle-Modell“ gebräuchlich.

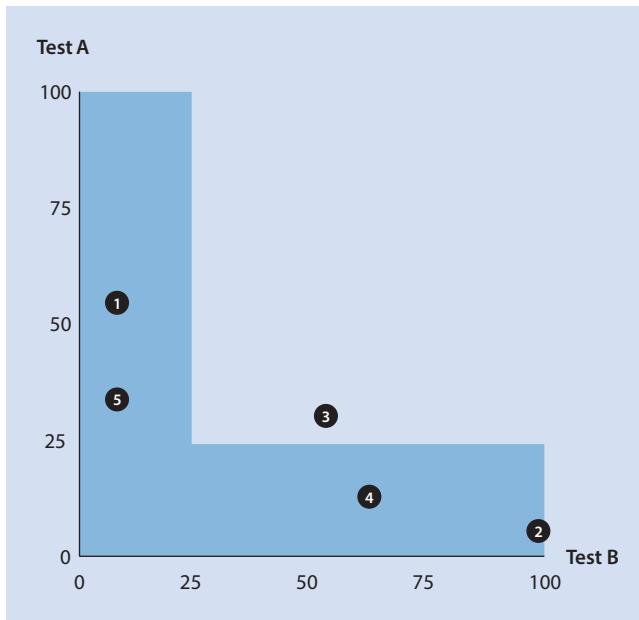


□ Abb. 5.3 Disjunktives Entscheidungsmodell. Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil

Wendet man die konjunktive Entscheidungsstrategie auf das Beispiel aus □ Abb. 5.2 und □ Abb. 5.3 an und geht davon aus, dass ein Wert von mindestens 25 in Test A und in Test B erreicht werden muss, erhält nur Person 3 ein positives Urteil (□ Abb. 5.4). In diesem Fall ist die konjunktive Strategie die strengste Entscheidungsstrategie, und das obwohl der Cut-off-Wert bei 25 statt bei 50 lag (s. „Multiple-Hurdle-Problem“).

Die Wahl angemessener Cut-off-Werte ist nicht trivial. Manchmal sind diese durch gesetzliche Regelungen oder Konventionen (z. B. Klassifikationssysteme für psychische Störungen) vorgegeben. In manchen Fällen liegen empirisch begründete Vorschläge für Cut-off-Werte vor (► Abschn. 5.1.3.3). In vielen Fällen ist dies jedoch nicht der Fall – dann müssen Cut-off-Werte für den vorliegenden diagnostischen Zweck angemessen gewählt werden. Ziel muss es dabei sein, dass ein vorliegender positiver Zustand (z. B. geeignet im Sinne der Stellenanforderung) auch als solcher erkannt wird. Das heißt, diejenigen Personen, die den Cut-off-Wert übertreffen, sollten auch tatsächlich „Positive“ sein. Weiterhin muss gewährleistet sein, dass ein vorliegender negativer Zustand (z. B. nicht geeignet im Sinne der Stellenanforderung) ebenfalls als solcher erkannt wird – eben dadurch, dass solche Personen den Cut-off-Wert nicht überschreiten. Dies wird in ► Abschn. 5.1.3.3 näher erläutert.

Cut-off-Werte sinnvoll wählen



□ Abb. 5.4 Konjunkтивes Entscheidungsmodell. Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil

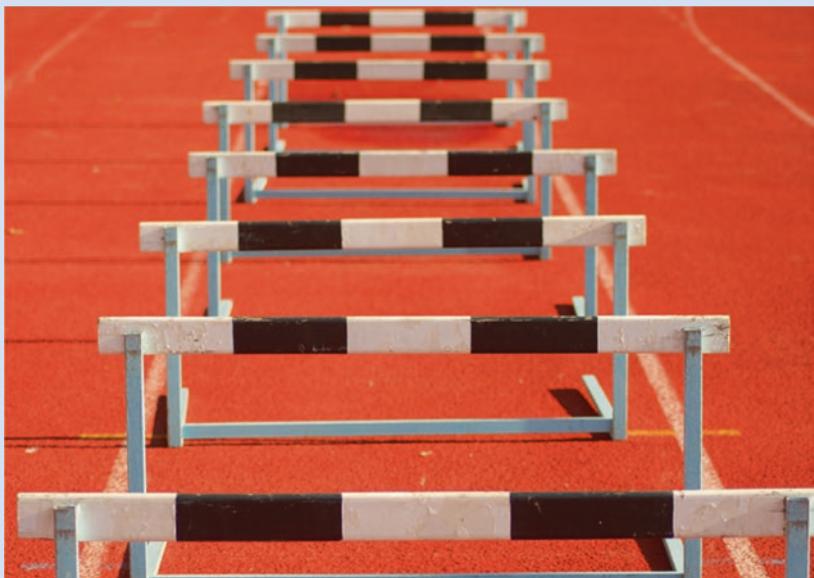
Das Multiple-Hurdle-Problem

Ziegler und Bühner (2012) legen dar, dass selbst bei geringen Hürden (z. B. nicht schlechter als 1 Standardabweichung unter dem Mittelwert der repräsentativen Vergleichsgruppe) die Wahrscheinlichkeit, alle Hürden zu bewältigen, gering sein kann. Dies ist insbesondere dann der Fall, wenn die Hürden aus unabhängigen, d. h. unkorrelierten Messungen bestehen. Die Basiswahrscheinlichkeit, einen Wert zu erreichen, der größer ist als der Mittelwert minus 1 Standardabweichung der repräsentativen Vergleichsgruppe, ist zwar hoch ($= 0,84$; da bei normalverteilten Messwerten 84 % der Normgruppe diesen Wert übertreffen), für 5 unabhängige Hürden würden diese Wahrscheinlichkeiten jedoch multipliziert. Damit liegt die Basiswahrscheinlichkeit, 5 Hürden zu „überspringen“ nur noch bei $0,84 \times 0,84 \times 0,84 \times 0,84 \times 0,84 \approx 0,42$ (Ziegler und Bühner 2012, S. 52).

Diese Wahrscheinlichkeit erhöht sich, wenn die den Hürden zugrunde liegenden Messungen korreliert sind, Sprünge über die Hürden also keine unabhängigen Ereignisse sind. Mit der Formel

$$P_{A \cap B} = P_A \times P_B + r_{XA;XB} \times \sqrt{P_A \times (1 - P_A) \times P_B \times (1 - P_B)}$$

kann die Wahrscheinlichkeit, dass die Ereignisse A und B gemeinsam auftreten, berechnet werden – unter Berücksichtigung der Korrelation zwischen den Indikatorvariablen für A und B (S. Pohl, persönliche Mitteilung, 10. September 2019). Belässt man die „Höhe“ der 5 Hürden so, wie bei Ziegler und Bühner (2012) beschrieben, und nimmt eine Korrelation von .30 zwischen jeder nachfolgenden und den jeweils vorherigen Hürden an, so liegt die Basiswahrscheinlichkeit, alle 5 Hürden zu überspringen (wobei als Hürde nach wie vor gilt: Der Wert soll größer sein als der Mittelwert minus 1 Standardabweichung der repräsentativen Vergleichsgruppe), bei 0,57. Somit ist die Wahrscheinlichkeit, alle Hürden zu überspringen, zwar gestiegen, aber das Multiple-Hurdle-Problem ist auch hier (bei korrelierten Messungen) noch virulent.



Multiple Hürden. (© cxvalentina/stock.adobe.com).

Neben dem „reinen“ kompensatorischen, disjunktiven und konjunktiven Vorgehen sind Kombinationen dieser Strategien möglich. So könnte man das Vorgehen anhand des kompensatorischen Modells (Abb. 5.2) um einen Cut-off-Wert für Test A ergänzen. Man könnte beispielsweise alle Personen, deren Wert in Test A < 25 ist, als „negativ“ klassifizieren und nur für Personen mit einem Wert ≥ 25 in Test A die kompensatorische Verrechnung der Ergebnisse aus den Tests A und B anwenden (Abb. 5.5). Anders als bei dem rein kompensatorischen Vorgehen würde nun Person 2 nicht mehr als „positiv“ beurteilt werden.

Verlangt man nun noch einen Cut-off-Wert für Test B von ≥ 25 , würde nur noch Person 3 ein positives Urteil erhalten (Abb. 5.6).

Kombinationen der Entscheidungsstrategien

5.1.3.3 Empirische Festlegung der Cut-off-Werte

Der Sinn von Cut-off-Werten besteht darin, zu spezifizieren, ab wann ein zu diagnostizierender Zustand (eine Erkrankung, Eignung im beruflichen Sinne etc.) wahrscheinlich vorliegt. Wird der Cut-off-Wert (in einem diagnostischen Verfahren oder in einem über mehrere diagnostische Verfahren gebildeten Gesamtwert) überschritten, gehen Diagnostikerinnen und Diagnostiker davon aus, dass der fragliche Zustand vorliegt. Wird der Cut-off-Wert nicht überschritten, ist davon auszugehen, dass der fragliche Zustand nicht vorliegt.

Cut-off-Werte spezifizieren, wann ein zu diagnostizierender Zustand wahrscheinlich vorliegt

Natürlich muss der Anspruch bei diesem Vorgehen sein, dass diese Annahmen möglichst auch der Realität entsprechen. Das bedeutet: Wenn Diagnostikerinnen und Diagnostiker anhand des Cut-off-Wertes und auf Basis der diagnostischen Instrumente sagen, ein fraglicher Zustand läge vor, dann sollte das idealerweise möglichst oft der Realität entsprechen – also der fragliche Zustand tatsächlich vorliegen. Dann würde man von einem „Treffer“ sprechen – was etwas umgangssprachlich anmutet, aber tatsächlich als Begriff in der Signalentdeckungstheorie etabliert ist. Ebenso wünschenswert wäre es, wenn eine Aussage von Diagnostikerinnen und Diagnostikern, dass der fragliche Zustand nicht vorliegt, ebenfalls möglichst oft der Realität entspricht. Hierbei spricht man von „korrekten Zurückweisungen“. Zwei andere Fälle sollten möglichst selten auftreten:

Ziel: Anhand des Cut-off-Wertes möglichst viele korrekte Diagnosen stellen

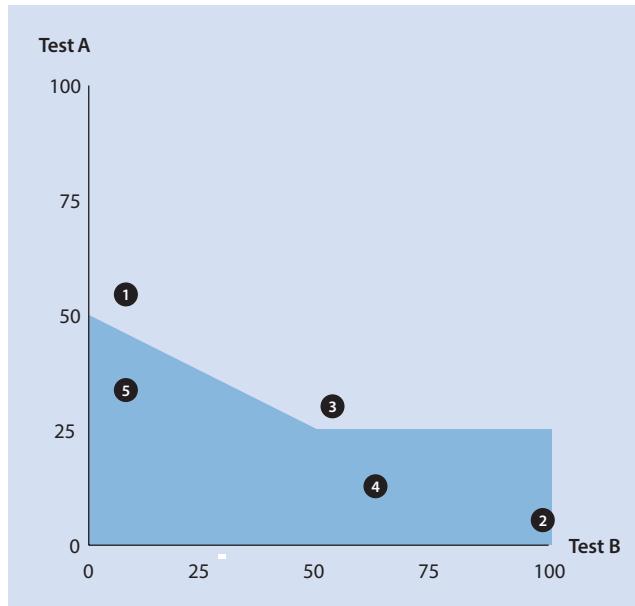
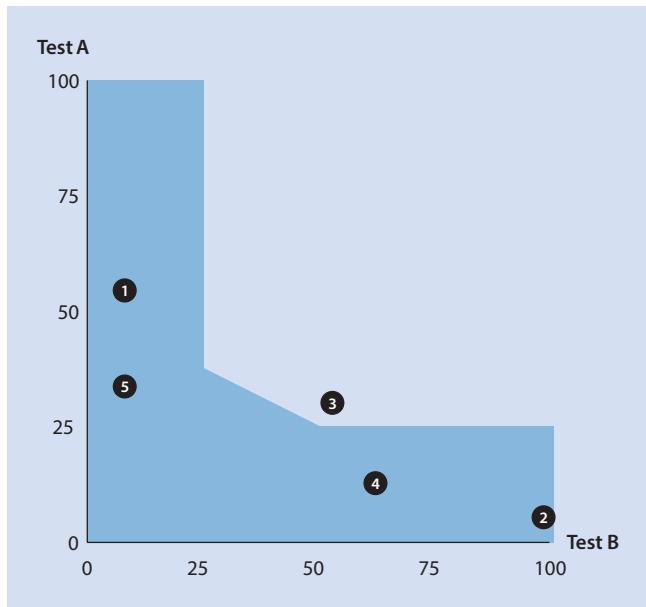


Abb. 5.5 Kombination der Verrechnungsstrategien. Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil



■ Abb. 5.6 Kombination der Verrechnungsstrategien. Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil

- Diagnostizierende denken, ein Zustand läge vor, was aber gar nicht so ist (Alpha-Fehler).
- Diagnostizierende denken, ein Zustand läge nicht vor, in der Realität liegt er aber vor (Beta-Fehler).

Die 4 möglichen Kombinationen des dichotomen diagnostischen Urteils mit einem realen, ebenfalls dichotom ausgeprägten Zustand sind in ■ Abb. 5.7 dargestellt.

Im besten Falle, der aber nicht realistisch ist, macht man keine Fehler – alle diagnostischen Aussagen sind Treffer oder korrekte Zurückweisungen. Der verwendete Cut-off-Wert und das zugehörige diagnostische Instrument bzw. die zugehörigen diagnostischen Instrumente würden perfekt trennen zwischen positiven und negativen Fällen. Das kann nur gelingen, wenn sich die Verteilungen (der mit den relevanten Instrumenten generierten Werte) der beiden Gruppen (tatsächlich „Positive“ und „Negative“) nicht überschneiden. In ■ Abb. 5.8 sieht man, dass rechts des Cut-off-Wertes nur tatsächlich „Positive“ (blaue Balken) und links des Cut-off-Wertes nur tatsächlich „Negative“ (graue Balken) liegen. Bestünden beide Gruppen aus 100 Personen, so würde sich ein Vier-Felder-Schema ergeben wie in ■ Abb. 5.8 dargestellt.

Leider gelingt es fast nie, 2 Gruppen anhand eines oder mehrerer diagnostischer Instrumente perfekt zu unterscheiden. Fast immer muss man akzeptieren, dass Personen den Cut-off-Wert überschreiten, aber dennoch zur Verteilung der „Negativen“ gehören. Auch der umgekehrte Fall ist in Kauf zu nehmen – Personen können den Cut-off-Wert nicht überschreiten, aber dennoch zur Verteilung der „Positiven“ gehören. Wie häufig die daraus fast zwangsläufig resultierenden Fehlklassifikationen vorkommen, ist eine Frage der Überlappung der Verteilungen. Anders gesagt: Es kommt darauf an, wie gut das oder die verwendeten Messinstrumente Verteilungen für die beiden Gruppen produzieren, die sich nicht überlappen. In dem in ■ Abb. 5.9 präsentierten Beispiel überlappen sich die Verteilungen leicht. Dadurch werden bei dem gewählten Cut-off-Wert 11 Personen als nicht „positiv“ klassifiziert, obwohl sie es sind. Weitere 12 Personen werden als „positiv“ beurteilt, sind es aber nicht.

Perfekte Diagnosen nur bei nicht überlappenden Verteilungen

Fehlklassifikationen fast unvermeidlich

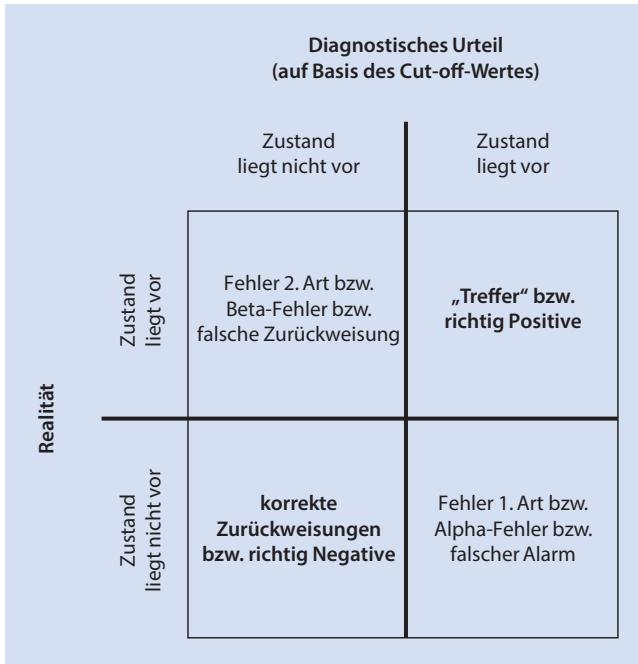


Abb. 5.7 Vier-Felder-Schema zur Konvergenz zwischen diagnostischem Urteil und realer Gegebenheit

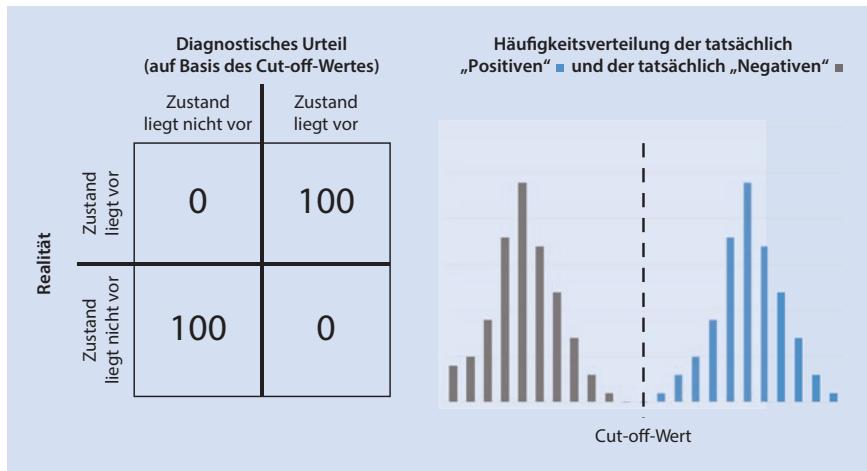


Abb. 5.8 Perfekte Unterscheidung anhand eines Cut-off-Wertes

Es kann auf den ersten Blick durchaus akzeptabel erscheinen, bei 200 Entscheidungen, die in dem in Abb. 5.9 dargestellten Beispiel vorgenommen werden, insgesamt 23 Fehlentscheidungen zu treffen. Man stelle sich jedoch vor, in dem Beispiel würden 11 Personen, bei denen eine prinzipiell behandelbare psychische Störung vorliegt, als gesund klassifiziert („positiv“ bedeutet in diesem Fall, dass eine zu entdeckende Gegebenheit vorliegt, analog zu „HIV positiv“). Dies wäre für die betreffenden Personen sicherlich alles andere als „akzeptabel“. Gleiches gilt für die 12 Personen, die eigentlich gesund sind, aber als krank klassifiziert werden. Sie werden nach dem diagnostischen Urteil in der Wahrnehmung leben, unter einer psychischen Störung zu leiden, und sich unnötigerweise einer Behandlung unterziehen, was ebenfalls im Einzelfall problematisch ist.

Fehlklassifikation im Einzelfall problematisch

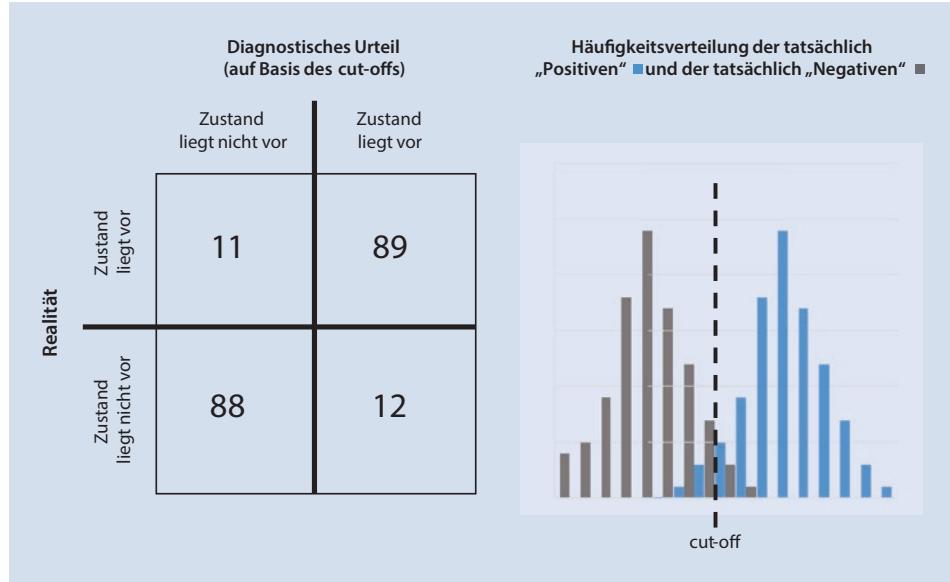


Abb. 5.9 Unterscheidung anhand eines Cut-off-Wertes bei überlappenden Verteilungen

! Es ist jedoch auch zu betonen, dass bei überlappenden Verteilungen – wie in diesem Beispiel und in vielen Bereichen der realen diagnostischen Beurteilung – eine für alle diagnostizierten Personen fehlerfreie Aussage nicht möglich ist. Sofern keine bessere Form der Diagnostik verfügbar ist, muss ein gewisser Fehleranteil hingenommen werden. Keinesfalls darf aufgrund solcher Fehlerquoten gänzlich von wissenschaftlich fundierter Diagnostik abgewichen und stattdessen auf Laienurteile oder pseudowissenschaftliche Vorgehensweisen vertraut werden.

Durch Wahl des Cut-off-Wertes einen Fehler zulasten des anderen minimieren

Es besteht jedoch die Möglichkeit, eine der beiden Fehlerarten zu minimieren. Möchte man einen vorliegenden Zustand auf keinen Fall übersehen – also auf keinen Fall eine falsche Zurückweisung (Beta-Fehler) vornehmen – muss man lediglich den Cut-off-Wert niedrig genug ansetzen. Im Zweifel könnte man den Cut-off-Wert so weit verringern, dass alle zu beurteilenden Personen als „positiv“ klassifiziert würden. Damit wäre sichergestellt, dass alle tatsächlich „Positiven“ erkannt würden. Natürlich geht dieses Vorgehen zulasten der anderen Fehlerart – viele Personen würden nun als „positiv“ beurteilt, obwohl sie es tatsächlich nicht sind (Alpha-Fehler). Im vorliegenden Beispiel (Abb. 5.10) könnte man den Cut-off-Wert so weit nach unten verschieben, dass keine Beta-Fehler vorkommen, aber dann in 42 Fällen der Alpha-Fehler vorliegt.

In diesem Fall wäre das Vorgehen sehr *sensitiv*, da alle als „positiv“ vorliegenden Fälle erkannt werden. Es wäre aber wenig *spezifisch*, da nur etwas mehr als die Hälfte der nicht „positiv“ vorliegenden Fälle auch als solche erkannt würden.

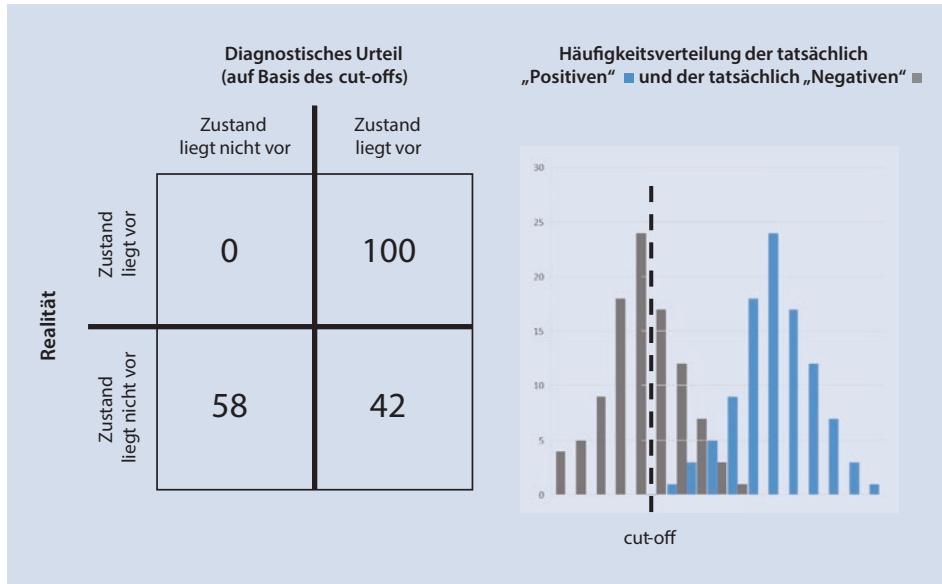


Abb. 5.10 Vollständige Vermeidung des Beta-Fehlers

Definition

Sensitivität ist definiert als die Wahrscheinlichkeit, mit der ein vorliegender positiver Zustand als solcher erkannt wird. Sie beschreibt also den Anteil der „Treffer“ an allen tatsächlich Positiven.

$$\text{Sensitivität} = \frac{\text{Treffer}}{\text{Treffer} + \text{falsche Zurückweisungen}}$$

Spezifität bezeichnet die Wahrscheinlichkeit, mit der ein vorliegender negativer Zustand als solcher erkannt wird. Das heißt, sie beschreibt den Anteil der korrekten Zurückweisungen an allen tatsächlich zurückzuweisenden Personen.

$$\text{Spezifität} = \frac{\text{korrekte Zurückweisungen}}{\text{korrekte Zurückweisungen} + \text{falsche Alarme}}$$

Natürlich lassen sich Sensitivität und Spezifität auch als bedingte Wahrscheinlichkeiten formulieren: $\text{Sensitivität} = p(U^+ | R^+)$, also die Wahrscheinlichkeit eines positiven diagnostischen Urteils, gegeben die reale Ausprägung, ist positiv; $\text{Spezifität} = p(U^- | R^-)$, also die Wahrscheinlichkeit eines negativen diagnostischen Urteils, gegeben die reale Ausprägung, ist negativ.

In dem in Abb. 5.10 skizzierten Fall läge die Sensitivität bei 1.00 und die Spezifität bei .58. Sofern es tatsächlich sehr wichtig wäre, falsche Zurückweisungen zu vermeiden, und falsche Alarne zu verschmerzen wären, könnte man den Cut-off-Wert in der Tat so wählen, dass die Sensitivität maximal ist. Läge beispielsweise ein absoluter Mangel an geeigneten Bewerberinnen und Bewerbern vor, würde man eigentlich geeignete auf keinen Fall übersehen wollen und sich für ein Vorgehen mit hoher Sensitivität entscheiden. Die daraus ggf. resultierende geringe Spezifität wäre ggf. unkritisch, wenn eigentlich ungeeignete Eingestellte beispielsweise zunächst eine betreute Probearbeitsphase durchlaufen und umfangreich geschult werden.

Eine geringe Spezifität ist also notfalls hinnehmbar, wenn das diagnostische Urteil oder auch dessen Konsequenzen nachträglich revidiert werden

Vorgehen mit sehr hoher Sensitivität (und niedriger Spezifität)

Früherkennung von Suizidalität erfordert hohe Sensitivität

können. Ein weiteres Beispiel macht deutlich, dass eine zunächst geringe Spezifität sogar im Rahmen eines stufenweisen Vorgehens sehr sinnvoll sein kann: Bei Patientinnen und Patienten, die in eine psychosomatische Klinik aufgenommen werden, wird mithilfe eines ökonomischen Fragebogens geprüft, ob sie suizidgefährdet sind. Die Interpretation des „Suizidwertes“ ist aber wenig valide. Der Cut-off-Wert wird deshalb so niedrig gelegt, dass sehr wahrscheinlich alle suizidgefährdeten Personen entdeckt werden. Von 100 mit dem Fragebogen untersuchten Personen sind deshalb 30 als positiv (suizidgefährdet) zu diagnostizieren. Diese werden mithilfe eines aufwendigen diagnostischen Interviews, das weitaus valide Interpretationen erlaubt, erneut auf ihre Suizidneigung untersucht. Nicht alle der 5 nun positiv diagnostizierten Personen würden während des Klinikaufenthalts einen Suizidversuch unternehmen; sie werden dennoch alle unter besondere Beobachtung gestellt oder einer speziellen Psychotherapie zugeführt. Die 3 Personen, die innerhalb von 2 Wochen keinerlei Anzeichen für eine Suizidneigung erkennen lassen, können nun in den normalen Klinikalltag integriert werden; es verbleiben 2 Personen unter Beobachtung.

In vielen Fällen wird es weniger klar sein, ob Sensitivität oder Spezifität zu priorisieren sind. Dann gilt es, einen Cut-off-Wert zu wählen, der beides – Sensitivität und Spezifität – optimiert. Als grafische Entscheidungshilfe für den bestmöglichen Cut-off-Wert kann eine ROC-Kurve (von Receiver Operating Characteristic aus der Signalentdeckungstheorie; Schäfer 1989) erstellt werden. Dazu trägt man für jeden Cut-off-Wert die resultierende Sensitivität und Spezifität in ein Koordinatensystem ein. Aus Gründen der besseren Darstellung verwendet man statt der Spezifität jedoch den inversen Wert (also $1 - \text{Spezifität}$).

Verlauf der ROC-Kurve

Je weiter die beiden Verteilungen auseinanderliegen, desto eher verlaufen ROC-Kurven so wie in dem Beispiel in Abb. 5.11, also sehr nahe an der linken oberen Ecke. Würden die Verteilungen vollständig übereinanderliegen, verlief die ROC-Kurve auf der gestrichelt gezeichneten Diagonalen. In diesem sehr ungünstigen Fall würde jede Verbesserung der Sensitivität linear und in gleichem Maße zu einer Verschlechterung der Spezifität führen. Anders gesagt: Eine Sensitivität nahe 1 wäre nur durch eine Spezifität nahe 0 zu erreichen (und umgekehrt).

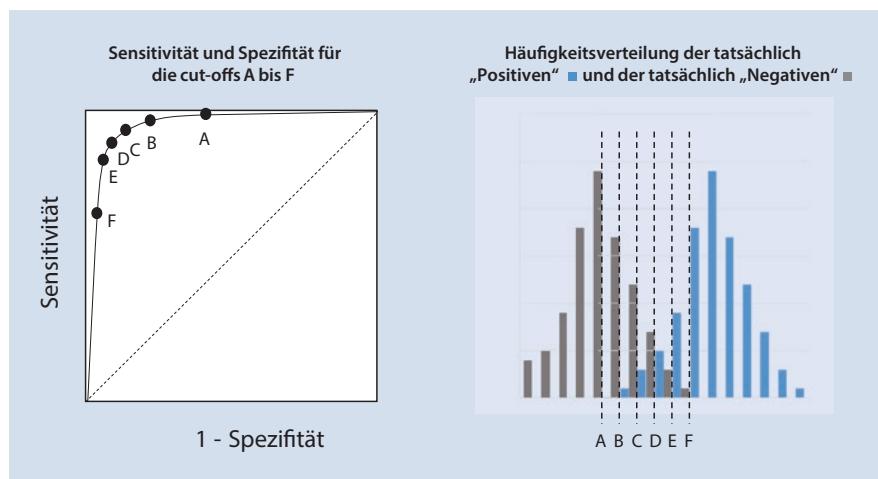


Abb. 5.11 ROC-Kurve der Sensitivität und Spezifität auf Basis der Cut-off-Werte A bis F

Der Cut-off-Wert mit dem besten Kompromiss aus Sensitivität und Spezifität ist jener, der am weitesten in der linken oberen Ecke des Quadrats liegt. In unserem Beispiel wäre das der Cut-off-Wert D. Allerdings liegt auch C recht weit in der oberen Ecke, und es ist dem bloßen Augenschein nach nicht einfach zu entscheiden, welcher weiter in der Ecke liegt. In diesen Fällen empfiehlt sich eine rechnerische Bestimmung des optimalen Cut-off-Wertes anhand des Youden-Index (Goldhammer und Hartig 2012).

Identifikation des besten Cut-off-Wertes

Youden-Index

Zur rechnerischen Bestimmung des optimalen Cut-off-Wertes kann der Youden-Index herangezogen werden. Dabei wird für jeden Cut-off-Wert der Youden-Index berechnet:

$$\text{Youden - Index} = \text{Sensitivität} + \text{Spezifität} - 1$$

Der Cut-off-Wert mit dem höchsten Youden-Index wird gewählt.

Mit der Sensitivität und der Spezifität sind jedoch noch nicht alle anhand des Vier-Felder-Schemas (vgl. Abb. 5.7) bestimmbaren Wahrscheinlichkeiten eingeführt. Sensitivität und Spezifität beschreiben lediglich die Wahrscheinlichkeiten, dass reale Gegebenheiten durch diagnostische Instrumente richtig erkannt werden. Bislang nicht besprochen wurden die Wahrscheinlichkeiten, dass Aussagen aus diagnostischen Instrumenten der Realität entsprechen. Also: Wie wahrscheinlich ist es, dass bei einem positiven Ergebnis eines Tests eine Person tatsächlich zu den „Positiven“ gehört? Wie wahrscheinlich ist es, dass bei einem negativen Ergebnis eines Tests eine Person tatsächlich zu den „Negativen“ gehört? Die erste Frage wird durch den positiven Prädiktionswert, die zweite durch den negativen Prädiktionswert beantwortet.

Positiver und negativer Prädiktionswert

Definition

Der **positive Prädiktionswert** ist definiert als die Wahrscheinlichkeit, mit der eine positive Diagnose zutreffend ist. Sie beschreibt also den Anteil der „Treffer“ an allen als positiv laut Test diagnostizierten Personen. In der beruflichen Eignungsdiagnostik beschreibt der Begriff „Trefferquote“ das Gleiche. Diese Trefferquote gibt an, wie viel Prozent der Personen, die aufgrund eines positiven diagnostischen Urteils eingestellt wurden (also Treffer + falsche Alarme), sich tatsächlich als geeignet (als Treffer) erweisen.

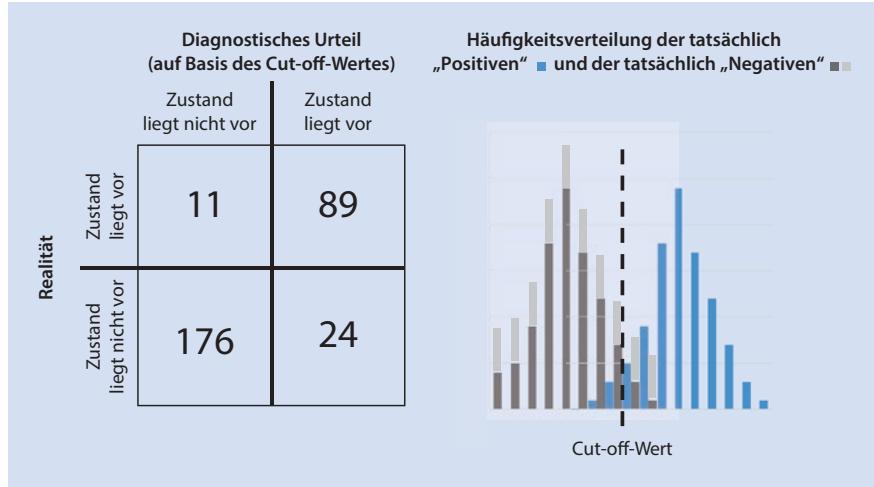
$$\text{Positiver Prädiktionswert} = \frac{\text{Treffer}}{\text{Treffer} + \text{falsche Alarme}}$$

Der **negative Prädiktionswert** bezeichnet die Wahrscheinlichkeit, mit der eine negative Diagnose zutreffend ist. Sie beschreibt also den Anteil der korrekten Zurückweisungen an allen laut Test negativ diagnostizierten Personen.

$$\text{Negativer Prädiktionswert} = \frac{\text{korrekte Zurückweisungen}}{\text{alle Zurückweisungen}}$$

Positiver und negativer Prädiktionswert können ebenfalls als bedingte Wahrscheinlichkeiten verstanden werden, wobei gilt: positiver Prädiktionswert = $p(R^+ | U^+)$, also die Wahrscheinlichkeit einer real vorliegenden Ausprägung, gegeben das diagnostische Urteil war positiv; negativer Prädiktionswert = $p(R^- | U^-)$, also die Wahrscheinlichkeit einer real negativen Ausprägung, gegeben das diagnostische Urteil war negativ.

5



■ Abb. 5.12 Veränderung der Grundrate

Positiver und negativer
Prädiktionswert abhängig von der
Grundrate

Sensitivitt und Spezifitt lassen sich unabhangig von den Grundraten oder der Pravenz bestimmen; hingegen unterliegen die Pradiktionswerte sehr stark deren Einfluss. Warum dies so ist, lsst sich leicht an dem bislang verwendeten Beispiel verdeutlichen. Nehmen wir dazu einfach an, dass zu der Verteilung in Abb. 5.9 bei den „Negativen“ nochmals 100 Personen dazukommen, die sich gleichmig auf alle Testwerte der ursprnglich „Negativen“ verteilen (Abb. 5.12).

Wie man sieht, ändert sich an der Sensitivität nichts, da die Zahl der „Positiven“ gleichgeblieben ist. Auch die Spezifität ändert sich nicht, da das Verhältnis konstant geblieben ist (vorher 88 zu 100, nun 176 zu 200). Das Verhältnis von Treffern zu Treffern plus falschen Alarmen (= positiver Prädiktionswert) hat sich allerdings verändert (vorher 89 zu 101, nun 89 zu 113). Durch die Verringerung der Grundrate des als „positiv“ deklarierten, d. h. des zu entdeckenden Zustands verschlechtert sich also der positive Prädiktionswert. Der negative Prädiktionswert hingegen verbessert sich in diesem Fall (vorher 88 zu 99, nun 176 zu 187).

Wendet man das Bayes-Theorem auf die zuvor genannten bedingten Wahrscheinlichkeiten an, wird nochmals der Einfluss der Grundrate deutlich. Der positiver Prädiktionswert, $p(R^+ | U^+)$, lässt sich gemäß Bayes-Theorem (s. z. B. Eid et al. 2015) formulieren als:

$$p(R^+|U^+) = \frac{p(U^+|R^+) \times p(R^+)}{p(U^+|R^+) \times p(R^+) + p(U^+|R^-) \times p(R^-)}$$

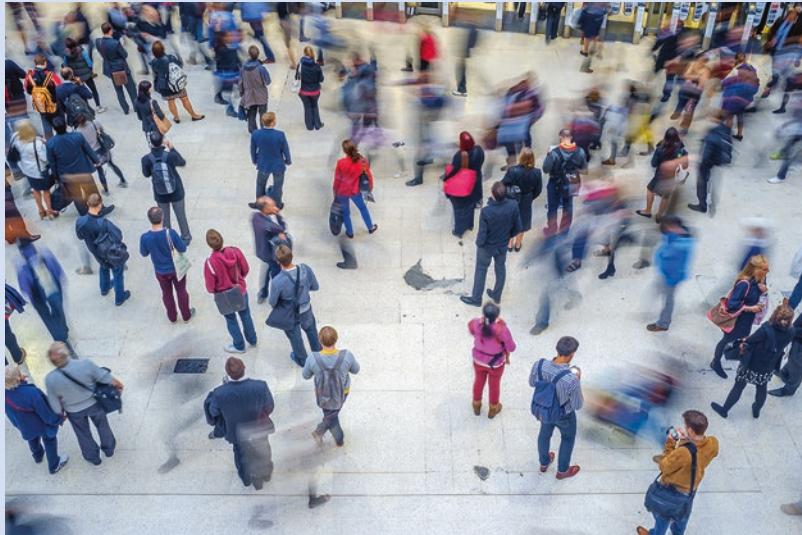
Wie man sieht, wird $p(R^+ | U^+)$ klein, wenn die Grundrate des zu erkennenden realen Zustands, $p(R^+)$, gering ist.

! Es wird umso schwieriger, einen vorliegenden Zustand zu entdecken, je seltener dieser in der untersuchten Population vorkommt.

Es ist festzuhalten, dass die Treffsicherheit von diagnostischen Urteilen nicht nur von der Validität der Interpretation eines Testwertes abhängt (► Abschn. 2.6.3). Diese würde in den hier diskutierten Beispielen prinzipiell zu gut unterscheidbaren Verteilungen der beiden Gruppen führen. Allerdings haben wir gezeigt, dass die Grundrate der zu identifizierenden Gegebenheit ebenfalls entscheidend ist. Darüber hinaus beeinflusst die Wahl des Cut-off-Wertes die Qualität der Entscheidungen.

Modellversuch zur automatischen Gesichtserkennung

In ihrer „Unstatistik des Monats“ (► <https://www.rwi-essen.de/unstatistik>) besprechen Gigerenzer et al. (2018) im Oktober 2018 die Aussage des Bundesinnenministeriums, dass der Modellversuch zur automatischen Gesichtserkennung, der an einem großen Bahnhof durchgeführt wurde, erfolgreich gewesen sei.



Gesichtserkennung an Bahnhöfen. (© Chris Mann/stock.adobe.com)

Die Autoren zeigen mithilfe von Sensitivität und Spezifität Folgendes auf: Wenn man von 10 gesuchten Testpersonen 8 durch die Gesichtserkennung richtig erkennt, ist dies noch lange nicht als Erfolg der Gesichtserkennung zu werten. Zwar liegt damit die Sensitivität der Gesichtserkennung (beim Erkennen der gesuchten Testpersonen) bei .80 – aber wie sieht es mit der Spezifität aus? Dazu berichten Gigerenzer et al. (2018), dass pro 1000 nicht gesuchter Personen eine fälschlicherweise als Zielperson klassifiziert wurde. Die Spezifität ist also ebenfalls hoch, sie liegt bei .999. Das Problem wird erst offenkundig, wenn man die Grundraten mit einbezieht. Dazu nehmen die Autoren der „Unstatistik des Monats“ an, dass sich an deutschen Bahnhöfen täglich etwa 100 Zielpersonen (z. B. sog. „Gefährderinnen“ oder „Gefährder“) aufhalten, von denen 80 erkannt werden. Dazu kommen 11,9 Mio. Reisende, von denen 0,1 % zu Unrecht „erkannt“ werden. Das System würde also 11.900 plus 80 Personen als Zielpersonen identifizieren. Die Wahrscheinlichkeit, dass es sich tatsächlich um eine gesuchte Person handelt, wenn das System anschlägt (= positiver Prädiktionswert), liegt also bei 80 zu 11.980, also bei 0,7 %!

Die Berechnung anhand des Bayes-Theorems sähe dann so aus:

$$p(G^+|ERK^+) = \frac{p(ERK^+|G^+) \times p(G^+)}{p(ERK^+|G^+) \times p(G^+) + p(ERK^+|G^-) \times p(G^-)}$$

$$p(G^+|ERK^+) = \frac{0,8 \times \frac{100}{11.900.100}}{0,8 \times \frac{100}{11.900.100} + 0,001 \times \frac{11.900.000}{11.900.100}} \approx 0,007$$

ERK^+ = durch System als Gefährderin bzw. Gefährder erkannt

G^+ = tatsächliche(r) Gefährderin bzw. Gefährder

G^- = tatsächlich keine Gefährderin bzw. kein Gefährder

Cut-off-Wert nicht immer frei wählbar

Bislang sind wir davon ausgegangen, dass der Cut-off-Wert beliebig – und damit optimal – gewählt werden kann. In manchen diagnostischen Kontexten, insbesondere in der Eignungsdiagnostik, ist der Cut-off-Wert allerdings nicht frei wählbar. Wenn z. B. 20 freie Stellen zu besetzen sind und sich darauf 100 Personen bewerben, muss der Cut-off-Wert zwangsläufig so gewählt werden, dass genau 20 Personen diesen übertreffen (es sei denn, man entscheidet sich dafür, einige der Stellen mangels geeigneter Bewerberinnen und Bewerber unbesetzt zu lassen). Es existiert nur eine Möglichkeit, den Cut-off-Wert weiter nach oben zu schieben und sich damit sicherer zu sein, dass die ausgewählten Personen tatsächlich geeignet sind (= positiver Prädiktionswert, dies ist in unserem Vier-Felder-Schema die für Organisationen zumeist einzig relevante Wahrscheinlichkeit): Man muss den Anteil der auszuwählenden Personen an allen Bewerberinnen und Bewerbern verringern – in der Eignungsdiagnostik wird dies als „Selektionsrate“ bezeichnet. Kann man aus 200 statt aus 100 Personen 20 auswählen, verschiebt sich der Cut-off-Wert nach oben (da nun die besten 10 % und nicht die besten 20 % ausgewählt werden). Überträgt man die vorab genannte Trias für gute Urteile (gegebene Validität der Interpretation, hohe Grundrate der interessierenden Gegebenheit, hoher Cut-off-Wert) auf die Eignungsdiagnostik, so ist dort optimal: hohe Validität der Interpretation, hohe Grundrate an geeigneten Bewerberinnen und Bewerbern, geringe Selektionsrate. Die genauen Auswirkungen dieser Größen auf die Trefferquoten (d. h. den positiven Prädiktionswert) in der Eignungsdiagnostik haben bereits Taylor und Russell (1939) in umfangreichen Tabellen zusammengefasst. Heute stehen zahlreiche Kalkulatoren im Internet zur Verfügung (z. B. ► <https://psychometrics.shinyapps.io/utility/>).

Übertrag auf die Eignungsdiagnostik

Die bisher gewählten Darstellungen (► Abb. 5.8, 5.9, 5.10, 5.11 und 5.12) sind in der Eignungsdiagnostik unüblich. Es hat sich dort eine andere, aber auf die bisher gewählten Darstellungen übertragbare Form der Visualisierung eingebürgert (► Abb. 5.13). Um die Übertragbarkeit zu verdeutlichen, wurde für die Ausgangskonfiguration (links oben) das Zahlenbeispiel aus ► Abb. 5.9 gewählt. Man sieht, dass sich eine Verringerung der Selektionsrate (rechts oben) förderlich auf den positiven Prädiktionswert auswirkt (nun $45/(45+5)=0,9$, statt vorher $89/(89+12)=0,88$). Gleiches gilt, wenn sich die Basisrate erhöht (links unten), was z. B. durch gezieltere Rekrutierung von Bewerberinnen und Bewerbern erreicht werden kann. Verschiebt man die Ellipse in der Abbildung links unten gedanklich noch weiter nach oben, so wird deutlich, dass sie auch kreisrund sein könnte – also ein invalides Instrument kennzeichnen könnte – und dennoch der positive Prädiktionswert sehr hoch wäre.

Strategien zur Verringerung der Selektionsrate und zur Erhöhung der Basisrate sind also förderlich für den positiven Prädiktionswert. Organisationen, die Eignungsdiagnostik betreiben, dürfen damit zufrieden sein. In ► Abb. 5.13b–c wird aber auch deutlich, dass dies zulasten des negativen Prädiktionswertes geht – viele eigentlich geeignete Personen werden abgelehnt. Aus Sicht von Bewerberinnen und Bewerbern ist dies unbefriedigend (s. auch die Ausführungen zu Fairness als Nebengütekriterium; ► Abschn. 2.6.5.5). Die einzige Möglichkeit, sowohl den positiven als auch den negativen Prädiktionswert gleichermaßen zu optimieren, besteht in der Wahl von Instrumenten, die für den vorliegenden diagnostischen Zweck validere Aussagen machen (► Abb. 5.13d).

Verringerung der Selektionsrate und Erhöhung der Basisrate

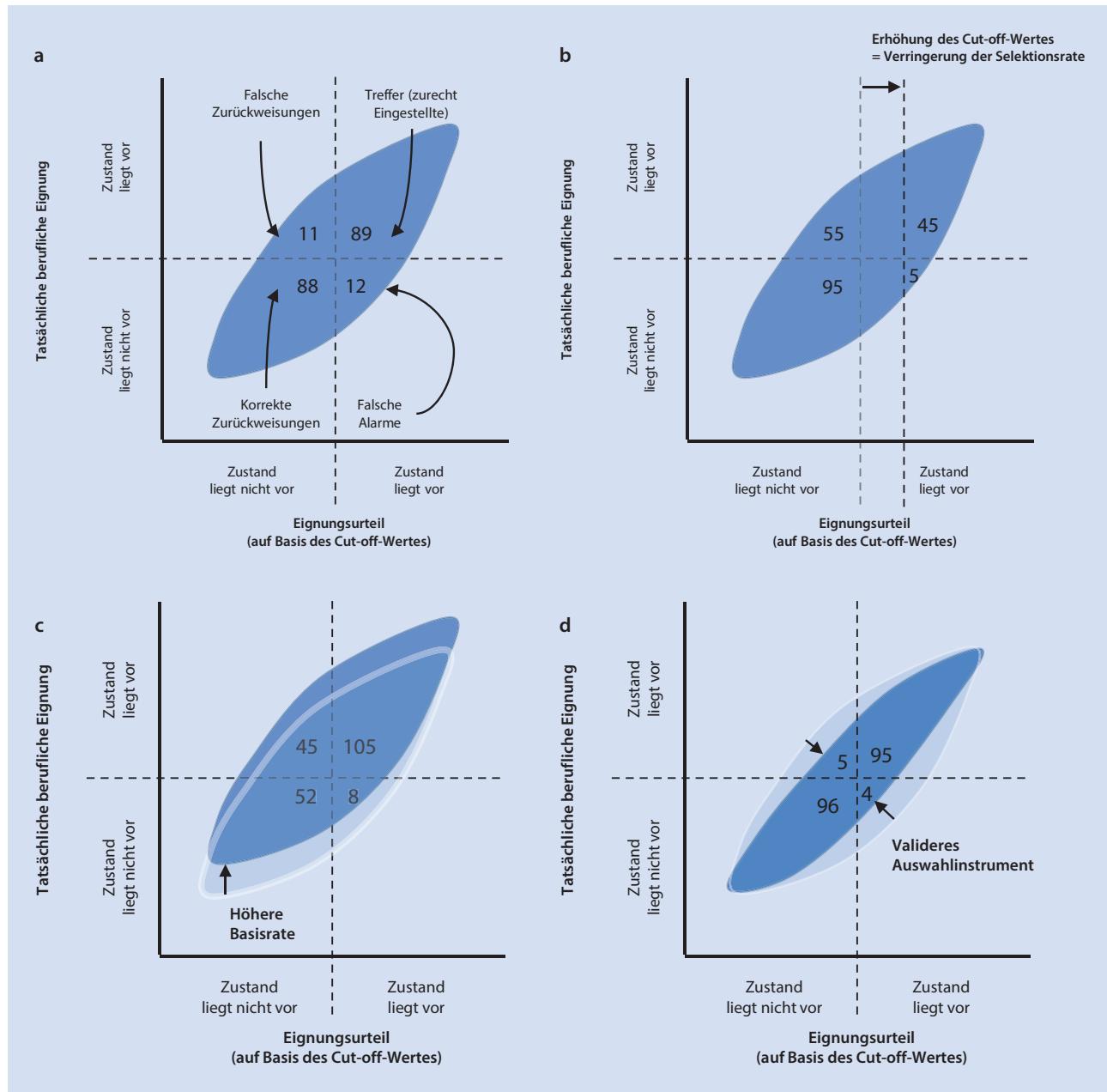


Abb. 5.13 a–d Grundrate, Selektionsrate und Validität der Interpretation des Auswahlinstruments

Auswahl geeigneter Personen

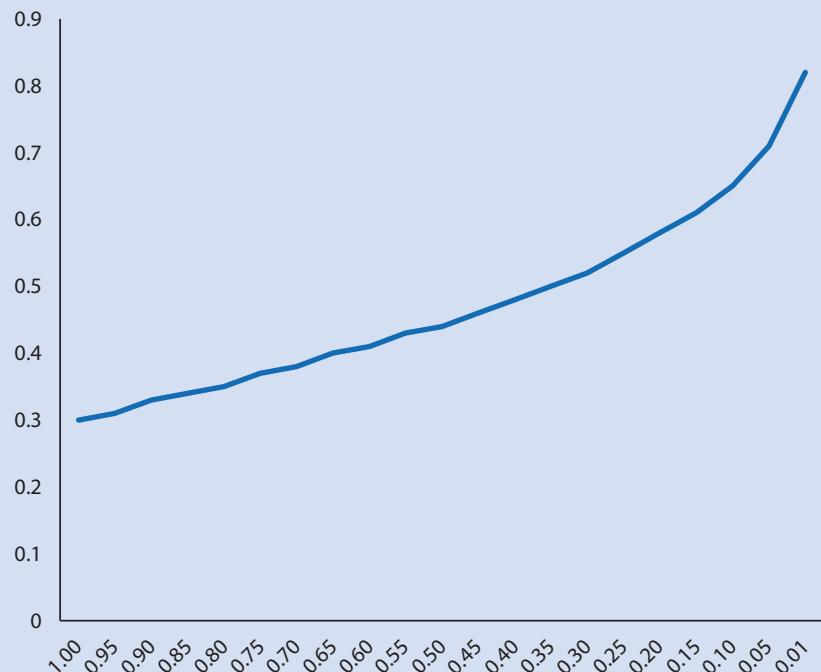
Wenn geeignete Personen für anspruchsvolle Tätigkeiten gesucht werden, ist oftmals von einer eher geringen Basisrate auszugehen. Dies wirkt sich, wie zuvor dargestellt, negativ auf den positiven Prädiktionswert (d. h. die Trefferquote) aus. Auch den Zusammenhang zwischen den verwendeten Auswahlverfahren und der vorherzusagenden beruflichen Leistung kann man nicht beliebig erhöhen. Dann besteht die einzige Möglichkeit, die Wahrscheinlichkeit für hohe Trefferquoten zu optimieren, darin, die Selektionsrate möglichst gering zu halten. Folgende Tabelle stellt die Auswirkungen der Selektionsrate auf die Trefferquote bei konstanter und geringer Basisrate (0,30) sowie für einen Zusammenhang des Eignungsurteils zu beruflicher Leistung von $R = .50$ dar (die

Trefferquoten wurde mithilfe des Kalkulators unter ► <https://psychometrics.shinyapps.io/utility/> berechnet).

Selektionsrate	Positiver Prädiktionswert bzw. Trefferquote
0,35	0,50
0,30	0,52
0,25	0,55
0,20	0,58
0,15	0,61
0,10	0,65
0,05	0,72
0,01	0,82

Wenn 35 % der Bewerberinnen und Bewerber ein Stellenangebot erhalten, so kann man unter den gegebenen Randbedingungen davon ausgehen, dass nur jede 2. Bewerberin bzw. jeder 2. Bewerber erfolgreich sein wird. Zu beachten ist, dass in diesen Fällen die geringen Trefferquoten nicht den Auswahlinstrumenten anzulasten sind – was in der Praxis jedoch gerne gemacht wird –, sondern durch die geringe Basisrate bedingt sind.

Erst bei deutlich niedrigeren Selektionsraten ist von einer Trefferquote $> 0,70$ auszugehen. Grafisch lässt sich dieser Zusammenhang wie folgt darstellen:



Trefferquote in Abhängigkeit der Selektionsrate. Die Trefferquote ist auf der Ordinate (y-Achse), die Selektionsrate auf der Abszisse (y-Achse) abgetragen. Annahme einer Basisrate = 0,30 und eines Zusammenhangs des Eignungsurteils zu beruflicher Leistung von $R = .50$.

5.1.4 Einstufige vs. mehrstufige diagnostische Entscheidungen

Aufmerksame Leserinnen und Leser haben bereits in ▶ Abschn. 5.1.3 bemerkt, dass bei manchen Strategien das Gesamтурteil bereits nach einer Testung feststeht. So erreicht unter der kompensatorischen Verrechnungsstrategie Person 1 (► Abb. 5.2) bereits in Test A einen so hohen Wert, dass sie insgesamt mehr als 100 Punkte erhält und damit als „positiv“ beurteilt wird (da Test A doppelt gewichtet wurde). Das Ergebnis von Test B ändert also nichts an dem Urteil für Person 1. Genau anders herum verhält es sich für Person 5 im konjunktiven Entscheidungsmodell (► Abb. 5.4). Sie erreicht für Test A nicht die geforderte Schwelle – damit ist es irrelevant, wie sie in Test B abschneidet. Im Sinne eines, für alle Beteiligten, ökonomischen Vorgehens sollte auf Messungen, die das Gesamтурteil nicht mehr beeinflussen, verzichtet werden.

Einstufige Entscheidungen

- ! Man stelle sich nur vor, Test B sei kein Test, sondern eine Arbeitsprobe (z. B. ein Lehrvortrag bei der Auswahl von Professorinnen und Professoren), auf die man sich als Kandidatin oder Kandidat lange vorbereiten würde, für die Reisezeit und -kosten anfallen würden, ein Auswahlkomitee zugegen sein würde o. Ä. – und das Abschneiden in der Arbeitsprobe sei aufgrund des Ergebnisses in Test A irrelevant. Spätestens in solchen Fällen ist ein sequenzielles Vorgehen angebracht.

Folgende ein- oder mehrstufigen Entscheidungsstrategien lassen sich unterscheiden:

- Nichtsequenzielle Einzelmessung
- Nichtsequenzielle Messbatterie
- Sequenzielle konjunktive Strategie (Pre-reject-Strategie)
- Sequenzielle disjunktive Strategie (Pre-accept-Strategie)
- Sequenzielle Kombinationsstrategie (Pre-reject- und Pre-accept-Strategie)

Nichtsequenzielle Messungen

Erfolgt nur eine einzelne Messung, ist natürlich kein sequenzielles Vorgehen möglich. Auch mehrere diagnostische Informationen können innerhalb einer Sitzung erhoben werden – sofern die Dauer der Messungen zumutbar ist. Man spricht dann von einer nichtsequenziellen Messbatterie. Der Vorteil dieses Vorgehens besteht darin, dass für alle Probandinnen und Probanden stets die gesamte diagnostische Information vorliegt und zur Entscheidungsfundung herangezogen wird. Besteht das Ziel der Diagnostik in der Modifikation (▶ Abschn. 5.1.2), wird man – sofern möglich – alle relevanten Informationen erheben und nicht im Sinne eines sequenziellen Vorgehens die Informationsammlung für einige Klientinnen und Klienten abbrechen.

Pre-reject-Strategie

Sequenzielle Vorgehensweisen bieten sich an, wenn die Selektion als diagnostisches Ziel im Vordergrund steht. Bei einem konjunktiven Entscheidungsmodell erfolgt keine positive Selektionsentscheidung, wenn Testpersonen bereits im 1. Test die geforderte Hürde nicht überspringen. Diese Personen können vorzeitig von weiteren Messungen ausgeschlossen werden (Pre-reject-Strategie). Beispielsweise kann eine Analyse der Bewerbungsunterlagen ergeben, dass einige Bewerberinnen und Bewerber für eine Professur keine eingeworbenen Forschungsmittel vorweisen können. Es würde keinen Sinn ergeben, diese Personen dennoch zu einem Berufsvortrag und einem Interview einzuladen und sie erst danach – eventuell trotz hervorragendem Berufsvortrag – aufgrund der fehlenden Forschungsmittel abzulehnen.

Pre-accept-Strategie

Im Rahmen einer disjunktiven Strategie können Entscheidung ebenfalls sequenziell getroffen werden. Wenn für die Zulassung zu einem Studium entweder exzellente Schulnoten oder ein sehr gutes Ergebnis in einem kognitiven

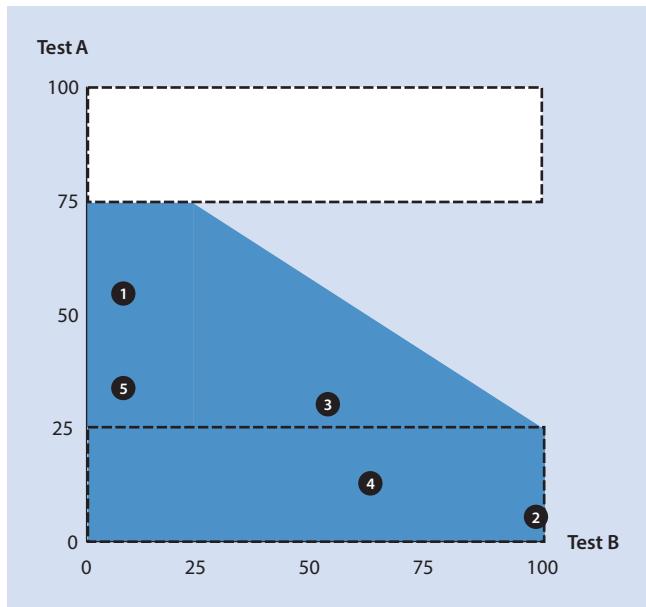


Abb. 5.14 Sequenzielle Kombinationsstrategie (Pre-reject- und Pre-accept-Strategie). Innerhalb des blau gekennzeichneten Bereichs erfolgt kein positives Urteil. Weitere Erläuterungen im Text

Kombination aus Pre-accept- und Pre-reject-Strategien

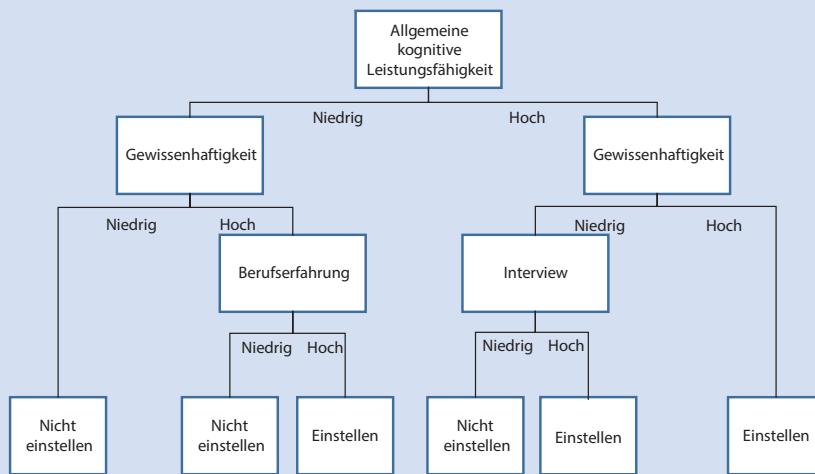
Leistungstest zur Zulassung gefordert würden, so könnte man Schülerinnen und Schülern mit exzellenten Schulnoten den Leistungstest ersparen. Sie würden nur aufgrund ihrer Noten zugelassen (Pre-accept-Strategie).

Für dieses Beispiel wäre auch eine sequenzielle Kombinationsstrategie (Pre-reject- und Pre-accept-Strategie) denkbar. Dabei würden Schülerinnen und Schüler mit exzellenten Schulnoten initial zugelassen, solche mit durchschnittlichen oder schlechteren Schulnoten abgelehnt und alle verbleibenden Schülerinnen und Schüler (mit guten, aber eben nicht exzellenten Noten) müssten einen kognitiven Leistungstest machen. In dem bisher genutzten grafischen Beispiel könnte sich eine sequenzielle Kombinationsstrategie wie in **Abb. 5.14** gezeigt gestalten: Im ersten Schritt würde Test A durchgeführt. Personen mit einem Testergebnis <25 würden als „negativ“ beurteilt und nicht weiter berücksichtigt (s. blauer gestrichelter Kasten) – dies trifft auf die Personen 2 und 4 zu. Personen mit einem Wert >75 würden direkt als „positiv“ beurteilt und ebenfalls keiner weiteren Testung unterzogen. Für die Personen 1, 3 und 5 in unserem Beispiel ist der Ausgang nach Test A noch offen. Für Test B wird ebenfalls ein Mindestwert >25 erwartet. Person 1 und 5 erfüllen dies nicht. Für alle Personen mit einem Mindestwert <25 wird eine kompensatorische Strategie angewandt, d. h., die Ergebnisse aus Test A und B werden zu einem Gesamtwert verrechnet, der über der blauen Diagonalen liegen sollte. Auch die verbleibende Person 3 erfüllt dieses Kriterium nicht.

Sequenzielle Entscheidungen anhand von Entscheidungsbäumen

Um sequenzielle Entscheidungen zu systematisieren, können Entscheidungsbäume genutzt werden. Dazu werden – ausgehend von einer initialen Entscheidung – nacheinander Verzweigungen eingebaut, die weitere (Teil-)Entscheidungen darstellen. Dies macht man so lange, bis eine finale Entscheidung getroffen ist (z. B. in der Eignungsdiagnostik „einstellen“ vs. „nicht einstellen“).

Marks (2018) stellt ein anschauliches Beispiel vor, das wir hier zur Illustration (in angepasster Form) ebenfalls nutzen. Idealerweise beginnt man mit dem für die zu treffende Entscheidung besten Prädiktor – im Falle der Eignungsdiagnostik wäre das die allgemeine kognitive Leistungsfähigkeit. Im nächsten Schritt wird die Gewissenhaftigkeit der Bewerbernden erfasst, gefolgt von – in manchen Fällen – einem Interview und – in anderen Fällen – einer Bewertung der Berufserfahrung. Natürlich muss für jede Verzweigung geprüft werden, welche Cut-off-Werte (also ab wann man von „hoch“ vs. „niedrig“ spricht) angewendet werden. Marks (2018) kann zeigen, dass Entscheidungen anhand solcher Entscheidungsbäume gleichwertig zu einer anderen gängigen statistischen Methode (der logistischen Regressionsanalyse) sind. Natürlich ist der Anwendungsbereich dieser Methode nicht auf die Eignungsdiagnostik beschränkt.



Beispielhafter Entscheidungsbau. (Adaptiert nach Marks, 2018, S. 13, © Kyle C. Marks).

- ! Nur weil eine Selektion nach dem 1. Schritt einer sequenziellen Strategie erfolgt, bedeutet dies nicht, dass die diagnostischen Informationen aus Schritt 1 in der Folge nicht mehr verwendet werden sollten. Nehmen wir an, dass Bewerberinnen und Bewerber im 1. Schritt nach ihrer kognitiven Leistungsfähigkeit ausgewählt werden. Die besten 50 von insgesamt 200 Bewerberinnen und Bewerber werden zu einem diagnostischen Interview eingeladen werden. Es ist wahrscheinlich, dass es auch zwischen diesen 50 Personen noch Unterschiede der kognitiven Leistungsfähigkeit gibt, die zur finalen Auswahlentscheidung – neben dem Ergebnis des Interviews – genutzt werden können und sollten (vgl. Sackett und Roth 1996).

Genaue Gestaltung sequenzieller Strategien ist komplex

- Die Wahl der besten sequenziellen Strategie ist komplex (De Corte et al. 2006). Für die Frage, welche Merkmale zuerst gemessen werden sollen und welche in darauffolgenden Schritten, muss Folgendes abgewogen werden:
- die Relevanz der jeweiligen Merkmale für das diagnostische Ziel (z. B. für die Vorhersage des Kriteriums „Studienleistung“),
 - die mit den jeweiligen Messungen verbundene Kosten (ggf. möchte man teure Messungen nicht mit allen Probandinnen und Probanden im 1. Schritt durchführen, sondern erst anhand günstigerer Methoden selektieren),
 - die Höhen der Hürden, also die Selektionsrate, die bei jedem Schritt angewandt werden (► Abschn. 5.1.3.3), sowie
 - Auswirkungen der Strategie auf die Zusammensetzung der ausgewählten Population.

Punkt d lässt sich an folgendem Beispiel illustrieren: Wenn Personen mit hohem Sozialstatus bessere Schulnoten aufweisen und eine Auswahlstrategie die Nutzung von Noten betont, d. h., diese in einem frühen Schritt heranzieht und eine hohe Hürde dafür ansetzt, werden in der ausgewählten Population mehr Personen mit hohem Sozialstatus sein (vgl. z. B. Sackett und Roth 1996). Darüber hinaus muss geprüft werden, ob Gesetze oder Verfahrensregeln den Einsatz von sequenziellen Strategien allgemein oder einer bestimmten sequenziellen Strategie verbieten.

5.2 Evaluation des Vorgehens

Evaluation ≠ Diagnostik

In ► Abschn. 1.1 haben wir betont, dass Evaluation von Psychologischer Diagnostik abgrenzen ist. Wir haben dort ebenfalls bereits eine Definition präsentiert, die wie hier noch einmal wiederholen.

Definition

„Evaluation ist die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes. Solche Evaluationsgegenstände können z. B. Programme, Projekte, Produkte, Maßnahmen, Leistungen, Organisationen, Politik, Technologie oder Forschung sein“ (Gesellschaft für Evaluation 2008, S. 15).

Formative und summative Evaluation der Psychologischen Diagnostik

Damit ist klar, dass auch der Nutzen oder Wert der Psychologischen Diagnostik im konkreten Anwendungsfall bewertet werden kann. Es sind sowohl der Prozess der Psychologischen Diagnostik (formative Evaluation) als auch das Ergebnis (summative Evaluation) zu bewerten.

5.2.1 Prozessevaluation der Psychologischen Diagnostik

Evaluation des diagnostischen Prozesses

Zur Prozessevaluation stellt sich allgemein die Frage: Sind wissenschaftliche Standards beim Vorgehen – von der Formulierung der Fragestellung bis hin zu einem diagnostischen Urteil und dessen Dokumentation – eingehalten worden? Dies inkludiert die Ableitung der diagnostischen Kriterien, die Formulierung psychologischer Fragen, die Wahl der diagnostischen Instrumente, deren Durchführung und Auswertung, die Interpretation der Ergebnisse und deren Integration zu einem Gesamтурteil (► Abschn. 1.5) sowie die nachvollziehbare und korrekte Darstellung all dessen im Rahmen eines Gutachtens (► Abschn. 4.6).

Für viele Teilschritte der Psychologischen Diagnostik sowie das Vorgehen in spezifischen Anwendungsfeldern existieren Qualitätsstandards sowie Checklisten, die zur Bewertung des diagnostischen Vorgehens herangezogen werden. Diese werden, soweit vorliegend, in den jeweiligen Kapiteln dieses Buchs besprochen (z. B. für Gutachten in ▶ Abschn. 4.6 oder die DIN-Norm 33430 für berufsbezogene Eignungsbeurteilung in ▶ Abschn. 6.4).

Internationale Richtlinien für die Testanwendung

Die von der International Test Commission (ITC) herausgegebenen „Internationalen Richtlinien für die Testanwendung“ (ITC 2001) spezifizieren in vielen Punkten den angemessenen Gebrauch von Tests. Die Richtlinien liegen in vielen Sprachen, darunter auch Deutsch, vor. Mit der Evaluation betraute Personen können diese Richtlinien heranziehen (verfügbar unter ▶ <https://www.int-testcom.org/>) und prüfen, ob im konkreten Fall das Vorgehen den geforderten Qualitätsmaßstäben der International Test Commission entspricht. Diese fordert z. B. bei der „Auswahl technisch einwandfreier und für die Situation angemessener Tests“ (ITC 2013; Auszüge aus der deutschen Fassung: ZPID 2001, S. 15; gendergerechte Formulierungen wurden durch die Autoren dieses Lehrbuchs vorgenommen):

Fachkompetente Testanwendende...

- Überprüfen vor der Testauswahl alle aktuellen Informationen im Hinblick auf die infrage kommenden Tests (z. B. Musterzusammenstellungen, unabhängige Beurteilungen, Expertinnen und Expertenmeinungen).
- Entscheiden, ob das technische Manual und Benutzerhandbuch eines Tests ausreichende Informationen liefert, um folgende Punkte zu beurteilen:
 - a) Geltungsbereich und Repräsentativität des Testinhalts, Angemessenheit der Normgruppen, Schwierigkeitsgrad des Inhalts usw.;
 - b) Genauigkeit der Messung und nachgewiesene Reliabilität im Hinblick auf die relevanten Populationen;
 - c) Validität (belegt im Hinblick auf die relevanten Populationen) und Bedeutsamkeit für die vorgesehene Verwendung;
 - d) Fehlen eines systematischen Fehlers im Hinblick auf die vorgesehene Probandengruppen,
 - e) Annehmbarkeit für die an der Testanwendung Beteiligten, unter anderen die von diesen wahrgenommene Fairness und Bedeutsamkeit;
 - f) Praktikabilität, unter anderem hinsichtlich des notwendigen Zeit-, Kosten- und anderen Ressourcenaufwands.
- Vermeiden die Anwendung von Tests, die nur eine unzureichende oder unklare technische Dokumentation aufweisen.
- Verwenden Tests nur für solche Zwecke, für die bedeutsame und angemessene Validitätsbelege vorliegen.
- Vermeiden es, einen Test nur auf der Grundlage des Augenscheins, den Berichten anderer Anwenderinnen und Anwender oder den Empfehlungen von Personen zu beurteilen, die ein nachvollziehbares kommerzielles Interesse daran haben.
- Stellen interessierten Personen und Personengruppen (z. B. Probandinnen und Probanden, deren Eltern, Vorgesetzten) auf Anfrage ausreichende Informationen zur Verfügung, damit diese die Gründe für die Auswahl eines Tests nachvollziehen können.

Es soll an dieser Stelle nicht unterschlagen werden, dass man natürlich auch die betroffenen Personen, d. h. diejenigen, die beurteilt wurden, nach ihrer Meinung über das diagnostische Vorgehen fragen kann. Idealerweise erfolgt diese Befragung standardisiert und nicht bloß im persönlichen Austausch zwischen

Standardisierte Befragung von Teilnehmenden

Klientin bzw. Klient und Diagnostikerin bzw. Diagnostiker. Für den Anwendungsbereich der Arbeits- und Organisationspsychologie werden mögliche Methoden zur Messung der Akzeptanz in ▶ Abschn. 6.2.1.4 näher besprochen.

5.2.2 Ergebnisevaluation der Psychologischen Diagnostik

Abgleich des diagnostischen Urteils mit der Realität

5

Realität jedoch häufig unbekannt

Evaluation auch ohne Kriteriumsdaten möglich

Korrekturen in Vergleichsstudien beachten

Kleine Zusammenhänge können von großer Relevanz sein

Eine Ergebnisevaluation des diagnostischen Vorgehens kann vorgenommen werden, wenn neben dem Ergebnis der Psychologischen Diagnostik das „wahre Ergebnis“ vorliegt. Beispielsweise kann die Beurteilung, ob Personen ein Studium erfolgreich absolvieren, mit dem tatsächlichen Abschneiden dieser Personen im Studium verglichen werden. Ebenso kann der vorhergesagte Therapieerfolg am tatsächlichen Ausmaß der Verbesserung evaluiert werden. Oder es kann geprüft werden, ob aus der Gruppe der als geeignet beurteilten Personen später tatsächlich mehrheitlich erfolgreiche Mitarbeiterinnen und Mitarbeiter geworden sind. Oder man prüft, ob das Ergebnis einer Paardiagnostik mit der Zufriedenheit als Paar zu einem späteren Zeitpunkt einhergeht.

An dieser Stelle muss man konstatieren, dass solche Daten in der Praxis selten vorliegen. Idealerweise wird bei der Planung des diagnostischen Vorgehens – wenn immer möglich – bereits berücksichtigt, dass eine Evaluation erfolgen soll. Dann kann die Erhebung der relevanten (Kriteriums-) Daten bereits fest eingeplant werden. In manchen Fällen lassen sich die benötigten Daten aus der Realität nachträglich noch aus Akten oder Datenbanken entnehmen und mit dem Ergebnis der Psychologischen Diagnostik in Verbindung bringen. Doch selbst wenn solche Daten vorliegen, kann eine nachträgliche Kombination von diagnostischem Ergebnis und realer Ausprägung praktisch nicht möglich oder aus Gründen des Datenschutzes unzulässig sein.

Aber auch wenn keine Kriteriumsdaten vorliegen, muss nicht vollständig auf eine Ergebnisevaluation verzichtet werden. Auch innerhalb einzelner diagnostischer Verfahren lassen sich relevante Kennwerte ermitteln. So kann im Rahmen von Verhaltensbeobachtungen oder diagnostischen Interviews die Übereinstimmung zwischen 2 oder mehr Beurteilenden geprüft und verglichen werden – mit für solche Verfahren typischen Übereinstimmungsmaßen. Ebenso kann geprüft werden, ob die Faktorenstruktur eines Fragebogens in der spezifischen Stichprobe stabil, d. h. unverändert zu dem im Fragebogenmanual berichteten Ergebnis bleibt.

! Die Resultate einer Ergebnisevaluation müssen im Kontext von Resultaten vergleichbarer Studien beurteilt werden. Es wäre fatal, eine Korrelation zwischen dem diagnostischen Ergebnis und einem Kriterium von .30 gemäß den Konventionen von Cohen (1988) als niedrig zu bezeichnen und damit das diagnostische Vorgehen als gescheitert anzusehen. Entscheidend ist vielmehr, welche (korrigierten!) Korrelationen im betreffenden Kontext zu erwarten sind und ob die aktuell zu evaluierende diagnostische Prozedur an diese Erwartungen heranreicht bzw. sie übertrifft bzw. dahinter zurückfällt.

Zu beachten ist, dass Vergleichswerte aus anderen Studien einer Korrektur unterzogen wurden. In vielen Metaanalysen werden für Unreliabilität und Streuungseinschränkung korrigierte Werte berichtet. Unterliegen die eigenen Ergebnisse den gleichen mindernden Einflüssen, wurde aber keine Korrektur vorgenommen, so ist kein fairer Vergleich gegeben. Entweder man nutzt die unkorrigierten Werte aus den Vergleichsstudien oder man wendet für die eigenen Werte eine analoge Korrektur an.

Es muss auch betont werden, dass selbst kleine Zusammenhänge zwischen dem diagnostischen Urteil und einem real vorliegenden Kriterium einen

hohen gesellschaftlichen und/oder monetären Nutzen haben können. Wir wenden uns daher nun einer wesentlichen Form der Evaluation zu: der Schätzung des Nutzens Psychologischer Diagnostik.

5.2.3 Schätzung des Nutzens Psychologischer Diagnostik

Institutionelle und individuelle Entscheidungen werden getroffen, weil sich die jeweiligen Organisationen bzw. Personen im Fall richtiger Entscheidungen etwas davon versprechen, nicht zuletzt positive ökonomische Auswirkungen, also Gewinne, während bei falschen Entscheidungen die Gefahr von Verlusten droht. So mögen sich für ein Unternehmen die erheblichen Investitionen für das Einstellen einer fähigen Führungskraft um ein Vielfaches auszahlen, wenn es die richtige Wahl war. Umgekehrt kann eine krasse Fehlbesetzung an wichtiger Stelle den Konzern an den Rand des Ruins bringen, wie viele Beispiele aus der Gegenwart anschaulich vor Augen führen. Der Börsenwert von großen Unternehmen hat sich schon von einem Tag auf den anderen um Milliardenbeträge erhöht, wenn ein nicht erfolgreicher Vorstandsvorsitzender nun „endlich“ abgelöst wurde. Auch individuelle Entscheidungen für Ausbildung und Beruf können sich in „Euro und Cent“ bemerkbar machen, weil bei richtigen Entscheidungen unter sonst gleichen Voraussetzungen ein vergleichsweise höherer Erfolg als bei falschen zu erwarten ist.

Eine einfache und daher leicht nachzuvollziehende Nutzenschätzung kann mithilfe des Brogden-Cronbach-Gleser-Nutzenmodells vorgenommen werden (Schmidt et al. 1979). Dieses Nutzenmodell geht – unter vereinfachenden Randbedingungen (s. u.) – davon aus, dass einerseits durch diagnostische Instrumente eine Prognose realer Kriterien (z. B. beruflicher Leistung) annähernd gelingt und andererseits realen Kriterien ein monetärer Wert beigemessen werden kann. Folglich sollte anhand des diagnostischen Ergebnisses auch eine Prognose des (dem realen Kriterium entsprechenden) monetären Wertes vorzunehmen sein. Anders gesagt: Wenn der Zusammenhang zwischen den diagnostischen Instrumenten und dem Kriterium bekannt ist (ausgedrückt als Korrelation) und die Umrechnung des Kriteriums in Geld ebenfalls feststeht, dann kann der monetäre Nutzen auch direkt anhand des diagnostischen Ergebnisses berechnet werden.

Nehmen wir zur Veranschaulichung an, ein diagnostisches Instrument korreliere zu $r=1.00$, also perfekt, mit dem Kriterium berufliche Leistung. Unter vereinfachenden Randbedingungen (kontinuierliche Ausprägung der diagnostischen Ergebnisse und der Kriteriumswerte, bivariate Normalverteilung, linearer Zusammenhang zwischen diagnostischem Ergebnis und Kriterium) kann dann davon ausgegangen werden, dass eine Person, die mit dem diagnostischen Instrument ein Ergebnis von 1 Standardabweichung über dem Mittelwert der relevanten Population erzielt, auch in Bezug auf das Kriterium ein Ergebnis zeigt, dass 1 Standardabweichung über dem Mittelwert liegt (schematische Darstellung in Abb. 5.15). Nun muss nur noch identifiziert werden, wie viel 1 Standardabweichung im Kriterium (z. B. eine um 1 Standardabweichung bessere berufliche Leistung) in Geld wert ist. Nehmen wir an, dieser Wert läge bei 30.000 EUR pro Jahr. Wenn es also gelingt, Personen zu identifizieren und einzustellen, die im Mittel mit dem verwendeten diagnostischen Instrument einen Wert erzielen, der 1 Standardabweichung über dem „Normalwert“ (ohne Auswahl bzw. ohne Psychologische Diagnostik) liegt, so „gewinnt“ man in diesem Fall pro Person und Jahr 30.000 EUR (abzüglich der für die Auswahlprozedur anfallenden Kosten).

Diagnostische Entscheidungen haben monetären Wert

Logik von Nutzenschätzungen

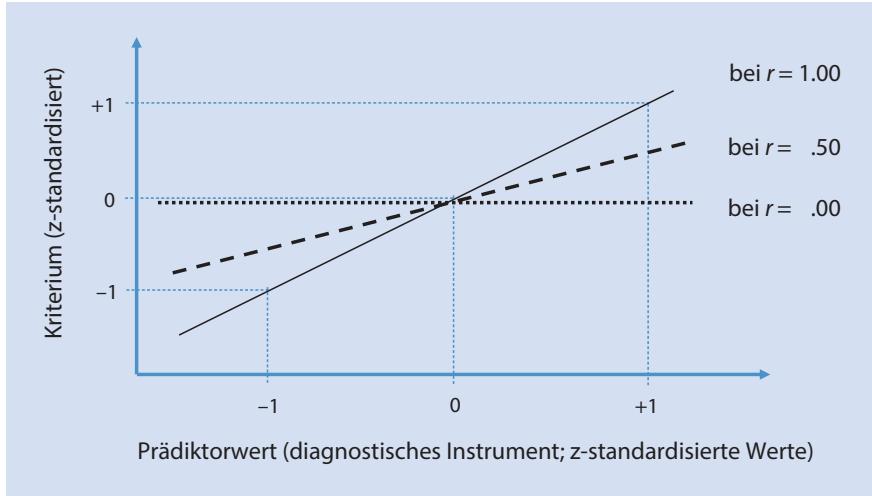


Abb. 5.15 Vorhersage von Kriteriumswerten durch Prädiktorwerte

Natürlich ist nie von einer Korrelation zwischen diagnostischem Instrument und Kriterium von $r=1.00$ auszugehen. Das Ausmaß, in dem 1 Standardabweichung im Ergebnis des diagnostischen Instruments mit einer Zu- oder Abnahme des prognostizierten Kriteriumswertes einhergeht, reduziert sich dann entsprechend (unter den bereits genannten, vereinfachenden Annahmen). Um dies zu illustrieren: Bei einer Korrelation von $r=.50$ würde man davon ausgehen, dass 1 Standardabweichung im diagnostischen Instrument einer halben Standardabweichung im Kriterium entspricht. In dem obigen Beispiel läge der geschätzte Nutzen pro Person und Jahr dann „nur“ bei 15.000 EUR (wovon wiederum noch die Kosten für die Auswahl abzuziehen wären).

Für eine vollständige Nutzenschätzung nach Brogden-Cronbach-Gleser ist das bisherige Vorgehen noch mit der Zahl der eingestellten Bewerberinnen bzw. Bewerber und mit deren mittlerer Verbleibedauer im Unternehmen zu multiplizieren; zusätzlich sind – wie bereits erwähnt – die Kosten der Auswahl abzuziehen.

Formel zur Nutzenschätzung nach Brogden-Cronbach-Gleser

$$\text{Nutzen} = SD_{\text{Krit}} \times r_{xy} \times z_y \times N_{\text{Ausw}} \times T - K \times N$$

SD_{Krit} = Standardabweichung des Kriteriums in Euro

r_{xy} = Korrelation zwischen Prädiktor (Auswahlverfahren) und Kriterium

z_y = standardisierter mittlerer Prädiktorwert der Ausgewählten

N_{Ausw} = Zahl ausgewählter Personen

T = Mittlere Verweildauer im Unternehmen

K = Kosten pro Bewerberin bzw. Bewerber

N = Gesamtzahl Bewerberinnen und Bewerber

Standardabweichung des Kriteriums
in Geldwerteinheiten schätzen

Der Charme dieses Vorgehens liegt in seiner Einfachheit. Die Logik dahinter wurde bereits erläutert; die konkrete Berechnung kann von Hand, mit gängiger Software oder über diverse, im Internet zu findende Kalkulatoren erfolgen (z. B. ► <https://psychometrics.shinyapps.io/utility/>).

Allerdings mag die Standardabweichung des Kriteriums in Geldwerteinheiten nicht immer bekannt sein. Zu deren Schätzung gibt es mehrere Möglichkeiten, die ausführlich von Görlich und Schuler (2014) beschrieben werden. Als Daumenregel kann die sog. 40 %-Regel herangezogen werden, bei der davon ausgegangen wird, dass 1 Standardabweichung des Kriteriums

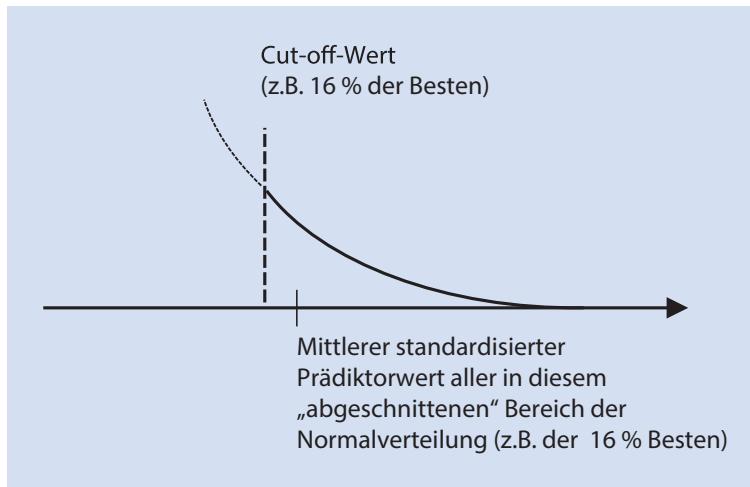


Abb. 5.16 Mittlerer standardisierte Prädiktorwert

in Geldwerteinheiten bei 40 % des Bruttojahresgehalts der betreffenden Stelleninhaberinnen und -haber liegt.

Weitere Erwähnung verdient der standardisierte durchschnittliche Prädiktorwert der Ausgewählten. Möchte man den erwarteten Nutzen einer Auswahl a priori berechnen, liegen die Prädiktorwerte der Ausgewählten nicht vor. Stattdessen müssen sie aus der Selektionsrate abgeleitet werden. Möchte man beispielsweise die besten 16 % auswählen, so ist damit zwar auch der standardisierte Cut-off-Wert bekannt: Er liegt bei 1 (vgl. ▶ Abschn. 2.6.4). Den mittleren standardisierten Prädiktorwert von diesen 16 % muss man jedoch erst ermitteln (Abb. 5.16).

Da wir von normalverteilten Werten ausgehen und wir eine Verteilung mit bekannten Parametern (Mittelwert = 0, Standardabweichung = 1) an einer bestimmten Stelle abschneiden, kann der Mittelwert der daraus resultierenden abgeschnittenen Normalverteilung (truncated normal distribution) – in diesem Fall also der Mittelwert der besten 16 % – errechnet werden. Dazu liest man den Ordinatenwert der Dichtefunktion der Standardnormalverteilung für den Cut-off-Wert (in diesem Fall = 1) in gängigen Tabellen ab und teilt diesen durch die Selektionsrate (Ström 2014). Alternativ kann man den standardisierten durchschnittlichen Prädiktorwert für eine gegebene Selektionsrate auch direkt in den von Naylor und Shine (1965) erstellten Tabellen nachschlagen. In Tab. 5.1 haben wir für 5 Selektionsraten die zu erwartenden mittleren standardisierten Prädiktorwerte errechnet (diese sind aufgrund von Rundungsdifferenzen als ungefähre Werte zu verstehen).

Soll der Nutzen geschätzt werden, nachdem die Auswahl stattgefunden hat, kann der tatsächlich vorliegende gemittelte standardisierte Prädiktorwert der Ausgewählten genutzt werden. Es sollte dennoch geprüft werden, ob dieser ungefähr dem Prädiktorwert entspricht, den man nach obiger Prozedur a priori

Standardisierte durchschnittlicher Prädiktorwert

Voraussetzung: Normalverteilung der Prädiktorwerte

Tab. 5.1 Standardisierte Prädiktorwerte für 5 Selektionsraten

Selektionsrate	Standardisierte durchschnittlicher Prädiktorwert
0,25	1,275
0,20	1,402
0,15	1,549
0,10	1,758
0,05	2,046

erwartet hätte. Weichen der empirisch ermittelte Prädiktorwert und der auf Basis der Normalverteilung erwartete Prädiktorwert deutlich voneinander ab, darf vermutet werden, dass die vorliegenden Werte nicht normalverteilt sind und damit eine Voraussetzung für diese Art der Nutzenschätzung verletzt ist.

5

Bestehen konstante Zusammenhänge zwischen Prädiktoren und Kriterien über lange Zeiträume?

Dem Brogden-Cronbach-Gleser-Nutzenmodell zufolge kann der Nutzen pro Person und Jahr mit der mittleren Verweildauer der ausgewählten Personen im Unternehmen multipliziert werden. Mit anderen Worten: Es wird ein konstanter Zusammenhang zwischen Prädiktor und Kriterium über die Zeit angenommen. Während diese Annahme für manche Prädiktoren (wie allgemeine Intelligenz) empirisch haltbar erscheint, muss dies für andere Prädiktoren (z. B. berufliche Erfahrung) angezweifelt werden. Schmidt und Hunter (2004) berichten von einem abnehmenden Zusammenhang zwischen der beruflichen Erfahrung bei der Einstellung und der Leistung im Beruf von $r=.49$ (bei Personen bis zu 3 Jahren Berufserfahrung) auf $r=.15$ (bei Personen mit mehr als 12 Jahren Berufserfahrung). Würde man $r=.49$ in die obige Formel einsetzen und eine Verweildauer von 12 Jahren annehmen, käme es zu einer deutlichen Überschätzung des Nutzens.

Darüber hinaus ist zu beachten, dass mehrere andere Randbedingungen ausgeklammert werden. So ist es beispielsweise naheliegend, dass die Auswahl von besser geeigneten Kandidatinnen und Kandidaten dazu führen kann, dass man diesen höhere – und über die Zeit stärker ansteigende – Gehälter zahlen muss als nur durchschnittlich geeigneten Kandidatinnen und Kandidaten. Für den Nutzen einer validen Auswahl dürfte dies jedoch keine besondere Einschränkung darstellen.

► Beispiel

Beispiel einer Nutzenschätzung

In ► Abschn. 5.1.3.1 haben wir beschrieben, dass eine mechanische Integration von Daten in der Eignungsdiagnostik zu einer deutlichen Steigerung der Prognosegüte gegenüber einer klinischen Integration führen sollte (Kuncel et al. 2013). Der ermittelte Unterschied lag bei $r=.44$ vs. $r=.28$. Wie lässt sich diese Verbesserung von $\Delta r=.16$ in Nutzen umrechnen? Nehmen wir an, es gelänge, für die Auswahl von Professorinnen und Professoren eine mechanische Urteilsbildung zu etablieren – anstelle der Besprechung und intuitiven Verrechnung von Erkenntnissen in Kommissionssitzungen. Zur Berechnung des zu erwartenden Nutzens bei 10 zu besetzenden Professuren gehen wir von folgenden Daten aus:

Durchschnittliches Bruttojahreseinkommen	90.000 EUR (zur Beispielrechnung verwendeter, fiktiver Wert)
40 % des Bruttojahreseinkommens (SD_{Kri})	36.000 EUR
Δr_{xy}	.16
Zahl ausgewählter Personen (N_{Ausw})	10
Gesamtzahl Bewerberinnen und Bewerber (N)	300
Selektionsrate	$10/300 = 0,0333$
Standardisierter mittlerer Prädiktorwert der Ausgewählten (z_y)	2,23
Mittlere Verweildauer (T)	10 Jahre
Kosten pro Bewerberin bzw. Bewerber (K)	Nicht in Berechnung aufgenommen, da wir davon ausgehen, dass die Kosten für klinische und mechanische Urteilsbildung in etwa gleich hoch sind

Daraus resultiert eine Nutzenschätzung für 10 Professuren von
 $\text{Nutzen} = 36.000 \times 0,16 \times 2,23 \times 10 \times 10 = 1.284.480 \text{ EUR}$

- !** Es ist zu betonen, dass in verschiedenen Anwendungsfeldern der Psychologischen Diagnostik rein monetäre Nutzenerwägungen unangebracht sind. Ethische und auf das jeweilige Individuum bezogene Erwägungen über die Notwendigkeit einer Maßnahme sind monetären Aspekten voranzustellen.

5.3 Zusammenfassung

Diagnostische Strategien beschreiben unterscheidbare Prinzipien des Diagnosizierens. Die Wahl der Strategie richtet sich danach, wie das diagnostische Ziel unter den jeweiligen Gegebenheiten am besten zu erreichen ist. Wird die aktuelle Ausprägung eines Merkmals zu einem Zeitpunkt gemessen, spricht man von Statusdiagnostik. Im Gegensatz dazu dient die Veränderungsdiagnostik dazu, einen Unterschied zwischen 2 oder mehreren Messungen zu identifizieren. Es kann auch unterschieden werden, ob Diagnostik zum Zweck der Selektion oder Modifikation erfolgt. Des Weiteren kann sich die diagnostische Strategie dahingehend unterscheiden, wie die vorliegenden Daten zu einem Gesamturteil integriert werden. Dazu stehen als prinzipielle Strategien die klinische und die mechanische Urteilsbildung zur Verfügung, wobei sich Letztere als überlegene Strategie erwiesen hat. Im Detail ist zu klären, wie die Informationen über eine Person miteinander verrechnet werden sollen: Bei einem kompensatorischen Modell kann eine geringe Ausprägung eines Merkmals durch eine hohe Ausprägung eines anderen Merkmals kompensiert werden. Das gegenteilige Vorgehen dazu bestünde in einem disjunktiven Modell: Dann würden Mindestausprägungen in jedem Merkmal definiert, deren Unterschreitung nicht durch andere Merkmale kompensiert werden kann. Werden Mindestausprägungen oder kritische Testtrennwerte (Cut-off-Werte) festgelegt, ist genau zu prüfen, wie sich diese auf die Sensitivität und Spezifität der diagnostischen Entscheidung auswirken. Das heißt, es ist zu prüfen, welche Formen der Fehlentscheidungen in welcher Häufigkeit gemacht werden und ob dies zu tolerieren ist. Hierzu sind auch ökonomische Nutzenüberlegungen anzustellen.

Weiterführende Literatur und Internetressourcen

Für weitere Lektüre zur mechanischen und klinischen Urteilsbildung können die im Text zitierten Metaanalysen empfohlen werden. Darüber hinaus existiert ein sehr lesenswerter Beitrag von Highhouse (2008) mit dem Titel „Stubborn reliance on intuition and subjectivity in employee selection“. Weiterführende Erläuterungen zu Quoten in der Eignungsdiagnostik sowie zu Nutzenanalysen finden sich im *Lehrbuch der Personalpsychologie* von Schuler und Kanning (2014).

Trefferquoten und Nutzen lassen sich mit der folgenden Webapplikation berechnen:
▶ <https://psychometrics.shinyapps.io/utility/>. Für mehrstufige Auswahlverfahren steht unter
▶ <https://users.ugent.be/~wdecorte/software.html> eine Software zur Verfügung, anhand derer die optimale Konfiguration der Auswahlquoten über die Auswahlstufen ermittelt werden kann (s. De Corte et al. 2011). Richtlinien zur Testanwendung und damit auch zur Evaluation einer sachgerechten Testanwendung finden sich unter ▶ <https://www.intestcom.org/>.

?

Übungsfragen

— Abschn. 5.1:

- Was versteht man unter diagnostischen Strategien?
- Wie unterscheiden sich diagnostische Strategien zur Selektion bzw. zur Modifikation?
- Welche Erkenntnisse liefert der Vergleich mechanischer vs. klinischer Urteilsbildung?

- Unter welchen Randbedingungen sind die klinische und die mechanische Urteilsbildung etwa gleich gut?
 - Was versteht man unter kompensatorischer, konjunktiver und disjunktiver Entscheidungsstrategie?
 - Was bezeichnet das Multiple-Hurdle-Problem?
 - Welche Fehlerarten sind beim Festlegen eines Cut-off-Wertes zu beachten?
 - Erläutern Sie, warum der positive und der negative Prädiktionswert anfällig für die Grundquote sind, Sensitivität und Spezifität jedoch nicht!
 - Wie kann die Wahl des optimalen Cut-off-Wertes grafisch vollzogen werden?
 - Welche Quoten sind in der Eignungsdiagnostik interessant und wie sollten diese ausgeprägt sein?
 - Was versteht man unter der Pre-reject-Strategie?
- **Abschn. 5.2:**
- Wie kann das diagnostische Vorgehen überprüft werden, wenn keine realen Kriterien vorliegen?
 Erläutern Sie die grundlegende Logik von Nutzenschätzungen!
 Was versteht man unter einer „truncated normal distribution“?

Literatur

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist* 34, 341–382.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology* 91, 523–537.
- De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology* 96, 907–926.
- Eid, M., Gollwitzer, M., & Schmitt, M. (2015). *Statistik und Forschungsmethoden* (4. Aufl.). Weinheim: Beltz.
- Gesellschaft für Evaluation (2008). *Standards für Evaluation* (4. Aufl.). Mainz: Gesellschaft für Evaluation.
- Gigerenzer, G., Krämer, W., & Bauer, T. K. (2018). Unstatistik des Monats: „Erfolgreiche“ Gesichtserkennung mit hunderttausenden Fehlalarmen. Unstatistik vom 30.10.2018. ► https://www.rwi-essen.de/media/content/pages/presse/downloads/181030_unstatistik_oktober.pdf. Zugegriffen: 24. März 2020.
- Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs: General and Applied* 79, 1–28.
- Goldhammer, F., & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 173–201). Berlin, Heidelberg: Springer.
- Görlich, Y., & Schuler, H. (2014). Personalentscheidung, Nutzen und Fairness. In H. Schuler, & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (3. Aufl., S. 1137–1199). Göttingen: Hogrefe.
- Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology* 115, 192–194.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment* 12, 19–30.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology* 92, 373–385.
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology* 1, 333–342.

- International Test Commission (ITC). (2013). ITC Guidelines on Test Use. 8th October, 2013, Version 1.2. Final Version. Document reference: ITC-G-TU-20131008. ► https://www.intest-com.org/files/guideline_test_use.pdf. Zugegriffen: 26. März 2020.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59, 12–19.
- Jäger, R. S. (1988). *Psychologische Diagnostik – Ein Lehrbuch*. München: Psychologie Verlags Union.
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology* 98, 1060–1072.
- Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID). (2001). Internationale Richtlinien für die Testanwendung, Version 2000. Deutsche Fassung. ► https://www.zpid.de/pub/tests/itc_richtlinien.pdf. Zugegriffen: 15. April 2020.
- Marks, K. (2018). Comparing the accuracy of decision trees and logistic regression in personnel selection. [Dissertation]. Murfreesboro, TN: Middle Tennessee State University.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Monahan, J. (2003). Violence risk assessment. In A. M. Goldstein, & I. B. Weiner (Eds.), *Handbook of psychology: Forensic psychology* (Vol. 11, pp. 527–540). New York: Wiley.
- Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology* 3, 33–42.
- Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology* 49, 549–572.
- Schäfer, H. (1989). Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine* 8, 1381–1391.
- Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology* 86, 162–173.
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology* 64, 609–626.
- Schuler, H., & Kanning, U. P. (2014). *Lehrbuch der Personalpsychologie* (3. Aufl.). Göttingen: Hogrefe.
- Ström, M. (2014). Selection methods: Precision and return on investment. ► <https://psychometrics.shinyapps.io/utility/>. Zugegriffen: 19. Oktober 2020.
- Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology* 23, 565–578.
- Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice* 40, 525–531.
- Westmeyer, H. (2006). Wissenschaftstheoretische und erkenntnistheoretische Grundlagen. In F. Petermann, & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 35–45). Göttingen: Hogrefe.
- Ziegler, M., & Bühner, M. (2012). *Grundlagen der Psychologischen Diagnostik*. Wiesbaden: VS Verlag.



Diagnostik in der Arbeits-, Organisations- und Wirtschaftspsychologie

Stefan Krumm, Lothar Schmidt-Atzert und Manfred Amelang

Inhaltsverzeichnis

- 6.1 Organisationsdiagnostik – 570**
 - 6.1.1 Fragebögen zur Beschreibung der Arbeit und des Klimas in Teams bzw. Organisationen – 571
 - 6.1.2 Arbeits- und Anforderungsanalyse – 573
- 6.2 Diagnostik von Personenmerkmalen – 582**
 - 6.2.1 Selektion von Personen: Personalauswahl – 584
 - 6.2.2 Selektion von Bedingungen: Berufs- und Ausbildungswahl – 609
 - 6.2.3 Modifikation von Personen: Personalentwicklung – 613
 - 6.2.4 Modifikation von Bedingungen – 617
- 6.3 Evaluation der Psychologischen Diagnostik in der Arbeits-, Organisations- und Wirtschaftspsychologie – 618**
 - 6.3.1 Evaluation von Arbeits- und Anforderungsanalysen – 619
 - 6.3.2 Evaluation von Personalauswahlverfahren – 621
 - 6.3.3 Evaluation von diagnostischen Verfahren zur Berufs- und Ausbildungswahl – 624
 - 6.3.4 Evaluation von diagnostischen Verfahren zur Feststellung des Personalentwicklungsbedarfs – 625
 - 6.3.5 Evaluation der Diagnostik von individuellen Merkmalen zum Zwecke der Modifikation von Arbeitskontexten/Arbeitsgestaltung – 626
 - 6.3.6 Messung der Auswirkung von „Passung“ – 627
- 6.4 Ein Qualitätsstandard für berufsbezogene Eignungsbeurteilungen – die DIN 33430 – 632**
- 6.5 Zusammenfassung – 636**
- Literatur – 637**

Die Fachgruppe für Arbeits-, Organisations- und Wirtschaftspsychologie (AOW) der Deutschen Gesellschaft für Psychologie grenzt ihre Teildisziplin wie folgt ein:

- » AOW-Psychologinnen und -Psychologen erforschen und gestalten Wechselbeziehungen zwischen Arbeits-, Organisations- und Marktbedingungen einerseits und menschlichem Erleben und Verhalten in Organisationen andererseits. Ziel ist es, mit Blick auf Gesundheit, Leistung und Effizienz, die Passung zwischen Individuum und Arbeitskontext zu erhöhen. (Deutsche Gesellschaft für Psychologie, 2020)

6

Passung (Fit) zwischen Individuum und Merkmalen des Arbeitskontextes

Merkmale der Person und des Tätigkeitsumfelds

Selektion und Modifikation

Gleichzeitige Modifikation von Personen und Arbeitsumfeldern

In dem hier genannten Ziel kommt ein für die Psychologische Diagnostik wesentlicher Aspekt zum Ausdruck: Eine hohe Passung (Fit) zwischen Individuum und verschiedenen Merkmalen des Arbeitskontextes wirkt sich positiv auf Leistung, Gesundheit und Zufriedenheit von Individuen aus und ist daher von Organisationen anzustreben. Eine geringe Passung (Misfit) wirkt sich hingegen negativ auf Leistung, Gesundheit und Zufriedenheit von Individuen aus und hat daher eine negative Wirkung auf Organisationen (für Evidenz aus einer Längsschnittstudie und einer Metaanalyse s. Cable und DeRue 2002; Kristof-Brown et al. 2005).

Der Diagnostik in der Arbeits-, Organisations- und Wirtschaftspsychologie kommt somit nicht nur die Aufgabe der Beurteilung von Personenmerkmalen (wie Allgemeiner Intelligenz oder sozialer Fertigkeiten) zu, sondern auch der Beurteilung von korrespondierenden Merkmalen des Tätigkeitsumfelds (wie der Anforderung einer Tätigkeit oder des Teamklimas). Schließlich ist eine Gegenüberstellung von Merkmalen der Person einerseits und der Organisation/des Tätigkeitsumfelds andererseits nötig, um deren Passung einzuschätzen.

Die Beurteilung der Passung kann zum Zwecke der Selektion oder der Modifikation durchgeführt werden (► Abschn. 5.1.2). Dabei können sowohl Personen als auch Bedingungen (hier die Organisation/das Tätigkeitsumfeld) selegiert bzw. „modifiziert“ werden. Je nach Anlass der Diagnostik gilt eine der beiden folgenden Optionen:

- a) Die Merkmale der Organisation/des Tätigkeitsumfelds werden vorab identifiziert und als feststehend angesehen, danach werden die dazu passenden Personen selegiert oder entsprechend trainiert.
- b) Die Merkmale einer oder mehrerer Personen werden vorab identifiziert und als feststehend angesehen, danach werden die dazu passenden Organisationen/Arbeitsumfelder identifiziert oder modifiziert.

Je nach Ausmaß der Passung können dann verschiedene Entscheidungen zur Selektion bzw. Modifikation von Personen bzw. Bedingungen getroffen werden (► Tab. 6.1).

In einigen Fällen können Personen und Organisationen/Arbeitsumfelder gleichzeitig modifiziert werden. Beispielsweise könnten Mitarbeitende bei geringer Passung zu den Anforderungen eines vorgesehenen Auslands-einsatzes zunächst ein spezielles (z. B. interkulturelles) Training durchlaufen, aber gleichzeitig die Gegebenheiten am ausländischen Arbeitsplatz angepasst werden (z. B. durch Zuweisung interkultureller Mentorinnen und Mentoren vor Ort). Auch eine gleichzeitige Selektion und Modifikation ist denkbar. Beispielsweise könnten Mitarbeitende aufgrund ihrer interkulturellen Fertigkeiten ausgewählt und gleichzeitig die interkulturellen Anforderungen am ausländischen Arbeitsplatz reduziert werden.

Tab. 6.1 Selektion und Modifikation in der Eignungsdiagnostik

	Selektion		Modifikation	
	Personen	Bedingungen	Personen	Bedingungen
Erläuterung/Art der Passung	Eine Selektion von Personen erfolgt im Rahmen der Personalauswahl; in diesem Fall wird für viele Bewerber geprüft, wie gut diese zu einer Organisation passen	Eine Selektion von Bedingungen erfolgt z. B. im Zuge der Arbeitsplatzsuche von Bewerbenden oder im Rahmen der Berufsberatung; in diesem Fall beurteilt eine Person, wie gut sie zu vielen verschiedenen Organisationen bzw. Berufen bzw. Ausbildungen passt	Eine Modifikation von Personen erfolgt im Zuge der Personalentwicklung; in diesem Fall wird geprüft, wie gut eine Person zu einer (zukünftig zu verrichtenden) Tätigkeit passt und welche Weiterbildungmaßnahmen nötig sind	Eine Modifikation von Bedingungen erfolgt im Zuge von Arbeitsgestaltungs- oder Organisationsentwicklungsmaßnahmen; in diesem Fall wird geprüft, wie gut das aktuelle Arbeitsumfeld zu den Mitarbeitenden passt und welche Anpassungen des Arbeitsumfelds nötig sind, um diese Passung zu erhöhen
Entscheidung/Maßnahme bei...				
Guter Passung	Jobangebot wird ausgesprochen	Entscheidung für Bewerbung oder Aufnahme einer Tätigkeit (z. B. eines Studiums)	Keine unmittelbare Maßnahme	Keine unmittelbare Maßnahme
Schlechter Passung	Ablehnung der Bewerbenden	Entscheidung gegen Bewerbung oder Aufnahme einer Tätigkeit	Ausbildung bzw. Weiterentwicklung der Person mit dem Ziel, die Passung zur Tätigkeit zu verbessern	Veränderung des Arbeitsumfelds mit dem Ziel, die Passung zu den Mitarbeitenden zu verbessern

Fazit Es soll zunächst festgehalten werden, dass eine gute Passung zwischen Personen und Organisationen/Arbeitsbedingungen nur dann aktiv – durch Selektion oder Modifikation – herbeigeführt werden kann, wenn zuvor die relevanten Merkmale von Personen und Organisationen diagnostiziert wurden. Letzteres wird unter dem Begriff Organisationsdiagnostik zusammengefasst.

6.1 Organisationsdiagnostik

Verhalten und Erleben in Organisationen

6

Organisationen wie Betriebe, Behörden, Schulen, Universitäten, Krankenhäuser etc. können aus betriebswirtschaftlicher, organisationssoziologischer, verwaltungswissenschaftlicher und eben auch aus psychologischer Perspektive beschrieben und analysiert werden. Die psychologisch ausgerichtete Organisationsdiagnostik befasst sich mit dem Verhalten und Erleben der Mitglieder in Organisationen.

Definition

Die **Organisationsdiagnostik** „dient dazu, die psychologischen Aspekte des Erlebens und Verhaltens von Mitgliedern in Organisationen zu diagnostizieren, um Regelhaftigkeiten im Erleben, im Verhalten und in den Interaktionen zu beschreiben, zu erklären und zu prognostizieren“ (Büssing 2007, S. 558).

Datenquellen der Organisationsdiagnostik

Nach Büssing (2007) werden zur Organisationsdiagnostik vorwiegend die folgenden Datenquellen herangezogen, wobei sich Nr. 2 bis 7 mithilfe von standardisierten psychologischen Verfahren wie diagnostischen Interviews, Beobachtungsplänen oder Fragebögen erheben lassen.

Datenquellen in der Organisationsdiagnostik

1. Analyse von Dokumenten (z. B. Organigramme)
 2. Organisations- und betriebswirtschaftliche Statistiken (z. B. Fluktuation)
 3. Befragung von Schlüsselpersonen sowie Expertinnen und Experten
 4. Befragung von Mitarbeitenden
 5. Beobachtungen am Arbeitsplatz
 6. Gruppengespräche
 7. Analyse von Interaktionen (z. B. Soziometrie)
 8. physikalische Methoden (z. B. Messung von Lärm oder Beleuchtung)
 9. physiologische Methoden (zur Messung von Beanspruchung/Stress)
- (Liste erweitert nach Büssing, 2007, Tab. 3, S. 574).

Die folgenden Merkmale einer Organisation/eines Tätigkeitsumfelds sollen beispielhaft erläutern, welche Inhalte anhand der voranstehend genannten Datenquellen beurteilt werden können.

Für die Organisationsdiagnostik relevante Merkmale einer Organisation/eines Tätigkeitsumfelds (Beispiele)

- Berufliche Aufgaben
- Berufliche Anforderungen
- Team-/Organisationsklima
- Werte einer Organisation
- Vorherrschende Führungsstile
- Befriedigungspotenziale einer Tätigkeit (z. B. persönliche Entfaltungs- oder Weiterbildungsmöglichkeiten)
- Gesundheitsförderliche bzw. gesundheitsschädliche Aspekte

Wenngleich zuvor der Gedanke der Passung (Fit) betont wurde, so zeigt diese Auflistung auch, dass einige Merkmale einer Organisation auch ohne einen Abgleich mit Merkmalen von Personen als Ausgangspunkt von Veränderungen genutzt werden können. So geben hohe Belastungen durch das Arbeitsumfeld, beispielsweise ungünstige Arbeits- und Ruhezeiten oder ein vorherrschend destruktiver Führungsstil (vgl. Rothe et al. 2017) sowie ein schlechtes Team- bzw. Organisationsklima, auch losgelöst von der Frage der Passung Anlass für Veränderungen.

Veränderungen auch unabhängig vom Fit

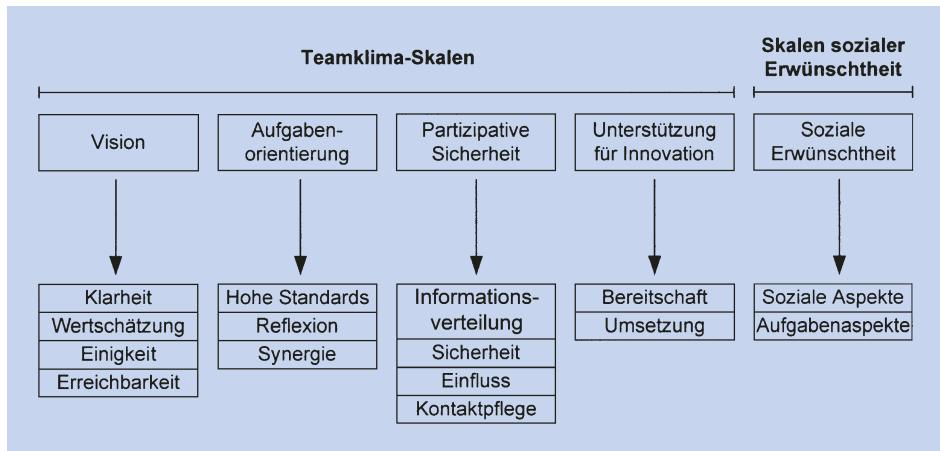
6.1.1 Fragebögen zur Beschreibung der Arbeit und des Klimas in Teams bzw. Organisationen

Zur Beschreibung der Arbeit und des Klimas in Teams bzw. Organisationen wurden spezielle Fragebögen konstruiert, von denen hier nur das *Teamklima-Inventar (TKI)* von Brodbeck et al. (2000) ausführlich dargestellt wird (► Abschn. 6.1.1). Der *Fragebogen zur Erfassung des Organisationsklimas (FEO)* von Daumenlang et al. (2004) wurde konstruiert, um mehrere Dimensionen des Organisationsklimas (Vorgesetztenverhalten, Kollegialität, Bewertung der Arbeit, Arbeitsbelastung, Organisation, berufliche Perspektiven, Entgelt, Handlungsräum, Einstellung zum Unternehmen, Interessenvertretung, Mitarbeiterbewertung) durch Fremd- oder Selbstbeurteilungsskalen zu erfassen. Speziell mit organisationsinterner Kommunikation befasst sich der *Fragebogen zur Erfassung der Kommunikation in Organisationen (KOM-MINO)* von Sperka und Rózsa (2007). Die Skalen liefern Informationen über 7 Merkmale der Kommunikation mit dem direkten Vorgesetzten, mit Kolleginnen und Kollegen und (bei Führungskräften) mit den unterstellten Mitarbeiterinnen und Mitarbeitern. Der *Fragebogen zur Arbeit im Team (FAT)* von Kauffeld (2004) soll mit den Skalen „Zielorientierung“, „Aufgabenbewältigung“, „Zusammenhalt“ und „Verantwortungsübernahme“ Stärken und Schwächen von Teams darstellen und damit auch Hinweise auf eventuell notwendige Teamentwicklungsmaßnahmen liefern.

■ Teamklima-Inventar (TKI)

Theoretischer Hintergrund und Aufbau Ausgehend von dem englischen Original, dem „Team Climate Inventory“ (Anderson und West 1998), haben Brodbeck et al. (2000) deutschsprachige Adaptationsarbeiten vorgenommen. Der Fragebogen enthält 44 Items und zielt auf das Klima für Innovation und Leistung in Arbeitsgruppen ab. Darunter verstehen Brodbeck et al. (2000, S. 8) die „subjektive Wahrnehmung von Individuen über ihre soziale Umgebung in Organisationen oder Arbeitsgruppen, die mehr oder weniger sozial geteilt sind“. Den theoretischen Rahmen stellt eine Vier-Faktoren-Theorie dar, der zufolge sich bei der Teameffektivität und Innovation die beiden Aspekte „Qualität“ und „Quantität“ von Innovationen unterscheiden lassen.

» Qualität bezieht sich auf die Neuartigkeit von Ideen, deren Bedeutsamkeit, gemessen an den jeweils relevanten Kriterien, und deren Nutzen. Quantität bezieht sich auf die Anzahl neuer Ideen, die vorgeschlagen und umgesetzt werden. (Brodbeck et al. 2000, S. 10)



■ Abb. 6.1 Dimensionen und Subskalen des TKI. (Nach Brodbeck et al. 2000, S. 9, © Hogrefe)

Für die Qualität sind die beiden Faktoren „Vision“ und „Aufgabenorientierung“, für die Quantität „partizipative Sicherheit“ und „Unterstützung für Innovationen“ maßgeblich. Diese 4 Faktoren sollen mit insgesamt 13 Skalen gemessen werden, zu denen noch 2 kurze Skalen zur sozialen Erwünschtheit kommen. Die Zugehörigkeit der einzelnen Skalen zu den Faktoren ist der □ Abb. 6.1 zu entnehmen.

Wie aus den Itemtexten (□ Tab. 6.2) ersichtlich ist, zielen manche Formulierungen auf individuelle Einstellungen unter selbst- oder teambezogener Perspektive, andere verlangen eine Einschätzung der atmosphärischen Gegebenheiten im Team, und einige Fragen erfordern ein gedankliches Hin- einersetzen in andere Mitglieder des Teams. Das Klima eines Teams ergibt sich aus der Mittelung der (gewöhnlich anonym abgelieferten) individuellen Punktwerte.

Teamklima als Aggregat der individuellen Skalenwerte

Hohe interne Konsistenz und hohe Interkorrelation der Skalen

Konstruktionsprinzipien und psychometrische Kennwerte Insgesamt 810 Personen aus 149 Teams stellten die Analyse- bzw. Normierungsstichprobe dar. Obwohl die Autoren betonen, dass das TKI „zur Messung von Merkmalen auf Teamebene konstruiert“ worden sei (Brodbeck et al. 2000, S. 39), wurden die internen Konsistenzen doch anhand der individuellen Daten ermittelt (die Werte für Cronbachs Alpha liegen für die 4 Skalen zwischen .84 und .89, für die Subskalen zwischen .61 und .82). Desgleichen beruhen die relativ hohen Interkorrelationen zwischen den Skalen (um .60) auf individuellen Werten, im Weiteren auch die konfirmatorischen Faktorenanalysen, die die Vier-Faktoren-Struktur bestätigen. Die Validierung erfolgte allerdings auf der Aggregatebene des Teams; als Kriterien wurden auf das Team bezogene Fremdeinschätzungen des Projektmanagements und der Teammoderation herangezogen. Mit fremdeingeschätzter Innovation korrelierte die TKI-Dimension „Vision“ zu $r = .64$, „Unterstützung für Innovation“ zu $r = .62$; die Korrelationen für „Aufgabenorientierung“ und „Partizipative Sicherheit“ lagen bei $r = .32$ bzw. .48 (alle Werte bis auf den vorletzten hoch signifikant; jeweils $N = 29$ Teams). Darüber hinaus ließ sich zeigen, dass die Übereinstimmung zwischen den Mitgliedern eines Teams mit Korrelationen über .90 sehr hoch ist und das Verfahren zwischen verschiedenen Teams (Industrie, Pflege, Entwicklung und Planspiel) signifikant diskriminiert (wobei aber der letztgenannte Vergleich erneut auf Individualdaten beruht, was nicht ganz der Logik des Verfahrens entspricht).

Tab. 6.2 Einige Itembeispiele für die Skalen und Subskalen des TKI. (Aus Brodbeck et al. 2000, S. 22 f., © Hogrefe)

Skala	Subskala	Itembeispiel(e)
Vision	Klarheit	„Wie genau sind Sie sich im Klaren über die Ziele Ihres Teams?“ „Was denken Sie, inwieweit sind die Ziele Ihres Teams den anderen Teammitgliedern klar und deutlich gegenwärtig?“
	Wertschätzung	„Was denken Sie, inwieweit sind diese Ziele nützlich und angemessen?“
	Einigkeit	„Inwieweit stimmen Sie mit diesen Zielen überein?“ „Was denken Sie, inwieweit fühlen sich die Mitglieder Ihres Teams diesen Zielen verpflichtet?“
	Erreichbarkeit	„Was denken Sie, inwieweit sind diese Ziele realistisch und erreichbar?“
Aufgabenorientierung	Hohe Standards	„Ist es den Teammitgliedern ein echtes Anliegen, dass das Team den höchstmöglichen Leistungsstandard erreicht?“
	Reflexion	„Sind die Teammitglieder bereit, die Grundlagen der eigenen Arbeit infrage zu stellen?“
	Synergie	„Bauen die Teammitglieder gegenseitig auf ihren Ideen auf, um das bestmögliche Ergebnis zu erhalten?“
Partizipative Sicherheit	Informationsverteilung	„Wir halten uns über arbeitsrelevante Themen gegenseitig auf dem Laufenden.“
	Sicherheit	„Die Teammitglieder fühlen sich gegenseitig akzeptiert und verstanden.“
	Einfluss	„Jede Ansicht wird angehört, auch wenn es die Meinung einer Minderheit ist.“
	Kontaktpflege	„Wir stehen in häufigem, gegenseitigem Austausch.“
Unterstützung für Innovation	Bereitschaft (artikulierte Normen)	„Das Team ist Veränderungen gegenüber aufgeschlossen und empfänglich.“
	Umsetzung (im Handeln erkennbare Normen)	„In unserem Team nehmen wir uns die Zeit, die wir brauchen, um neue Ideen zu entwickeln.“

Bewertung Das TKI ist ein theoretisch verankertes und mit einer ca. 15-minütigen Bearbeitungszeit sehr ökonomisches Instrument, mit dem das in der Gruppe herrschende Arbeitsklima durch Mittelung der individuellen Teammitglieder-Scores erfasst wird. Auf individueller Ebene weisen die 4 Skalen eine für Fragebögen hohe interne Konsistenz auf. Für die aggregierten Skalenwerte des Teams ließ sich die Validität gegenüber fremdeingeschätzten Maßen für Innovation demonstrieren. Insgesamt handelt es sich um ein für die Team- und die Organisationsentwicklung sehr nützliches Instrument.

6.1.2 Arbeits- und Anforderungsanalyse

Das Vorgehen zur Diagnostik verschiedener Arbeitsbedingungen (Tätigkeitssinhalte, Werte, Klima, etc.) wird auch als Arbeitsanalyse bezeichnet.

Definition

„Mit **Arbeitsanalyse** wird ganz allgemein die systematische Erfassung und Bewertung von Informationen über die Interaktion von Mensch und Arbeitsbedingungen bezeichnet [...]. In der psychologischen Arbeitsanalyse geht es um die Analyse und Bewertung der Arbeitsaufgabe(n) und der Arbeitsbedingungen“ (Dunckel und Resch 2010, S. 1111 f.).

Davon abzugrenzen ist die Anforderungsanalyse.

Abgrenzung von Arbeitsanalyse und Anforderungsanalyse

Definition

Eine **Anforderungsanalyse** ist definiert als „systematische Analyse der Anforderungen und der Motivations-/Demotivationspotenziale der beruflichen Tätigkeit mit dem Ziel der Ermittlung derjenigen Eignungsmerkmale von Personen, die bedeutsam dafür sind, dass sie die erforderlichen Leistungen erbringen oder mit dem zu besetzenden Arbeitsplatz, dem Aufgabenfeld, der Ausbildung bzw. dem Studium oder dem Beruf zufrieden sind sowie die Festlegung der dafür erforderlichen Ausprägungsgrade dieser Eignungsmerkmale“ (DIN 2016, S. 6).

Kompetenzmodellierung

Wenngleich Arbeitsanalyse und Anforderungsanalyse häufig synonym oder als kombinierter Begriff (Arbeits- und Anforderungsanalyse) verwendet werden, so wird deren Abgrenzung anhand der beiden Definitionen deutlich: Die Arbeitsanalyse bezieht sich auf die Arbeit bzw. Tätigkeiten; die Anforderungsanalyse leitet aus der Kenntnis der Tätigkeiten ab, welche Merkmale Menschen aufweisen sollten, um diese Tätigkeiten erfolgreich erledigen zu können.

Der ebenfalls häufig verwendete Begriff der Kompetenzmodellierung kann weniger klar definiert und damit auch schlecht von Arbeits- und Anforderungsanalysen abgegrenzt werden. Im Jahr 2000 veröffentlichte die „Job Analysis and Competency Modeling Task Force“ eine Studie zu dem Unterschied zwischen Arbeits- und Anforderungsanalysen (im Englischen „job analysis“) und Kompetenzmodellierungen (Schippmann et al. 2000). Zum Zwecke dieser Studie hatte die Task Force umfangreiche Befragungen von Fachkundigen vorgenommen und Literaturrecherchen durchgeführt. Das Ergebnis zeigt, dass die meisten Fachkundigen von Arbeits- und Anforderungsanalysen höhere wissenschaftliche Standards erwarten (etwa bezüglich der Genauigkeit der Beschreibung von Anforderungsdimensionen oder bezüglich der Entwicklung der Instrumente zur Arbeits- und Anforderungsanalyse). Einen Vorteil der Kompetenzmodellierung sahen die Befragten in deren engerer Verzahnung mit strategischen Zielen der Organisation. Nachfolgend soll der Fokus auf Arbeits- und Anforderungsanalysen als (im Idealfall) wissenschaftlichen Standards genügenden Prozessen zur Diagnostik von Tätigkeitsinhalten und beruflichen Anforderungen liegen.

In der Literatur werden häufig die folgenden 3 grundsätzlichen Zugänge der Arbeits- und Anforderungsanalyse beschrieben.

Grundsätzliche Zugänge der Arbeits- und Anforderungsanalyse

- Erfahrungsgeleitet-intuitiv
- Personenbezogen-empirisch
- Arbeitsplatzanalytisch-empirisch

Erfahrungsgeleitet-intuitive Methode

Um die angestrebte Trennung zwischen Arbeitsanalyse und Anforderungsanalyse beizubehalten, sei an dieser Stelle erwähnt, dass die arbeitsplatzanalytisch-empirische Methode der Arbeitsanalyse zuzuordnen ist. Die personenbezogen-empirische Methode dient der Anforderungsanalyse. Arbeits- und Anforderungsanalyse lassen sich gleichermaßen erfahrungsgeleitet-intuitiv durchführen.

Unter der erfahrungsgeleitet-intuitiven Methode versteht man die freie, nicht formalisierte Beurteilung der Tätigkeitsmerkmale oder Anforderungen an einen Beruf. Sie verlangt gründliche Kenntnisse der Stelle und ihrer organisatorischen Einbettung. Meist wird diese Methode in Form von wenig standardisierten Interviews oder Workshops mit (langjährigen) Stelleninhaberinnen bzw. Stelleninhabern oder deren Vorgesetzten umgesetzt.

Bei der personenbezogen-empirischen Methode werden empirisch ermittelte Zusammenhänge zwischen Personenmerkmalen (wie beispielsweise Intelligenz) und Kriterien des Ausbildungs- oder des Berufserfolgs genutzt. Idealerweise stammen die ermittelten Zusammenhänge aus dem eigenen Unternehmen. Aber auch metaanalytisch aufbereitete Ergebnisse der (internationalen) Forschung oder Befunde aus Einzelstudien mit vergleichbaren Berufsgruppen können für eine personenbezogen-empirische Anforderungsanalyse genutzt werden. Als relevante Anforderungen gelten solche Personenmerkmale, für die ein möglichst hoher Zusammenhang mit den relevanten Erfolgsmaßen (Leistung, Zufriedenheit etc.) nachgewiesen wurde (s. □ Abb. 6.2 für ein Beispiel aus der Führungsforschung).

Bei der arbeitsplatzanalytisch-empirischen Methode werden mithilfe von standardisierten Verfahren Informationen über die Stelle erhoben. Dazu stehen verschiedene Instrumente zur Verfügung, die sich u. a. hinsichtlich der theoretischen Grundlage und der Auswahl der zu beschreibenden Arbeitsmerkmale unterscheiden.

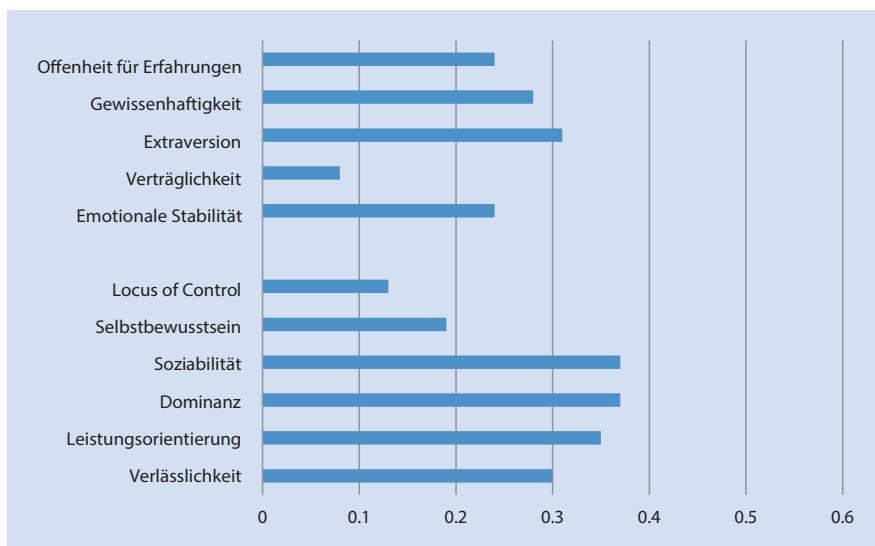
Personenbezogen-empirische Methode

Arbeitsplatzanalytisch-empirische Methode

6.1.2.1 Fragebogen zur Arbeitsanalyse (FAA)

Der FAA von Frieling und Hoyos (1978), der aus dem im angloamerikanischen Raum gebräuchlichen *Position Analysis Questionnaire (PAQ)* von McCormick, Jeanneret und Mecham (1969) hervorgegangen ist, umfasst annähernd 200 Items. Die Items beziehen sich auf Variablengruppen wie kognitive Prozesse (u. a. Informationsaufnahme), den Arbeits-Output, die Beziehung zu anderen Personen oder die Arbeitsumgebung. Der FAA soll explizit die Bedingungen der Arbeit erfassen (□ Tab. 6.3). Die Beantwortung der Items geschieht teilweise durch Befragungen, teilweise durch Beobachtungen. Das Verfahren kann relativ breit eingesetzt werden, also etwa zur Beschreibung und Bewertung manueller Tätigkeiten in der Produktion wie auch geistiger Arbeit bei Bürotätigkeiten oder Verwaltungsaufgaben. Der Fragebogen wird hauptsächlich zur Klassifikation und zum Vergleich von Arbeitstätigkeiten eingesetzt.

Verfahren zur Klassifikation und zum Vergleich von Arbeitstätigkeiten



■ Abb. 6.2 Metaanalytische Zusammenhänge zwischen Persönlichkeitsmerkmalen und Führungsverhalten. (Aus Judge et al. 2002, mit freundlicher Genehmigung der American Psychological Association). Die metaanalytisch „wichtigsten“ Eigenschaften sind Sozialität, Dominanz und Leistungsorientierung sowie, als breites Persönlichkeitsmerkmal, Extraversion. Als Kriterium „Führungsverhalten“ diente ein kombiniertes Maß des Führungserfolgs und der Führungs emergenz. Angegeben sind mittlere Korrelationen, korrigiert für Messfehler in Prädiktor und Kriterium.

Tab. 6.3 Instruktion und Beispielitems für die Bereiche des FAA. (Frieling und Hoyos 1978, © Hogrefe)

Bereich	Beispielitem(s)
Informationsaufnahme und Gefährdungsarten	„Stufen Sie die Arbeitselemente danach ein, wie häufig sie als Informationsquellen vom Stelleninhaber benutzt werden, um die Aufgaben erfolgreich erledigen zu können. Die Häufigkeit soll dabei auf die Gesamtheit aller am Arbeitsplatz auftretenden Arbeitsprozesse bezogen werden.“
Optische Quellen der Arbeitsinformation	<ul style="list-style-type: none"> – „Wie häufig dient gedrucktes, maschinengeschriebenes oder in Druckschrift geschriebenes Material (z. B. Bücher, Zeitschriften, Zeitungen, Berichte, Dienstschriften, Texte oder Briefe) als Quelle der Arbeitsinformation?“ – „Wie häufig dient handgeschriebenes Material (z. B. Entwürfe für Briefe, Notizen, handschriftliche Anweisungen oder Stenogramme) als Quelle der Arbeitsinformation?“ – „Wie häufig dient Zahlenmaterial (Material, das aus Zahlen oder Beträgen besteht; z. B. numerische Angaben, Rechnungen, technische Daten oder Zahlentabellen) als Quelle der Arbeitsinformation?“
Gefährdungsarten	<ul style="list-style-type: none"> – Gefährdung durch Werkzeuggebrauch (Stelleninhaberinnen und Stelleninhaber benutzen unfallträchtige Werkzeuge; z. B. Schnitt- und Stechwerkzeuge, Sägen oder Skalpelle) – Gefährdung durch sich bewegende oder fallende Objekte (Stelleninhaberinnen und Stelleninhaber steuern oder bedienen Fahrzeuge und/oder Transportgeräte, oder arbeiten an Transporteinrichtungen, Hebezeugen oder Hochregalen; z. B. Anschläger, Gabelstaplerfahrer, Kranführer oder Lagerist) – Gefährdung durch „Arbeit an erhöhten Plätzen“ (Stelleninhaberinnen und Stelleninhaber arbeiten auf Leitern, Gerüsten, Dächern, Kaminen usw. Berücksichtigen Sie bei der Einstufung, dass die Unfallgefährdung durch Wettereinflüsse noch gesteigert werden kann) – Gefährdung durch Hitze bzw. Feuer (Stelleninhaberinnen und Stelleninhaber sind bei der Arbeit der Gefahr von Verbrennungen ausgesetzt; z. B. beim Schweißen, beim Kochen oder beim Löschen von Bränden)

6**6.1.2.2 Das Tätigkeitsbewertungssystem (TBS)****Verfahren zur Arbeitsgestaltung**

Ein dem handlungsregulations- bzw. tätigkeits theoretischen Ansatz verpflichtetes Verfahren ist das TBS von Hacker et al. (1995). Es wird vor allem zur Arbeitsgestaltung eingesetzt und ist für verschiedene Bereiche wie Produktion oder Büro und Verwaltung geeignet.

Die 52 Skalen (Items) sollen 5 Merkmalsbereiche erfassen:

- Organisatorische und technische Bedingungen, die die Vollständigkeit bzw. die Unvollständigkeit von Tätigkeiten determinieren (z. B. körperliche Abwechslung)
- Kooperation und Kommunikation
- Verantwortung, die aus dem Arbeitsauftrag folgt
- Erforderliche geistige (kognitive) Leistungen
- Qualifikations- und Lernerfordernisse

6.1.2.3 Work Design Questionnaire (WDQ)**Personenbezogenes Verfahren**

Als Beispiel für ein personenbezogenes Verfahren der Arbeitsanalyse kann der WDQ von Morgeson und Humphrey (2006; deutsche Übersetzung und Validierung: Stegmann et al. 2010) genannt werden.

Die Inhalte des WDQ sind das Ergebnis einer Analyse der Autoren des englischen Originals, die verschiedene Datenbanken hinsichtlich relevanter Arbeitsplatzmerkmale gesichtet haben. Es wurden insgesamt 21 relevante Merkmale identifiziert, die in 4 großen Merkmalsbereichen zusammengefasst sind (Aufgaben-, Wissens-, soziale und kontextuelle Arbeitsplatzmerkmale; **Tab. 6.4**). Diese Struktur konnte durch eine Faktorenanalyse bestätigt werden. Ebenfalls ließen sich theoriekonforme Zusammenhänge zu Kriteriumsvariablen wie Zufriedenheit und Arbeitsmotivation nachweisen (Stegmann et al. 2010).

Tab. 6.4 Inhalte und Beispielitems des Work Design Questionnaire. (WDQ, Morgeson und Humphrey 2006, © American Psychological Association, Stegmann et al. 2010, © Hogrefe)

Merkmalsbereich	Arbeitsplatzmerkmal	Beispielitem ^a
Aufgabenmerkmale	Autonomie: Planung	Ich kann meine Arbeit so planen, wie ich es möchte
	Autonomie: Entscheidungen	Ich kann bei meiner Arbeit viele Entscheidungen selbstständig treffen
	Autonomie: Methode	Ich habe viele Freiheiten in der Art und Weise, wie ich meine Arbeit verrichte
	Aufgabenvielfalt	Meine Aufgabe ist sehr abwechslungsreich
	Wichtigkeit	Das Ergebnis meiner Arbeit hat einen großen Einfluss auf andere Menschen
	Ganzheitlichkeit	Bei meiner Arbeit habe ich die Möglichkeit, Produkte/Dienstleistungen, die ich beginne, fertigzustellen
	Rückmeldung durch die Tätigkeit	Durch die Tätigkeit selbst erhalte ich automatisch Rückmeldung über meine Leistung
Wissensmerkmale	Komplexität	Meine Arbeit kann fast jeder ohne große Einarbeitung machen
	Informationsverarbeitung	Ich verarbeite bei meiner Arbeit sehr viele Informationen
	Problemlösen	Meine Arbeit verlangt ungewöhnliche Ideen oder Problemlösungen
	Anforderungsvielfalt	Meine Arbeit erfordert eine Fülle von Fertigkeiten
	Spezialisierung	Die Werkzeuge, Prozeduren, Materialien, etc., die ich verwende, sind speziell auf meine Tätigkeit zugeschnitten
Soziale Merkmale	Soziale Unterstützung	Meine Kolleginnen und Kollegen interessieren sich für mich
	Initiierte Interdependenz	Bevor meine Arbeit nicht fertig ist, können andere ihre Arbeit nicht erledigen
	Rezipierte Interdependenz	Meine Arbeitsaufgaben sind stark von der Arbeit anderer Personen abhängig
	Interaktion außerhalb der Organisation	In meiner Arbeit kommuniziere ich häufig mit Personen, die nicht in meiner Organisation arbeiten
	Rückmeldung durch andere	Ich erhalte von Kolleginnen und Kollegen Rückmeldung über meine Arbeitsleistung
Kontextuelle Merkmale	Ergonomie	An meinem Arbeitsplatz ist die Art der Sitzgestaltung angemessen (z. B. ausreichend viele Sitzgelegenheiten, bequeme Stühle, gute Haltung wird unterstützt)
	Physische Anforderungen	Die Arbeit erfordert starke körperliche Anstrengung
	Arbeitsbedingungen	Bei meiner Arbeit ist das Unfallrisiko gering
	Technikgebrauch	Meine Arbeit beinhaltet die Benutzung vieler verschiedener Geräte/Werkzeuge/Instrumente

^aAls Beispielitem wurde jeweils das Item ausgewählt, das die höchste Trennschärfe in Studie 2 bei Stegmann et al. (2010) aufwies

6.1.2.4 Critical-Incident-Technik (CIT)

Neben den hier dargestellten Fragebögen kann auch eine etablierte Interviewtechnik, die CIT von Flanagan (1954), genutzt werden. Die bereits 1954 entwickelte Methode verlangt von Führungskräften, dass sie Verhaltensweisen ihrer Mitarbeitenden beschreiben, die zu Erfolg oder Misserfolg geführt haben, beispielsweise zu einer Erhöhung oder einer Verringerung der Produktion. Die Instruktion kann lauten:

Critical-Incident-Interviews

Beispielhaftes Vorgehen im Rahmen der Critical-Incident-Technik

„Denken Sie an ein Beispiel für das Arbeitsverhalten einer Mitarbeiterin oder eines Mitarbeiters, das besonders effektive oder besonders ineffektive Verhaltensweisen veranschaulicht. Beschreiben Sie die Situation und das fragliche Verhalten möglichst konkret. Stellen Sie sich dazu die folgenden Fragen:

- Was waren die Umstände oder Hintergrundbedingungen, die zu einem Verhalten führten?
- Beschreiben Sie das konkrete Verhalten der Mitarbeiterin bzw. des Mitarbeiters. Was war besonders effektiv oder ineffektiv an diesem Verhalten?
- Was waren die Konsequenzen dieses Verhaltens?“

(Schuler, 2006, S. 55).



Critical Incident: Wütende Kundin und wütender Kunde. (© Antonioguillem/stock.adobe.com).

Diese Fragen können mündlich oder schriftlich gestellt werden. Die als Antworten resultierenden Incidents können von Führungskräften oder Stelleninhaberinnen bzw. Stelleninhabern hinsichtlich ihrer Wichtigkeit bewertet und anschließend gruppiert werden; ähnliche Incidents bilden eine Kategorie (z. B. „Umgang mit kritischen Kundinnen und Kunden“). Dabei ist das Vorgehen sehr genau zu dokumentieren, da sonst die Gefahr besteht, dass die Ergebnisse in hohem Maße fehlerbehaftet sind (Koch et al. 2009). Zudem empfiehlt es sich, möglichst verschiedene Urteilende, also beispielsweise Führungskräfte und Stelleninhaberinnen bzw. Stelleninhaber, einzubeziehen (Koch et al. 2012).

Übersetzung der Arbeitsplatzmerkmale oder Verhaltensweisen in Anforderungsmerkmale

Es wird deutlich, dass Verfahren zur Arbeitsanalyse keineswegs automatisch Erkenntnisse darüber liefern, welche Merkmale die Personen aufweisen müssen, die eine bestimmte Arbeit erledigen. Auch Situationen, die im Rahmen eines Critical-Incident-Interviews ermittelt wurden, erlauben keinen trivialen Rückschluss auf die zugrunde liegenden Anforderungen. Dies ist erst nach einer theoriegeleiteten Übersetzung der Arbeitsplatzmerkmale oder Verhaltensweisen in Anforderungsmerkmale möglich. Alternativ können Anforderungsmerkmale auch direkt erhoben werden, beispielsweise durch die Befragung relevanter Mitarbeiterinnen und Mitarbeiter anhand von standardisierten Fragebögen.

6.1.2.5 EXPLOJOB – Werkzeug zur Beschreibung von Berufsanforderungen und -tätigkeiten

EXPLOJOB (Joerin Fux und Stoll 2006) beansprucht die Beschreibung von Berufen anhand der 6 Dimensionen des RIASEC-Modells (Holland 1997; Abb. 6.3).

Anhand eines in 10–15 min auszufüllenden Fragebogens beantworten Stelleninhaberinnen bzw. -inhaber oder Fachkundige insgesamt 84 Fragen auf einer 3-stufigen Skala („oft“, „manchmal“, „selten/nie“), die sich zu folgenden Bereichen zusammenfassen lassen.

6 Dimensionen zur Beschreibung von Berufen

Inhaltsbereiche des EXPLOJOB

1. Zu verrichtende Tätigkeiten (z. B. „etwas erforschen“)
2. Erforderliche Begabungen oder Eigenschaften (z. B. „Geschick im Umgang mit Menschen“)
3. Zu verwirklichende Interessen, Bedürfnisse, Werte oder Vorlieben (z. B. „Gemeinschaftssinn, Teamwork“)
4. Passung zu Berufssektoren (z. B. „Handwerk, Technik, Produktion“)

Anhand der Aggregation der Rohwerte aller Items, die zu einer der 6 Dimensionen nach Holland (1997) gehören, wird eine Rangreihe der vorherrschenden Anforderungen eines Berufs gebildet. Die Buchstaben der wichtigsten Anforderungen werden zur Beschreibung des Berufs genutzt (z. B. für Architektin ACE, d. h. Artistic, Conventional, Enterprising).

Rangreihe der Anforderungen eines Berufs

Die 84 Items laden nach Extraktion von 6 Dimensionen auf den postulierten Faktoren, wenngleich die berichteten Extraktionskriterien eine Extraktion von weniger als 6 Faktoren nahelegen. Die im Manual berichteten Reliabilitätsschätzungen (interne Konsistenz) liegen größtenteils bei .90 und sind als sehr gut zu bezeichnen. Angaben zur Beurteilungsübereinstimmung sind dem Manual nicht zu entnehmen. Als erste Belege für die Konstruktvalidität dienen größtenteils übereinstimmende Buchstabenkombinationen, wenn

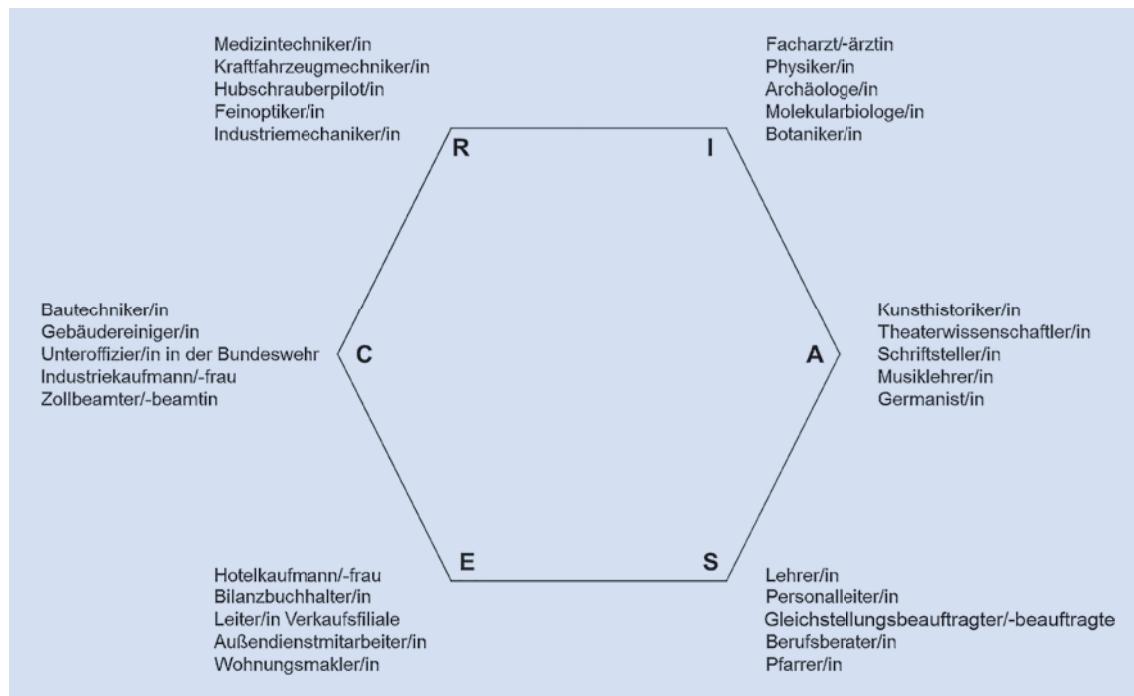


Abb. 6.3 Beispielhafte Berufsklassifikationen nach dem Holland-Modell. Nennung von Berufen, deren Codes aus dem Buchstaben des jeweiligen Ecks sowie der angrenzenden Buchstaben bestehen

diese mit dem EXPLOJOB und dem Allgemeinen Umwelt-Struktur-Test (in der revidierten Fassung, AIST-R; Bergmann und Eder 2005) verglichen wurden. Normen werden bewusst nicht bereitgestellt, es erfolgt nur ein intrabefruchtlicher Abgleich der 6 Dimensionen (Abb. 6.3).

6.1.2.6 Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – Anforderungsmodul (BIP-AM und BIP-6 F-AM)

Beschreibung von beruflichen Positionen

6

Anforderungsprofil

Neben der Selbst- und Fremdbeschreibungsversion des in ► Abschn. 3.3.3.4 dargestellten Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) bzw. des Bochumer-Inventars zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren (BIP-6 F; Hossiep et al. 2018) existieren weitere Versionen, das BIP-AM (Hossiep und Weiß 2020) bzw. BIP-6 F-AM (Hossiep et al. 2020, in Vorbereitung). Diese Varianten werden ebenfalls in ► Abschn. 3.3.3.4 dargestellt. Sie ermöglichen eine Beschreibung von beruflichen Positionen anhand der gleichen 14 (bzw. 6) Dimensionen (z. B. im BIP-AM: Leistungsmotivation, emotionale Stabilität, Durchsetzungsfähigkeit), die auch in der Selbst- und Fremdbeschreibungsversion enthalten sind. Ebenfalls analog zur Selbst- und Fremdbeschreibungsversion können die Dimensionen in breiten Merkmalsbereichen zusammengefasst werden (berufliche Orientierung, Arbeitsverhalten, soziale Kompetenzen, psychische Konstitution). Die Skalen zur Erfassung der enthaltenen Dimensionen zeigten in Studien der Testautoren Reliabilitätschätzungen, die als gut zu beurteilen sind. Zur Validität berichten die Testautorinnen und -autoren Übereinstimmungen mit Experteneinschätzungen (BIP-AM) bzw. der Selbstbeschreibung (BIP-6 F-AM).

Unabhängig vom methodischen Vorgehen liefert eine Anforderungsanalyse in der Regel mehrere Anforderungen. Durch Befragung von Stelleninhaberinnen und Stelleninhabern oder deren Vorgesetzten lässt sich feststellen, welche Ausprägung jedes einzelnen Anforderungsmerkmals erwünscht oder optimal ist. So kann je nach Anforderungsmerkmal eine andere Mindestausprägung festgelegt werden. Für manche Merkmale kann es zudem sinnvoll sein, neben einer Mindest- auch eine Höchstausprägung festzulegen. Denn nicht immer gilt: je mehr, desto besser. Beispielsweise kann eine mindestens durchschnittliche Sorgfalt beim Arbeiten erwünscht sein, eine sehr hohe Sorgfalt wäre aber ein Zuviel des Guten. Listet man die Anforderungsmerkmale nacheinander auf, versieht sie mit einer Skala zur Ausprägung und kennzeichnet auf jeder Skala den optimalen Bereich, erhält man ein Anforderungsprofil (Abb. 6.4). In dieses Anforderungsprofil können die festgestellten Merkmalsausprägungen einer Person eingetragen werden. Man erkennt dann meist leicht, wie gut diese Person zu den Anforderungen passt.

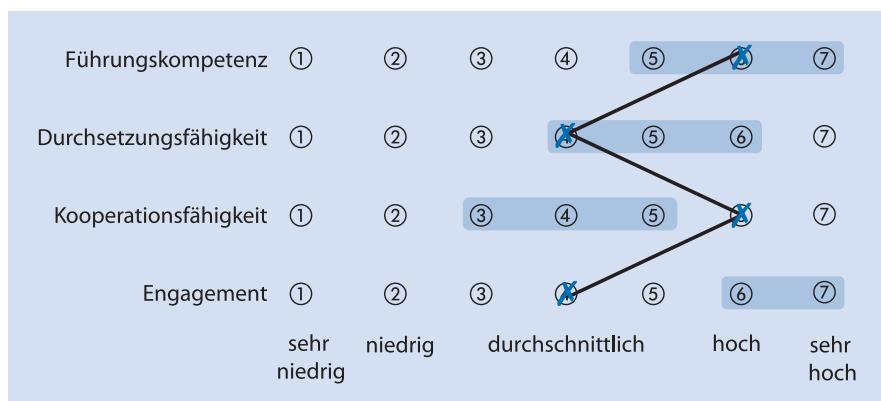
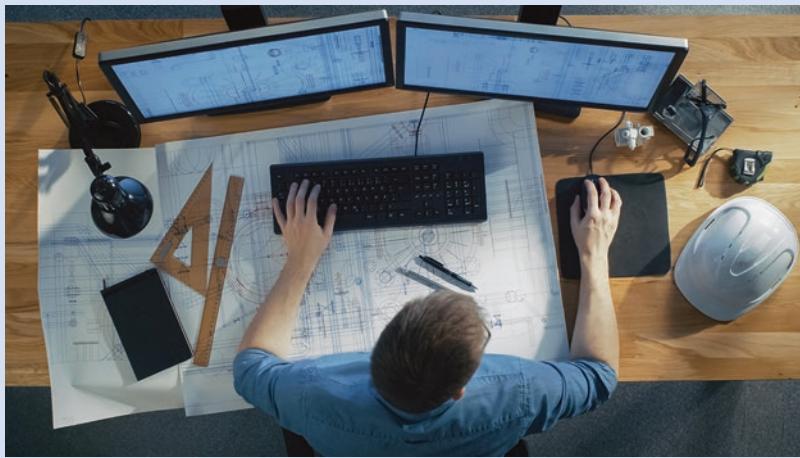


Abb. 6.4 Anforderungsprofil (Ausschnitt)

Occupational Information Network (O*NET)

Die wahrscheinlich größte und systematischste Sammlung arbeits- und anforderungsanalytischer Ergebnisse bietet das Occupational Information Network (O*NET), das vom US Department of Labor, Employment and Training Administration gefördert wird (O*NET 2020). Der frei im Internet zugänglichen Datenplattform (► <https://www.onetonline.org/>) liegt ein Inhaltsmodell zugrunde, in dem Anforderungen an Personen (u. a. Fähigkeiten, Interessen, Erfahrungen, Kenntnisse und Fertigkeiten, Ausbildung) und Beschreibungen von Berufen (u. a. Tätigkeiten, Arbeitsumfeld, Arbeitsmarktinformationen, benötigte Hilfsmittel und Arbeitswerkzeuge) unterschieden werden – es besteht also aus einer Anforderungsanalyse und einer Arbeitsanalyse. Insgesamt werden 277 Beschreibungsmerkmale erhoben. Diese lagen im Sommer 2017 für über 974 Berufe vor. Diese reichhaltige Informationsbasis kann von Berufsinteressierten und Berufsberaterinnen und -beratern im Vorfeld der Berufswahl genutzt werden. Aber auch Personalentwicklerinnen und -entwickler können O*NET zur Identifikation von notwendigen Fertigkeiten und Kompetenzen, die ggf. geschult werden müssen, nutzen.

Die Inhalte von O*NET basieren auf Befragungen US-amerikanischer Expertinnen und Experten sowie Stelleninhaberinnen und Stelleninhaber. Einige Berufsbilder im deutschsprachigen Raum mögen analog zu ihren US-amerikanischen Pendants ausgestaltet sein und sind daher angemessen im O*NET beschrieben. Es ist jedoch anzunehmen, dass einige Berufsbilder divergieren. Auf dem Portal „berufenet.de“ der Bundesagentur für Arbeit (► <https://berufenet.arbeitsagentur.de/>) sind für viele Berufe die dem deutschen Arbeitsmarkt entsprechenden Informationen hinterlegt. Neben psychologischen Anforderungen (z. B. werden für den Beruf „Bäcker/in“ genannt: Geschicklichkeit und Sinn für Ästhetik, Verantwortungsbewusstsein, gute körperliche Konstitution) sind dort Tätigkeitsbeschreibungen, erwarteter Schulabschluss und Verdienstmöglichkeiten aufgeführt. Ähnliche Datenbanken existieren für Österreich (► <https://www.berufslexikon.at/>) und die Schweiz (► <https://www.berufsberatung.ch/>).



Beispielbild zum Beruf Architektin bzw. Architekt. Die Tätigkeit, die hier illustriert wird, lautet: eine detaillierte Ausführungszeichnung erstellen. (© Gorodenkoff/stock.adobe.com).

Weitere Tätigkeiten dieses Berufsbilds sind laut Berufenet: im Büro eine Leistungsbeschreibung erstellen; den Plan für ein Bauprojekt im Besprechungsraum erörtern; auf der Baustelle die Bauausführung dokumentieren; die Statik eines Gebäudeteils überprüfen; Entwurfsunterlagen zusammenfassen; einen geplotteten Bauplan auf korrekte Darstellung prüfen; Maße für bauphysikalische Berechnungen abnehmen; mit einem CAD-Programm Ansichtszeichnungen erstellen; einem Handwerker Planänderungen erläutern.

US-amerikanische Datenplattform mit Anforderungen an Personen und Berufsbeschreibungen

Deutschsprachige Plattformen

Diagnostische Verfahren der Personalauswahl

6.2 Diagnostik von Personenmerkmalen

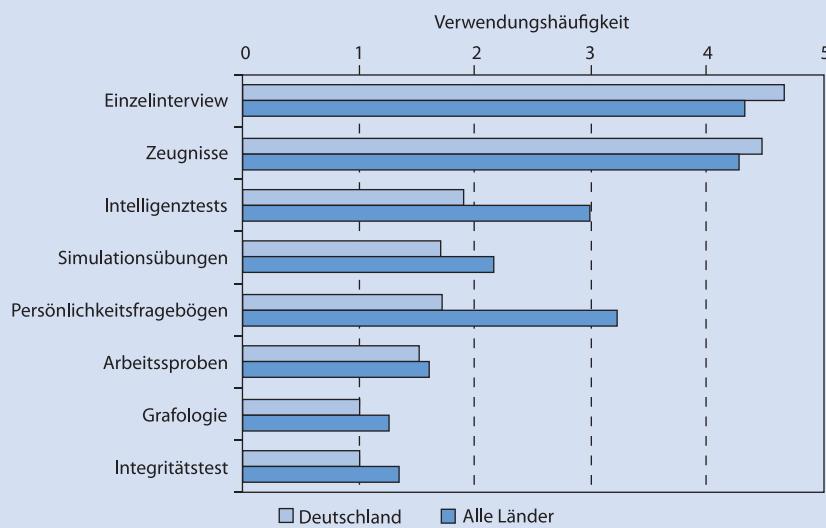
Möglichkeiten zur Diagnostik von Personenmerkmalen sind in ▶ Kap. 3 ausführlich dargestellt. Viele der dort vorgestellten diagnostischen Verfahren kommen auch in der betrieblichen Praxis zum Einsatz. Dies soll am Beispiel der Personalauswahl etwas genauer ausgeführt werden. Welche Verfahrenstypen von Unternehmen zur Personalauswahl häufig eingesetzt werden, ist aus Befragungen bekannt. Schuler et al. (2007) sowie Armoneit, Schuler und Hell (2020) haben Unternehmen in Deutschland befragt; die Ergebnisse sind in □ Tab. 6.5 aufgeführt. Das Interview stellt das mit Abstand am häufigsten verwendete diagnostische Verfahren dar. Gegenüber einer früheren Befragung (Schuler et al. 1993) hat sich das Verhältnis von strukturierten zu unstrukturierten Interviews deutlich zugunsten der strukturierten verschoben. Dieser Trend hat sich in der Befragung aus 2017/2018 fortgesetzt. Auch strukturierte Telefoninterviews kommen mittlerweile häufig zum Einsatz. Während der Einsatz von Assessment-Centern deutlich abgenommen hat, erfreuen sich online durchführbare Verfahren zunehmender Beliebtheit.

□ **Tab. 6.5** Einsatzhäufigkeit ausgewählter eignungsdiagnostischer Verfahren in den befragten Unternehmen. (Nach Armoneit et al. 2020, mit freundlicher Genehmigung des Hogrefe Verlages)

Eignungsdiagnostisches Verfahren	Einsatzhäufigkeit 2007 (Angaben in %)	Einsatzhäufigkeit 2017/2018 (Angaben in %)
Arbeitsproben/Fallstudien	44,8	45,7
Kognitive Leistungstests	35,6	23,6
Online-Leistungstests	0	16,4
Interview, strukturiert	72,8	72,9
Interview, unstrukturiert	42,4	33,6
Telefoninterview, strukturiert	32,0	40,7
Telefoninterview, unstrukturiert	24,0	25,7
Videointerview, strukturiert	–	12,1
Videointerview, unstrukturiert	–	8,6
Persönlichkeitsfragebögen	20,0	19,3
Online-Persönlichkeitsfragebögen	1,6	23,6
Assessment-Center	57,6	37,1
Situational-Judgment-Tests	–	5,0
Referenzen	56,8	41,4
Grafologie	2,4	2,1

Einsatz von Auswahlverfahren im internationalen Vergleich

Ältere Befragungsergebnisse zeigen deutliche Unterschiede bei der Verwendung von Personalauswahlinstrumenten zwischen Deutschland und anderen Ländern. Ryan et al. (1999) befragten fast 1000 Unternehmen in 20 Ländern über die Verfahren, die sie zur Personalauswahl einsetzen. In folgender Abbildung sind ausgewählte Verfahren geordnet nach ihrer damaligen Verwendungshäufigkeit in Deutschland aufgeführt.



Verwendungshäufigkeit verschiedener Verfahren zur Personalauswahl im internationalen Vergleich (Ryan et al. 1999). Antwortskala: 1 = nie, 2 = selten (1–20 %), 3 = gelegentlich (21–50 %), 4 = oft (51–80 %), immer oder 5 = fast immer (81–100 %).

Übereinstimmend mit der internationalen Praxis wurden in Deutschland fast immer ein Interview mit Bewerbenden durchgeführt und deren Berufsqualifikation anhand von Zeugnissen etc. festgestellt. Grafologie und Integritätstests spielten in Deutschland praktisch keine Rolle; diese Verfahren wurden auch in anderen Ländern selten verwendet. Ebenso wurden Arbeitsproben sowohl in Deutschland wie auch in anderen Ländern eher selten durchgeführt. Bei den übrigen Verfahren sind zum Teil erhebliche Unterschiede festzustellen: Intelligenztests, Simulationen (z. B. Assessment-Center-Übungen) und Persönlichkeitsfragebögen fanden in Deutschland seltener Verwendung als in anderen Ländern.

Eine Übersicht von Steiner (2012), die teilweise auch neuere Daten inkludiert, bestätigt diese Ergebnisse weitgehend. Allerdings findet sich hier kein Unterschied zwischen Deutschland und anderen Ländern in Bezug auf die Nutzung von Assessment-Centern; eine unterschiedliche Nutzung von Tests und Persönlichkeitsfragebögen besteht weiterhin. In einer weiteren Studie, in der Ryan et al. (2017) über 1000 Personen aus 23 verschiedenen Ländern befragten, konnte jedoch nicht bestätigt werden, dass es bedeutsame Einflüsse der Kultur auf Testpraktiken gibt.

Diagnostische Interviews, Intelligenztests und Persönlichkeitsfragebögen werden in ▶ Kap. 3 beschrieben. Hier soll vorwiegend auf wichtige Aspekte und Methoden in der Arbeits-, Organisations- und Wirtschaftspsychologie eingegangen werden, die nicht in ▶ Kap. 3 enthalten sind. Dabei wird die

Tab. 6.6 Vor- und Nachteile biografieorientierter Verfahren

Vorteile	Nachteile
Verfügbarkeit der Informationen (z. B. Zeugnisse)	Nicht leicht vergleichbare Leistungen über Bewerbende hinweg (z. B. die Abiturnote aus verschiedenen Bundesländern)
Hohe Akzeptanz (z. B. eines biografischen Interviews)	Keine ausgeprägte Biografien junger Bewerber, denen die notwendigen Informationen entnommen werden könnten
Geringe Verfälschbarkeit (z. B. von Zeugnissen)	
Aggregation von Informationen über einen längeren Zeitraum (z. B. Noten aus der gesamten Studienzeit)	

Systematik aus **Tab. 6.1** verwendet. Das heißt, es wird die Diagnostik von Personenmerkmalen im Zuge der Selektion von Personen (Personalauswahl), der Selektion von Bedingungen (Berufs- und Ausbildungswahl) sowie der Modifikation von Personen (Personalentwicklung) und der Modifikation von Bedingungen (Arbeitsgestaltungs- oder Organisationsentwicklung) dargestellt.

6.2.1 Selektion von Personen: Personalauswahl

Schuler und Höft (2007) haben mit dem sog. „trimodalen Ansatz“ vorgeschlagen, Inhalte der Personalauswahl auf 3 grundsätzlich verschiedene Arten zu operationalisieren. Sie unterscheiden biografie-, simulations- und konstruktorientierte Verfahren der Personalauswahl.

Werden Informationen herangezogen, die in der Vergangenheit von Bewerberinnen und Bewerbern entstanden sind, spricht man vom biografieorientierten Ansatz (**Tab. 6.6**). Dies können zu einem früheren Zeitpunkt angeeignete Fertigkeiten, berufliche oder schulische Abschlüsse, absolvierte Fortbildungen, aber auch konkretes Verhalten aus der Vergangenheit sein (z. B. ein Konflikt, den man gelöst hat). Diese Informationen können genutzt werden, um die zukünftige Eignung und zukünftiges Verhalten zu prognostizieren.

Beispiele für biografieorientierte Verfahren sind Interviews (vgl. ► Abschn. 3.7), Kenntnisprüfungen, Analyse der Bewerbungsunterlagen (Nachweis fachlicher Qualifikationen durch Zeugnisse, Bescheinigungen über bestandene Ausbildungen und Fortbildungen, Zertifikate, Arbeitszeugnisse etc.) und biografische Fragebögen. Dabei stellt die Analyse der Bewerbungsunterlagen nicht nur den am häufigsten verwendeten biografischen, sondern insgesamt den am weitesten verbreiteten diagnostischen Zugang dar. Eine Befragung deutscher Unternehmen zeigte, dass nahezu alle Bewerbungsunterlagen heranzogen (Schuler et al. 2007).

Biografieorientierter Ansatz

Sichtung von
Bewerbungsunterlagen als
diagnostische Methode

Bewerbungsunterlagen

Insbesondere wenn viele Bewerbungen vorliegen, werden Bewerbungsunterlagen meist für ein initiales Screening genutzt, um zu entscheiden, welche Bewerberinnen und Bewerber in die nächste Runde eines Auswahlverfahrens kommen. Während für diese Gruppe weitere diagnostische Informationen anfallen, die das Ergebnis der Bewerbungsunterlagenprüfung korrigieren könnten, ist die Entscheidung für abgelehnte Bewerberinnen und Bewerber endgültig. Somit stellt sich auch für den oftmals ersten Schritt der Personalauswahl die Frage nach deren Zusammenhang mit relevanten Kriterien. Diese ist allerdings nicht leicht, manchmal sogar gar nicht zu beantworten. Einerseits enthalten Bewerbungsunterlagen viele potenziell nützliche Informationen. So ist bekannt, dass biografische Daten – wenn sie anhand eines Fragebogens erhoben werden – einen substanziellen Zusammenhang mit Kriterien des Berufs- und Ausbildungserfolgs aufweisen (vgl. Schmidt und Hunter 1998). Es ist ebenfalls bekannt, dass aggregierte Schulabschlussnoten oder die Mathematiknote einen deutlichen Zusammenhang mit Ausbildungs- und Berufserfolgskriterien zeigen (z. B. Baron-Boldt et al. 1988; Trapmann et al. 2007) – wenngleich Abschlussnoten über Regionen hinweg deutlich variieren und damit nur bedingt vergleichbar sind (Müller-Benedict 2010).

Andererseits ist auch bekannt, dass beim Sichten von Bewerbungsunterlagen auch irrelevante, d. h. invalide Informationen zur Entscheidung herangezogen werden. So berichten Cole et al. (2009), dass Beurteilende auf Basis schriftlicher Bewerbungsunterlagen eine Einschätzung von Persönlichkeitsmerkmalen der Bewerbenden vornehmen und diese bei ihrer Auswahlentscheidung heranziehen. Allerdings konnten Cole et al. (2009) auch zeigen, dass die Einschätzung der Persönlichkeitsmerkmale nicht valide war, d. h. nicht mit Selbsteinschätzungen der Bewerberinnen und Bewerber korrelierte ($r = -.09$ bis .13). Möglicherweise nutzen Beurteilende nicht die richtigen Hinweise bzw. es kam zu einer Überinterpretation der verfügbaren Informationen. So berichten Cole et al. (2009), dass die Information aus den Bewerbungsunterlagen, eine Kapitänin oder ein Kapitän einer Sportmannschaft gewesen zu sein, keinerlei Korrelationen zu breiten Persönlichkeitsmerkmalen (Big Five) aufweist. Ähnliche Ergebnisse berichten auch Kanning et al. (2018), die nach einer Analyse von Bewerbungsunterlagen von 127 Bewerbenden zeigen können, dass Hinweise wie Tippfehler, Umfang des Anschreibens, persönliche Anrede u. a. in nahezu keinem Zusammenhang zu breiten Persönlichkeitsmerkmalen stehen. Nichtsdestotrotz gibt es vielleicht Personen, die im Rahmen der Personalauswahl ein Persönlichkeitsmerkmal mit der Rolle als Kapitänin oder Kapitän einer Sportmannschaft in Verbindung bringen. Kanning (2015) berichtet ebenfalls, dass Beurteilende häufig Informationen nutzen, deren Validität unklar ist (z. B. Lücken im Lebenslauf; Frank und Kanning 2014). Auch im Lebenslauf enthaltene Bilder üben einen Einfluss auf Beurteilende aus – grundsätzlich gilt, dass sich attraktive Bewerberinnen und Bewerber dadurch einen Vorteil verschaffen können (Marlowe et al. 1996). Allerdings gelingt auch bei Bewerbungsunterlagen in „reichhaltigerem“ Format (z. B. Videobewerbungen) keine akkurate Einschätzung der Persönlichkeit (Apers und Derous 2017).

Insgesamt dürfte daher die Analyse von Bewerbungsunterlagen nur eine geringe Prognosekraft für berufliche Leistungen haben. Bedauerlicherweise liegen

Konstruktorientierter Ansatz

unseres Wissens nach keine Studien vor, die genau dies prüfen, nämlich inwiefern Beurteilungen von Bewerbungsunterlagen mit späteren Leistungen in Beruf oder Ausbildung korrespondieren. Hierzu sind Studien erforderlich, die häufig nicht im Feld und unter Realbedingungen durchgeführt werden können. Werden beispielsweise alle Bewerbenden anhand eines Kriteriums ausgeschlossen, so können alle verbleibenden und später eingestellten Personen dieses Kriterium (z. B. den Abschluss eines Psychologiestudiums) vorweisen; es variiert damit nicht innerhalb dieser Personengruppe und kann – zumindest in diesem Setting – nicht auf seine Prognosekraft für berufliche Leistungen geprüft werden.

Simulationsorientierter Ansatz

Beim konstruktorientierten Ansatz stehen psychologische Eigenschaften im Fokus – daher wird er häufig auch als eigenschaftsorientierter Ansatz bezeichnet (► Tab. 6.7). Die zentrale Annahme ist, dass Verhalten über verschiedene Situationen hinweg von stabilen Eigenschaften determiniert wird. Beispielsweise sollten Personen mit einer hohen Ausprägung in der Eigenschaft „Verträglichkeit“ über verschiedene Situationen hinweg tendenziell verträglich handeln. Beispiele für konstruktorientierte Verfahren sind Intelligenztests, andere Leistungstests und Persönlichkeitsfragebögen (► Kap. 3).

Wenn diagnostische Verfahren ein möglichst realistisches Abbild zukünftiger beruflicher Gegebenheiten präsentieren (simulieren) und anhand des Verhaltens von Bewerberinnen und Bewerbern in dieser Simulation erheben, wie geeignet sie sind, dann zählt man diese zum simulationsorientierten Ansatz (► Tab. 6.8). Dabei wird davon ausgegangen, dass Bewerberinnen und Bewerber, die sich in einer oder mehreren simulierten beruflichen Situationen bewähren, dies auch später in der Realität tun werden (Bruk-Lee et al. 2013). Beispiele für simulationsorientierte Verfahren sind Assessment-Center (► Abschn. 6.2.1.1), Situational-Judgment-Tests (► Abschn. 6.2.1.2), Computersimulationen und Arbeitsproben.

In der Arbeits-, Organisations- und Wirtschaftspsychologie gut beforschte und auch in der Praxis verbreitete simulative Verfahrenstypen sind Assessment-Center und Situational-Judgment-Tests. Diese werden nachfolgend näher vorgestellt.

► Tab. 6.7 Vor- und Nachteile konstruktorientierter Verfahren

Vorteile	Nachteile
Existenz etablierter Messmethoden für viele Eigenschaften	Eventuell keine Messinstrumente für sehr spezifische, von Organisationen identifizierte Anforderungen
Erfassung von Eigenschaften, die auch dann relevant sind, wenn sich die berufliche Situation ändert	Bezug zur organisationalen Realität für Bewerber nicht leicht ersichtlich (d. h., es liegt eine geringe Augenscheinvalidität vor)
Ökonomisch	Leichte Verfälschbarkeit eines Teils der konstruktorientierten Verfahren, nämlich der (meisten) Persönlichkeitsfragebögen

Tab. 6.8 Vor- und Nachteile simulationsorientierter Verfahren

Vorteile	Nachteile
Hohe Augenscheininvalidität (d. h., der Bezug zur organisationalen Realität ist für Bewerbende leicht ersichtlich)	Geringe Übertragbarkeit der Erkenntnisse einer simulierten Situation auf andere bzw. sich ändernde Gegebenheiten im Beruf
Hohe Akzeptanz	Teilweise unklare Zusammenhänge zu anderen Konstrukten (► Abschn. 6.2.1.1 und ► Abschn. 6.2.1.2)
Möglichkeit, komplexe berufliche Gegebenheiten abzubilden	Teilweise unzureichende Reliabilität (► Abschn. 6.2.1.2)

6.2.1.1 Assessment-Center-Methodik

Definition

„Ein **Assessment-Center** (AC) ist ein eignungsdiagnostisches Verfahren zur Potenzial- und Eignungsbeurteilung im Rahmen von Personalauswahl- oder Entwicklungsfragestellungen, bei dem mehrere Methoden kombiniert und die Teilnehmer von mehreren Assessoren beobachtet sowie bewertet werden.[...] Herauszustellen ist die Nutzung von Simulationen als spezifisches Charakteristikum des AC-Ansatzes (Simulationsprinzip), die einen unmittelbaren Zugang zu komplexen berufsbezogenen Verhaltenskompetenzen ermöglichen.“ (Arbeitskreis Assessment Center 2016, S. 3).

Kern der Assessment-Center-Methodik ist dabei die Simulation des Arbeitsplatzes, auf den sich Teilnehmerinnen und Teilnehmer beworben haben (► Abb. 6.5). Diese Simulation erfolgt durch Rollenspiele, Fallstudien, Präsentationsaufgaben und Gruppendiskussionen, in denen diese mit Rollenspielern interagieren, miteinander diskutieren oder Inhalte präsentieren. Das heißt, berufsbezogene Situationen werden „live“ vor Beobachtenden durchgeführt; „echtes“ Verhalten wird bewertet. Daher werden Assessment-Center als sog. „High-Fidelity-Simulationen“, also sehr realitätsgerechte Simulationen, bezeichnet. Dabei ist wichtig, dass es sich um eine grundsätzliche Methodik handelt – d. h., die konkrete Ausgestaltung der Übungen muss an dem zu simulierenden Arbeitsplatz und den damit verbundenen Anforderungen erfolgen. Daher gibt es nicht *das* Assessment-Center, vielmehr wird in der Regel für jede zu besetzende Position und für jede Organisation ein eigenes Assessment-Center entwickelt.

Simulation des Arbeitsplatzes

Neben verhaltensbezogenen Übungen können ergänzend Testverfahren und Interviews zum Einsatz kommen. Allerdings ist zu beachten, dass in der Forschung unter dem Assessment-Center meist nur eine Methode verstanden wird, die auf standardisierten Übungen basiert und sich der Verhaltensbeobachtung bedient. Ergebnisse, etwa zu Tests, sollten daher separat berichtet werden. In der Praxis trifft man oft auf die „breite“ Auslegung des Begriffs, der zufolge die klassischen Übungen mit beliebigen anderen eignungsdiagnostischen Verfahren kombiniert werden.

Testverfahren und Interviews als Ergänzung

Nachfolgend werden wichtige Gestaltungsprinzipien von Assessment-Centern erläutert. Dabei gehen wir zunächst auf 5 grundlegende Prinzipien nach Obermann (2009) und dann auf die Standards des Arbeitskreises Assessment Center ein.



Abb. 6.5 Simuliertes Teammeeting als Assessment-Center-Übung. (© fizkes/stock.adobe.com)

5 grundlegende Prinzipien für ein Assessment-Center (Obermann 2009, S. 10)

- Anforderungsbezug
- Simulation
- Methodenvielfalt
- Einsatz mehrerer Beobachterinnen und Beobachter
- Transparenz

Anforderungsdimensionen

Anforderungsbezug Mit dem Assessment-Center sollen die Merkmale oder Verhaltensweisen erfasst werden, die für die zu besetzende Stelle bzw. für mögliche künftige Aufgaben relevant sind. Übungen und Fallstudien werden ausgewählt oder neu entwickelt, sodass sie sich genau auf die Anforderungsmerkmale beziehen.

In der Fachliteratur wird über sehr viele Anforderungsdimensionen berichtet, die in Assessment-Center-Übungen abgebildet werden. Arthur et al. (2003) fanden insgesamt 168 Bezeichnungen, die von Expertinnen und Experten zu 7 globalen Dimensionen zusammengefasst werden konnten (die Obergriffe wurden bewusst nicht wörtlich übersetzt, sondern anhand der Erläuterungen der Autorinnen und Autoren teilweise umbenannt; in Klammern sind die Bezeichnungen der Autorinnen und Autoren genannt).

7 globale Dimensionen von Assessment-Centern nach Arthur et al. (2003)

- Kommunikationsfähigkeit (communication)
- Bewusstsein für die Bedürfnisse und Gefühle anderer und Rücksichtnahme (consideration/awareness)
- Aktivität und Motivation („drive“)
- Führungskompetenz und Durchsetzungsfähigkeit (influencing others)
- Organisations- und Planungsfähigkeit (organizing and planning)
- Problemlösefähigkeit (problem solving)
- Belastbarkeit (tolerance for stress/uncertainty)

Es kann für jeden Teilnehmenden eine Bewertung jeder einzelnen Dimension, gemessen in verschiedenen Assessment-Center-Übungen, vorgenommen werden (dimensionsbezogene Bewertung) und/oder eine globale Bewertung der Leistung im gesamten Assessment-Center (overall performance rating). Allerdings ist eine dimensionsbezogene Bewertung aufgrund mangelnder diesbezüglicher Validitätsbelege nicht zu empfehlen (s. nachfolgender Absatz zu Zusammenhängen mit anderen Konstrukten).

Empfehlenswert: globale Leistungsbewertung

Simulation Die Übungen und Fallstudien werden so konzipiert, dass sie dem später erwarteten Arbeitsverhalten möglichst ähnlich sind. Dies kann jedoch auch bedeuten, dass Aufgaben nicht exakt die Realität abbilden, sondern lediglich die gleichen Anforderungen an die Teilnehmenden stellen, wie sie für die Stelle erforderlich sind. In der konkreten Ausgestaltung können Unterschiede zwischen einer Assessment-Center-Übung und der Tätigkeit im Beruf bestehen. Beispielsweise könnten Bewerber, die später Autos verkaufen sollen, in einer Assessment-Center-Übung Zeitschriften „verkaufen“.

Identische Anforderungen an die Teilnehmenden stellen

Methodenvielfalt In einem Assessment-Center wird jedes Anforderungsmerkmal in verschiedenen Übungen erfasst. Durch die Aggregation der Beurteilungen über mehrere Messgelegenheiten gleichen sich Vor- und Nachteile einzelner Übungen für gewisse Teilnehmende aus. Zusätzlich erhöht sich die Reliabilität durch die Aggregation.

Aggregation der Beurteilungen

Einsatz mehrerer Beobachtender Alle Teilnehmenden werden von mehreren zuvor geschulten Personen beobachtet und beurteilt. Beobachtungs- und Urteilsfehler (► Abschn. 3.6.4) einer einzelnen Person sollen damit kompensiert werden. Beobachtende können aus der Fach- und der Personalabteilung des Unternehmens stammen, es können aber auch externe Personen mitwirken – meist kommen diese aus einem anderen, beratenden Unternehmen, das mit der Konzeption und Durchführung des Assessment-Centers betraut wurde. Am Ende eines Assessment-Centers findet eine „Beobachtendenkonferenz“ statt. Darin können sich die Beobachtenden austauschen und zu einer gemeinsamen Beurteilung der Teilnehmenden gelangen.

Beobachtendenkonferenz

Transparenz Die Teilnehmenden werden vor einem Assessment-Center über die Übungen und die dabei bewerteten Anforderungsdimensionen informiert. Dies gebietet das Prinzip der informierten Einwilligung. Nach dem Assessment-Center werden Teilnehmende über das Ergebnis informiert. Insgesamt trägt eine große Transparenz auch dazu bei, dass das Verfahren bei Bewerbenden akzeptiert wird. Allerdings konnte gezeigt werden, dass Assessment-Center einen höheren Zusammenhang mit beruflichen Kriterien aufweisen, wenn die Anforderungsdimensionen nicht transparent sind (Ingold et al. 2016; Kleinmann 2013). Dies wird dadurch erklärt, dass die Fähigkeit von Teilnehmenden, intransparente Anforderungsdimensionen richtig zu erkennen, nicht nur positiv zur Leistung in Assessment-Centern beiträgt, sondern auch in vielen beruflichen Tätigkeiten relevant ist (König et al. 2007).

Hohe Akzeptanz durch Transparenz

Die Standards der Assessment-Center-Technik, die vom Arbeitskreis Assessment Center herausgegeben werden, gehen über die 5 Prinzipien von Obermann (2009) hinaus. In diesen Standards werden Leitlinien zu 9 Bereichen genannt.

Standards der Assessment-Center-Technik gemäß Arbeitskreis Assessment Center (2016, S. 6 ff.)

1. Auftragsklärung und Einbindung: Vor der Entwicklung und Durchführung eines Assessment-Centers sind die Ziele und die Rahmenbedingungen des Auftrages sowie die Konsequenzen für die Teilnehmenden und andere Stakeholder verbindlich zu klären und zu kommunizieren.
2. Festlegung eines Anforderungsprofils: Eine gültige Eignungsbeurteilung lässt sich nur mit einer exakten Analyse der konkreten Anforderungen sinnvoll gestalten.
3. Verfahrensauswahl und -entwicklung: Ein Assessment-Center ist eine für die jeweilige eignungsdiagnostische Fragestellung optimierte Kombination aus mindestens drei Verfahrenselementen, darunter mindestens eine Verhaltenssimulation sowie wenigstens ein Verfahrenselement, das auf einem anderen Methodenansatz beruht, z. B. ein Testverfahren.
4. Auswahl und Vorbereitung der Durchführungsmitglieder: Nur gut vorbereitete Prozessbeteiligte, die die Organisation angemessen repräsentieren, gewährleisten fundierte und treffsichere Eignungsbeurteilungen.
5. Vorauswahl und Vorbereitung der Teilnehmenden: Systematische Vorauswahl und realistische Vorinformation sind die Grundlage für eine hohe Trefferquote und Akzeptanz.
6. Vorbereitung und Durchführung des Verfahrens: Eine gute Vorbereitung und Moderation des Verfahrens gewährleisten einen transparenten und zielführenden Ablauf.
7. Datengewinnung und Bewertung: Datengewinnung und Bewertung im Rahmen von Verhaltensbeobachtungen, Interviews sowie Test- bzw. Fragebogenverfahren sind zentrale Schritte zur Ableitung belastbarer Eignungsaussagen.
8. Datenintegration und Ergebnisermittlung: Die Integration der Ergebnisse der Verfahrenselemente und Beobachterurteile erfolgt im Rahmen eines vorab definierten regelgeleiteten Prozesses.
9. Feedback und Folgemaßnahmen: Alle Assessment-Center-Teilnehmenden haben das Recht auf individuelles Feedback, um so das Ergebnis nachvollziehen und daraus lernen zu können. Nach dem Assessment-Center sind konkrete Folgemaßnahmen festzulegen, deren Umsetzung regelmäßig überprüft wird.
10. Evaluation: Eine regelmäßige Güteprüfung und Qualitätskontrolle stellen den Nutzen des Verfahrens sowie das nachhaltige Erreichen der ange strebten Ziele sicher.

(© Forum Assessment e. V.)

■ Reliabilität

Trainings zur Verbesserung der Beobachterübereinstimmung

Die meisten Befunde liegen zur Übereinstimmung zwischen den Beurteilenden vor. In den meist älteren Studien wurde eine große Bandbreite an Übereinstimmungskoeffizienten gefunden. Obermann (2009, S. 278) weist auf eine ältere Übersichtsarbeit von Howard (1974) hin, der zufolge die Spanne von .60 bis .98 reicht – bei einem Median von .75 (Howard gibt interne Konsistenzen als Maß der Übereinstimmung an, korrigiert für die Anzahl der Beurteilenden). Diese Übereinstimmungskoeffizienten scheinen nach wie vor zutreffend zu sein – auch neuere Studien berichten ähnliche Spannen (Vanhove et al. 2016). Durch Trainings kann die Übereinstimmung der Beobachtenden verbessert werden (Lievens 2001). Zur Retest-Reliabilität von Assessment-Centern liegen nur wenige Studien vor. Kelbetz und Schuler (2002) analysierten Daten von 47 Personen, die ein Assessment-Center nach (im Mittel) 2 Jahren wiederholten und fanden eine Retest-Reliabilität

der Gesamtbewertung im Assessment-Center (dort als AC-Durchschnitt bezeichnet) von $r_{tt} = .41$.

■ Zusammenhänge mit beruflichen Kriterien

Die Angaben zur Korrelation von Assessment-Centern mit beruflichen Kriterien fallen sehr unterschiedlich aus. In der sog. „AT&T-Studie“ von Bray et al. (1974) lag der Zusammenhang des Assessment-Center-Ergebnisses mit einer Beförderung innerhalb der ersten 8 Jahre bei $r = .46$. Gaugler et al. (1987) legten die erste große Metaanalyse vor. Die Autorinnen und Autoren ermittelten eine für Varianzeinschränkung und Kriteriumsreliabilität korrigierte mittlere Korrelation von .37 zwischen Assessment-Center-Ergebnissen und Berufserfolgsmaßen. Hardison und Sackett (2007) analysierten nur Studien, die in den darauffolgenden 20 Jahren erschienen waren und fanden nur noch eine korrigierte mittlere Korrelation von .26. Die Höhe des Zusammenhangs korrelierte mit dem Alter der Studie ($r = -.25$); je neuer die Studien waren, desto niedriger fielen die Zusammenhänge mit beruflichen Leistungsmaßen aus. In einer weiteren Metaanalyse betrachteten Hermelin et al. (2007) den Zusammenhang zwischen Assessment-Center-Ergebnissen und der Vorgesetztenbeurteilung, der bei .28 lag (ebenfalls korrigiert für Varianzeinschränkung und Unreliabilität des Kriteriums). Becker et al. (2011) werteten in ihrer Metaanalyse ausschließlich im deutschsprachigen Raum durchgeführte Validierungsstudien aus. Über 19 Studien hinweg ermittelten sie einen mittleren Zusammenhang (korrigiert für Unreliabilität des Kriteriums und Streuungseinschränkung) von .40. Dabei wurden allerdings auch Studien aufgenommen, in denen ein Intelligenztest Bestandteil des Assessment-Centers war. In diesen Studien war der Zusammenhang zwischen Assessment-Center-Ergebnissen und beruflichen Leistungsmaßen höher als bei den anderen, nämlich .56. Der Mittelwert der korrigierten Korrelation über alle Studien für Assessment-Center ohne Intelligenztest betrug hingegen .25. Nicht in dieser Metaanalyse enthalten ist eine Untersuchung an 451 schweizer Offizierinnen und Offizieren, für die jährliche Beurteilungen durch ihre Vorgesetzten vorlagen (Melchers und Annen 2010). Die Assessment-Center-Beurteilung und spätere Beurteilung durch die Vorgesetzten korrelierten (unkorrigiert) zu $r = .32$ (.31 bei Studienleistungen als Kriterium). Zusätzlich kam ein Test zur allgemeinen Intelligenz zum Einsatz. Wurde die Intelligenz in einer multiplen Regression als erster Prädiktor eingesetzt und damit auspartialisiert, konnte für das Assessment-Center immerhin noch eine zusätzliche Varianzaufklärung von 15 % (Vorgesetztenbeurteilung, $N = 311$) bzw. 7 % (Studienleistungen, $N = 246$) erreicht werden. Dies spricht dafür, dass ein Assessment-Center über die Intelligenz hinaus eine bedeutende Vorhersagekraft haben kann. Arthur et al. (2003) fanden in einer Metaanalyse für die Gesamtbeurteilung eine (korrigierte) Korrelation von .36, was dem Ergebnis von Gaugler et al. (1987) entspricht. Allerdings fielen die Koeffizienten für die einzelnen Beurteilungsdimensionen sehr unterschiedlich aus; sie reichten von .25 (Bewusstsein für die Bedürfnisse und Gefühle anderer und Rücksichtnahme) bis .39 (Problemlösefähigkeit). Es liegt nahe, dass bei einer optimalen Gewichtung der einzelnen Dimensionen höhere Zusammenhänge zu beruflichen Kriterien als mit einem einfachen Gesamtmaß erreicht werden. In einer schrittweisen multiplen Regression schätzten Arthur et al. (2003) eine multiple Korrelation von $R = .45$. Als Prädiktoren wurden dabei nur Problemlösefähigkeit, Führungskompetenz und Durchsetzungsfähigkeit, Organisations- und Planungsfähigkeit sowie Kommunikationsfähigkeit mit signifikanten Beiträgen zur Varianzaufklärung aufgenommen.

Meriac et al. (2008) befassten sich mit dem inkrementellen Beitrag von Assessment-Centern über andere eignungsdiagnostische Kriterien hinaus. Sie analysierten 38 Studien, die zwischen 1969 und 2006 erschienen waren

Umfangreiche Validitätsbelege

Inkrementelle Varianzaufklärung

und Angaben zu den 7 Assessment-Center-Dimensionen nach Arthur et al. (2003) enthielten (s. o.). Die einzelnen Dimensionen korrelierten (korrigiert) zwischen .16 (Aktivität und Motivation) und .35 (Organisations- und Planungsfähigkeit) mit Berufserfolgsmaßen. In einer hierarchischen Regression setzten sie die 7 Dimensionen zusammen nach der Intelligenz (Schritt 1) und den Big-Five-Persönlichkeitsmerkmalen (Schritt 2) als Prädiktoren für Berufserfolg ein. Die zusätzliche Varianzaufklärung betrug 9,7 %. Mit einzelnen Dimensionen konnten sie im Vergleich dazu zwischen 3 und 13 % zusätzlicher Kriteriumsvarianz aufklären. Einige Ergebnisse dieser Metaanalysen sind in □ Tab. 6.9 zusammenfassend dargestellt.

Fazit Als Fazit kann man festhalten, dass Assessment-Center zwar nicht in der gleichen Höhe wie strukturierte Interviews oder Intelligenztests mit beruflichen Kriterien korrelieren, aber im Durchschnitt dennoch (korrigierte) Zusammenhänge zu beruflichen Kriterien in einer Größenordnung von etwa .30 aufweisen. Allerdings ist der von Hardison und Sackett (2007) beobachtete Abfall dieser Zusammenhangsmaße besorgniserregend. Ebenso fällt die große Streubreite der Koeffizienten in den einzelnen Metanalysen auf (bei Gaugler et al. 1987, immerhin -.25 bis .78).

■ Zusammenhänge mit anderen Konstrukten

Meriac et al. (2008) haben in einer Metaanalyse zahlreiche Korrelationen zwischen typischen Assessment-Center-Dimensionen einerseits und Persönlichkeitsmerkmalen sowie Intelligenz andererseits zusammengetragen (□ Tab. 6.10). Dabei wird deutlich, dass diese Assessment-Center-Dimensionen nur gering bis moderat mit Persönlichkeitsmerkmalen und Intelligenz in Zusammenhang stehen – die maximale Korrelation beträgt .29. Dies ist einerseits erfreulich: Mit dem Assessment-Center steht folglich ein von diesen Merkmalen weitgehend unabhängiger, ergänzender Prädiktor für berufliche Erfolgsmaße zur Verfügung. Die negative Erkenntnis ist,

Uneindeutige Zusammenhänge zu anderen Methoden

□ **Tab. 6.9** Ergebnisse von Metaanalysen zur Kriteriumsvalidität von Assessment-Centern

Autorinnen und Autoren	r	ρ	Anmerkungen
Gaugler et al. (1987)	0.29	0.37	–
Arthur et al. (2003)	.28 (.20–.30)	.36 (.25–.39)	In Klammern Angaben zu einzelnen Assessment-Center-Dimensionen
Hardison und Sackett (2007)	0.22	0.26	Nur Studien aus den letzten 20 Jahren
Meriac et al. (2008)	.13–.28	.16–.35	Nur Angaben zu einzelnen Assessment-Center-Dimensionen (z. B. Organisieren und Planen)
Becker et al. (2011)	0.21	0.25	Nur Studien aus dem deutschen Sprachraum; hier nur Assessment-Center ohne Intelligenztest

r = mittlere Korrelation; ρ = korrigierte Korrelation (weitere Erläuterungen im Text)

Tab. 6.10 Korrelation der Beurteilungsdimensionen in Assessment-Centern mit Persönlichkeit und Intelligenz

Assessment-Center Dimension	Persönlichkeitsmerkmal					Intelligenz
	Neurotizismus	Extraversion	Offenheit	Verträglichkeit	Gewissenhaftigkeit	
Kommunikationsfähigkeit	-.08	.11	.12	.09	.09	.28
Bewusstsein und Rücksichtnahme	-.07	.07	.06	.05	.09	.17
Aktivität und Motivation	-.04	.21	.06	.09	.10	.21
Führungskompetenz und Durchsetzungsfähigkeit	.01	.15	.08	.08	.09	.22
Organisations- und Planungsfähigkeit	-.07	.09	.09	.02	.05	.29
Problemlösefähigkeit	-.07	.08	.11	.06	.13	.28
Belastbarkeit	-.07	.12	.11	.06	.12	.22

Quelle: Aus Meriac et al. (2008, Tab. 1, mit freundlicher Genehmigung der American Psychological Association). Unkorrigierte Korrelationen; N pro Zelle zwischen 310 (Gewissenhaftigkeit und Belastbarkeit) und 12.599 (Intelligenz und Problemlösefähigkeit)

dass damit unklar bleibt, was Assessment-Centern eigentlich messen. Beispielsweise korrelierte Belastbarkeit (gemessen im Assessment-Center) nicht mit Neurotizismus (gemessen mit einem Fragebogen), obwohl theoretisch eine moderat negative Korrelation zu erwarten wäre. Ebenso konnten für Rücksichtnahme (gemessen im Assessment-Center) keine Zusammenhänge zu Verträglichkeit (gemessen mit einem Fragebogen) gezeigt werden – auch hier wären diese theoretisch zu erwarten gewesen. Eine neuere Metaanalyse von Hoffman et al. (2015) bestätigt die Ergebnisse der früheren: So sind die Zusammenhänge zwischen ähnlichen Konstrukten, gemessen in einem Assessment-Center und mit einem Fragebogen, bestenfalls moderat.

Lievens (2017) führt 4 Gründe für die geringe Konvergenz zwischen Assessment-Center-Bewertungen und Fragebogenergebnissen an.

Gründe für die geringe Konvergenz zwischen Assessment-Center-Bewertungen und Fragebogenergebnissen (Lievens, 2017, S. 432; Übersetzung durch die Autoren dieses Buchs)

1. In einem Assessment-Center werden keine Persönlichkeitsdimensionen bewertet, sondern die Angemessenheit von Verhalten in spezifischen Situationen.
2. In einem Assessment-Center wird maximales Verhalten erfasst (d. h., Teilnehmende zeigen ihre bestmögliche Leistung), Persönlichkeitsfragebögen messen typisches Verhalten.
3. Ausprägungen der Persönlichkeit von Assessment-Center-Kandidatinnen und -Kandidaten könnten einer größeren Streuungseinschränkung unterliegen, da Teilnehmende ggf. bereits bezüglich der für einen Beruf relevanten Merkmale (selbst-)selegiert sind.
4. Manche Merkmale manifestieren sich weniger gut in beobachtbarem Verhalten und können daher in Assessment-Centern schlechter gemessen werden.

(Abdruck mit freundlicher Genehmigung von John Wiley and Sons)

6

Annahme: Übungsunspezifisches Teilnehmendenverhalten

Neben der geringen Konvergenz von gleichen oder ähnlichen Merkmalen, die in Assessment-Centern oder mit Persönlichkeitsfragebögen gemessen wurden, wurde ein weiteres Phänomen lange als Problem von Assessment-Centern diskutiert (vgl. Lance 2008). Ursprünglich wurde davon ausgegangen, dass Bewertungen der gleichen Assessment-Center-Dimension über mehrere Übungen hinweg deutlich miteinander korrelieren sollten. Eine Person, der etwa in einer Präsentation eine sehr gute Kommunikationsfähigkeit bescheinigt wird, sollte in einer Gruppendiskussion auch eine hohe Kommunikationsfähigkeit zeigen. Hingegen sollten Bewertungen unterschiedlicher Dimensionen innerhalb einer Übung (z. B. Belastbarkeit, Kommunikationsfähigkeit und Rücksichtnahme in einem Rollenspiel) eher gering ausfallen, da diese Dimensionen theoretisch weitgehend unabhängig voneinander sein sollten. Zugrunde gelegt wird dabei die Annahme, dass das Verhalten von Teilnehmenden in einem Assessment-Center weitgehend übungsunspezifisch ist, d. h., sich über verschiedene Übungen hinweg konsistent zeigt.

Eine Multitrait-Multimethod-Analyse (► Abschn. 2.6.3.4) sollte Aufschluss geben, ob dies der Fall ist. Bowler und Woehr (2006) haben dazu eine Metaanalyse durchgeführt. Wurde das gleiche Merkmal (Monotrait) in unterschiedlichen Situationen/Übungen (Heteromethod) gemessen, betrug die durchschnittliche Korrelation .25. Die unterschiedlichen Dimensionen (Heterotrait) korrelierten im Vergleich dazu innerhalb einer Übung (Monomethod) mit durchschnittlich .53. Melchers et al. (2007) kamen anhand anderer metaanalytischer Daten zu der gleichen Erkenntnis: Die Monotrait-Heteromethod-Korrelationen lagen im Durchschnitt bei .33, während die Heterotrait-Monomethod-Korrelationen durchschnittlich .62 betrugen – was nicht dem eigentlich zu erwartenden Muster der Zusammenhänge entspricht.

Eine naheliegende Erklärung für diese Befunde ist, dass die Beurteilenden sehr schlecht zwischen den Merkmalen differenzieren, die sie beurteilen sollen. Sind sie vielleicht Opfer eines Halo-Effekts, lassen sich also von ihrem positiven oder negativen Gesamteindruck der Person zu sehr beeinflussen? Aufbauend auf dieser Vermutung versuchten verschiedene Forschende herauszufinden, wie Assessment-Center gestaltet werden sollten, um die Beurteilenden zu entlasten und damit differenziertere Urteile zu ermöglichen. Das heißt, man hat sich auf die Suche nach Moderatoren der Konstruktvalidität gemacht. Einige Ergebnisse dessen sind bei Lievens (1998) zusammengefasst, der die bis dato verfügbare Literatur gesichtet und daraus praktische Empfehlungen abgeleitet hat.

Ergebnisse von Multitrait-Multimethod-Analysen

Schlechte Differenzierung zwischen Merkmalen

Zusammenfassung der forschungsbasierten Empfehlungen zur Verbesserung der Konstruktvalidität von Assessment-Centern nach Lievens (1998, S. 147; Übersetzung durch die Autoren dieses Buchs)

- Dimensionen
 - Nutzung weniger Dimensionen, besonders bei Assessment-Centern, die auf Einstellungen abzielen.
 - Auswahl von Dimensionen, die konzeptuell distinkt sind (d. h. solche, die relativ unabhängig voneinander sind).
 - Verwendung konkreter und jobbezogener Definition der Dimensionen.
- Durchführende
 - Psychologinnen und Psychologen sollten eine Schlüsselfunktion im Beobachtungsteam innehalten (z. B. als Coach der beurteilenden Vorgesetzten).
 - Fokus auf die Qualität des Trainings der Beobachtenden (statt auf die Länge der Trainings).
 - Einbinden eines Bezugsrahmentrainings in das Programm (neben weiteren Trainingsansätzen). Dafür sollten die Beobachtenden mit den Dimensionen und Leistungsstufen vertraut gemacht sowie eine konsistente Kategorisierung implementiert werden.
- Situative Aufgaben
 - Entwicklung von Assessment-Center-Aufgaben, die bezüglich der Dimensionen möglichst klar abgegrenzt sind. Folglich sollten die Aufgaben so ausgewählt werden, dass sie möglichst viel dimensionsrelevantes Verhalten bei den Teilnehmenden hervorrufen. „Schwammige“ Aufgaben, die für mehrere Dimensionen relevante Verhaltensweisen hervorrufen sollten vermieden werden.
 - Training und Standardisierung von Rollenspielerinnen und -spielern, um die Varianz bei der Durchführung möglichst gering zu halten.
 - Einsatz von Rollenspielerinnen und -spielern, die aktiv anstreben, dimensionsbezogenes Verhalten bei den Teilnehmenden hervorzurufen.
 - Offenlegung der Dimensionen (und dazugehörigen Verhaltensweisen) gegenüber den Teilnehmenden, besonders wenn das Assessment-Center auf Entwicklung abzielt.
- Systematische Beobachtungen, Evaluation und Integrationsverfahren
 - Bereitstellen von Beobachtungshilfen für die Durchführenden (z. B. Verhaltens-Checklisten, die die dimensionsbezogenen Verhaltensweisen für jede Aufgabe aufführen).
 - Operationalisierung jeder Dimension in den Checklisten mit minimal 2 und maximal 12 Verhaltensweisen; somit sollte sich auf die Schlüsselverhaltensweisen konzentriert werden.
 - Gruppierung der Verhaltensweisen in den Checklisten zu natürlich auftretenden Clustern.
 - Nutzung eines Rotationssystems, das Ratingverzerrungen minimiert, z. B. das von Andres und Kleinmann (1993).
 - Nutzung von Videotechnologien und konsensuellen Diskussionsformaten, die die Konstruktvalidität von Assessment-Centern wenig zu beeinflussen scheinen.

(Abdruck mit freundlicher Genehmigung von John Wiley and Sons).

Übungsspezifisches Verhalten statt Beurteilungsfehler?

Lance (2008) hinterfragte in einem viel beachteten theoretischen Beitrag die der Forschung zur Konstruktvalidität von Assessment-Centern zugrunde liegende Annahme, dass Verhalten von Teilnehmenden über Übungen hinweg konsistent sein sollte und Beurteilende dies (nur) nicht akkurat genug beurteilen können. Er kommt zu dem Schluss, dass diese Annahme nicht zutreffend ist. Als Beleg führt er beispielsweise Forschung an, die zeigt, dass Beurteilende sehr wohl akkurate Urteile fällen können: Lievens (2002) führte eine experimentelle Studie durch, in der Beurteilende Videos von an Assessment-Centern Teilnehmenden sahen. Deren Verhalten variierte hinsichtlich zweier Aspekte (2×2 -Design) – es war entweder über die Übungen hinweg konsistent oder inkonsistent (1. Aspekt) sowie innerhalb der Übungen hinsichtlich der zu beurteilenden Dimensionen differenziert konsistent oder inkonsistent (2. Aspekt). Seine Ergebnisse zeigen, dass (trainierte) Beurteilende sowohl innerhalb von Übungen zwischen Dimensionen differenzieren als auch konsistentes Verhalten über Übungen hinweg akkurat beurteilen können. Lance (2008) führt zudem an, dass auch globale Übungsbeurteilungen signifikante Zusammenhänge zu externen Leistungsindikatoren aufweisen und kommt insgesamt zu dem Schluss, dass nicht von konsistenten Verhaltensweisen über Assessment-Center-Übungen ausgegangen werden kann. Vielmehr argumentiert er, dass Teilnehmende übungsspezifisches Verhalten zeigen, was mit ihrem Verhalten in ähnlichen beruflichen Situationen korrespondiert.

Unähnlichkeit abgebildeter Situationen

Dass insbesondere das über mehrere unterschiedliche Assessment-Center-Übungen beobachtete Verhalten spätere Leistungen im Beruf vorhersagen kann, zeigten Speer et al. (2014). Sie analysierten Assessment-Center, die von einer Unternehmensberatung im US-amerikanischen Raum durchgeführt wurden. Sie ordneten Assessment-Center-Übungen zu Paaren und sortierten diese Paare nach „Unähnlichkeit“ der darin abgebildeten Situationen. Dieser Einteilung stellten sie die mittlere Konstruktvalidität, also die Korrelation der gleichen Dimensionen in den beiden Übungen, und die Kriteriumsvalidität der Übungen gegenüber. Es zeigte sich, dass „Unähnlichkeit“ der Übungen negativ mit der Konstruktvalidität korreliert war: Je weniger sich die Übungen ähnelten, desto geringer war die Konvergenz der gleichen Dimensionen. Wenn also beispielsweise Planungsfähigkeit in 2 Übungen bewertet wurde, dann war die Korrelation zwischen Planungsfähigkeit in der einen Übung und Planungsfähigkeit in der anderen Übung höher, wenn sich die in diesen Übungen abgebildeten Situationen ähnelten.

Bei der Kriteriumsvalidität verhielt es sich genau andersherum: Je unähnlicher die Übungen waren, desto höher fiel deren gemeinsamer (gemittelter) Zusammenhang mit berufsbezogenen Kriterien (in diesem Fall Vorgesetztenbeurteilungen) aus. Der Zusammenhang zwischen Unähnlichkeit der Übungen und der Kriteriumsvalidität lag in 2 der 3 untersuchten Stichproben bei $r \geq .55$. Speer et al. (2014) argumentieren, dass ähnliche Übungen eine geringere situative Spezifität mit sich brächten und es daher wahrscheinlicher sei, dass Teilnehmende in diesen Übungen ein ähnliches Verhalten zeigen. Andererseits führt eine große „Unähnlichkeit“ der Übungen eher dazu, dass die Breite des Kriteriums (effektives Agieren in verschiedenen beruflichen Kontexten) besser abgebildet wird.

Fazit Insgesamt muss also der Anspruch, mit Assessment-Centern einzelne Verhaltensdimensionen konsistent über Übungen hinweg messen zu können, hinterfragt werden. Vor dem Hintergrund der verfügbaren Evidenz muss von übungs- bzw. situationsspezifischen Verhaltensweisen ausgegangen werden, deren globale Beurteilung zur Prädiktion beruflicher Leistungen taugt (sofern die im Assessment-Center enthaltenen Übungen ein hinreichend gutes Abbild der beruflichen Realität darstellen).

6.2.1.2 Situational-Judgment-Tests

Mit Situational-Judgment-Tests versucht man – wie mit Assessment-Centern –, wesentliche Aspekte einer beruflichen Tätigkeit zu simulieren. Allerdings erfolgt die Simulation bei Situational-Judgment-Tests weniger realitätsgetreu, weshalb man auch von Low-Fidelity-Simulationen spricht. Berufsbezogene relevante Situationen werden nicht real durchgeführt, sondern lediglich in Form eines kurzen Textes oder eines Videos präsentiert. Teilnehmende geben nach Sichtung des Textes oder Videos an, wie sie sich in dieser Situation verhalten würden (oder alternativ: wie sie sich idealerweise verhalten sollten). Die Reaktion auf die geschilderte oder als Video gezeigte berufliche Situation erfolgt in der Regel in Form eines geschlossenen Antwortformats – Teilnehmende wählen aus vorgegebenen Antwortalternativen eine aus oder bewerten mehrere Alternativen hinsichtlich ihrer Angemessenheit.

Low-Fidelity-Simulationen

Beispielitem des Team-Role-Tests (Mumford et al. 2008; Übersetzung durch die Autoren dieses Buchs)

Situationsschilderung (Itemstamm): „Sie sind Mitglied eines Vertriebsteams in einem lokalen Buchladen dessen Verkaufszahlen in letzter Zeit aufgrund von Kundenschwund substanziell gesunken sind. In einer Teambesprechung diskutieren Sie Lösungen für den rückläufigen Umsatz. Die Diskussion wird ein wenig hitzig, als das älteste Teammitglied suggeriert, dass die Verkaufszahlen der neuen Mitarbeiterinnen und Mitarbeiter recht niedrig sind. Eines der jüngeren Teammitglieder entgegnet schnell, dass immer, wenn er um Hilfe mit Kundenschaft fragt, der ältere Mitarbeiter den Verkauf für sich beansprucht. Die andere neue Mitarbeiterin schaut zu Boden und sagt nichts.“

Antwortformat: „Bitte bewerten Sie die Effektivität der folgenden Antworten.“

	Sehr ineffektiv	Eher ineffektiv	Neutral	Eher effektiv	Sehr effektiv
Sie beziehen die ruhigere der neuen Teammitglieder ein, indem Sie fragen, ob der ältere Mitarbeiter auch schon Verkäufe von ihr für sich beansprucht hat					
Sie erinnern die beiden Mitarbeiter daran, dass persönliche Angriffe unangemessen sind und dass das Team sich auf zukünftige Lösungen konzentrieren sollte					
Sie unterstützen die neuen Teammitglieder, indem Sie ihre Seite ergreifen und sicherstellen, dass sie nicht als „Sündenböcke“ für die Probleme im Team ausgenutzt werden					
Sie erinnern das Team daran, dass kritische Bemerkungen die betreffenden Personen in die Defensive bringen und dadurch jeglicher Erfolg im Team verhindert wird					

(Abdruck mit freundlicher Genehmigung der American Psychological Association)

Beispielitem des Situational-Judgment-Tests zu Facetten der Big-Five-Persönlichkeitsdimensionen – hier: soziale Befangenheit (Mussel et al. 2016)

„Sie nehmen an einer öffentlichen Vorlesung mit ungefähr 100 anderen Zuhörern teil. Sie finden den Vortrag sehr interessant und würden gerne eine Frage stellen. Was würden Sie tun?“ (Eine Antwortoption soll angekreuzt werden).

Ich stelle die Frage nicht, denn ich fühle mich nicht wohl dabei, vor so vielen Menschen zu sprechen	
Falls überhaupt, stelle ich meine Frage nach der Vorlesung, sofern ich die Lehrkraft alleine treffe	
Ich stelle die Frage, sobald ich sicher bin, dass andere ebenfalls Fragen stellen	
Ich stelle meine Frage, sobald die Lehrkraft kurz zwischen zwei Sätzen pausiert	

(Translated/Used with permission from European Journal of Psychological Assessment (2018), 34(5), 328–335, ©2016 Hogrefe Publishing, ► www.hogrefe.com, ► <https://doi.org/10.1027/1015-5759/a000346>)

Die geschilderten Situationen entstammen typischerweise Anforderungsanalysen nach der Critical-Incident-Technik (► Abschn. 6.1.2), d. h. verschiedene Experten bzw. Expertinnen (das können z. B. Stelleninhaber/-innen oder Vorgesetzte sein) werden danach gefragt, in welchen Situationen sie Personen beobachtet haben, die besonders gute oder schlechte Leistungen gezeigt haben. Diese Schilderungen werden dann gesichtet. Beispielsweise möchte man ein ausreichendes Level an Spezifität in den Schilderungen wiederfinden. Bei positiver Begutachtung werden die Formulierungen der Schilderungen ggf. angepasst und diese als Itemstämme (s. obige Beispielitems) verwendet.

In manchen Fällen ist es sinnvoll, die generierten Situationen nicht als Text, sondern in Form von kurzen Videosequenzen in den Test aufzunehmen. Christian et al. (2010) verglichen im Rahmen ihrer Metaanalyse die Kriteriumsvalidität von Situational-Judgment-Tests im video- oder textbasierten Format. Der deutlichste Unterschied zwischen den beiden Formaten zeigte sich für Situational-Judgment-Tests zur Messung interpersoneller Fertigkeiten – hier waren videobasierte Verfahren den textbasierten deutlich überlegen (metaanalytischer Zusammenhang von .47 vs. .27; korrigiert für die Unreliabilität des Kriteriums); dabei gingen jedoch nur 2 videobasierte Situational-Judgment-Tests in die Analyse ein. Etwas aussagekräftiger ist daher der Vergleich von Situational-Judgment-Tests, die nicht klar einem Konstruktbereich zuzuordnen waren. Hier konnten immerhin 5 videobasierte mit 40 textbasierten Situational-Judgment-Tests verglichen werden. Auch bei diesem Vergleich war die Kriteriumsvalidität der videobasierten Verfahren besser als die der textbasierten (.36 vs. .25). Neben einer höheren Kriteriumsvalidität führen videobasierte Situational-Judgment-Tests in der Regel zu positiveren Reaktionen bei Bewerbenden (Chan und Schmitt 1997; Chan et al. 1997; Kanning et al. 2006).

Itemgenerierung: Critical-Incident-Technik

Video- vs. textbasierte Varianten

Situatives Urteilen ohne Situation?

Eine Kernannahme von Situational-Judgment-Tests ist, dass die präsentierte Situation die Basis für eine Entscheidung von Testpersonen, wie sie sich verhalten würden oder sollten, darstellt. Schon der Name *Situational-Judgment-Tests* unterstreicht die Bedeutung der Situationen. Diese Kernannahme wird jedoch von verschiedenen Studien infrage gestellt (Krumm et al. 2015; Schäpers et al. 2019). In diesen Studien wurden Situational-Judgment-Test-Items entweder mit oder ganz ohne Situationsschilderungen appliziert. Das heißt, ein Teil der Versuchspersonen war völlig „blind“ für die Situationen und mussten trotzdem angeben, wie sie sich verhalten würden. Erstaunlicherweise fanden Krumm et al. (2015), dass über mehrere Situational-Judgment-Tests hinweg und in mehreren Stichproben (Berufstätige und Studierende) zwischen ungefähr 50 und 70 % der Items ohne Situationsbeschreibungen genauso gut gelöst wurden wie mit. Diese Ergebnisse deuten darauf hin, dass auch andere als die bislang angenommenen Prozesse beim Beantworten von Situational-Judgment-Tests eine Rolle spielen. In einer darauf aufbauenden Studie konnten Freudenstein et al. (2020) zeigen, dass Testpersonen in Situational-Judgment-Tests (mit und ohne Situationsbeschreibungen) Situationsbewertungen vornehmen, diese aber vorwiegend auf Basis der Antwortoptionen erfolgen.

Neben den Situationsschilderungen muss ein geschlossenes Antwortformat realisiert werden. Dies beginnt mit dem Generieren von effektiven und ineffektiven Antwortalternativen, was keineswegs eine leichte Aufgabe ist.

Effektive und ineffektive Antwortalternativen generieren

Zum einen ist in vielen Situationen nicht völlig klar, was effektiv und inefektiv ist. Selbst Expertinnen und Experten können hier bisweilen unterschiedliche Auffassungen haben. Zum anderen müssen ineffektive Antwortalternativen so formuliert sein, dass sie nicht auf den ersten Blick als inefektiv erkannt werden. Neben dem Generieren von Antwortalternativen sind weitere Entscheidungen zu treffen. Idealerweise wird dazu ein mehrschrittiges Vorgehen gewählt (vgl. Corstjens et al. 2017).

6

Generieren von Antwortalternativen

1. Expertinnen und Experten generieren effektive Antwortalternativen; Testautorinnen und -autoren sollten prüfen, ob sich die Expertinnen und Experten darüber einig sind, welche Antwortalternativen effektiv sind. Sind die Expertinnen und Experten uneinig, sollte ggf. das gesamte Item verworfen werden.
2. Novizinnen und Novizen beschreiben, wie sie sich in einer geschilderten Situation verhalten würden. Dies ermöglicht es Testautorinnen und -autoren, plausible inefektive Verhaltensweisen zu generieren.
3. Testautorinnen und -autoren erstellen ein vorläufiges Set an Antwortalternativen zu jeder Situationsschilderung.
4. Testautorinnen und -autoren entscheiden, welche Fragestellungen sie verwenden möchten: „Was würden Sie tun?“ (Verhaltenstendenz-Instruktion) oder „Was sollten Sie tun?“ (Wissensinstruktion). Diese Wahl hat Auswirkungen darauf, ob der entwickelte Situational-Judgment-Test eher höhere Korrelationen zu Persönlichkeitsdimensionen (was bei der Verhaltenstendenz/Instruktion der Fall ist) oder zu kognitiven Fähigkeiten (bei einer Wissensinstruktion) aufweist (McDaniel et al. 2007).
5. Testautorinnen und -autoren entscheiden sich für ein Antwortformat. Typische Formate sind Single-Choice- und Multiple-Choice-Antwortformate (z. B. beste und schlechteste Alternative), das Ranking der Antwortoptionen (von gut nach schlecht) und die Bewertung aller Antwortoptionen (s. vorheriges Beispielitem aus dem Team-Role-Test).

Reliabilitätsschätzungen durch Retest-Reliabilität oder parallele Testversionen

■ Reliabilität

Mittlerweile liegen 3 Metaanalysen zur mittleren Reliabilität von Situational-Judgment-Tests vor (Campion et al. 2014; Catano et al. 2012; Kassten und Freund 2016). Diese bescheinigen Situational-Judgment-Tests nur geringe Reliabilitäten (zwischen .46 und .68), sofern diese in Form von internen Konsistenzmaßen (Cronbachs Alpha) geschätzt werden. Allerdings erfüllen Situational-Judgment-Tests eine wesentliche Voraussetzung für interne Konsistenzschätzungen, nämlich die Eindimensionalität, nicht. Vielmehr sind die Items zumeist sehr heterogen; sie bestehen nicht selten aus ganz unterschiedlichen, komplexen Situationen, zu deren effektivem Bewältigen mehrere Konstrukte beitragen. Somit sind Reliabilitätsschätzungen durch Testwiederholung (Retest-Reliabilität) oder parallele Testversionen angemessener. In der Tat fallen die so ermittelten Reliabilitätskoeffizienten deutlich höher aus (zwischen .66 und .82; vgl. Clause et al. 1998).

Vorhersagegüte von berufsbezogenen Kriterien

■ Zusammenhänge mit beruflichen Kriterien

Verschiedenen Metaanalysen attestieren Situational-Judgment-Tests eine Vorhersagegüte von berufsbezogenen Kriterien, die sich mit anderen etablierten Auswahlverfahren messen kann. Die erste Metanalyse erschien 2001. McDaniel et al. (2001) analysierten 102 Validitätsbefunde und ermittelten einen Zusammenhang zwischen Situational-Judgment-Tests und beruflicher Leistung von .34 (korrigiert für Messfehler im Kriterium). McDaniel et al. (2007)

aggregierten 118 empirische Zusammenhänge zwischen Situational-Judgment-Tests und beruflicher Leistung. Der für den Messfehler des Kriterium korrigierte Zusammenhang lag bei .26. Wenn ein sorgfältiges Matching zwischen dem Inhalt des Situational-Judgment-Tests und dem Kriterium erfolgt ist, steigt der metaanalytisch ermittelte Zusammenhang auf .35 (Christian et al. 2010) – wenn der Situational-Judgment-Test-Entwicklung eine Anforderungsanalyse vorausging, sogar auf .38 (McDaniel et al. 2001).

Zudem stellt sich die Frage nach dem inkrementellen Beitrag von Situational-Judgment-Tests über etablierte andere Prädiktoren: Werden Situational-Judgment-Tests irrelevant, wenn Persönlichkeit und Intelligenz als weitere Prädiktoren genutzt werden? Die Antwort auf diese Frage lautet: Nein. McDaniel et al. (2001) konnten zeigen, dass Situational-Judgment-Tests zwischen 3 und 5 % der Varianz berufsbezogener Kriterien zusätzlich zu allgemeiner kognitiver Leistungsfähigkeit aufklären. Wenn Situational-Judgment-Tests zusätzlich zu Persönlichkeitstests eingesetzt wurden, klärten sie 6–7 % zusätzliche Varianz berufsbezogener Kriterien auf.

■ Zusammenhänge mit anderen Konstrukten

Da eine Entscheidung, wie man sich in einer berufsbezogenen Situation verhalten sollte oder würde, selten auf ein einziges Konstrukt zurückzuführen ist, weisen die meisten Situational-Judgment-Tests keine überzeugenden (d. h. erwartungskonformen) Zusammenhänge zu verwandten Konstrukten auf, die mittels Fragebögen erhoben wurden. Eine erwähnenswerte Ausnahme findet man bei Mussel et al. (2016), die einen Situational-Judgment-Tests entwickelt haben, dessen Korrelationen mit konvergenten und diskriminanten Selbstberichtsmaßen den intendierten Messanspruch klar unterstreichen (► Tab. 6.11). Dieser Situational-Judgment-Test misst Persönlichkeitsfacetten statt breiter Persönlichkeitsmerkmale. Dies sind „Befangenheit“, „Geselligkeit“, „Offenheit für Ideen“, „Entgegenkommen“ sowie „Selbstdisziplin“. Jede der Facetten wird mit jeweils 22 (also vergleichsweise vielen) Items erfasst. Neben dem Fokus auf Facetten und der großen Itemzahl wurden die enthaltenen Situationen theoriegeleitet anhand von Definitionen der Facetten entwickelt – und nicht, wie sonst üblich, durch Expertinnen bzw. Experten mit dem Ziel generiert, die Breite eines Berufs abzubilden.

Korrelation zwischen Persönlichkeitsfacetten und -merkmalen

► **Tab. 6.11** Bivariate Korrelationen zwischen den 5 Facetten des NEO-Persönlichkeitssinventars und des Situational-Judgment-Tests zur Erhebung der Persönlichkeit

	SJT-N	SJT-E	SJT-O	SJT-A	SJT-C
NEO-N	.60	-.23	-.22	.12	.07
NEO-E	-.26	.66	.07	.01	-.08
NEO-O	-.31	.10	.70	-.01	.19
NEO-A	.08	-.01	.11	.41	.02
NEO-C	.02	-.07	.19	.03	.52

Quelle: Adaptiert von Mussel et al. (2016, S. 329). Used with permission from European Journal of Psychological Assessment (2018), 34(5), 328–335, ©2016 Hogrefe Publishing, ► www.hogrefe.com, ► <https://DOI.org/10.1027/1015-5759/a000346>

SJT = Situational-Judgment-Test; NEO = NEO-Persönlichkeitssinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R, deutsche Version: Ostendorf und Angleitner 2004; ► Abschn. 3.3.3.5); N = Befangenheit (Neurotizismus); E = Geselligkeit (Extraversion); O = Offenheit für Ideen (Offenheit für Erfahrungen); A = Entgegenkommen (Verträglichkeit); C = Selbstdisziplin (Gewissenhaftigkeit)

Christian et al. (2010) identifizierten 3 Konstruktbereiche, denen die bislang verfügbaren Situational-Judgment-Tests zuzuordnen sind. Als Beleg für deren Konstruktvalidität kann gelten, dass konstruktorientierte Situational-Judgment-Tests theoretisch nahestehende Kriterien besser vorhersagten als nicht konstruktorientierte, heterogen zusammengestellte Situational-Judgment-Tests.

Konstruktbereiche von Situational-Judgment-Tests (nach Christian et al. 2010)

- Berufsspezifisches Wissen und Fertigkeiten (z. B. Wissen über richtiges Verhalten im Cockpit von Flugzeugen)
- Soziale Fertigkeiten (z. B. beim Führen von Mitarbeitenden)
- Persönlichkeitsmerkmale (z. B. Gewissenhaftigkeit)

6

Verfügbare Tests In deutscher Sprache sind vor allem Situational-Judgment-Tests zur Diagnostik von Führungskompetenzen bzw. -stilen verfügbar, z. B. der *Leadership Judgment Indicator* von Neubauer et al. (2012) und das *Leadership Style Assessment* von Peus et al. (2016). Zur Teamarbeit liegt ein Situational-Judgment-Test, der *SJT-TA*, von Gatzka und Volmer (2017) vor, der über das Portal des Leibniz-Instituts für Sozialwissenschaften unter ► <https://www.gesis.org/home> zu beziehen ist.

Fazit Situational-Judgment-Tests stellen vielversprechende Methoden der Eignungsdiagnostik dar. Es bleibt aktuelle Forschung abzuwarten, um die Funktionsweise dieser Methode noch besser zu verstehen.

6.2.1.3 Kombination von Verfahren und Integration der Ergebnisse zu einer Auswahlentscheidung

Multimethodales Vorgehen

Zur Feststellung der Eignung für eine bestimmte Tätigkeit ist grundsätzlich ein multimethodales Vorgehen vorteilhaft. Jeder der 3 genannten Ansätze (konstrukt-, simulations- und biografieorientierte Ansätze) sowie jedes einzelne diagnostische Verfahren haben Vor- und Nachteile. Werden mit mehreren Methoden Informationen erhoben, kommt eine umfassendere Beurteilung der Eignung einer Person zustande, als wenn man sich auf eine Methode (z. B. das Interview) verlässt. Multimethodales Vorgehen bedeutet, dass verschiedene Arten von Verfahren zum Einsatz kommen. Ein Persönlichkeitsfragebogen sollte also nicht durch einen weiteren Persönlichkeitsfragebogen zur Erfassung des gleichen Merkmals ergänzt werden, sondern beispielsweise durch ein Interview oder ein Fremdbeurteilungsverfahren (z. B. ein Persönlichkeitsfragebogen, der von einer Bezugsperson bearbeitet wird).

Bei der Kombination mehrerer Verfahren stellt sich natürlich auch die Frage nach deren inkrementeller Prognosekraft (► Abschn. 2.6.3.4). Die 5 Verfahren, die laut einer Metaanalyse von Schmidt et al. (2016) die höchsten inkrementellen Beiträge über allgemeine kognitive Leistungsfähigkeit hinaus erwarten lassen, sind (absteigend geordnet nach inkrementellem Beitrag zur statistischen Vorhersage des Berufserfolgs):

Inkrementelle Beiträge über allgemeine kognitive Leistungstests hinaus (Schmidt et al. 2016)

- Integritätstests
- Interviews (auch Telefoninterviews)
- Interessentests
- Gewissenhaftigkeitsfragebögen
- Eingeholte Referenzen

Wenn mehrere Verfahren zum Einsatz kommen, müssen die daraus resultierenden diagnostischen Informationen zu einem Gesamтурteil „verrechnet“, also integriert werden. Dies stellt das sog. „abschließende Eignungsurteil“ dar. Zur Integration von einzelnen Informationen zu einem Urteil können die in ▶ Abschn. 5.1.3 genannten Formen der diagnostischen Urteilsbildung genutzt werden. Allerdings fordern Qualitätsstandards der Eignungsdiagnostik (DIN 2016), dass die Vorgehensweisen zur Integration von Informationen vorab, d. h. bei der Planung von Eignungsuntersuchungen, festzulegen sind. Es wird also eine mechanische Urteilsbildung gefordert. So wird sichergestellt, dass alle Bewerbenden nach den gleichen Regeln beurteilt werden.

Es werden nur die eignungsdiagnostischen Informationen erhoben, die sich im Rahmen der Arbeits- und Anforderungsanalysen als relevant herausgestellt haben. Am Ende einer eignungsdiagnostischen Untersuchung kann daher geprüft werden, ob eine Bewerberin oder ein Bewerber hinsichtlich der Anforderungsdimensionen die vorab festgelegten Mindestwerte erfüllt (und ggf. zusätzlich, ob vorab festgelegte Maximalwerte nicht überschritten werden). Wenn dies der Fall ist, kann das Eignungsurteil „geeignet“ ausgestellt werden. Erfüllen Bewerberinnen bzw. Bewerber eine vorab festgelegte Mindestmenge an Anforderungen nicht, erhalten diese das Urteil „nicht geeignet“. Bewerberinnen bzw. Bewerber, die zwar die Mindestmenge an Anforderungen erfüllen, aber nicht die Gesamtheit der Anforderungen, erhalten das Urteil „bedingt geeignet“.

Im Idealfall erweisen sich nach einer Eignungsuntersuchung mehr Bewerberinnen und Bewerber als geeignet als freie Positionen zu besetzen sind. Dann reicht eine Integration der eignungsdiagnostischen Informationen zu Kategorien wie „geeignet“, „bedingt geeignet“ und „ungeeignet“ nicht mehr aus. Vielmehr muss nun innerhalb der Gruppe der geeigneten Bewerberinnen und Bewerber eine Differenzierung vorgenommen werden. Wie dies geschieht, ist ebenfalls vorab festzulegen. In der Regel werden die Personenmerkmale zu einem (nicht kategorialen, sondern kontinuierlichen) Gesamtwert verrechnet, für den gilt: je mehr, desto besser. Beispielsweise könnte vorab festgelegt werden, dass Bewerberinnen bzw. Bewerber nicht „zu teamfähig“ und nicht „zu intelligent“ sein können. Daher könnten die Werte aus den zugehörigen Messverfahren summiert werden. Bewerberinnen und Bewerbern, die die höchsten Summenwerte erzielt haben, werden die ausgeschriebenen Positionen offeriert (s. aber auch andere Strategien in ▶ Abschn. 5.1.3).

Gesamтурteil nach feststehenden Regeln bilden

Rangreihe der Bewerbenden bilden

6.2.1.4 Akzeptanz von Personalauswahlverfahren

Im Idealfall nehmen Bewerberinnen bzw. Bewerber, die aufgrund ihres Abschneidens in verschiedenen Verfahren der Personalauswahl ein Job- oder Ausbildungsangebot erhalten, dieses auch an. Allerdings formen sie

Bewerbende bilden sich Eindruck von Unternehmen

während des Personalauswahlprozesses auch Eindrücke über das Unternehmen, für das sie zukünftig arbeiten könnten. Im schlimmsten Fall entscheiden sich geeignete Bewerberinnen bzw. Bewerber aufgrund negativer Erfahrungen im Bewerbungsverfahren gegen ein Jobangebot, teilen ihre negativen Erfahrungen mit anderen potenziellen Bewerbenden (z. B. in einschlägigen Foren) und kaufen ggf. die Produkte des Unternehmens nicht (mehr). Ebenso könnten sich ungeeignete Bewerberinnen bzw. Bewerber, die abgelehnt wurden, aufgrund ihrer negativen Eindrücke während der Auswahlsituation dazu entschließen, das Eignungsurteil juristisch anzufechten.

Gerechtigkeitsmodell nach Gilliland (1993)

6

Daher sind Unternehmen gut beraten, die für sie eigentlich positive Entscheidung von Bewerberinnen bzw. Bewerbern (nämlich sich überhaupt zu bewerben) im Zuge der Personalauswahl nicht nachteilig zu beeinflussen. Was Unternehmen tun können, um möglichst positive „applicant reactions“, „candidate experiences“ oder „fairness perceptions“ (diese Begriffe werden in der Literatur verwendet) zu erzeugen, beschreiben mehrere Modelle. Das wohl umfangreichste stammt von Gilliland (1993). Er unterscheidet Gerechtigkeitswahrnehmungen bezüglich der Entscheidung und bezüglich des Prozesses. Erstere wird als *distributive Gerechtigkeitswahrnehmung* bezeichnet; es geht also darum, ob die Entscheidung, dass man ein Jobangebot bekommt oder nicht, als fair wahrgenommen wird. Letzteres wird als *prozedurale Gerechtigkeit* bezeichnet, d. h. inwiefern das Vorgehen, um zu einer Entscheidung zu gelangen, als fair wahrgenommen wird. Prozedurale Gerechtigkeit unterteilt Gilliland (1993) nochmals in Aspekte, die die Auswahlmethode, das grundsätzliche Vorgehen und die handelnden Personen betreffen.

Prozedurale Gerechtigkeit nach Gilliland (1993)

Als fair wahrgenommenes Vorgehen/Verhalten bezogen auf ...

- die Auswahlmethode:
 - Bezug zur ausgeschriebenen Position ist erkennbar.
 - Bewerbende haben Gelegenheit, unmittelbar Einfluss auf das Ergebnis des Auswahlverfahrens nehmen zu können.
 - Bewerbende können das Ergebnis hinterfragen.
 - Auswahlverfahren werden konsistent administriert.
- das grundsätzliche Vorgehen:
 - Bewerbende erhalten ein Feedback.
 - Bewerbende erhalten eine Erklärung zu der getroffenen Entscheidung.
 - Die Organisation ist offen und ehrlich gegenüber Bewerbenden.
- die für das Unternehmen handelnden Personen:
 - Bewerbende werden freundlich und respektvoll behandelt.
 - Bewerbende haben die Möglichkeit, ihre Sicht einzubringen sowie Feedback zu geben.
 - Bewerbenden werden angemessene Fragen gestellt.

Relevanz prozeduraler Gerechtigkeitswahrnehmungen

Eine längsschnittliche Studie von Bauer et al. (1998), die in einem realen Auswahlkontext durchgeführt wurde, belegt die Relevanz prozeduraler Gerechtigkeitswahrnehmungen (genauer der freundlichen und respektvollen Behandlung sowie zusätzlich der Information über die Auswahlverfahren) für die Einstellung der Bewerbenden zum Unternehmen. Zudem sagte der

Bezug der Auswahlverfahren zur ausgeschriebenen Position die Attraktivität des Unternehmens auch nach Kontrolle des Ergebnisses der Auswahl (Einstellung vs. Ablehnung) vorher.

Im deutschen Sprachraum haben Schuler und Stehle (1983) erstmals den Begriff der sozialen Validität geprägt. Mit diesem Begriff wird betont, dass neben dem klassischen und in der Wissenschaft häufiger beachteten Gütekriterium der Validität auch die subjektiven Bewertungen von Bewerbenden beachtet und zur Evaluation von Auswahlverfahren herangezogen werden sollten. Möglichkeiten der Einflussnahme auf solche Bewertungen sehen Schuler und Stehle (1983) in 4 Bereichen: Information (z. B. über die Organisation und die ausgeschriebene Position), Partizipation (z. B. Einflussnahme auf die diagnostische Situation, „opportunity to perform“ nach Gil-liland, 1993), Transparenz (z. B. hinsichtlich des Bewertungsprozesses) und Urteilskommunikation (z. B. respektvoll, offen).

Zur Messung der Akzeptanz können die AKZEPT!-Fragebögen (für eine Übersicht s. Kersting 2008b) verwendet werden. Es stehen für die Verfahrensgruppen Leistungstests, Persönlichkeitstests, Assessment-Center und Interviews eigene AKZEPT!-Versionen zur Verfügung (► Tab. 6.12).

Am Beispiel des AKZEPT!-L (für Leistungstests) berichtet Kersting (2008b) akzeptable bis gute Reliabilitätskoeffizienten. Bei Anwendung des AKZEPT!-L auf verschiedene Leistungstests erwies sich vor allem die Skala „Augenscheininvalidität“ (d. h. der wahrgenommene Bezug des Tests zu den Anforderungen der ausgeschriebenen Position) als relevant für die globale Akzeptanzbewertung der Teilnehmenden (Kersting 2008b).

Ein Vorteil der unterschiedlichen Akzeptanzfragebogenversionen liegt in ihrer Anwendbarkeit für verschiedene eignungsdiagnostische Instrumente. In ► Abb. 6.6 sind mittlere Akzeptanzeinschätzungen von einem Intelligenztest, einem breiten Persönlichkeitstest (beide aus Beermann et al. 2013) und einem Assessment-Center (aus Kersting 2010) gegenübergestellt. Dabei schneidet der Intelligenztest in Bezug auf die Bewertung der Augenscheininvalidität deutlich schlechter ab als der Persönlichkeitstest und das Assessment-Center. Natürlich sind die hierzu herangezogenen Stichproben nicht

Soziale Validität

AKZEPT!-Fragebögen

Vergleichbarkeit von
Akzeptanzeinschätzungen über
Messinstrumente hinweg

► Tab. 6.12 Akzept!-Fragebogenskalen mit Beispielitems. (Abdruck mit freundlicher Genehmigung von Prof. Dr. Martin Kersting)

Skala	AKZEPT!-Version	Beispielitems
Kontrollierbarkeit	L	Bei der Bearbeitung der Testaufgaben wusste ich jederzeit, was ich tun muss
Messqualität	I	Das Interview ermöglicht es, die zwischen verschiedenen Menschen bestehenden Unterschiede in den vom Interview gemessenen Merkmalen exakt zu messen
Augenscheininvalidität	P	Dass man mit den Fragen/Aussagen wie denen des Verfahrens geeignete Personen für einen Job herausfinden kann, ist zu bezweifeln
Belastungsfreiheit	AC	Die Teilnahme am AC war belastend.
Wahrung der Privatsphäre	P	Was ich auf solche Fragen/Aussagen antworte, geht diejenigen, die die Verfahrensergebnisse erhalten, nichts an
Antwortfreiheit	P	Aufgrund der vorgegebenen Antwortmöglichkeiten des Verfahrens hatte ich nicht die Freiheit, so zu antworten, wie es für meine Person zutreffend ist
Gute Organisation	AC	Ich wusste jederzeit, wann und wo die nächste Aufgabe für mich beginnt
Positive Atmosphäre	I	Der Umgang der Interviewer mit den Bewerbern war jederzeit freundlich und wertschätzend
Gesamtbeurteilung	AC	Welche Schulnote würden Sie dem soeben bearbeiteten AC geben?

AKZEPT!-Versionen: L = Leistungstests, P = Persönlichkeitsfragebögen, I = Interview, AC = Assessment-Center

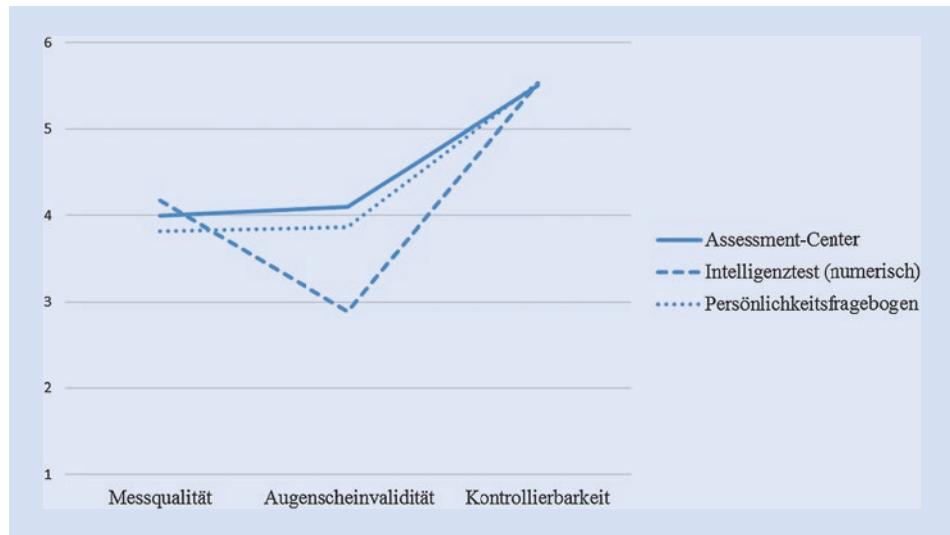


Abb. 6.6 Vergleich von Auswahlmethoden hinsichtlich der Akzeptanzbewertungen (Bewertungen auf einer Skala von 1 bis 6)

Personalauswahl so gestalten, dass Bewerbende ein Jobangebot auch annehmen

direkt vergleichbar. Ebenso stellen die verwendeten Verfahren nur ein Beispiel aus der Gesamtmenge der in der jeweiligen Verfahrensgruppe enthaltenen Verfahren dar. Allerdings berichtet der Autor (Kersting 2008b) ähnliche Bewertungen der Augenscheininvalidität bei anderen Leistungstests.

Es ist insgesamt festzuhalten, dass Unternehmen auch durch die Gestaltung der Personalauswahl beeinflussen können, ob Bewerberinnen bzw. Bewerber ein Jobangebot annehmen. Dies sollte nicht so weit gehen, dass prognostisch valide Auswahlverfahren gegen besser akzeptierte, aber wenig valide Verfahren ausgetauscht werden. Ist bekannt oder zu befürchten, dass valide Auswahlverfahren geringe Akzeptanz bei Bewerbenden finden, kann man auch versuchen, alle Randbedingungen so zu gestalten, dass gemäß gängiger Akzeptanzmodelle (s. o.) eine insgesamt positive Reaktion der Bewerbenden wahrscheinlich wird. Schließlich ist Unternehmen nicht geholfen, wenn Auswahlverfahren gute Eignungsprognosen liefern, aber geeignete Bewerberinnen bzw. Bewerber die ihnen angebotenen Positionen nicht annehmen.

Fazit Als Fazit und Abschluss dieses Abschnitts zur Selektion von Personen für die Personalauswahl soll einer der prominentesten Forscher im Bereich der Eignungsdiagnostik, Prof. Dr. Heinz Schuler, zu Wort kommen.

Interview mit Prof. Dr. Heinz Schuler (Auszug)

Prof. Dr. Heinz Schuler (Foto: Rolf Schulten).

Dieses Interview wurde von Dr. Anne Klostermann, der Pressereferentin der Deutschen Gesellschaft für Psychologie, anlässlich der Verleihung des Psychologiepreises 2017 mit Prof. Dr. Heinz Schuler geführt. Es wird hier in gekürzter Form wiedergegeben (Abdruck mit freundlicher Genehmigung von Prof. Schuler und Dr. Klostermann).

Wie findet ein Unternehmen geeignete Mitarbeiter? Zum „Finden“ gehört natürlich auch das Personalmarketing. Je besser ein Unternehmen die Anforderungen seiner

Arbeitsplätze kennt und weiß, welche Personenmerkmale hierzu passen, desto gezielter kann es potenzielle Bewerber ansprechen und desto chancenreicher kann es sie auswählen. Auch das Image eines Unternehmens spielt dafür eine große Rolle. Je größer die Zahl der Bewerber ist, desto strenger kann ein Unternehmen auswählen und desto geringer ist die Gefahr, ungeeignete Personen einzustellen. Auch das Zusammenpassen von Menschen und Organisationen ist ein wichtiges Thema. Es sind andere Personen, die sich in der öffentlichen Verwaltung bewerben, als diejenigen, die eine Event-Agentur als Arbeitgeber suchen.

Wie können eignungsdiagnostische Verfahren bei der Auswahl helfen? Die eingesetzten Verfahren sollten den Anforderungen der Arbeitsplätze möglichst gut entsprechen. Das kann in sehr konkreter Form stattfinden, indem man z. B. eine technische Arbeitsprobe für einen Mechaniker einsetzt oder ein simuliertes Verkaufsgespräch für einen Verkäufer. Die Verfahren können aber auch abstrakter Natur sein wie z. B. Fähigkeits- und Persönlichkeitstests. Einstellungsinterviews können sehr flexibel gestaltet werden. Zum Beispiel kann man nach ganz konkreten Arbeitserfahrungen und Ergebnissen fragen, man kann aber auch

Fragen zu Interessen und Motiven stellen. Bei allen Typen von Auswahlverfahren findet man in der Praxis eine große Spannweite in der Qualität – und auch große Unterschiede in der Qualifikation der Auswählenden.

Was macht denn gute Einstellungsverfahren aus? Für einzelne Einstellungsverfahren lässt sich das einfach bestimmen: verschiedene Diagnostiker oder Auswählende sollten zum gleichen Ergebnis kommen, dann ist das Verfahren objektiv. Diese Objektivität lässt sich als Kennwert bestimmen. Für konventionelle Auswahlgespräche ist sie niedrig, für die meisten Tests ist sie hoch. Der zweite Kennwert bezeichnet die Reliabilität, das ist das Maß der Übereinstimmung der Testergebnisse einer Person zu verschiedenen Zeitpunkten. Dieser Kennwert ist hoch für Intelligenztests, mittel für Persönlichkeitstests und niedrig für die meisten Aufgaben im Assessment-Center. Der wichtigste Kennwert ist die Validität. Er bezeichnet, ob ein Verfahren tatsächlich das misst, was es messen soll, und deshalb helfen kann, eine zutreffende Prognose der Leistung, Zufriedenheit oder Gesundheit zustande zu bringen.

Bedeutet das, dass Assessment-Center per se nicht tauglich sind? Glücklicherweise nicht, im Gegenteil gehört ein sachkundig konstruiertes und durchgeführtes Assessment-Center zu den besten und „sozial validesten“ Verfahren zur Auswahl und zur Personalentwicklung. Leider hat die Prognosekraft der in der Praxis durchgeföhrten Assessment-Center in den letzten zwei Jahrzehnten kontinuierlich abgenommen.

Woran liegt das? Das hat meines Erachtens zwei Ursachen: Erstens wird

zunehmend darauf verzichtet, multimodal vorzugehen. Stattdessen werden fast nur Simulationsaufgaben wie Rollenspiele, Präsentationen und Gruppendiskussionen aneinandergereiht. Dessen inkrementelle Validität ist gering, das bedeutet, sie haben zusammen genommen kaum mehr Aussagekraft als eine dieser Aufgaben allein. Zweitens werden die meisten Praxis-Assessment-Center heute nicht mehr von qualifizierten Eignungsdiagnostikern konzipiert und gemeinsam mit erfahrenen Führungskräften durchgeführt.

Viele Unternehmen setzen inzwischen Online-Assessment-Center ein und beurteilen sie aufgrund ihrer Ökonomie sehr positiv. Was ist aus eignungsdiagnostischer Perspektive davon zu halten? Online-Diagnostik hat sich als probates Mittel erwiesen, die Vorauswahl zu verbessern. Assessment-Center allerdings sind, wie gesagt, zu einem Teil durch interaktive Aufgaben wie Rollenspiele, Präsentationen und Gruppendiskussionen gekennzeichnet. Dies lässt sich online nicht gut gestalten. Aber Tests und biografische Fragebögen lassen sich sehr gut auf diese Weise durchführen und ersparen beiden Seiten erheblichen Aufwand. Man darf dabei allerdings nicht vergessen, dass diese Tests den gleichen psychometrischen Ansprüchen zu genügen haben wie schriftlich durchgeführte Verfahren und dass es späterer Überprüfung bedarf, ob die Antworten tatsächlich von dem betreffenden Kandidaten kommen und nicht z. B. von seiner großen Schwester.

Kurz nach der Jahrtausendwende haben wir eine der ersten sehr groß angelegten Aktionen der internetgestützten Personalauswahl durchgeführt und dabei für einen Automobilhersteller die Mitarbeiter für ein neues Werk ausgewählt. Die politische Vorgabe war, dass nur arbeitslose Perso-

nen ausgewählt werden sollten und dass die Vorbildung keine Rolle spielen darf. Das Ergebnis war, dass dieses neue Werk nach wenigen Jahren ein neues Automobil produzierte und ausgezeichnete Ergebnisse lieferte. Allerdings war die internetgestützte Auswahl nur der erste Schritt in einer Abfolge diagnostischer Phasen.

Zusammengefasst ist damit zu rechnen, dass derartige internetgestützte Vorgehensweisen weitere Verbreitung finden. Die Möglichkeiten, den Datenschutz zu gewährleisten, haben sich zwar verbessert, allerdings findet eine laufende Eskalation des Aufrüstens auch auf Seiten derer statt, die ihn unterlaufen wollen.

6.2.2 Selektion von Bedingungen: Berufs- und Ausbildungswahl

Sozial-kognitive Theorien der Berufs- und Ausbildungswahl betonen das eingangs dieses Kapitels erwähnte Prinzip der Passung zwischen Merkmalen der Organisation und der Person in besonderer Weise. Sie gehen von einer Interaktion zwischen Personenmerkmalen wie Selbstwirksamkeit (= die Einschätzung des Individuums über seine Fähigkeiten, die notwendigen Verhaltensweisen zu zeigen, die nötig sind, um ein bestimmtes Ziel zu erreichen; Bandura 1986), Ergebniserwartungen (= Erwartungen, ob Verhalten zu gewünschten Zielen führt), Interessen und Zielen einerseits und der beruflichen Umwelt andererseits aus.

So können initiale Erfahrungen, welches relevante Verhalten

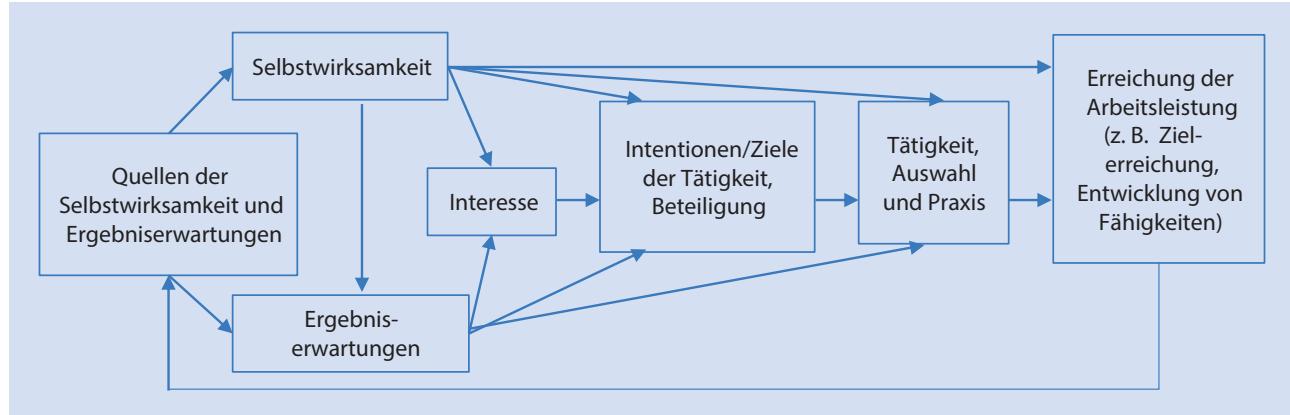
- a) eine Person in einer bestimmten Umwelt gut ausüben kann (z. B. über mehrere Wochen ein Lehrbuch lesen und verstehen) und
- b) zu den gewünschten Ergebnissen führt (z. B. eine Klausur bestehen),

dazu beitragen, dass sich Interessen ausbilden und berufliche Ziele formuliert werden (z. B. eine Promotion in der Psychologie), was sich wiederum in zielorientiertem Verhalten (z. B. Aufnahme studentischer Hilfskrafttätigkeiten, Verfassen der Masterarbeit im gewünschten Bereich der Psychologie, Bewerbung auf eine Promotionsstelle) und ggf. dem Erreichen der Ziele niederschlägt. Werden die gewünschten Ziele nicht erreicht oder kann das dazu notwendige Verhalten nicht in ausreichendem Maß ausgeübt werden (z. B. Unterbrechungen des Lernens aufgrund anderer Verpflichtungen), so erfolgt ggf. eine Anpassung der Interessen und der Berufs- bzw. der Ausbildungsziele (vgl. Lent et al. 1994).

Das in Abb. 6.7 dargestellte Modell der kognitiven und behavorialen Einflüsse auf die Berufswahl stellt persönliche Interessen in den Mittelpunkt. Dies spiegelt sich in der aktuellen Praxis der Berufs- und Studienberatung wider. Zumeist erfolgt eine Diagnostik der berufs- und

Sozial-kognitive Theorien der Berufs- und Ausbildungswahl

Persönliche Interessen im Mittelpunkt der Berufs- und Studienberatung



6

Abb. 6.7 Modell zur Entwicklung von grundlegenden Karriereinteressen im Laufe der Zeit. (Aus Lent et al. 1994, S. 88, © 1994, with permission from Elsevier; Übersetzung durch die Autoren dieses Buchs)

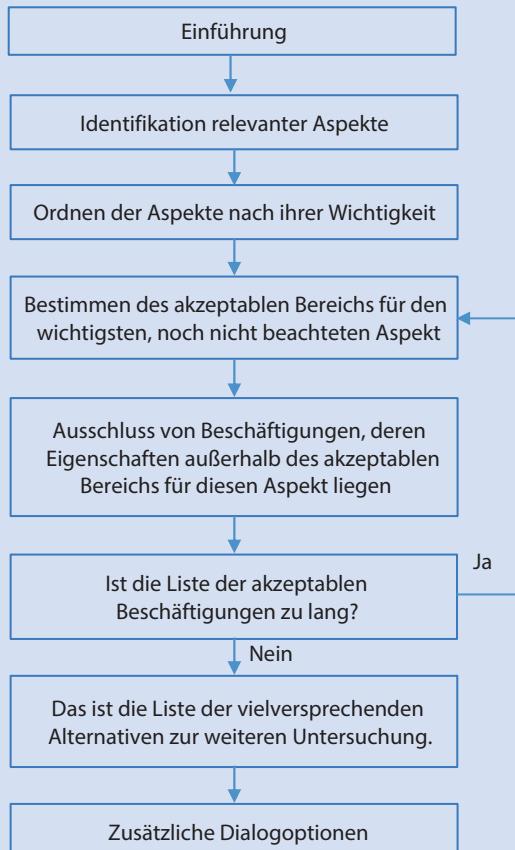
studienbezogenen Interessen. Diagnostische Verfahren, die dies leisten, sind in ▶ Kap. 3 beschrieben. Das Ergebnis dieser Diagnostik wird damit abgeglichen, in welchem Ausmaß es Berufe und Ausbildungen ermöglichen, entsprechende Interessen auszuüben. Viele diagnostische Verfahren verwenden für diesen Abgleich das Modell nach Holland (1997; □ Abb. 6.3). Beispielsweise kann eine Person vorwiegend Interessen haben, die im Bereich I (investigativ, d. h. intellektuell, untersuchend-forschend) anzusiedeln sind. Auf Basis dieser Interessen wären manche Berufe eher zu empfehlen (z. B. in der Forschung) als andere (z. B. aus dem Bereich E=enterprising, in dem u. a. verkäuferische Tätigkeiten angesiedelt sind).

Eine Abklärung von Interessen fällt in den konstruktorientierten Ansatz. Darüber hinaus können im Rahmen dieses Ansatzes auch die für ein berufliches Ziel erforderlichen Fähigkeiten sowie relevante Persönlichkeitseigenschaften abgeklärt werden. Der Abgleich zwischen Merkmalen der Person und Anforderungen des Berufs/der Ausbildung erfolgt hier analog zur Personalauswahl – mit dem Unterschied, dass der Abgleich nicht vom *einem* Unternehmen oder *einer* Ausbildungsstätte erfolgt, sondern von Berufsinteressierten und für *viele* mögliche Unternehmen bzw. Ausbildungen. Ein weiterer Unterschied zur Personalauswahl besteht darin, dass die Beurteilung der Passung unverbindlich ist. Das heißt, Bewerberinnen und Bewerber können trotz vorher diagnostizierter geringer Passung einen Beruf oder eine Ausbildung aufnehmen.

Unverbindliche Einschätzung der Passung

Berufsberatung durch den Computer

Computer- oder Internetbasierte Karriereplanungssysteme haben den Vorteil, schnell und automatisiert auf eine große Menge an Informationen über Berufe zurückgreifen zu können und diese mit Präferenzen von Berufssuchenden abgleichen zu können. Gati et al. (2006) beschreiben den Aufbau eines solchen Systems mit dem folgenden Schema:



Aufbau eines Karriereplanungssystems (verkürzte Darstellung) nach Gati et al. (2006, S. 207, © 2006, with permission from Elsevier).

Das Schema macht deutlich, dass das wesentliche Prinzip darin besteht, die für Berufssuchende wichtigen Aspekte des zukünftigen Berufs so lange zu priorisieren und einzugrenzen, bis eine angemessene (d. h. nicht zu umfangreiche) Auswahl potenziell passender Berufe zur Verfügung steht. Anschließend stehen zusätzliche Optionen zur Verfügung, z. B. eine Liste von „beinahe passenden Berufen“ und weitere Informationen über die potenziell passenden Berufe.

Die Studie von Gati et al. (2006) ist ein seltes Beispiel für die Evaluation eines computerbasierten Karriereplanungssystems. Die Autorinnen und Autoren riefen dazu 6 Jahre nach erfolgter Berufsberatung durch dieses System Teilnehmende an und fragten diese u. a. danach, ob sie sich für einen vom System empfohlenen Beruf entschieden haben und wie zufrieden sie mit ihrer Berufswahl waren. Interessanterweise waren 84 % der Befragten, die dem System gefolgt waren, nach 6 Jahren mit ihrer Berufswahl sehr zufrieden. Unter denjenigen, die dem System nicht gefolgt waren, lag der Anteil der sehr zufriedenen Befragten nur bei 38 %. Leider sind keine Rückschlüsse darauf möglich, inwiefern die hohe Zufriedenheit in der ersten Gruppe durch die Berufsberatung an sich oder durch die Möglichkeiten der computerbasierten Durchführung oder durch beides gemeinsam zu erklären ist.



■ Abb. 6.8 Vielleicht bald gängige Praxis? Simulationsbasierte Berufswahl mithilfe von Virtual Reality. (© WavebreakmediaMicro/stock.adobe.com)

Virtuelles Probearbeiten

Simulationsorientierte Ansätze zur Berufs- und Ausbildungswahl sind aufgrund des damit verbundenen Aufwands weniger verbreitet. Seit einiger Zeit bieten jedoch einige Unternehmen eine Art „virtuelles Probearbeiten“ (■ Abb. 6.8) an, in dessen Rahmen man einige typische Gegebenheiten der interessierenden Tätigkeit im Rahmen einer Internetpräsenz durchlaufen kann. Diese Ansätze firmieren auch unter der Bezeichnung „Recruitment“. Evidenz zur Wirksamkeit dieser Methode der Berufs- und Ausbildungswahl liegt derzeit noch nicht vor.

Im Bereich der Beratung von Studierenden sind sog. „Online-Self-Assessments“ verbreitet. Neben einer Erfassung der Merkmale, die im Studium gefordert werden, erfolgt oft auch eine Darstellung der Studieninhalte und wesentlicher Rahmenbedingungen. Für eine ausführliche Beschreibung dieses Ansatzes sei auf ▶ Abschn. 7.1.3 verwiesen.

Aber nicht nur aus Sicht von Bewerbenden ergeben Online-Self-Assessments Sinn. Unternehmen und Universitäten wollen damit einerseits potentielle Bewerbende bei ihrer Studien- bzw. Berufswahl unterstützen und anderseits geeignete Personen zu einer Bewerbung anregen, während sie hoffen, dass ungeeignete von einer Bewerbung Abstand nehmen. Durch diese Selbstselektion verbessert sich die Basisrate (▶ Abschn. 5.1.3.3): Je höher der Anteil der geeigneten Kandidatinnen bzw. Kandidaten unter allen Bewerbenden ist, desto größer ist die Chance, durch das Auswahlverfahren geeignete zu entdecken. Sind nur wenige der Kandidatinnen bzw. Kandidaten geeignet (Basisrate niedrig), sucht man sprichwörtlich die Nadel im Heuhaufen – ist der Anteil geeigneter Kandidatinnen bzw. Kandidaten dagegen groß (Basisrate hoch), werden wahrscheinlich auch viele geeignete ausgewählt.

Biografieorientierte Ansätze der Berufs- und Ausbildungswahl gewinnen mit der Erfahrung der zu beratenden Personen an Bedeutung. Beratungen zur Berufswahl von Personen mit umfangreicher Berufserfahrung finden bei freiwilligen oder (durch drohende oder bestehende Arbeitslosigkeit) erzwungenen beruflichen Neuorientierungen statt. Je länger eine Person bereits im Berufsleben stand, desto wichtiger wird die Frage, ob für eine neue Position notwendige (formale) Qualifikationen erworben wurden bzw. die vorausgesetzten Erfahrungen gemacht wurden. Zudem ändern sich berufsbezogene Bedürfnisse über die Lebensspanne. Veränderte Bedürfnisse sollten – neben früheren Bedürfnissen, die eventuell zur vorherigen Berufswahl geführt haben – berücksichtigt werden (Kooij et al. 2011).

Online-Self-Assessments

Selbstselektion herbeiführen

Biografieorientierte Ansätze der Berufs- und Ausbildungswahl

Fazit Personen, die sich fragen, welche Berufe oder Ausbildungen für sie infrage kommen, können also konstrukt-, simulations- und biografieorientierte Verfahren nutzen, um Informationen über sich selbst und ihre beruflichen Optionen zu erhalten. Es werden dabei solche Informationen erhoben, die mit Merkmalen des Berufs oder der Ausbildung abgeglichen werden und damit zu einer Einschätzung der Passung genutzt werden können.

6.2.3 Modifikation von Personen: Personalentwicklung

Der Ausdruck „Modifikation von Personen“ mag ein wenig befremden. Wie bei der Berufs- bzw. Ausbildungswahl sowie der Personalauswahl steht allerdings auch im Rahmen der Personalentwicklung die Passung von Personen- zu Organisationsmerkmalen im Vordergrund. Folglich ist keine grundsätzliche Modifikation einer Person intendiert – vielmehr zielt Personalentwicklung darauf ab, solche Merkmale von Personen gezielt weiterzuentwickeln, die sich im Rahmen von Anforderungsanalysen als essenziell für eine aktuelle oder zukünftige berufliche Position erwiesen haben.

Merkmale der Person entwickeln, die tätigkeitsrelevant, aber nicht ausreichend vorhanden sind

Definition

Personalentwicklung bezeichnet planmäßige, beim Individuum ansetzende Maßnahmen der Verbesserung/Erweiterung von beruflichen Handlungskompetenzen (Holling und Liepmann 2004).

Maßnahmen der Verbesserung bzw. Erweiterung von beruflichen Handlungskompetenzen zu entwickeln und durchzuführen (in Form von wissensorientierten Schulungen, verhaltensorientierten Trainings, durch Coaching, Mentorenprogramme etc.) ist allerdings keine Aufgabe der Psychologischen Diagnostik. Die Psychologische Diagnostik nimmt dennoch im Rahmen der Modifikation von Personen eine entscheidende Rolle ein. Einerseits dient sie der Identifikation von Anforderungen eines Berufs oder einer Ausbildung. Des Weiteren gibt sie Antworten auf die Frage: „Welche Personen sollen hinsichtlich welcher Merkmale weiterentwickelt bzw. trainiert werden?“ Denn nicht alle Personen haben bezüglich der relevanten Merkmale „Nachholbedarf“ – einige verfügen bereits über ausreichend hohe Ausprägungen relevanter Merkmale, um aktuelle und zukünftige Aufgaben bewältigen zu können. Darüber hinaus wäre es ineffektiv, alle zu trainierenden Personen dem gleichen Training zuzuweisen. Während manche Personen bei den Merkmalen A und B „Nachholbedarf“ haben, benötigen andere nur für die Merkmale C und D ein Training. Somit dient die Psychologische Diagnostik im Rahmen der Personalentwicklung der zielgerichteten Identifikation von spezifischen Entwicklungsbedarfen von Personen. In der Trainingsliteratur wird dies als Teil einer Trainingsbedarfsanalyse verstanden.

Diagnostik zur Identifikation des Entwicklungsbedarfs

Definition

Eine **Trainingsbedarfsanalyse** ist eine systematische Methode, um zu identifizieren, weshalb Leistungen in Organisationen geringer waren als erwartet oder erforderlich (Blanchard und Thacker 2013, S. 108).

Soll-Ist-Diskrepanz als Ausgangspunkt

Ausgangspunkt einer Trainingsbedarfsanalyse ist eine tatsächliche oder subjektiv wahrgenommene Minderleistung der Organisation oder von Teilen der Organisation (Soll-Ist-Diskrepanz). Statt – wie leider noch weitverbreitet – „auf Verdacht“ und unsystematisch Veränderungen im Unternehmen

Erfolgskontrolle

6

vorzunehmen, sollte systematisch vorgegangen werden. Neben der Klärung, ob organisationale Randbedingungen für die Soll-Ist-Diskrepanz verantwortlich sind (organisationale Analyse) erfolgt eine Anforderungsanalyse (in der Trainingsliteratur als „operationale Analyse“ bezeichnet), in deren Rahmen analysiert wird, was genau Personen tun müssen, um in ihrer beruflichen Position effektiv und erfolgreich zu sein. Zudem wird eine sog. „Personenanalyse“ vorgenommen, d. h. eine Identifikation der Personen, die nicht oder nicht in ausreichendem Maße über die notwendigen Fähigkeiten oder Fertigkeiten verfügen, um effektiv und erfolgreich zu sein. Diese Personen erhalten dann passgenaue Entwicklungsmaßnahmen.

Aber auch nachdem Personalentwicklungsmaßnahmen durchgeführt wurden, kommt der Psychologischen Diagnostik noch eine wichtige Aufgabe zu: die erneute Messung von Personenmerkmalen zur Beurteilung des Erfolgs der Maßnahmen. Denn trotz Teilnahme an einer Maßnahme (beispielsweise an einem Training) kann das damit adressierte Merkmal unverändert bleiben (vgl. ▶ Abschn. 5.1.1). Zudem ist es möglich, dass zwar eine Verbesserung des intendierten Merkmals (z. B. der sozialen Kompetenz) eintritt, sich dies aber nicht im beruflichen Alltag manifestiert (d. h. kein sozial kompetentes Verhalten gezeigt wird). In beiden Fällen wäre die durchgeführte Maßnahme ineffektiv – die gewünschten Ergebnisse für das Unternehmen (z. B. Verringerung der Soll-Ist-Diskrepanz beim Umsatz) würden ausbleiben und der Prozess müsste von vorne beginnen (Abb. 6.9).

Zur Identifikation von Anforderungen kommen auch im Rahmen der Personalentwicklung Methoden zum Einsatz, die bereits in ▶ Abschn. 6.1.2 dargestellt wurden. Zur Diagnostik von Personenmerkmalen können, wie in ▶ Abschn. 6.2.1 erläutert, biografie-, konstrukt- und simulationsorientierte Verfahren eingesetzt werden. Instrumente, die nicht zur Personalauswahl, sondern vorrangig zur Diagnostik des Entwicklungsbedarfs von Personen eingesetzt werden, sind sog. „Entwicklungs-Assessment-Center“ (häufig auch „Development-Center“ bezeichnet).

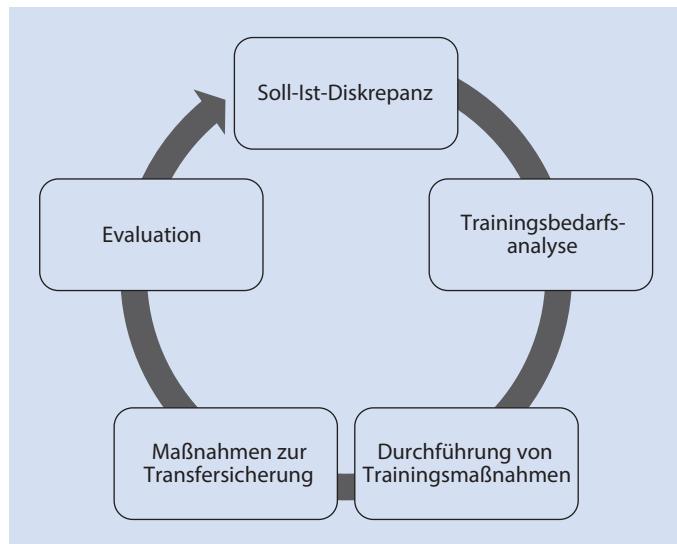


Abb. 6.9 Systematischer Prozess der Personalentwicklung. (Angelehnt an Solga et al. 2011, S. 23)

Entwicklungs-Assessment-Center

Für Entwicklungs-Assessment-Center existieren in der wissenschaftlichen Literatur viele Definitionen und in der Praxis viele unterschiedliche Auffassungen, wie diese gestaltet werden sollten (vgl. Rupp et al. 2006; Tillema, 1998). Im Kern stellen Entwicklungs-Assessment-Center jedoch die Anwendung der Assessment-Center-Methode (vor allem simulationsorientierter Übungen wie Rollenspiele, Gruppendiskussionen) zum Zwecke der Feststellung des Trainingsbedarfs einer Person dar). Darüber hinaus werden Entwicklungs-Assessment-Center in der Literatur auch als eigene Interventionsmethode beschrieben, da Teilnehmende am Ende ein ausführliches Feedback erhalten und daraufhin eine eigene Weiterentwicklung anstreben sollen. Allerdings konnten Jones und Whitmore (1995) zeigen, dass Teilnehmende eines Entwicklungs-Assessment-Centers sich nicht intensiver um ihre eigene Weiterentwicklung kümmerten als die Personen in einer Kontrollgruppe. Andererseits berichten Rupp et al. (2006) moderate (selbst- und fremdberichtete) Leistungsverbesserungen der Teilnehmenden eines Entwicklungs-Assessment-Centers.

Eine weitere Form der Diagnostik in der Personalentwicklung kommt in Form von Leistungsbeurteilungen zum Einsatz. Das Ziel von Leistungsbeurteilungen ist es, zu quantifizieren, in welchem Ausmaß eine Mitarbeiterin bzw. ein Mitarbeiter zum Erfolg des Unternehmens beigetragen hat (Lohaus und Schuler 2014). Dies kann aus vielerlei Gründen erfolgen. Beispielsweise können Unternehmen einen Teil des Gehalts an die Erreichung von vorab festgelegten Individualzielen (z. B. Höhe des Umsatzes, den eine Mitarbeiterin bzw. ein Mitarbeiter erzielt) knüpfen. Dieser Gehaltsbestandteil wird dann anteilig in dem Umfang ausgezahlt, in dem das Ziel erreicht wurde. Eine Leistungsbeurteilung dient jedoch auch als Feedback für Mitarbeitende und damit als Basis für zukünftige Personalentwicklungsschritte (wie die Teilnahme an Trainingsprogrammen, individuell gesteuerte Verhaltensänderungen im Beruf, Karriereplanung). Entscheidend ist auch hierbei, dass die Leistungsbeurteilung anhand vorab festgelegter, anforderungsanalytisch begründeter Kriterien erfolgt. Die Methoden, anhand derer Leistungsbeurteilungen erfolgen, werden nachfolgend aufgelistet und bei Lohaus und Schuler (2014) ausführlich dargestellt.

Leistungsbeurteilungen

Methoden der Leistungsbeurteilung (nach Lohaus und Schuler 2014, S. 369 ff.)

- Freie Eindrucksschilderung
- Einstufungsverfahren (Beurteilung anhand von idealerweise verhaltensverankerten Ratingskalen)
- Kennzeichnungs- und Auswahlverfahren (Einschätzung, ob vorgegebene Aussagen auf den Beurteilten zutreffen)
- Normorientierte Verfahren (Vergleich von Beurteilten)
- Zielorientierte Verfahren (Beurteilung anhand der Erreichung eines Unternehmens[teil]ziels; z. B. Umsatz)

In vielen Fällen erfolgt die Leistungsbeurteilung durch unmittelbare Vorgesetzte (Hell et al. 2006). Eine besonders elaborierte Form der Leistungsbeurteilung besteht darin, eine Person von unterschiedlichen Personengruppen gleichzeitig bewerten zu lassen. Dies erfolgt beispielsweise im Rahmen von 360-Grad-Feedbacks.

Wahrnehmung der Veränderbarkeit von Merkmalen wichtig

Definition

Als **360-Grad-Feedback** bezeichnet man die systematische Sammlung und Rückmeldung von Leistungseinschätzungen durch verschiedene, relevante Personengruppen (Ward 1997; zitiert nach Morgan et al. 2005). Es erfolgt eine Art „Feedback-Rundumschlag“ – daher der Name 360-Grad-Feedback. Zumeist werden Einschätzungen von Mitarbeitenden, Kolleginnen und Kollegen, Vorgesetzten und Externen (z. B. Kundinnen und Kunden) eingeholt. Zusätzlich geben die Beurteilten in der Regel auch eine Selbsteinschätzung der Leistung ab. Der Nutzen einer Beurteilung aus verschiedenen Perspektiven besteht darin, dass jede Beurteilendengruppe unterschiedliche Eindrücke über die Beurteilte bzw. den Beurteilten beisteuern kann und somit die Beurteilung insgesamt valider wird (vgl. Oh und Berry 2009).

Wie bereits betont, liegt der Fokus von diagnostischen Methoden in der Personalentwicklung sinnvollerweise auf dem Teil der anforderungsanalytisch relevanten Merkmale, der sich innerhalb eines realistischen Zeitraums verändern lässt, also erlernbar ist. Rupp et al. (2006) gehen sogar noch einen Schritt weiter und postulieren (am Beispiel von Entwicklungs-Assessment-Centern), dass Merkmale nicht nur relevant und veränderbar sein sollten, sondern von den zu entwickelnden Personen auch als veränderbar wahrgenommen werden sollten, da dies eine tatsächliche Veränderung wahrscheinlicher macht (Abb. 6.10).

Merkmale, die als durch systematische Maßnahmen veränderbar gelten und von Mitarbeitenden in Unternehmen auch als solche erkannt werden, sind z. B. Kommunikationskompetenz, Planungs- und Organisationsfähigkeit sowie Problemlösen (vgl. Gibbons et al. 2006). Merkmale, die nur schwer zu verändern sind und daher keine Priorität in der Personalentwicklung haben, sind z. B. Gewissenhaftigkeit, Motive, Anpassungsfähigkeit oder allgemeine kognitive Leistungsfähigkeit (vgl. z. B. Rupp et al. 2006).

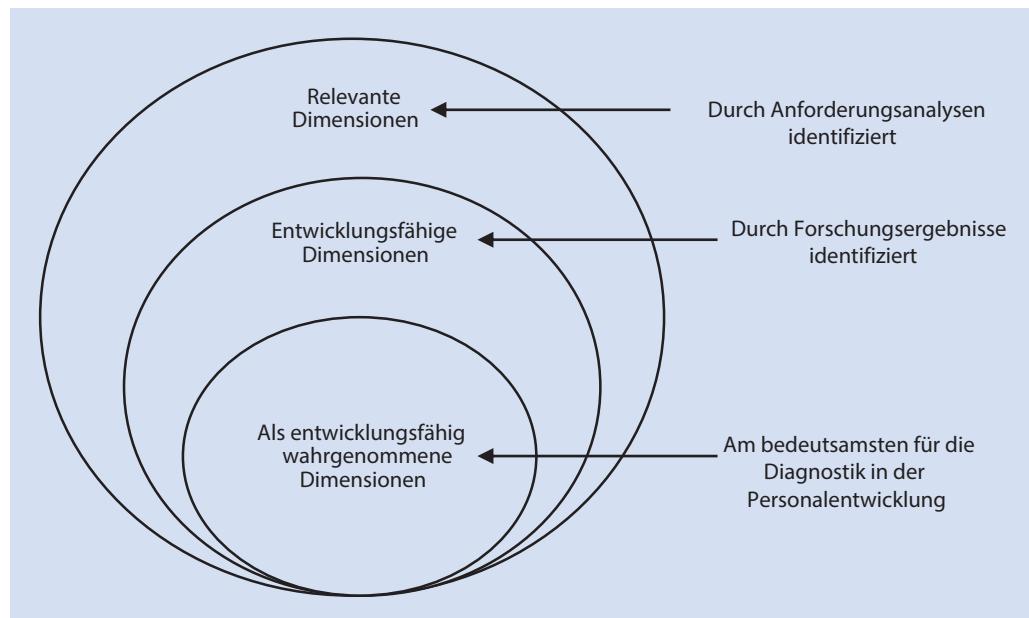


Abb. 6.10 Eingrenzung der für Diagnostik im Rahmen der Personalentwicklung besonders relevanten Dimensionen. (Aus Rupp et al. 2006, S. 86, mit freundlicher Genehmigung der American Psychological Association; Übersetzung durch die Autoren dieses Buchs)

6.2.4 Modifikation von Bedingungen

Der in ► Abschn. 6.2.3 skizzierte Prozess der systematischen Personalentwicklung lässt sich auch zur Modifikation von Bedingungen/Arbeitskontexten anwenden. Dies wird häufig als Organisationsentwicklung bezeichnet. Eine tatsächlich vorhandene oder subjektiv wahrgenommene Soll-Ist-Diskrepanz ist auch hier Ausgangspunkt eines systematischen Abgleichs zwischen Merkmalen der Organisation/des Arbeitskontexts einerseits (s. auch ► Abschn. 6.1) und Merkmalen der Personen andererseits. Im Falle einer Nichtpassung werden nun jedoch, statt Personen zu entwickeln, organisationale Gegebenheiten verändert. Beispielsweise könnten bei zu vielen Fehlern in der Abwicklung von Aufträgen zunächst die Anforderungen dieser Tätigkeit mit den Fähigkeiten und Fertigkeiten der Mitarbeitenden abgeglichen werden. Statt jedoch die Mitarbeitenden zu schulen, könnte sich das Unternehmen auch dazu entschließen, die Tätigkeit so umzugestalten, dass einige der zuvor vorhandenen Anforderungen abgemildert werden. Beispielsweise könnte die Komplexität der verwendeten Software verringert werden (durch Hilfemenüs und intuitive Gestaltung), sodass auch mit weniger EDV-Kenntnissen eine fehlerfreie Abwicklung von Aufträgen möglich ist. Dabei muss natürlich geprüft werden, ob Modifikationen der Arbeitsbedingungen überhaupt möglich sind. Beispielsweise könnten Abläufe einer Tätigkeit durch Gesetze geregelt sein (z. B. der Ablauf einer Kreditberatung bei einer Bank).

Organisationsentwicklung

Beispiele für organisationale Bedingungen, von denen positive Effekte auf die psychische Gesundheit von Arbeitenden zu erwarten sind (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin 2017), sind folgende:

- Handlungsspielraum bei der Arbeit
- Möglichkeit, sich soziale Unterstützung zu holen
- Ganzheitliche Aufgabenbearbeitung
- Angemessene Pausen
- Gleichermaßen mitarbeitenden- und aufgabenorientierte Führung
- Vermeidung von destruktiver Führung
- Strukturierte und transparente Kommunikation durch die Unternehmensleitung
- Begrenzung der wöchentlichen Arbeitszeit auf < 50 h
- Einführung regelmäßiger Pausen
- Möglichst planbare Arbeitszeiten

An dieser Stelle soll auch betont werden, dass berufliche Tätigkeiten nicht nur Anforderungen an Arbeitende stellen, sondern auch Ressourcen bereitstellen. Dies können Möglichkeiten sein, mit konkreten Anforderungen umzugehen (z. B. ein Helpdesk für technische Probleme), oder sie können grundsätzlich eine belastungsmindernde Wirkung haben (z. B. soziale Unterstützung durch Kolleginnen und Kollegen). Das *Job-Demands-Resources-Modell* (s. z. B. Demerouti und Bakker 2011, □ Abb. 6.11) integriert Arbeitsanforderungen und -ressourcen in einem Modell. Dem Modell zufolge können hohe Anforderungen auf Dauer zu Überlastung und gesundheitlichen Problemen führen. Ressourcen, die im Rahmen der beruflichen Tätigkeit bereitgestellt werden, führen jedoch zu höherem Commitment und Arbeitsengagement. Zudem wird angenommen, dass arbeitsbezogene Ressourcen den negativen Effekt von hohen Arbeitsanforderungen abmildern können. Gleichermaßen gilt für hohe Arbeitsanforderungen – sie können den positiven Effekt von arbeitsbezogenen Ressourcen verringern.

Arbeitsbezogene Ressourcen

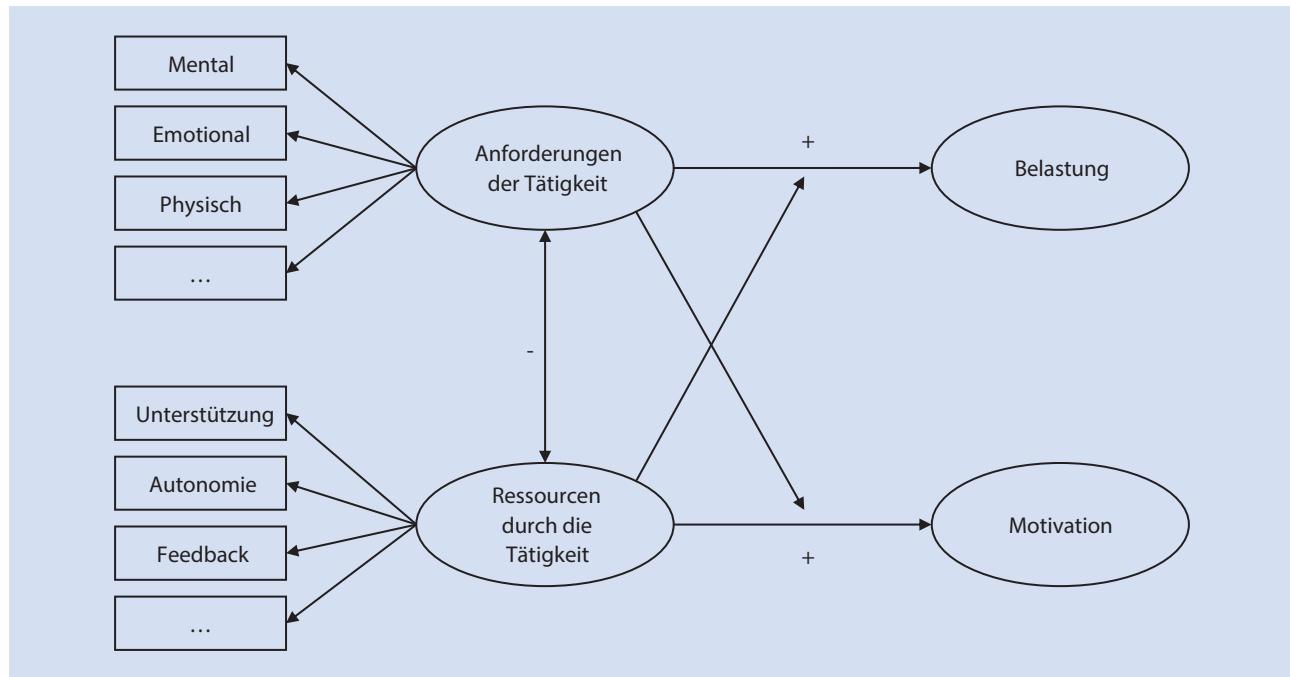


Abb. 6.11 Job-Demands-Resources-Modell. (Aus Demerouti und Bakker 2011, S. 3, lizenziert unter CC BY 4.0; Übersetzung durch die Autoren dieses Buchs)

Allerdings gilt auch hier das Prinzip der Passung zwischen organisationalen Gegebenheiten und Merkmalen der Person. Manche Anforderungen der beruflichen Tätigkeit werden alleine deshalb nicht als beanspruchend erlebt, weil eine Person über die notwendigen Fertigkeiten und Fähigkeiten verfügt. Ebenso fungieren manche Ressourcen nicht als solche, da sie für manche Personen keine Priorität haben (z. B. Arbeitsplatzsicherheit, wenn Personen finanziell abgesichert sind).

6.3 Evaluation der Psychologischen Diagnostik in der Arbeits-, Organisations- und Wirtschaftsprüfung

War die Diagnose der Passung zwischen Merkmalen der Person und der Organisation valide?

Wie in anderen Teildisziplinen der Psychologie muss auch in der Arbeits-, Organisations- und Wirtschaftsprüfung geprüft werden, ob die von ihr proklamierten Maßnahmen zu dem intendierten Resultat führen. Für die hier beschriebenen, diagnostisch relevanten Teilgebiete der Arbeits-, Organisations- und Wirtschaftsprüfung bedeutet dies: War die Diagnose der Passung zwischen Merkmalen der Person und der Organisation zutreffend? Denn nur dann können die darauf aufbauenden Maßnahmen und Entscheidungen zu den gewünschten Resultaten führen.

Dies lässt sich gut anhand unzutreffender Prognosen veranschaulichen: Werden im Rahmen einer Anforderungsanalyse Merkmale als anforderungsrelevant identifiziert, die tatsächlich irrelevant sind, so kann eine nachfolgende Personalauswahl – in der diese Merkmale erfasst werden – kaum dazu führen, dass geeignete Personen ausgewählt werden. Gleichsam wird eine Auswahl anhand von Methoden, die in keinem belegbaren Zusammenhang mit beruflichen Erfolgskriterien stehen (z. B. Grafologie), nicht zu der gewünschten „Bestenauswahl“ führen. Unzutreffende Empfehlungen im Rahmen der Berufs- und Ausbildungswahl führen möglicherweise zur Wahl von

unpassenden Berufen und Ausbildungen. Im Bereich der Personalentwicklung werden sich keine Erfolge einstellen, wenn Personen aufgrund einer nicht zutreffenden Trainingsbedarfsanalyse an einem Training teilnehmen, das sie gar nicht benötigen. Gleiches gilt für Modifikationen von Arbeitskontexten, die eigentlich gar keiner Modifikation bedürfen.

Anhand dieser Negativbeispiele wird auch deutlich, dass unzutreffende Schlüsse in der Diagnostik nicht nur den damit intendierten Erfolg (z. B. Auswahl geeigneter Mitarbeiterinnen und Mitarbeiter) verhindern, sondern auch substantielle negative Folgen haben kann. Falls im Rahmen der Personalauswahl eigentlich geeignete Personen nicht eingestellt werden, werden diese ihre Fähigkeiten möglicherweise bei einem Konkurrenzunternehmen gewinnbringend einsetzen. Trainings und andere Personalentwicklungsmaßnahmen sind meist recht teuer. Noch kostspieliger dürfte es jedoch für viele Unternehmen sein, wichtige Aufgaben (z. B. mit Führungsverantwortung) an nicht genügend geschulte Personen zu übertragen. Werden Arbeitsbedingungen von der Unternehmensleitung aktiv verändert, obwohl sie eigentlich „passend“ waren, wird dies möglicherweise viele Personen in einem Unternehmen frustrieren und demotivieren.

Es folgt nun zunächst ein allgemeiner Überblick über Evaluationsergebnisse im Bereich der Arbeits- und Anforderungsanalyse, der Personalauswahl, der Berufs- und Ausbildungswahl, der Personalentwicklung und der Arbeitsgestaltung bevor die Evaluation der Passung nochmals explizit thematisiert wird.

Invalide Diagnostik kann deutliche Konsequenzen nach sich ziehen

6.3.1 Evaluation von Arbeits- und Anforderungsanalysen

Arbeits- und Anforderungsanalysen werden selten losgelöst von der eigentlichen Personalauswahl evaluiert. Gelingt eine treffsichere Prognose der beruflichen Eignung durch die im Rahmen der Personalauswahl erfassten Merkmale, so wird davon ausgegangen, dass diese Merkmale zutreffende Anforderungen darstellen. Dieser Schluss ist zulässig und sinnvoll. Etwas schwieriger wird es, wenn die im Rahmen der Personalauswahl erfassten Merkmale nur eine mäßig gute oder unzutreffende Prognose der beruflichen Leistung erlauben. In diesem Fall sind keine klaren Schlüsse auf die Validität der Anforderungsanalyse mehr möglich, da suboptimale Prognosen aufgrund ungeeigneter Auswahlmethoden oder einer nicht zutreffenden Anforderungsanalyse (oder aufgrund von beidem) zustande gekommen sein können. Anders und etwas leger ausgedrückt: Entweder es wurden die richtigen Merkmale auf die falsche Art und Weise gemessen oder es wurden die falschen Merkmale gemessen (oder beides).

Arbeits- und Anforderungsanalysen werden formal evaluiert

- ! Viele Evaluationsstrategien im Rahmen der Personalauswahl trennen die Effekte der Anforderungsanalyse und der Personalauswahlmethoden nicht.

Dabei ist es praktisch bedeutsam, Arbeits- und Anforderungsanalysen genauer zu evaluieren. Welche Form der Arbeits- und Anforderungsanalyse liefert unter welchen Randbedingungen valide Ergebnisse? Welche Teilschritte im Rahmen einer Methode der Arbeits- und Anforderungsanalyse sind essenziell, welche optional? Sind personenbezogen-empirische Analysen, die im Rahmen einer publizierten Studie vorgenommen wurden, übertragbar auf

Analyse der Beurteilendenübereinstimmung

den eigenen Kontext? Antworten auf diese Fragen können nicht nur zu einer schrittweisen Optimierung von Arbeits- und Anforderungsanalysemethoden führen, sondern auch das oftmals in der Praxis als aufwendig erachtete Vorgehen verschlanken.

Eine relativ häufige (aber nicht ausreichende) Form der Evaluation von Arbeits- und Anforderungsanalysen ist die Analyse der Beurteilendenübereinstimmung. Die Ergebnisse dieser Analysen sind in 2 Metaanalysen zusammengefasst (Dierdorff und Wilson 2003; Voskuijl und van Sliedregt 2002). Die in der Metaanalyse von Dierdorff und Wilson (2003) berichteten Übereinstimmungen können als moderat bis gut bezeichnet werden (zwischen .61 und .77), die von Voskuijl und van Sliedregt (2002) als moderat (.59). Eine geringe Übereinstimmung der Beurteilenden muss aber nicht automatisch auf invalide Urteile hinweisen. Möglicherweise haben 2 oder mehr Beurteilende unterschiedliche Ausschnitte der gleichen beruflichen Tätigkeit vor Augen, wenn sie Urteile über Anforderungen abgeben. In solchen Fällen würde mehrere, nicht übereinstimmende Urteile gemeinsam ein besseres Bild der Anforderungen abgeben als übereinstimmende Urteile, die nur einen Teil der Anforderungen abbilden. So fanden Koch et al. (2012), dass Mitarbeitende und Vorgesetzte im Rahmen von Critical-Incident-Interviews (vgl. ► Abschn. 6.1.2 unterschiedliche Ereignisse beisteuerten und somit erst durch die Betrachtung aller Perspektiven ein vollständiges Bild der beruflichen Tätigkeit entstand.

Ein Teil der Evaluationsstudien zu Arbeits- und Anforderungsanalysen hat sich damit beschäftigt, inwiefern die Prognosekraft von Auswahlverfahren steigt, wenn diese auf zuvor durchgeföhrten Anforderungsanalysen basieren. McDaniel et al. (2001) untersuchten im Rahmen einer Metaanalyse zur Validität von *Situational-Judgment-Tests*, ob deren Kriteriumsvalidität sich in Abhängigkeit von der anforderungsanalytischen Methode unterschied. Dabei verglichen sie Situational-Judgment-Tests, deren Inhalte von der Testautorin bzw. dem Testautor oder von einer kleinen Gruppe von Fachleuten intuitiv festgelegt wurden, mit Situational-Judgment-Tests, denen sorgfältig durchgeföhrte Critical-Incident-Analysen zugrunde lagen. Die Ergebnisse der Metaanalyse zeigen einen deutlichen Vorteil des anforderungsanalytischen Vorgehens (für Streuungseinschränkung und Unreliabilität des Kriteriums korrigierte Validität von .38 vs. .29). Eine etwas ältere Studie von Wiesner und Cronshaw (1988) kann den Vorteil einer formalen Anforderungsanalyse gegenüber einem intuitiven Vorgehen auch für *strukturierte Interviews* bestätigen – hier liegt die für Streuungseinschränkung und Unreliabilität des Kriteriums korrigierte Validität bei .87 (bei vorheriger, formaler Anforderungsanalyse) bzw. bei .59 (bei intuitiver Anforderungsanalyse). Die bei Wiesner und Cronshaw (1988) berichteten Validitätskoeffizienten sind erstaunlich hoch – aussagekräftig dürfte hier vor allem der Vergleich der beiden Bedingungen (formale vs. intuitive Anforderungsanalyse) sein.

Eine weitere Form der Evaluation von Arbeits- und Anforderungsanalysen besteht in der Identifikation möglicher Fehlerquellen. Morgeson und Campion (1997) fassen diese in einer umfangreichen Übersichtsarbeit zusammen:

Anforderungsanalyse erhöht Kriteriumsvalidität von Verfahren

Fehlerquellen bei Anforderungsanalysen nach Morgeson und Campion (1997) mit ihrer Beschreibung (Übersetzung durch die Autoren dieses Buchs)

- Soziale Quellen:
 - Prozesse der sozialen Beeinflussung:
 - Konformitätsdruck: Tendenz, sich der Meinung der Mehrheit/Autoritäten anzuschließen
 - Gruppenpolarisierung: Tendenz, dass Gruppenentscheidungen extremer ausfallen
 - Motivationsverlust: Motivationsverlust aufgrund der Anwesenheit anderer
 - Prozesse der Selbstdarstellung:
 - Eindrucksmanagement: Tendenz, sich selbst positiv darzustellen
 - Soziale Erwünschtheit: Tendenz, sich selbst entsprechend der (angenommenen) sozialen Normen zu verhalten
 - Demand-Effekt: Tendenz, sich entsprechend der Erwartungen des Gegenübers zu verhalten
- Kognitive Quellen:
 - Einschränkungen in der Informationsverarbeitung:
 - Informationsüberlastung: verminderte Verarbeitung bei einem Zuviel an Informationen
 - Heuristiken: Verwendung vereinfachter Denkstrategien
 - Kategorisierung: Zusammenfassen von Informationen anhand markanter Merkmale
 - Verzerrungen in der Informationsverarbeitung:
 - Nachlässigkeit: vorsätzliches Nichtbeachten gewisser Informationen
 - Irrelevante Informationen: Beachten irrelevanter Informationen
 - Inadäquate Informationen: Beachten von inadäquaten, unvollständigen Informationen
 - Reihenfolge- und Kontrasteffekte: Urteilsverzerrungen bedingt durch die Reihenfolge der dargebotenen Informationen und den Vergleich von zufällig aufeinanderfolgenden Informationen
 - Halo-Effekt: Überbewertung einzelner Merkmale, die das Gesamтурteil verzerrn
 - Milde-/Strengefehler: eine generelle Tendenz zu milden/strengen Urteilen
 - Methodeneffekte: Verzerrungen durch die Verwendung von ausschließlich einer Methode

Fazit Arbeits- und Anforderungsanalysen identifizieren Merkmale der Organisation, die für alle hier dargestellten Varianten der Passungsbeurteilung genutzt werden. Somit nehmen sie eine zentrale Rolle ein. Kontinuierliche Evaluationen sollten vorgenommen werden, um der zentralen Bedeutung von Arbeits- und Anforderungsanalysen gerecht zu werden.

6.3.2 Evaluation von Personalauswahlverfahren

Auch für die Evaluation von Personalauswahlverfahren gilt, dass der Einfluss der Verfahren selten vom Einfluss der Anforderungsanalyse getrennt wird. So können manche Verfahren alleine deshalb als weniger relevant für berufliche Kriterien identifiziert werden, weil sie häufig ohne vorherige Anforderungsanalyse durchgeführt werden oder zumindest nicht maßgeschneidert für Ergebnisse einer spezifischen Anforderungsanalyse entwickelt wurden (z. B. breite Persönlichkeitstests). Für andere Verfahren ist es hingegen

Personalauswahlverfahren müssten eigentlich gemeinsam mit der Anforderungsanalyse evaluiert werden

■ **Tab. 6.13** Prognosegüte ausgewählter Verfahren zur Vorhersage von Ausbildungs- und Berufserfolg nach Schmidt und Hunter (1998, mit freundlicher Genehmigung der American Psychological Association)

Verfahren	Ausbildungserfolg	Berufserfolg
Kognitive Leistungstests	.56	.51
Arbeitsproben	—	.54
Integritätstests	.38	.41
Gewissenhaftigkeitsfragebögen	.30	.31
Berufswissenstests	—	.48
Interview, strukturiert	.35	.51
Interview, unstrukturiert	.35	.38
Persönlichkeitsfragebögen	.30	.31
Assessment-Center	—	.37
Biografische Fragebögen	.30	.35
Referenzen	.23	.26
Grafologie	—	.02

„Meta-Metaanalyse“ von Schmidt und Hunter (1998)

Zentrale Bedeutung kognitiver Leistungstests

Unter Fragebögen wiesen solche zu Integrität und Gewissenhaftigkeit die höchsten Korrelationen mit beruflicher Leistung auf

üblich, diese auf Basis einer vorherigen Anforderungsanalyse „maßgeschneidert“ zu entwickeln und anzuwenden (z. B. strukturierte Interviews).

Über viele berufliche Kontexte hinweg haben sich einige Auswahlverfahren als geeignet erwiesen, berufsbezogene Leistung und Ausbildungserfolg vorherzusagen. Die in ■ Tab. 6.13 aufgeführten Koeffizienten stammen aus zahlreichen Metaanalysen, die von Schmidt und Hunter (1998) zu einer „Meta-Metaanalyse“ zusammengeführt wurden. Es zeigt sich darin eine hohe Relevanz von kognitiven Leistungstests, Arbeitsproben, Integritäts- tests, strukturierten Interviews und Tests zur Erfassung von berufsrelevantem Wissen für die statistische Vorhersage von Berufserfolg. Kognitive Leistungstests erwiesen sich als besonders relevant zur statistischen Vorhersage des Ausbildungserfolgs. Arbeitsproben, Fragebögen zur Gewissenhaftigkeit und strukturierte Interviews können laut den Ergebnissen dieser Metaanalyse ergänzend zu kognitiven Leistungstests eingesetzt werden, sodass ein substanzialer Zuwachs der Prognosegüte erreicht wird.

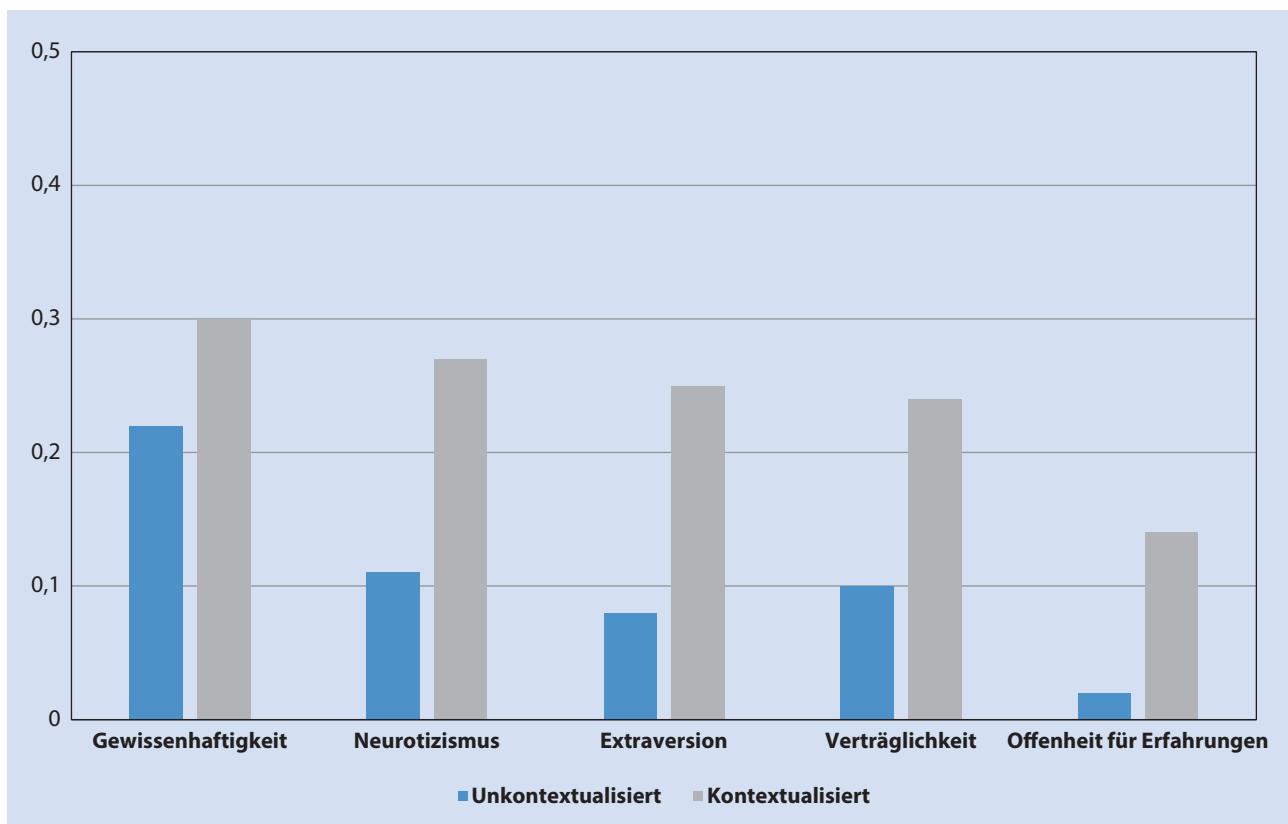
Übrigens: Die zentrale Bedeutung kognitiver Leistungstests konnte auch in Metaanalysen, die nur europäische (Salgado et al. 2003) oder nur deutsche Studien (Kramer, 2009) inkludierten, bestätigt werden (► Tab. 2.21). In einem neueren Working Paper präsentieren Schmidt et al. (2016) ein Update der Metaanalyse von Schmidt und Hunter (1998), in dem sie die meisten Erkenntnisse der ursprünglichen Metaanalyse bestätigen. Ein Fazit ist, dass Berufserfolgsmaße recht gut durch allgemeine kognitive Leistungstests und ein ergänzendes strukturiertes Interview vorhergesagt werden können. Als interessanten neuen Befund kann man dem Working Paper entnehmen, dass auch eine Kombination aus allgemeinen kognitiven Leistungstests und strukturiertem Telefoninterview gut „funktioniert“.

Andere Verfahrenstypen, die sich gemäß der Metaanalyse von Schmidt und Hunter (1998) als geeignet zur Prädiktion beruflicher Leistungen erwiesen haben, wurden bereits in ► Abschn. 3.7 und ► Abschn. 6.2.1.1 thematisiert. Daher soll hier nur noch kurz auf spezielle Aspekte des Einsatzes von Fragebögen in der Personalauswahl eingegangen werden. In der Metaanalyse von Schmidt und Hunter (1998) wurden nur die Fragebögen aufgenommen, die die Merkmale „Integrität“ und „Gewissenhaftigkeit“ adressieren. Dafür gab es einen einfachen Grund: Für diese Merkmale wurden

die höchsten Zusammenhänge zu beruflichen Erfolgskriterien gefunden. Eine weitere Metaanalyse befasste sich mit der Prädiktion von Berufs- und Ausbildungserfolg durch Persönlichkeitsfragebögen (Schmidt et al. 2008) – bei Anwendung neuerer Korrekturformeln für die Streuungseinschränkung (► Abschn. 2.6.3.4). Auch in dieser Metaanalyse erzielt Gewissenhaftigkeit die höchste Vorhersage von Ausbildungs- und Berufserfolg. Die berichteten Zusammenhangsmaße liegen für Fragebögen zu breiten Persönlichkeitsmerkmalen (z. B. Gewissenhaftigkeit) jedoch deutlich unter denen allgemeiner kognitiver Leistungstests (z. B. .31 vs. .51 bei Schmidt und Hunter 1998).

Eine Möglichkeit, die Prognosegüte von Persönlichkeitsfragebögen zu erhöhen, besteht darin, sie ein wenig mehr zu „kontextualisieren“. Im Vergleich zu kontextfreien Aussagen wie „Ich erledige Aufgaben gründlich“, die so typischerweise in Persönlichkeitsfragebögen enthalten sind, kann eine leicht kontextualisierte Variante wie folgt lauten: „Ich erledige Aufgaben *im Beruf* gründlich.“ Alternativ könnte über allen Aussagen stehen „Im Beruf...“ und dann die Fortsetzung z. B. lauten „... erledige ich Aufgaben gründlich“. Eine Metaanalyse von Shaffer und Postlethwaite (2012) zeigt, dass durch diesen Bezugsrahmen (frame of reference) die prädiktive Validität von Fragebögen zur Gewissenhaftigkeit (aber auch von allen anderen breiten Persönlichkeitsmerkmalen) substanzell steigt – z. B. von .22 auf .30 für Gewissenhaftigkeit (korrigiert für Streuungseinschränkung und Unreliabilität des Kriteriums). Für andere breite Persönlichkeitsmerkmale fiel der Anstieg sogar noch deutlicher aus (► Abb. 6.12).

Fragebögen mit Bezugsrahmen



► Abb. 6.12 Metaanalytische Unterschiede von kontextualisiert und unkontextualisiert erhobenen Persönlichkeitsdimensionen hinsichtlich ihres Zusammenhangs zu beruflichen Leistungskriterien. (Nach Shaffer und Postlethwaite 2012, S. 460, mit freundlicher Genehmigung von John Wiley and Sons)

Verfälschbarkeit beachten

Eine Besonderheit des Einsatzes von Fragebögen in der Personalauswahl ist zudem, dass Unternehmen von Bewerbenden keine „ehrlichen“ Antworten erwarten dürfen (vgl. Ziegler et al. 2011). Vielmehr werden Bewerbende versuchen, sich möglichst positiv darzustellen („faking good“, ▶ Abschn. 2.6.5.4). Dies ist bei Fragebögen leicht möglich und kaum durch Maßnahmen der Testgestaltung zu vermeiden. Eine Empfehlung besteht jedoch darin, ein Forced-Choice-Antwortformat (▶ Abschn. 2.4.2.6) zu verwenden (Brown und Maydeu-Olivares 2013). Darüber hinaus kann ein multimethodales Vorgehen helfen, bei dem die Ergebnisse eines oder mehrerer Fragebögen nicht allzu stark gewichtet werden und in der nächsten Runde des Auswahlverfahrens durch andere Methoden (z. B. ein strukturiertes Interview) abgesichert werden.

6

6.3.3 Evaluation von diagnostischen Verfahren zur Berufs- und Ausbildungswahl

Zusammenhänge von berufsbezogenen Interessenskalen zu beruflicher Leistung

In ▶ Abschn. 6.2.2 wurde besonders auf berufliche Interessen als diagnostischen Zugang zur Berufs- und zur Ausbildungswahl eingegangen. Hinsichtlich der Evaluation selbiger liegt mittlerweile eine einschlägige Metaanalyse von Van Iddekinge et al. (2011b) vor. In dieser Metaanalyse wurden insgesamt 74 Studien zusammengefasst und der Zusammenhang von Interessentests zu beruflicher Leistung, Ausbildungserfolg, Berufswechselabsichten und tatsächlichem Berufswechsel ermittelt. Es finden sich durchaus substantielle Zusammenhänge von berufsbezogenen Interessenskalen zu beruflicher Leistung ($\rho=.23$) und Ausbildungserfolg ($\rho=.29$; jeweils nur korrigiert für den Messfehler in den Kriteriumsmaßen). Schmidt et al. (2016) nutzen diese Metaanalyse, um den inkrementellen Beitrag von Interessentests über allgemeine kognitive Leistungstests hinaus zu prüfen. In der Tat finden sie einen inkrementellen Beitrag von $\Delta R=.06$. Zur Einordnung: Dies ist der viert-höchste inkrementelle Beitrag von insgesamt 30 Prädiktoren bei Schmidt et al. (2016).

Das metaanalytische Ergebnis deckt sich auch mit einer Einzelstudie von Van Iddekinge, Putka und Campbell (2011a). Sie untersuchten den Zusammenhang zwischen Berufsinteressen nach dem Holland-Modell (s. dazu die Ausführungen zum Interessentest EXPLORIX in ▶ Abschn. 3.3.3.7) und verschiedenen Kriterien des Berufserfolgs bei über 400 amerikanischen Soldatinnen und Soldaten. Am besten konnte technisches Wissen, das die Soldatinnen und Soldaten in ihrer Ausbildung erworben hatten, durch die 6 Interessenskomponenten erklärt werden; die multiple Korrelation betrug .46 (korrigiert für Varianzeinschränkung im Prädiktor und Reliabilität des Kriteriums). Aber auch die Leistungsbeurteilung durch die Vorgesetzten konnte mit einer (korrigierten) multiplen Korrelation von $R=.27$ erklärt werden. Bemerkenswert ist, dass bestimmte Interessen durch niedrige Werte zur Vorhersage beitragen; soziale und konventionelle Interessen (ordnend/verwaltend) korrelierten negativ mit dem technischen Wissen. Ausgeprägte Interessen in einigen und niedrige Interessen in anderen Bereichen ergeben zusammen ein Interessenprofil, das den Berufserfolg partiell erklärt. Ein weiterer Befund – in Einklang mit der Metaanalyse von Schmidt et al. (2016) – ist, dass die Interessen bei allen Kriterien des Berufserfolgs eine inkrementelle Validität gegenüber der kognitiven Leistungsfähigkeit und der Persönlichkeitsmerkmale aufwiesen. Die zusätzliche Varianzaufklärung durch Berufsinteressen lag zwischen 5 und 9 %.

Fazit Insgesamt sollten diese Ergebnisse dazu ermutigen, Berufsinteressen stärker in der Eignungsdiagnostik zu berücksichtigen. Besonders in der Beratung bei Berufsfeldentscheidungen besteht kaum die Gefahr der Verfälschung, sodass den Ergebnissen eines Interessentests in der Regel vertraut werden kann.

6.3.4 Evaluation von diagnostischen Verfahren zur Feststellung des Personalentwicklungsbedarfs

Der Fokus von Evaluationsstudien in der Personalentwicklung liegt auf der Einschätzung der Wirksamkeit von Interventionen (z. B. von verschiedenen Trainings; Aguinis und Kraiger 2009). Sofern diagnostische Verfahren zur Feststellung des Personalentwicklungsbedarfs (s. Trainingsbedarfsanalyse; ▶ Abschn. 6.2.3) eingesetzt werden, die auch in der Personalauswahl häufig zum Einsatz kommen, liegen oft auch umfangreiche Erkenntnisse zu deren Zusammenhang mit anderen Konstrukten vor. Diese Erkenntnisse sollten ebenfalls für den Kontext der Personalentwicklung nützlich sein.

Erkenntnisse für
Personalauswahlverfahren zum
Teil auf Personalentwicklung
übertragbar

! **Aber:** Für Verfahren, die im Rahmen der Personalentwicklung eingesetzt werden, reicht es nicht aus, dass sie in einem engen Zusammenhang mit beruflicher Leistung stehen. Vielmehr muss belegt sein, dass sie tatsächlich das intendierte Konstrukt messen. Andernfalls ist eine Entscheidung darüber, welche Fertigkeiten oder Kompetenzen trainiert werden sollen, nicht möglich.

Messanspruch ist zentral

Die Begründung für diese Aussage ist einfach: Man stelle sich vor, viele Mitarbeiterinnen und Mitarbeiter haben ein Verfahren zur Messung der Teamfähigkeit (beispielsweise einen Situational-Judgment-Test) absolviert. Personen mit niedrigen Werten in diesem Verfahren werden dann verpflichtet, ein Training zur Optimierung der Teamfähigkeit zu absolvieren. Wenn es aber nun so wäre, dass das Verfahren fälschlicherweise Intelligenz statt, wie beabsichtigt, Teamfähigkeit messen würde, so wäre der mit den Trainings verbundene Aufwand nutzlos. Es ist daher für diagnostische Verfahren zur Feststellung des Personalentwicklungsbedarfs stets abzusichern, dass diese ihrem Messanspruch gerecht werden.

Unzureichende Evaluation

Für diagnostischen Verfahren, die kaum oder nie in der Personalauswahl sondern hauptsächlich zur Feststellung des Personalentwicklungsbedarfs eingesetzt werden (z. B. Entwicklungs-Assessment-Center, 360-Grad-Feedbacks), stehen zumeist nur wenige einschlägige Validierungsstudien zur Verfügung. Dies verwundert insofern, als jede Intervention im Rahmen der Personalentwicklung nur dann von Nutzen für eine Organisation ist, wenn sie zum einen anforderungsanalytisch als relevant abgeleitete Merkmale fördert und zum anderen genau die Personen fördert, die nicht in ausreichendem Ausmaß über die trainierten Merkmale verfügen – was vorab treffsicher durch diagnostische Instrumente festgestellt werden sollte. Für Entwicklungs-Assessment-Center können Erkenntnisse der Assessment-Center-Forschung übertragen werden, sofern die allgemeinen Gestaltungsprinzipien gleich sind. Beispielsweise zeigte eine Metaanalyse von Woehr et al. (2007), dass auch für Entwicklungs-Assessment-Center die mittlere konvergente Validität unter der mittleren diskriminanten Validität liegt. Also bleibt auch bei Entwicklungs-Assessment-Centern unklar, was sie genau messen und – daraus folgend – welche spezifischen Trainings die Teilnehmenden absolvieren sollen.

Entwicklungs-Assessment-Center

Auch für Entwicklungs-Assessment-Center, die explizit von den üblichen Gestaltungsprinzipien von Assessment-Centern abweichen, fehlen derzeit einschlägige Erkenntnisse zu den damit gemessenen Konstrukten. Typische

Abweichung sind Übungswiederholungen, Feedback zwischen den Übungen und theoretischer Input im Rahmen der Durchführung. Diese Modifikationen spiegeln den Anspruch wider, dass Entwicklungs-Assessment-Center nicht „nur“ diagnostische Instrumente, sondern gleichzeitig auch Trainingsmaßnahmen darstellen sollten (Obermann 2009; Stangel-Meseke et al. 2005). Es ist zu befürchten, dass dieser Anspruch der klaren Messung eines oder mehrerer Konstrukte abträglich ist. Bislang stehen unseres Wissens nach keine Studien zur Verfügung, die sich mit der Auswirkung einzelner Trainingselemente auf die Validität von Entwicklungs-Assessment-Centern beschäftigen.

Eine weitere in der Personalauswahl beliebte Auswahlmethode muss hinsichtlich des Einsatzes im Rahmen der Personalentwicklung ebenfalls kritisch betrachtet werden. Strukturierte Eignungsinterviews (vgl. ▶ Abschn. 3.7) erwiesen sich in einer umfangreichen Studie von Van Iddekinge et al. (2004) als wenig übereinstimmend mit den intendierten Konstrukten. Die Autoren konnten im Rahmen von Multitrait-Multimethod-Analysen zeigen, dass nur ein geringer Anteil der Gesamtvarianz durch die zu messenden Merkmale erklärt wird. Auch Melchers et al. (2004) konnten zeigen, dass die mittlere konvergenten Korrelationen über mehrere Interviewteile mit gleichem Messanspruch geringer ausfielen (mittleres $r = .20$) als die mittleren diskriminanten Korrelationen (mittleres $r = .30$). Ähnlich wie bei der Assessment-Center-Methode bestehen also Zweifel an der Konstruktvalidität. Auch wenn wir wissen, dass mit strukturierten diagnostischen Interviews der Berufserfolg gut vorhergesagt werden kann, so wissen wir dennoch nicht genau, was dabei überhaupt gemessen wird.

Zur Validität von Leistungsbeurteilungen existiert eine ältere Metaanalyse von Dickinson et al. (1986), die Leistungsbeurteilungen moderate mittlere konvergente Validitätskoeffizienten ($r = .36$) bei diskriminanten Validitätskoeffizienten von $r = .13$ bescheinigt. Diese Metaanalyse zeigte auch, dass sich die konvergente Validität durch Maßnahmen wie die systematische Entwicklung der Beurteilungsskalen oder die Nutzung von verhaltensverankerten Ratingskalen (▶ Abschn. 3.6) steigern lässt. Conway (1996) konnte in einer etwas neueren Metaanalyse ebenfalls zeigen, dass ein substanzielles Anteil der Varianz der Leistungsbeurteilungen (abgegeben durch mehrere Beurteilende) auf die beurteilten Dimensionen zurückzuführen ist. Insgesamt ist sich die Literatur zu Leistungsbeurteilungen einig, dass sich das Zusammenführen verschiedener Beurteilendenperspektiven – beispielsweise im Rahmen von 360-Grad-Feedbacks – förderlich auf deren Validität auswirkt (vgl. Morgeson et al. 2005; Oh und Berry 2009).

6.3.5 Evaluation der Diagnostik von individuellen Merkmalen zum Zwecke der Modifikation von Arbeitskontexten/Arbeitsgestaltung

Die Diagnostik von individuellen Merkmalen zum Zwecke der Modifikation von Arbeitskontexten dient zur Prüfung, ob Belastungspotenziale (wie etwa Zeitdruck, Intensität der Tätigkeit), vorherrschende Führungsstile oder vom Unternehmen vermittelte Werte (z. B. Nachhaltigkeit) einen guten Fit zu den im Unternehmen arbeitenden Menschen aufweisen. Belege für die Sinnhaftigkeit eines Abgleichs von Personen- und Arbeitsmerkmalen – mit dem Ziel, durch Arbeitsgestaltungsmaßnahmen eine Passung herzustellen – gibt es viele. Beispielsweise konnten Cable und DeRue (2002) zeigen, dass die Passung der Werte von Personen zu denen der Organisation (Person-Organisation-Fit) in deutlichem Zusammenhang mit dem Commitment der Personen zur Organisation (gemessen 1 Jahr nach Erhebung der

Passung) steht. Diese Autoren fanden zudem, dass die Passung zwischen den Bedürfnissen einer Person und den Möglichkeiten der Bedürfnisbefriedigung durch die Arbeit (Needs-Supply-Fit) einen deutlichen Zusammenhang mit der Arbeitszufriedenheit aufwies (ebenfalls 1 Jahr später gemessen). Weitere Belege für die Relevanz einer Passung zwischen Personen- und Arbeitsmerkmalen liefern Kristof-Brown et al. (2005) im Rahmen einer umfangreichen Metaanalyse. Diese zeigt, dass verschiedene Formen der Passung (z. B. der Bedürfnisse und Belohnungen, der Werte von Personen und denen der Organisation) einen substanzuellen und positiven Zusammenhang mit Arbeitszufriedenheit und Commitment sowie einen negativen Zusammenhang mit Kündigungsabsichten und Belastung haben.

6.3.6 Messung der Auswirkung von „Passung“

Wie eingangs und über das gesamte Kapitel hinweg dargestellt, ist Passung (von Personen und Arbeitsbedingungen) die Grundlage der Diagnostik und darauf aufbauender Maßnahmen in der Arbeits- und Organisationspsychologie. Metaanalytische Ergebnisse zeigen, dass bei vorhandener Passung die Wahrscheinlichkeit steigt, dass Mitarbeitende gute Leistungen erbringen und zufrieden sind.

Dabei ist zu beachten, dass stets ein Zusammenhang zwischen 3 Variablen impliziert ist, nämlich zwischen

1. dem Merkmal der Organisation,
2. dem Merkmal der Person und
3. der abhängigen Variablen, z. B. Arbeitsleistung oder Zufriedenheit.

Überwiegend werden jedoch nur 2 Variablen betrachtet. Wie zuvor dargestellt erfolgt die Evaluation der Personalauswahl zumeist durch das Quantifizieren des Zusammenhangs zwischen dem Personenmerkmal und der abhängigen Variablen. Dabei wird davon ausgegangen, dass die zuvor erfolgte Anforderungsanalyse valide Ergebnisse erbracht hat und dass die Anforderungen über die inkludierte Stichprobe ohnehin nicht variieren, weil beispielsweise nur eine Gruppe Auszubildender (z. B. Gebäudereinigende) betrachtet wird.

Sobald Merkmale der Organisation und der Person innerhalb einer Studie bzw. einer Stichprobe Varianz aufweisen, ist eigentlich eine gemeinsame Betrachtung aller Variablen (Organisation, Person und abhängige Variable) erforderlich. Dies kann beispielsweise bei der Arbeitsgestaltung relevant sein: Personen in einer Organisation haben unterschiedliche Arbeitsplätze (Merkmale der Tätigkeit variieren zwischen Arbeitsplätzen) und unterschiedliche arbeitsbezogene Bedürfnisse (die Merkmalsausprägung variiert zwischen Personen). Wenn geprüft werden soll, ob die Passung zwischen beiden mit höherer Arbeitszufriedenheit einhergeht, reichen bivariate Analysen nicht aus.

Eine einfache, aber unangemessene Form, damit umzugehen, besteht in der Bildung von Differenzwerten. Das heißt, man könnte von der Ausprägung des Organisationsmerkmals die Ausprägung des Personenmerkmals abziehen und die so resultierenden Differenzwerte mit der abhängigen Variablen in Beziehung setzen. Allerdings ist dieses Vorgehen problematisch (vgl. Edwards 2002).

Bivariate Analysen reichen nicht aus

Probleme von Differenzwerten

Probleme von Differenzwerten (nach Edwards 2002)

- Differenzwerte weisen zumeist eine geringere Reliabilität als die Einzelwerte auf.
- Wenn ein Wert eine größere Varianz als der andere hat, so wird die Varianz des Differenzwertes stärker von ersterem beeinflusst. Damit wird weniger klar, was der Differenzwert konzeptuell repräsentiert.
- Das Ausmaß, in dem die Einzelwerte den Effekt des Differenzwertes auf eine abhängige Variable beeinflussen, ist nicht empirisch zu bestimmen.
- Ein dreidimensionaler Zusammenhang (also zwischen 2 unabhängigen und 1 abhängigen Variablen) wird zu einem zweidimensionalen (zwischen dem Differenzwert und der abhängigen Variablen) reduziert.

Edwards (1993) empfiehlt stattdessen, beide Werte getrennt zu nutzen und polynomiale Regressionsanalysen durchzuführen. Durch polynomiale Regressionsanalysen können alle Formen des dreidimensionalen Zusammenhangs zwischen dem Merkmal der Organisation, der Person und der abhängigen Variablen dargestellt werden.

Polynomiale Regressionsgleichung

$$Z = b_0 + b_1 \times X + b_2 \times Y + b_3 \times X^2 + b_4 \times X \times Y + b_5 \times Y^2$$

Z = abhängige Variable (z. B. Arbeitszufriedenheit)

X = Prädiktor 1 (z. B. Merkmal der Organisation)

Y = Prädiktor 2 (z. B. Merkmal der Person)

Dabei wird im Rahmen einer hierarchischen Regressionsanalyse geprüft, ob durch Hinzunahme des Interaktionsterms ($X \times Y$) eine signifikant höhere Varianzaufklärung des Kriteriums resultiert als durch Verwendung der beiden einzelnen Prädiktoren (Edwards und Parry 1993, S. 1579).

Response-Surface-Analysen

Eine Visualisierung des dreidimensionalen Zusammenhangs kann über Response-Surface-Analysen vorgenommen werden, in denen mithilfe der im Rahmen der polynomialen Regression ermittelten Regressionsgewichte und der Konstante b_0 Prädiktionswerte für Z in Abhängigkeit von X und Y ermittelt und grafisch abgetragen werden (Abb. 6.13).

Abb. 6.13 zeigt, wie eine Response-Surface-Analyse aussehen könnte, wenn Passung zu hohen Werten in der abhängigen Variable (in diesem Fall der Arbeitszufriedenheit) und Nichtpassung zu niedrigen Werten führt. In dem dreidimensionalen Koordinatensystem ist eine vollständige Passung zwischen dem Merkmal der Organisation und der Person entlang der mit ① und ② gekennzeichneten vertikalen Linie gegeben. Während an Punkt ① gleich hohe Werte der Organisation und der Person vorliegen, sind an Punkt ② die Werte von Organisation und Person gleichermaßen gering. Die Nichtpassung variiert entlang der mit ③ und ④ gekennzeichneten horizontalen Linie von innen nach außen. Das heißt, in der Nähe des Schnittpunkts der beiden Linien ist die Nichtpassung gering, an den mit ③ und ④ gekennzeichneten Enden groß. Bei ③ steht einer sehr hohen Ausprägung des Personenmerkmals eine sehr geringe Ausprägung des Organisationsmerkmals gegenüber. Bei ④ ist die Ausprägung des Personenmerkmals sehr gering, die

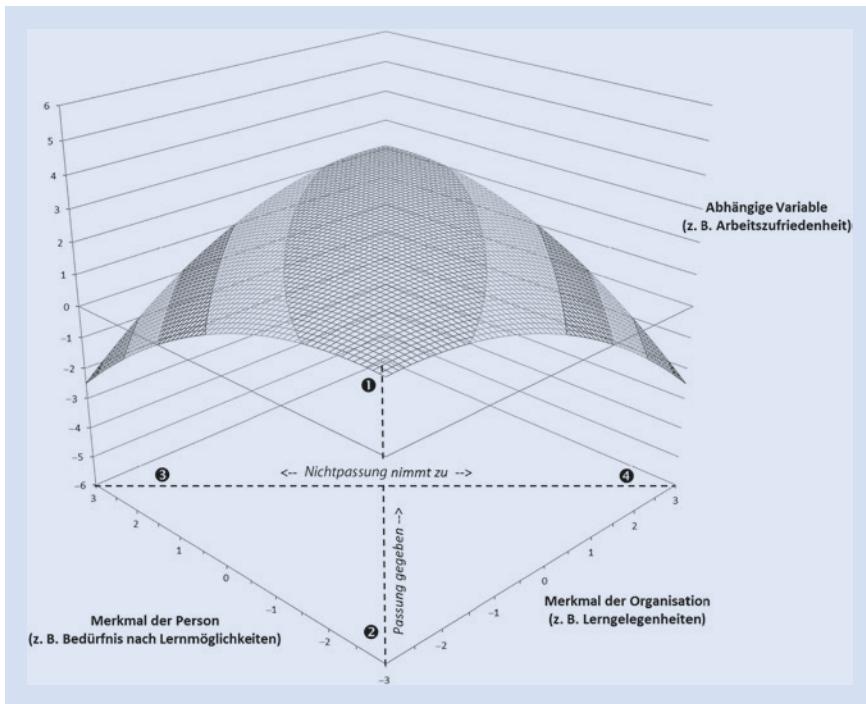


Abb. 6.13 Beispiel für eine Response-Surface-Analyse. (Modifiziert nach Krumm et al. 2013, mit freundlicher Genehmigung von Taylor & Francis Ltd, ► <https://www.tandfonline.com>)

des Organisationsmerkmals sehr groß. Die netzartig dargestellte Oberfläche im Koordinatensystem kennzeichnet die vorhergesagten Werte der abhängigen Variablen. Man sieht, dass sich die Oberfläche bei großer Nichtpassung, also bei ③ und ④, nach unten neigt. Bei vollständiger Passung, also entlang der vertikalen Linie, bildet sie ein Plateau.

Die Oberfläche nimmt eine völlig andere Gestalt an, wenn nicht die Passung der beiden Merkmale, sondern jedes Merkmal für sich unabhängig vom anderen die abhängige Variable prädiziert (Abb. 6.14). Nun sieht man, dass bereits hohe Ausprägungen des einen ③ oder des anderen ④ Merkmals zu hohen Werten in der abhängigen Variablen führen. Die höchsten Werte nimmt die abhängige Variable an, wenn beide Prädiktoren hoch ausgeprägt sind ①.

Die in Abb. 6.15 dargestellte Oberfläche würde man erwarten, wenn nur eine Form der Nichtpassung bedeutsam ist. In diesem Fall sind nur dann geringe Werte der abhängigen Variablen zu beobachten, wenn die Ausprägung des Organisationsmerkmals über der des Personenmerkmals liegt ④. Liegt die Ausprägung des Personenmerkmals über der des Organisationsmerkmals ③, hat dies keine Auswirkung auf die abhängige Variable. Dies könnte der Fall sein, wenn Anforderungen eines Arbeitsplatzes die Fähigkeiten von Personen übersteigen ④. Dann wären negative Auswirkungen auf die Arbeitsleistung zu erwarten. Verfügen Personen aber über größere Fähigkeiten als durch den Arbeitsplatz verlangt ③, was auch eine Form von Nichtpassung darstellt, wirkt sich dies nicht negativ auf die Leistung aus.

Eine weitere Möglichkeit, die statistische Vorhersage durch Merkmale der Organisation und der Person auf eine abhängige Variable gemeinsam zu

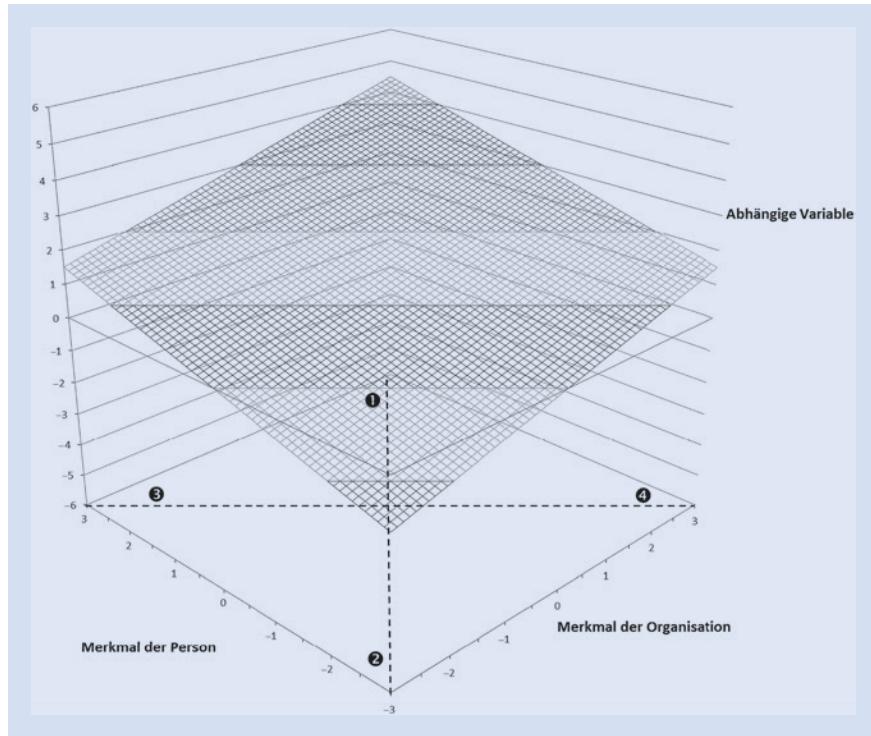


Abb. 6.14 Beispiel für eine Response-Surface-Analyse, wenn die Passung zwischen 2 Variablen nicht relevant ist

betrachten, besteht darin, sich einen Prädiktor (z. B. ein Merkmal der Personen) und dessen statistische Vorhersage auf die abhängige Variable mehrfach anzuschauen – und zwar aufgeteilt nach Abstufungen des jeweils anderen Prädiktors (z. B. eines Merkmals der Organisation). Da hier einer der Prädiktoren kategorial in die Analyse eingeht, ist dies als „gröbere“ Variante des zuvor dargestellten Ansatzes der polynomialen Regressionen zu verstehen.

Eine Metaanalyse von Judge und Zapata (2015) soll diesen Ansatz illustrieren. Die Autorin und der Autor aggregierten Zusammenhänge zwischen breiten Persönlichkeitsdimensionen (Big Five) und beruflicher Leistung. Dabei kodierten sie zusätzlich, welchen Bezug die jeweilige Tätigkeit, aus der die Daten stammten, zu den Big Five hatte. Es wurde also festgehalten, ob die jeweilige Persönlichkeitsdimension eine relevante Anforderung darstellt (in der Studie als Trait-Aktivierungs-Potenzial bezeichnet). Wir haben die Ergebnisse der Studie grafisch aufgearbeitet (Abb. 6.16). Auf der y-Achse in Abb. 6.16 sind von Judge und Zapata (2015) ermittelten Beta-Gewichte abgetragen. Sie beschreiben die individuelle Prognosekraft jeder Persönlichkeitsdimension. Diese sind insgesamt $6 \times$ nacheinander aufgeführt, getrennt nach den Anforderungsbereichen (1) bis (6). Wir haben an der x-Achse jeweils aufgeführt, welche Persönlichkeitsdimensionen in den 6 Anforderungsbereichen nach Judge und Zapata (2015) relevant sein sollten.

Aus Abb. 6.16 wird gut ersichtlich, dass in manchen Fällen die theoretisch relevanten Persönlichkeitsdimensionen auch die angenommene Prognosekraft zeigen. Dies gilt für Berufe der Anforderungsbereiche (3) bis (5).

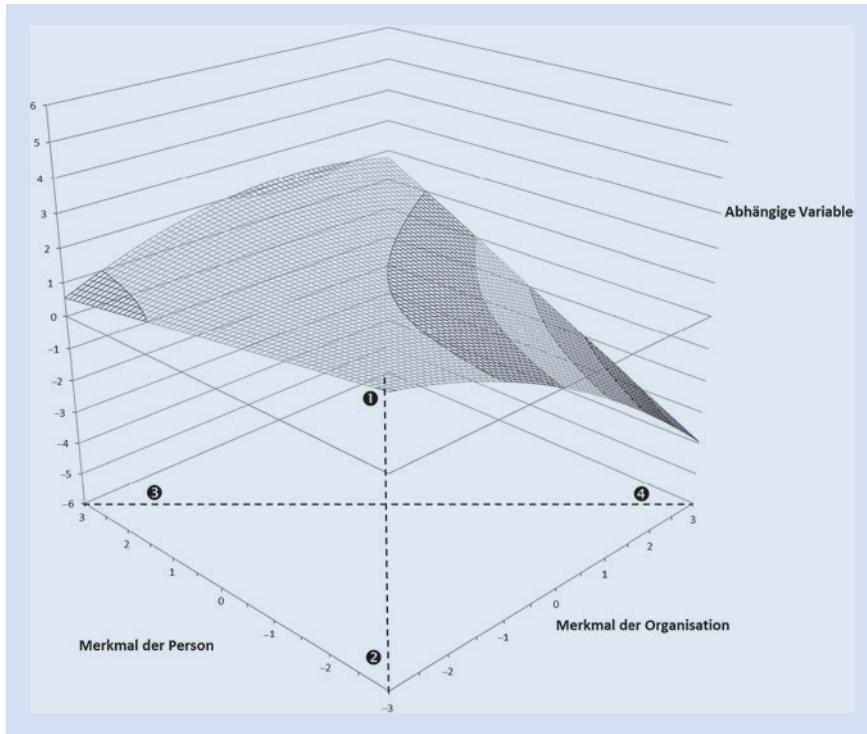


Abb. 6.15 Beispiel für eine Response-Surface-Analyse, wenn nur eine Form der Nichtpassung bedeutsam ist

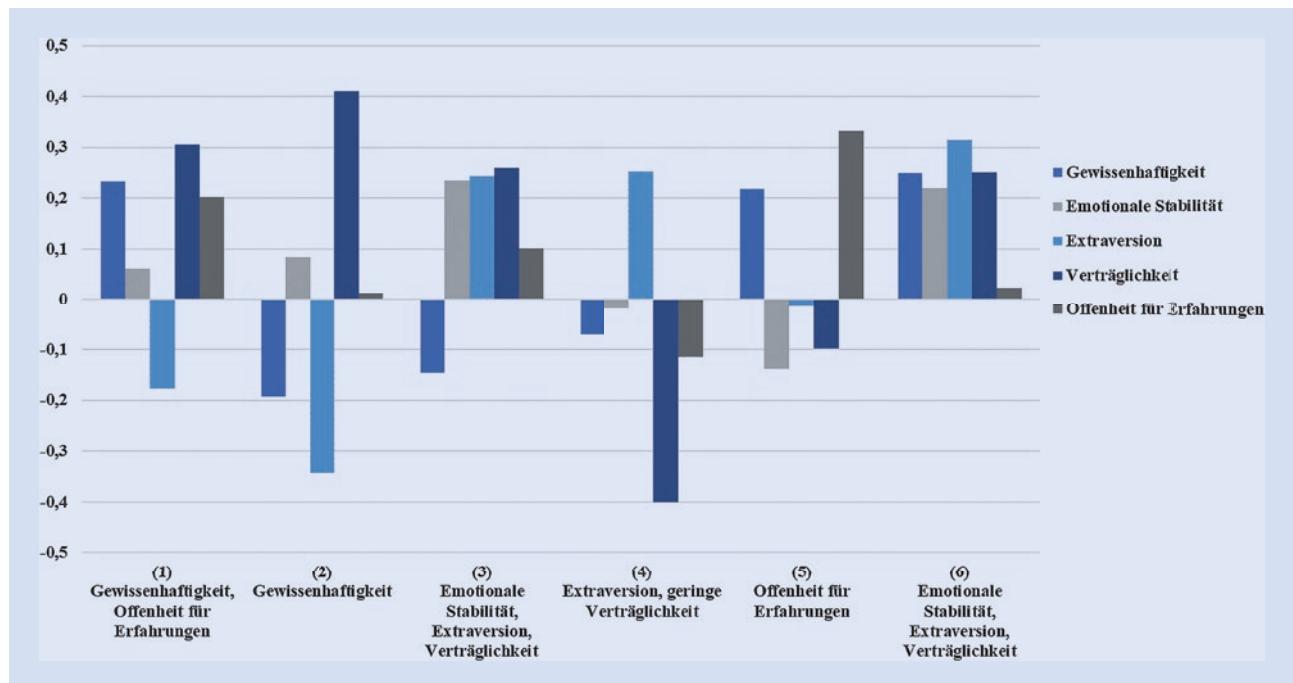


Abb. 6.16 Grafische Darstellung der Ergebnisse von Judge und Zapata (2015). Auf der y-Achse ist die Prognosegüte in Form von Beta-Gewichten abgetragen und auf der x-Achse die Anforderungsbereiche (1) bis (6). Weitere Erläuterungen s. Text

Für Berufe der Anforderungsbereiche (1), (2) und (6) gelingt dies nicht durchgängig. Die angenommenen Persönlichkeitsdimensionen zeigen nicht die höchsten Zusammenhänge. Beispielsweise wurde für Berufe des Anforderungsbereichs (6) angenommen, dass die Dimension „emotionale Stabilität“ relevant ist. Dies ist auch der Fall, allerdings ist die Prognosekraft dieser Dimension nicht höher als die der eigentlich nicht relevanten Dimension „Gewissenhaftigkeit“.

Fazit Bei der Evaluation im Bereich der Arbeits-, Organisations- und Wirtschaftspsychologie sollte berücksichtigt werden, dass die Beurteilung der Passung und deren Auswirkung auf abhängige Variablen eine gemeinsame Be trachtung von Organisations- und Personenmerkmalen erfordert.

6

6.4 Ein Qualitätsstandard für berufsbezogene Eignungsbeurteilungen – die DIN 33430

Standards für eine wissenschaftlich fundierte Vorgehensweise

Zur Feststellung der berufsbezogenen Eignung wurden in der Vergangenheit häufig Verfahren mit fraglicher Validität herangezogen. Aus der Unzufriedenheit mit der Praxis der beruflichen Eignungsdiagnostik entwickelte sich eine Initiative zur Etablierung eines Standards für eine wissenschaftlich fundierte Vorgehensweise. Der Berufsverband Deutscher Psychologinnen und Psychologen (BDP) e. V. stellte 1995 beim Deutschen Institut für Normung (DIN) e. V. den formalen Antrag, eine Norm zur beruflichen Eignungsdiagnostik zu erarbeiten. Unterstützung fand der BDP durch die Deutsche Gesellschaft für Psychologie (DGPs). Am 09. Juni 1997 nahm ein Ausschuss des Deutschen Instituts für Normung, der mit Vertretern aus Wissenschaft und Praxis, Unternehmen, Behörden, Verbänden und Verlagen besetzt war, unter Vorsitz von Prof. Hornke die Arbeit auf. Das Ergebnis dieser Arbeit ist letztlich ein Konsens, der auch von den Interessen der Beteiligten geprägt ist. Im Jahr 2002 erfolgte die erste Veröffentlichung der „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen: DIN 33430“. Eine Aktualisierung dieser Norm wurde 2016 vorgenommen (DIN 2016).

Checklisten vorhanden

Auf insgesamt 35 Seiten Text werden Qualitätsstandards sowohl für die Personen, die als Eignungsdiagnostikerinnen bzw. -diagnostiker oder Beobachtende mitwirken, als auch für die dabei eingesetzten Verfahren definiert. Personalverantwortliche können auch mithilfe von Checklisten prüfen, ob die von ihnen eingesetzten Verfahren sowie der gesamte Prozess der Eignungsbeurteilung mit der DIN 33430 konform sind. Entsprechende Checklisten hat Kersting (2018) vorgelegt. Die DIN 33430 ist nicht rechtsverbindlich (s. Kersting und Püttner 2018): Das Deutsche Institut für Normung ist ein privater Verein, der auf Antrag Dritter den Normerstellungsprozess koordiniert. Es steht Unternehmen daher frei, sich nach der Norm zu richten.

Zweck der DIN 33430

- Anbietende entsprechender Dienstleistungen erhalten Qualitätsstandards zur Planung und Durchführung von Eignungsbeurteilungen.
- Personen und Institutionen, die Eignungsbeurteilungen durchführen lassen, gewinnen Orientierung bei der Bewertung von externen Angeboten.
- Personalverantwortliche können Qualitätssicherung und -optimierung von Personalentscheidungen betreiben.
- Personen, deren Eignung beurteilt wird, werden vor unsachgemäßer oder missbräuchlicher Anwendung von Verfahren geschützt.

Bei den Personen, die an der Eignungsuntersuchung beteiligt sind, wird zwischen Dienstleistenden, Eignungsdiagnostikerinnen bzw. -diagnostikern und Beobachtenden unterschieden. Dienstleistende sind diejenigen (Personen oder Organisationen), die mit der Eignungsbeurteilung beauftragt wurden. Eignungsdiagnostikerinnen bzw. -diagnostiker sind hauptverantwortlich für den gesamten Prozess, der von der Planung und der Durchführung der Untersuchung über die Auswertung und die Interpretation der Ergebnisse bis zum Bericht an den Auftraggebenden reicht. Sie müssen die zur Verfügung stehenden Verfahren und Prozesse kennen. Dazu gehören auch Kenntnisse über die Konstrukte und über die Qualität und die Einsatzvoraussetzungen der Verfahren. Beobachtende sind geschulte Mitwirkende an Verhaltensbeobachtungen und Interviews.

Unterscheidung von handelnden Personen

Von den Verfahren wird verlangt, dass sie grundsätzlich einen Bezug zu den Anforderungen aufweisen. Ein Verfahren, das bei einer Fragestellung passend ist, kann bei einer anderen völlig unangemessen sein. Es werden jedoch auch allgemeingültige Auswahlkriterien genannt. So sollen in den Unterlagen zu einem Verfahren die Handhabung erklärt und für eine kritische Bewertung nötige Angaben gemacht werden. Konkrete Anforderungen werden an die Objektivität, die Zuverlässigkeit, die Gültigkeit und die Normen gestellt. In einem Anhang der DIN-Norm finden sich detaillierte Forderungen zu den Informationen, die über ein Verfahren verfügbar sein sollten (z. B. Zielsetzung, theoretische Grundlage, bestimmte Aspekte der Reliabilität).

Anforderungen an Verfahren

Auszug aus den Forderungen der DIN 33430

- Anforderungen einer Tätigkeit müssen festgelegt werden.
- Art und Weise der Integration von Einzelergebnissen zu einem Eignungsurteil müssen vorab spezifiziert werden.
- Bei gleichbleibendem eignungsdiagnostischem Vorgehen ist dieses alle 3 Jahre erneut zu begründen.
- Eignungsdiagnostische Verfahren müssen einen eindeutigen Bezug zu den festgelegten Anforderungen haben; dies gilt auch für Angaben, die vorliegenden Dokumenten (z. B. Lebensläufen) zu entnehmen sind.
- Durchführende von Interviews und Verhaltensbeobachtungen müssen dafür geschult sein.
- Interviews sind strukturiert durchzuführen.
- Verhaltensbeobachtungen müssen nach festgelegten Regeln bewertet werden.
- Tests und Fragebögen müssen hinsichtlich ihrer psychometrischen Qualität überprüfbar sein, d. h., die dazu notwendigen Informationen müssen verfügbar sein.
- Kandidatinnen und Kandidaten sind über den Ablauf einer Untersuchung zu informieren.
- Bei der Interpretation von Test- und Fragebogenergebnissen sind Konfidenzintervalle zu berücksichtigen.
- Das Vorgehen bei der Eignungsbeurteilung ist nachvollziehbar zu dokumentieren.
- Das eignungsdiagnostische Vorgehen ist zu evaluieren.

(© Beuth).

Grundsätzlich gibt es die folgenden beiden Möglichkeiten der Zertifizierung nach DIN 33430:

1. Personen können eine Lizenz erwerben. Dazu ist das Bestehen einer Prüfung notwendig. Lizensierte Personen dokumentieren dadurch, dass sie über die notwendigen eignungsdiagnostischen Kenntnisse verfügen. Die

Lizenzierung von Prozessen und Personen

Prüfungen zur Personenlizenzierung werden von der Deutschen Psychotherapeutenakademie (DPA) durchgeführt, einer Bildungseinrichtung des BDP, die auch ein öffentlich zugängliches Register lizenziierter Personen führt.

2. Unternehmen können einen bei sich etablierten Prozess der Eignungsbeurteilung und das damit verbundene Qualitätsmanagement zertifizieren lassen (s. Kersting 2008a; Reimann 2009). Die Zertifizierung einer Organisation kann ein speziell dazu qualifiziertes Zertifizierungsinstitut vornehmen (s. Reimann 2009).

Weiterführende Literatur und Internetressourcen

Informationen zur DIN-Prüfung, aber auch weitere aktuelle Informationen zur DIN 33430 finden sich unter ► <https://www.din33430portal.de/>. Das Grundwissen zur DIN 33430, das zugleich auch für den Lizenziererwerb nach DIN 33430 prüfungsrelevant ist, liegt in einem vom Diagnostik- und Testkuratorium (2018) herausgegebenen Band vor (s. auch Interview mit Prof. Dr. Martin Kersting).

Interview mit Prof. Dr. Martin Kersting



Prof. Dr. Martin Kersting (Foto: Anja Schaal).

Prof. Kersting, die DIN 33430 erschien erstmals 2002. Lässt sich sagen, wie sie die Praxis der Eignungsbeurteilung geprägt hat? In den letzten Jahren hat das Thema „Qualität in der Eignungsdiagnostik“ erheblich an Bedeutung gewonnen. Diese Entwicklung hat unterschiedliche Ursachen: Die Anforderungen sowohl der Organisationen als auch der (internen und externen) Bewerberinnen und Bewerber an die Eignungsdiagnostik sind deutlich gestiegen. Es geht längst nicht mehr nur um eine „irgendwie“ getroffene „Null-eins“-Aus-

wahlentscheidung, sondern um die transparente Erarbeitung von Kompetenzprofilen und Potenzialeinschätzungen sowie um Akzeptanz und Förderorientierung. Hinzu kommen strengere rechtliche Vorgaben. Die DIN 33430 war und ist sowohl eine Reaktion auf die gestiegenen Qualitätsanforderungen als auch Promotor dieser Entwicklung.

Welchen Stellenwert hat die DIN 33430 heute? Die DIN 33430 ist eine Richtlinie, an der sich diejenigen orientieren, die an einer möglichst treffsicheren und zugleich rechtlich einwandfreien und sozial hochgradig akzeptierten

Einschätzung der Potenziale und Kompetenzen von Kandidatinnen und Kandidaten sowie Mitarbeiterinnen und Mitarbeitern interessiert sind. Der Qualitätsstandard wurden von Expertinnen und Experten aus Wissenschaft und Praxis verfasst, um Anwenderinnen und Anwendern bei allen wesentlichen Prozessschritten Orientierung zu bieten: Von der Anforderungsanalyse über die Gestaltung des Ablaufs der Eignungsdiagnostik (einschließlich der Auswahl der richtigen Verfahren und der erforderlichen Qualifikation der beteiligten Personen) bis hin zur Evaluation.

Wem nützt die DIN 33430? Organisationen können durch eine Orientierung an der DIN 33430 ihre Eignungsdiagnostik optimieren. Sie können entweder ihre eigene Arbeit an diesem Qualitätsstandard ausrichten oder aber die DIN 33430 nutzen, um die Qualität der Angebote und der Arbeit von Externen, mit denen sie zusammenarbeiten, zu bewerten. Dabei besteht auch die Möglichkeit, die DIN 33430 zum verbindlichen Bestandteil eines Vertrags mit externen Dienstleisterinnen und Dienstleistern zu erklären und diese somit ohne großen Aufwand auf Qualität zu verpflichten. Mittlerweile wird in einschlägigen Verordnungen des öffentlichen Dienstes und in Ausschreibungen explizit auf die DIN 33430 Bezug genommen. Auch Sozialpartnerinnen und -partner können sich mithilfe der DIN 33430 auf konkrete Qualitätsstandards der Eignungsdiagnostik verständigen. Eignungsdiagnostische Dienstleisterinnen und Dienstleister wiederum können sich im Wettbewerb profilieren, indem sie nach außen hin darstellen und belegen, dass sie sich an der DIN 33430 orientieren. Wer will, kann eine Prüfung absolvieren und durch eine Lizenz nachweisen, dass sie/er über die nach DIN 33430 notwendigen Kenntnisse verfügt. Die internen und externen Bewerbenden profitieren dadurch von der DIN 33430, dass ihnen unangemessene Prozeduren er-

spart bleiben. Die Orientierung an der DIN 33430 erhöht zudem die Wahrscheinlichkeit, dass die Kompetenzen und Potenziale der Bewerbenden erkannt und gefördert werden und sie einen Ausbildungs- und Arbeitsplatz bekommen, der zu ihnen passt.

Und woran merken Bewerberinnen und Bewerber, dass die Eignungsbeurteilung nach DIN 33430 durchgeführt wurde? Organisationen, die sich an der DIN 33430 orientieren, führen die Verfahren transparent durch, sie sind gut vorbereitet und setzen qualifiziertes Personal ein. So schreibt die DIN 33430 beispielsweise vor, dass alle Verfahren einen Anforderungsbezug aufweisen müssen – dies verbietet Fragen, die einen unangemessenen Eingriff in die Privatsphäre darstellen.

Wie sieht Ihrer Meinung nach die Zukunft der DIN 33430 aus – insbesondere im Kontext internationaler Normen? Insbesondere Wirtschaftsorganisationen arbeiten zunehmend global. Trotz der Internationalisierung gibt es aber nationale Regelungen, die jeweils zu beachten sind. Es gibt daher zusätzlich zu den internationalen Standards auch nationale Qualitätsrichtlinien wie die DIN 33430, die die internationalen Überlegungen spezifisch ausführen. Normen müssen regelmäßig dahingehend überprüft werden, ob sie noch dem aktuellen Stand der Technik entsprechen. Die Fassung der Norm aus dem Jahr 2002 mussten wir überarbeiten, um den Veränderungen auf dem Arbeitsmarkt (z. B. der Globalisierung oder dem in vielen Bereichen vorherrschenden Personalmangel), den veränderten diagnostischen Praktiken (z. B. internetgestütztes Testen), den veränderten gesetzlichen Rahmenbedingungen sowie den aktuellen Erkenntnissen (z. B. Metastudien) gerecht zu werden. Ich gehe davon aus, dass sich die nächste Version der DIN 33430 verstärkt mit den Formen der Eignungsdiagnostik beschäftigen wird, die „Big Data“ mithilfe von künstlicher Intelligenz auswerten.

6.5 Zusammenfassung

In diesem Kapitel wurde betont, dass ein zentrales Anliegen der Arbeits-, Organisations- und Wirtschaftspsychologie darin besteht, die Passung (Fit) zwischen Personen und Arbeitsbedingungen zu optimieren. Es wurden Methoden erläutert, die dazu dienen, das Ausmaß der Passung zu quantifizieren und zielgerichtete Folgemaßnahmen anzustoßen. Es wurden diagnostische Herangehensweisen der Arbeits- und Anforderungsanalyse, der Personalauswahl, der Personalentwicklung, der Organisationsentwicklung und der Berufs- bzw. Organisationswahl dargestellt und deren Evidenzbasierung diskutiert.

6

Weiterführende Literatur und Internetressourcen

In den umfangreichen und umfassenden Lehrbüchern zur Organisationspsychologie (Schuler und Moser 2014) und zur Personalpsychologie (Schuler und Kanning 2014) finden sich weitere Ausführungen zu den Inhalten dieses Kapitels. Weiterführende Literatur zum Thema Assessment-Center bietet das Buch *Assessment-Center* (Kleinmann 2013) aus der Reihe „Praxis der Personalpsychologie“. Eine umfassende Behandlung der Forschung zu Situational-Judgment-Tests liefert das gleichnamige Buch von Weekley und Ployhart (2013). Zur DIN 33430 existiert ein Buch mit dem Titel *Personalauswahl kompetent gestalten* (Diagnostik- und Testkuratorium 2018), das das nach DIN 33430 relevante Wissen kompakt vermittelt.

Nützliche Links zum Thema Arbeits- und Anforderungsanalyse sind: O*Net (► <https://www.onetonline.org/>), das Portal „berufenet.de“ der Arbeitsagentur (► <https://berufenet.arbeitsagentur.de/>) und ähnliche Datenbanken für Österreich (► <https://www.berufslexikon.at/>) und die Schweiz (► <https://www.berufsberatung.ch/>). Weitere Informationen zur DIN 33430 finden sich unter ► <https://www.din33430portal.de/>.

?

Übungsfragen

- ► Abschn. 6.1:
 - Was sind die 3 grundsätzlichen Zugänge der Arbeits- und Anforderungsanalyse und wie unterscheiden sich diese voneinander?
 - Was versteht man unter der Critical-Incident-Technik?
 - Was versteht man unter einem Anforderungsprofil?
- ► Abschn. 6.2:
 - Nennen Sie biografieorientierte, simulationsorientierte und konstruktorientierte Verfahren der Personalauswahl!
 - Wie beurteilen Sie die Validität von Bewerbungsunterlagen für die Personalauswahl?
 - Welche Vor- und Nachteile bringt die Assessment-Center-Methode mit sich?
 - Was weiß man über die Abgrenzbarkeit der Dimensionen in Assessment-Centern?
 - Wie ist es um die Reliabilität von Situational-Judgment-Tests bestellt?
 - Welche Verfahren lassen den höchsten inkrementellen Beitrag – über allgemeine kognitive Leistungstests hinaus – bei der Prognose der beruflichen Leistung erwarten?
 - Was versteht man unter sozialer Validität?
 - Welche Rolle kommt der Psychologischen Diagnostik in der Personalentwicklung zu?
- ► Abschn. 6.3:
 - Welchen Fehlerquellen unterliegen Arbeits- und Anforderungsanalysen?
 - Was sind zentrale Erkenntnisse aus der Metaanalyse von Schmidt und Hunter (1998)?
 - Was weiß man über die Relevanz von beruflichen Interessen für die berufliche Leistung?
 - Warum ist bei Verfahren zur Feststellung des Entwicklungsbedarfs die Frage besonders zentral, ob die verwendeten Verfahren ihrem Messanspruch gerecht werden?

- Erläutern Sie, warum bivariate Analysen im Kontext der Diagnostik in der Arbeits- und Organisationspsychologie häufig nicht ausreichen!
- Was sind Probleme von Differenzwerten?
- ► Abschn. 6.4:
 - Welchen Nutzen hat die DIN 33430?
 - Welche Formen der Lizenzierung nach DIN 33430 existieren?

Literatur

- Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology* 60, 451–474.
- Anderson, N. R., & West, M. A. (1998). Measuring climate for work group innovation: development and validation of the team climate inventory. *Journal of Organizational Behavior* 19, 235–258.
- Andres, J., & Kleinmann, M. (1993). Die Entwicklung eines Rotationssystems für die Beobachtungssituation im Assessment-Center. *Zeitschrift für Arbeits- und Organisationspsychologie* 37, 19–24.
- Apers, C., & Derous, E. (2017). Are they accurate? Recruiters' personality judgments in paper versus video resumes. *Computers in Human Behavior* 73, 9–19.
- Arbeitskreis Assessment Center (2016). *AC-Standards: Standards der Assessment Center Technik*. ► https://www.forum-assessment-kongress.de/images/AKAC_AC_Standards_2016.pdf. Zugriffen: 24. März 2020.
- Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020: Fortführung einer Trendstudie. *Zeitschrift für Arbeits- und Organisationspsychologie* 64, 67–82.
- Arthur, W. J., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology* 56, 125–154.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*: Englewood Cliffs, NJ, US: Prentice-Hall, Inc.
- Baron-Boldt, J., Schuler, H., & Funke, U. (1988). Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie* 2, 79–90.
- Bauer, T. N., Maertz, C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology* 83, 892–903.
- Becker, N., Höft, S., Holzenkamp, M., & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology* 10, 61–69.
- Beermann, D., Kersting, M., Stegt, S., & Zimmerhofer, A. (2013). Vorurteile und Urteile zur Akzeptanz von Persönlichkeitsfragebogen. *PersonalQuarterly* 65, 41–45.
- Bergmann, C., & Eder, F. (2005). *AIST-R: Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) – Revision*. Göttingen: Beltz Test Gesellschaft.
- Blanchard, P. N., & Thacker, J. W. (2013). *Effective training: Systems, strategies, and practices*. Boston, MA: Pearson Education.
- Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology* 91, 1114–1124.
- Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives*. New York: Wiley.
- Brodbeck, F. C., Anderson, N., & West, M. (2000). *TKI: Teamklima Inventar*. Göttingen: Hogrefe.
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods* 18, 36–52.
- Bruk-Lee, V., Drew, E. N., & Hawkes, B. (2013). Candidate Reactions to Simulations and Media-Rich Assessments in Personnel Selection. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 43–60). New York, NY: Springer.
- Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. (2017). *Psychische Gesundheit in der Arbeitswelt – Wissenschaftliche Standortbestimmung*. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.
- Büssing, A. (2007). Organisationsdiagnose. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (4. Aufl., S. 557–599). Bern: Huber.
- Cable, D. M., & DeRue, D. S. (2002). The convergent and discriminant validity of subjective fit perceptions. *Journal of Applied Psychology* 87, 875–884.
- Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance* 27, 283–310.

- Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment* 20, 333–346.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology* 82, 143–159.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology* 82, 300–310.
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology* 63, 83–117.
- Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology* 51, 193–208.
- Cole, M. S., Feild, H. S., Giles, W. F., & Harris, S. G. (2009). Recruiters' inferences of applicant personality based on resume screening: do paper people have a personality? *Journal of Business and Psychology* 24, 5–18.
- Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management* 22, 139–162.
- Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational Judgement Tests for selection. In H. Goldstein, E. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 226–246). West Sussex, UK: Wiley-Blackwell.
- Daumenlang, K., Müskens, W., & Harder, U. (2004). *FEO: Fragebogen zur Erfassung des Organisationsklimas*. Göttingen: Hogrefe.
- Demerouti, E., & Bakker, A. B. (2011). The job demands-resources model: Challenges for future research. *SA Journal of Industrial Psychology* 37, 1–9.
- Deutsche Gesellschaft für Psychologie. (2020). Fachgruppe Arbeits-, Organisations- und Wirtschaftspsychologie. ► <https://www.dgps.de/index.php?id=156>. Zugegriffen: 18. Mai 2020.
- Deutsches Institut für Normung e. V. (DIN). (2016). *DIN 33430:2016-07: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.
- Diagnostik- und Testkuratorium. (2018). *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430*. Berlin, Heidelberg: Springer.
- Dickinson, T. L., Hassett, C. E., & Tannenbaum, S. I. (1986). *Work performance ratings: A meta-analysis of multitrait-multimethod studies*. San Antonio: Texas Maxima Corp.
- Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology*, 88, 635.
- Dunckel, H., & Resch, M. G. (2010). Arbeitsanalyse. In U. Kleinbeck, & K.-H. Schmidt (Hrsg.), *Arbeitspsychologie* (Enzyklopädie der Psychologie, Serie Wirtschafts-, Organisations- und Arbeitspsychologie, Bd. 1, S. 1111–1158). Göttingen: Hogrefe.
- Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology* 46, 641–665.
- Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression and response surface methodology. *Advances in Measurement and Data Analysis*, 350–400.
- Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal* 36, 1577–1613.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin* 51, 327–358.
- Frank, F., & Kanning, U. P. (2014). Lücken im Lebenslauf. *Zeitschrift für Arbeits- und Organisationspsychologie* 58, 155–162.
- Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. doi: ► <https://doi.org/10.1111/peps.12385>.
- Frieling, E., & Hoyos, C. G. (1978). *Fragebogen zur Arbeitsanalyse (FAA)*. Bern: Huber.
- Gati, I., Gadassi, R., & Shemesh, N. (2006). The predictive validity of a computer-assisted career decision-making system: A six-year follow-up. *Journal of Vocational Behavior* 68, 205–219.
- Gatzka, T., & Volmer, J. (2017). *Situational Judgement Test für Teamarbeit (SJT-TA)*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften.
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology* 72, 493–511.
- Gibbons, A. M., Rupp, D. E., Snyder, L. A., Holtub, A. S., & Woo, S. E. (2006). A preliminary investigation of developable dimensions. *The Psychologist-Manager Journal* 9, 99–123.
- Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review* 18, 694–734.
- Hacker, W., Fritzsche, B., Richter, P., & Iwanowa, A. (1995). *Tätigkeitsbewertungssystem (TBS): Verfahren zur Analyse, Bewertung und Gestaltung von Arbeitstätigkeiten*. Zürich: vdf.

- Hardison, C. M., & Sackett, P. R. (2007). Kriteriumsbezogene Validität des Assessment Centers: Lebendig und wohlau? In H. Schuler (Hrsg.), *Assessment Center zur Potenzialanalyse* (S. 192–202). Göttingen: Hogrefe.
- Hell, B., Schuler, H., Boramir, I., & Schaer, H. (2006). Verwendung und Einschätzung von Verfahren der internen Personalauswahl und Personalentwicklung im 10 Jahres-Vergleich. *German Journal of Human Resource Management* 20, 58–78.
- Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment* 15, 405–411.
- Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology* 100, 1143–1168.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.
- Holling, H., & Liepmann, D. (2004). Personalentwicklung. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (3. Aufl., S. 345–383). Bern: Huber.
- Hossiep, R., & Weiß, S. (2020). *BIP-AM: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – Anforderungsmodul*. Göttingen: Hogrefe.
- Hossiep, R., Paschen, M., & Krüger, C. (2018). *BIP plus BIP-6F: Bochumer Inventare zur berufsbezogenen Persönlichkeitsbeschreibung – Langform plus 6 Faktoren*. Göttingen: Hogrefe.
- Hossiep, R., Krüger, C., & Weiß, S. (2020). *BIP-6F-AM: Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren – Anforderungsmodul* (in Vorbereitung). Göttingen: Hogrefe.
- Howard, A. (1974). An assessment of assessment-centers. *Academy of Management Journal* 17, 115–134.
- Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2016). Transparency of assessment centers: Lower criterion-related validity but greater opportunity to perform? *Personnel Psychology* 69, 467–497.
- Joerin Fux, S., & Stoll, F. (2006). *EXPLOJOB – Das Werkzeug zur Beschreibung von Berufsanforderungen und -tätigkeiten. Deutschsprachige Adaptation und Weiterentwicklung des Position Classification Inventory (PCI) nach Garry D. Gottfredson und John L. Holland*. Bern: Huber.
- Jones, R. G., & Whitmore, M. D. (1995). Evaluating developmental assessment centers as interventions. *Personnel Psychology* 48, 377–388.
- Judge, T. A., & Zapata, C. P. (2015). The person-situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal* 58, 1149–1179.
- Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: a qualitative and quantitative review. *Journal of Applied Psychology* 87, 765–780.
- Kanning, U. P. (2015). Welche Aussagekraft besitzen biographische Daten bei der Sichtung von Bewerbungsunterlagen? Ein Überblick über aktuelle Studien. *Wirtschaftspsychologie* 17, 42–50.
- Kanning, U. P., Grawe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view. *European Journal of Psychological Assessment* 22, 168–176.
- Kanning, U. P., Budde, L., & Hülskötter, M. (2018). Wie valide ist die regelkonforme Gestaltung von Bewerbungsunterlagen? *Personal Quarterly* 70, 38–45.
- Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJT). *European Journal of Psychological Assessment* 32, 230–240.
- Kauffeld, S. (2004). *FAT: Fragebogen zur Arbeit im Team*. Göttingen: Hogrefe.
- Kelbetz, G., & Schuler, H. (2002). Verbessert Vorerfahrung die Leistung im Assessment Center? *Zeitschrift für Personalpsychologie* 1, 4–18.
- Kersting, M. (2008a). *Qualität in der Diagnostik und Personalauswahl – der DIN Ansatz*. Göttingen: Hogrefe.
- Kersting, M. (2008b). Zur Akzeptanz von Intelligenz- und Leistungstests. [The acceptance of intelligence and achievement tests]. *Report Psychologie* 33, 420–433.
- Kersting, M. (2010). Akzeptanz von Assessment Centern; Was kommt an und worauf kommt es an? *Wirtschaftspsychologie* 12, 58–65.
- Kersting, M. (2018). Zur Information über und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens – Die DIN SCREEN Checkliste 1, Version 3. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 223–244). Berlin, Heidelberg: Springer.
- Kersting, M., & Püttner, I. (2018). Einführung in die DIN 33430. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 1–25). Berlin, Heidelberg: Springer.
- Kleinmann, M. (2013). *Assessment-Center*. Göttingen: Hogrefe.
- Koch, A., Strobel, A., Kici, G., & Westhoff, K. (2009). Quality of the Critical Incident Technique in practice: Interrater reliability and users' acceptance under real conditions. *Psychology Science Quarterly* 51, 3–15.

- 6**
- Koch, A., Strobel, A., Miller, R., Garten, A., Cimander, C., & Westhoff, K. (2012). Never use one when two will do: The effects of a multi-perspective approach on the outcome of job analyses using the Critical Incident Technique. *Journal of Personnel Psychology* 11, 95–102.
- König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment* 15, 283–292.
- Kooij, D. T., De Lange, A. H., Jansen, P. G., Kanfer, R., & Dikkers, J. S. (2011). Age and work-related motives: Results of a meta-analysis. *Journal of Organizational Behavior* 32, 197–225.
- Kramer, J. (2009). Allgemeine Intelligenz und beruflicher Erfolg in Deutschland: Vertiefende und weiterführende Metaanalysen. *Psychologische Rundschau* 60, 82–98.
- Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individual's fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology* 58, 281–342.
- Krumm, S., Grube, A., & Hertel, G. (2013). No time for compromises: Age as a moderator of the relation between needs-supply fit and job satisfaction. *European Journal of Work and Organizational Psychology* 22, 547–562.
- Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in an situational judgment test? *Journal of Applied Psychology* 100, 399–416.
- Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology* 1, 84–97.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior* 45, 79–122.
- Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment* 6, 141–152.
- Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology* 86, 255–264.
- Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology* 87, 675–686.
- Lievens, F. (2017). Assessing Personality–Situation Interplay in Personnel Selection: Toward More Integration into Personality Research. *European Journal of Personality* 31, 424–440.
- Lohaus, D., & Schuler, H. (2014). Leistungsbeurteilung. In H. Schuler & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 357–411). Göttingen: Hogrefe.
- Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology* 81, 11–21.
- McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1969). *The development and background of the Position Analysis Questionnaire*. West Lafayette, IN: Purdue University: Occupational Research Center.
- McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology* 86, 730–740.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel Psychology* 60, 63–91.
- Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology* 69, 105–115.
- Melchers, K. G., Henggeler, C., & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? A meta-analytic reassessment. *Zeitschrift für Personalpsychologie* 6, 141–149.
- Melchers, K. G., Kleinmann, M., Richter, G. M., König, C. J., & Klehe, U. C. (2004). Messen Einstellungsinterviews das, was sie messen sollen? Zur Bedeutung der Bewerberkognitionen über bewertetes Verhalten. *Zeitschrift für Personalpsychologie* 3, 159–169.
- Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology* 93, 1042–1052.
- Morgan, A., Cannan, K., & Cullinane, J. (2005). 360 feedback: A critical enquiry. *Personnel Review* 34, 663–680.
- Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology* 82, 627–655.
- Morgeson, F. P., & Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology* 91, 1321–1339.

- Morgeson, F. P., Mumford, T. V., & Campion, M. A. (2005). Coming full circle: Using research and practice to address 27 questions about 360-degree feedback programs. *Consulting Psychology Journal: Practice and Research* 57, 196–209.
- Müller-Benedict, V. (2010). Grenzen von leistungsbasierten Auswahlverfahren. *Zeitschrift für Erziehungswissenschaft* 13, 451–472.
- Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology* 93, 250–267.
- Mussel, P., Gatzka, T., & Hewig, J. (2016). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment* 34, 328–335.
- Neubauer, A., Bergner, S., & Felfe, J. (2012). *LJI: Leadership Judgement Indicator. Deutschsprachige Adaptation des Leadership Judgement Indicator (LJI)* von M. Lock und R. Wheeler. Bern: Huber.
- Obermann, C. (2009). *Assessment-Center: Entwicklung, Durchführung, Trends. Mit originalen AC-Übungen* (4. Aufl.). Wiesbaden: Gabler.
- Occupational Information Network (O*NET). (2020). National Center for O*NET Development. O*NET OnLine. ► <https://www.onetonline.org/>. Zugegriffen: 24. März 2020.
- Oh, I.-S., & Berry, C. M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology* 94, 1498–1513.
- Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung*. Göttingen: Hogrefe.
- Peus, C., Braun, S., & Frey, D. (2016). *LSA: Leadership Style Assessment: Ein Situational Judgment Test zur Erfassung von Führungsstilen*. Göttingen: Hogrefe.
- Reimann, G. (2009). *Moderne Eignungsbeurteilung mit der DIN 33430*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Rothe, I., Adolph, L., Beermann, B., Schütte, M., Windel, A., Grewer, A., Grewer, A., et al. (2017). *Psychische Gesundheit in der Arbeitswelt: Wissenschaftliche Standortbestimmung*. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.
- Rupp, D. E., Snyder, L. A., Gibbons, A. M., & Thornton, G. C., III. (2006). What should developmental assessment centers be developing? *The Psychologist-Manager Journal* 9, 75–98.
- Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology* 52, 359–391.
- Ryan, A. M., Reeder, M. C., Golubovich, J., Grand, J., Inceoglu, I., Bartram, D., Derous, E., et al. (2017). Culture and testing practices: is the world flat? *Applied Psychology* 66, 434–467.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology* 56, 573–605.
- Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (2019). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*. doi: ► <https://doi.org/10.1037/apl0000457>.
- Schippmann, J., Ash, R., Battista, M., Carr, L., Eyde, L., Hesketh, B., Kehoe, J., et al. (2000). The practice of competency modeling. *Personnel Psychology* 53, 703–740.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* 124, 262–274.
- Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings. Working Paper. Tippie College of Business, University of Iowa. ► <https://testingtalent.com/wp-content/uploads/2017/04/2016-100-Yrs-Working-Paper-on-Selection-Methods-Schmit-Mar-17.pdf>. Zugegriffen: 27. März 2020.
- Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology* 61, 827–868.
- Schuler, H. (2006). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (2. Aufl., S. 45–68). Göttingen: Hogrefe.
- Schuler, H., & Höft, S. (2007). Diagnose beruflicher Eignung und Leistung. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (4. Aufl., S. 289–343). Bern: Huber.
- Schuler, H., & Kanning, U. P. (2014). *Lehrbuch der Personalpsychologie* (3. Aufl.). Göttingen: Hogrefe.
- Schuler, H., & Moser, K. (2014). *Lehrbuch Organisationspsychologie* (5. Aufl.). Bern: Hans Huber.
- Schuler, H., & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. *Zeitschrift für Arbeits- und Organisationspsychologie* 27, 33–44.

- Schuler, H., Frier, D., & Kauffmann, M. (1993). *Personalauswahl im europäischen Vergleich*. Göttingen: Verlag für Angewandte Psychologie.
- Schuler, H., Hell, B., Trapmann, S., Schaar, H., & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie* 6, 60–70.
- Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology* 65, 445–494.
- Solga, M., Ryschka, J., & Mattenkrott, A. (2011). Personalentwicklung: Gegenstand, Prozessmodell, Erfolgsfaktoren. In J. Ryschka, M. Solga, & A. Mattenkrott (Hrsg.), *Praxishandbuch Personalentwicklung* (S. 19–33). Wiesbaden: Gabler.
- Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology* 99, 282–295.
- Sperka, M., & Rózsa, J. (2007). *KOMMINO: Fragebogen zur Erfassung der Kommunikation in Organisationen*. Göttingen: Hogrefe.
- Stangel-Meseke, M., Akli, H., & Schnelle, J. (2005). Lernförderliches Feedback im modifizierten Lernpotenzial-Assessment Center: Umsetzung der Forschungsergebnisse in einer betrieblichen Studie. *Zeitschrift für Personalpsychologie* 4, 187–194.
- Stegmann, S., van Dick, R., Ullrich, J., Charalambous, J., Menzel, B., Eggold, N., & Wu, T. T.-C. (2010). Der Work Design Questionnaire. *Zeitschrift für Arbeits- und Organisationspsychologie* 54, 1–28.
- Steiner, D. D. (2012). Personnel selection across the globe. In N. Schmitt (Hrsg.), *The Oxford Handbook of Personnel Assessment and Selection* (S. 740–767). New York: Oxford University Press.
- Tillema, H. H. (1998). Assessment of potential, from assessment centers to development centers. *International Journal of Selection and Assessment* 6, 185–191.
- Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Voraussage des Studienerfolgs – eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie* 21, 132–151.
- Van Iddekinge, C. H., Raymark, P. H., Eidson, J., Carl E & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance* 17, 71–93.
- Van Iddekinge, C. H., Putka, D. J., & Campbell, J. P. (2011a). Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance, and continuance intentions. *Journal of Applied Psychology* 96, 13–33.
- Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011b). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology* 96, 1167–1194.
- Vanhove, A. J., Gibbons, A. M., & Kedarnath, U. (2016). Rater agreement, accuracy, and experienced cognitive load: Comparison of distributional and traditional assessment approaches to rating performance. *Human Performance* 29, 378–393.
- Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment* 18, 52–62.
- Ward, P. (1997). 360 degree feedback. London, UK: Charter House.
- Weekley, J. A., & Ployhart, R. E. (2013). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational and Organizational Psychology* 61, 275–290.
- Woehr, D., Arthur, W., & Meriac, J. (2007). Methodenfaktoren statt Fehlervarianz: eine Metaanalyse der Assessment Center-Konstruktvalidität. In H. Schuler (Hrsg.), *Assessment Center zur Potentialanalyse* (S. 81–108). Göttingen: Hogrefe.
- Ziegler, M., MacCann, C., & Roberts, R. (2011). *New perspectives on faking in personality assessment*. Oxford: Oxford University Press.



Diagnostik in der Pädagogischen Psychologie

Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang

Inhaltsverzeichnis

- 7.1 Diagnostik zur Schullaufbahnberatung – 644**
 - 7.1.1 Schuleingangsdiagnostik – 645
 - 7.1.2 Diagnostik zur Feststellung von sonderpädagogischem Förderbedarf – 648
 - 7.1.3 Diagnostik beim Übertritt in den tertiären Bildungsbereich – 654
- 7.2 Diagnostik bei Schulschwierigkeiten – 660**
 - 7.2.1 Diagnostik bei Lernschwierigkeiten – 660
 - 7.2.2 Diagnostik von Teilleistungsstörungen – 661
- 7.3 Hochbegabungsdiagnostik – 667**
- 7.4 Tests im Bildungsbereich – 673**
 - 7.4.1 Schultests – 673
 - 7.4.2 Tests zur Evaluierung des Bildungssystems – 677
- 7.5 Zusammenfassung – 682**
- Literatur – 684**

■ Vorbemerkungen

Die Pädagogische Psychologie befasst sich mit den psychologischen Grundlagen von Erziehung und Bildung.

- » Ein großes Forschungsfeld der Pädagogischen Psychologie mit unmittelbarer Bedeutung für die Bildungspraxis ist das Gebiet der pädagogisch-psychologischen Diagnostik und Intervention. Die Pädagogische Psychologie hat eine lange und erfolgreiche Tradition der Entwicklung von Testverfahren für den Bildungsbereich, etwa zur Diagnostik von schulischen Leistungen und den zugrunde liegenden kognitiven Fähigkeiten, von Lernstörungen und Lernbehinderungen, von Verhaltensauffälligkeiten bei Schüler(inne)n, von Hochbegabung, von lernrelevanten Persönlichkeitsmerkmalen oder beruflichen Interessen. (Richter et al. 2019, S. 111)

In diesem Kapitel fokussieren wir auf den Bildungsbereich Schule, der vom Schuleintritt bis zum Übergang in den tertiären Bildungsbereich (Hochschulen) reicht. Den Vorschulbereich klammern wir hier weitgehend aus, verweisen aber auf ► Abschn. 3.2.5, in dem Entwicklungstests vorgestellt werden. In ► Abschn. 7.1.1 befassen wir uns zunächst unter dem Begriff „Schullaufbahnberatung“ mit der Diagnostik, die der Navigation im Schulsystem dient. Anschließend gehen wir auf die Diagnostik bei speziellen Fragestellungen ein: Lernschwierigkeiten und Teilleistungsstörungen (► Abschn. 7.2) sowie Hochbegabung (► Abschn. 7.3). Diesbezüglich dient die Diagnostik den Individuen, also Schülerinnen und Schülern. Eine Perspektive, die zunehmend an Bedeutung gewinnt, ist die der Institutionen. Das Bildungssystem wird unter Einsatz von aufwendig in internationaler Kooperation entwickelten Tests evaluiert. Hier werden zwar auch Individuen getestet, die individuellen Ergebnisse sind jedoch nicht vorrangig von Interesse, sondern es kommt auf die für ein Land oder eine Region gemittelten Ergebnisse an. Der ► Abschn. 7.4 dieses Kapitels befasst sich mit Tests, die für die Individualdiagnostik geeignet sind, sowie mit Verfahren wie PISA (Programme for International Student Assessment), die ausschließlich der institutionellen Diagnostik dienen.

7.1 Diagnostik zur Schullaufbahnberatung

Der Einsatz diagnostischer Verfahren bei der Schullaufbahnberatung wird begründet durch den Wunsch nach Optimierung der Passung zwischen Lernvoraussetzungen von Kindern bzw. Jugendlichen und schulischen Anforderungen. Damit sollen auch frustrierende Erfahrungen durch schulische Überforderungen vermieden werden. Die entsprechende pädagogisch-psychologische Diagnostik findet vor allem beim Eintritt in die Schule, bei der Wahl der passenden Schulform und beim Übergang von der Schule zur Hochschule statt. Bei der Wahl der passenden Schulform beschränken wir uns hier auf die Frage, ob bei einer Schülerin oder einem Schüler ein sonderpädagogischer Förderbedarf besteht.

Auch für Schülerinnen und Schüler, die innerhalb der Regelschule einen Wechsel, beispielsweise von der Grundschule auf ein Gymnasium, in Erwägung ziehen, kann im Einzelfall Diagnostik zur Schullaufbahn hilfreich sein. Wenn Zweifel an der Eignung für eine bestimmte Schulform bestehen, können je nach Anlass für diese Zweifel Hypothesen formuliert und überprüft werden. So kann etwa der aktuelle Leistungsstand in einzelnen Schulfächern mithilfe von gut normierten Schulleistungstests objektiviert werden, wenn eine Einschätzung unabhängig von der Beurteilung durch Lehrerinnen und Lehrer erwünscht ist. Für eine Prognose künftiger Schulleistungen ist neben dem Vorwissen (Lernstand) die kognitive Leistungsfähigkeit und hier insbe-

Diagnostik auch bei Wechsel innerhalb der Regelschule

sondere die Intelligenz relevant (zu Intelligenztests s. ▶ Abschn. 3.2.3). Die Zweifel an der Eignung für die momentan besuchte oder die in Erwägung gezogene Schule können jedoch auch spezielle Merkmale im kognitiven Bereich sowie emotionale, motivationale und soziale Bedingungsfaktoren betreffen (das generelle diagnostische Vorgehen wird in ▶ Kap. 4 beschrieben). Als Beispiele für kognitive Merkmale sind Konzentrationsfähigkeit, sprachliche Kompetenzen und Lernstrategien zu nennen. Bei den emotionalen Faktoren sind besonders die Schul- und die Leistungsangst zu erwähnen. Im motivationalen Bereich kann etwa die Leistungsmotivation zu überprüfen sein. Beispiele für soziale Faktoren sind die Unterstützung durch die Eltern oder Geschwister, die Beziehung zu Mitschülern bzw. Mitschülerinnen und das Verhältnis zu den Lehrerinnen und Lehrern. Die auf einer gründlichen Diagnostik aufbauende Beratung muss nicht zwangsläufig zu einer Empfehlung für einen Wechsel der Schule oder Schulform führen; manche Probleme können auch anders als durch einen Schulwechsel gelöst werden.

7.1.1 Schuleingangsdiagnostik

Eine alte Vorstellung von Schulreife besagt, dass es sich dabei um einen biologischen Reifeprozess handelt. Die Beobachtung, dass Schülerinnen und Schüler, die den Anforderungen der 1. Schulklasse nicht gewachsen waren, 1 Jahr später dem Unterricht folgen konnten, ließ Kern (1963) vermuten, der Schulerfolg sei eine Funktion der Schulreife, die sich zu unterschiedlichen Zeitpunkten von selbst einstelle. Kinder sollten demnach in die Schule eintreten, wenn sie alt genug sind und die erwartete Reife erreicht haben (Snow 2006). Das vorrangige Ziel besteht nach dieser Vorstellung darin, das richtige Einschulungsalter zu finden (Abb. 7.1). International schwankt dieses zwischen 5 und 8 Jahren. In Deutschland werden Kinder mit 6 Jahren schulpflichtig bzw. wenn sie bis zu einem Stichtag das 6. Lebensjahr erreicht haben. Da die Schulreife interindividuell variiert, können Kinder sowohl vorzeitig als auch später eingeschult werden. Internationalen Studien zufolge werden 4 %

„Schulreife kommt mit der Zeit von selbst“



Abb. 7.1 Die Ergebnisse von Schuleingangstests helfen bei der Beantwortung der Frage, ob ein Kind jetzt oder besser ein Jahr später eingeschult werden sollte. (© Irina Schmidt/stock.adobe.com)

der Kinder vor und 17 % nach Erreichen des jeweiligen offiziellen Einschulungsalters eingeschult (Kammermeyer 2010).

Mit Schulreifetests zu frühe Einschulung verhindern

Durch den Einsatz geeigneter Schulreifetests (► Abschn. 7.4.1.1) wollte man verhindern, dass noch nicht schulreife Kinder zu früh eingeschult werden. Der von Kern (1963) für diese Zwecke vorgeschlagene Grundleistungs-test konnte die Aufgabe jedoch nur sehr unvollkommen erfüllen. Immerhin sank die Sitzenbleibendenquote, nachdem Kinder aufgrund von Testergebnissen zurückgestellt wurden (Kammermeyer 2010).

Regression zur Mitte

Gegen ein Screening zur Entdeckung fehlender Schulreife gibt es ein gewichtiges Argument, die „Regression zur Mitte“. Sie wird nun erläutert, weil sie auch andere diagnostische Urteile wie „Teilleistungsstörung liegt vor“ oder „hochbegabt“ betrifft. Noch allgemeiner ausgedrückt ist das nachfolgend beschriebene Phänomen der Regression zur Mitte für alle Diagnosen relevant, die bei extremem Abschneiden (hoch sowie niedrig) in diagnostischen Instrumenten gestellt werden.

7

Definition

Unter der **Regression zur Mitte** versteht man das Phänomen, dass sich extreme Messwerte bei einer Testwiederholung in Richtung Skalenmitte (dem Durchschnittswert) verschieben. Der Effekt ist umso größer, je extremer der Messwert bei der ersten Messung ist.

Einfache Schätzung des Testwerts bei Testwiederholung

Messwerte sind aufgrund der begrenzten Retest-Reliabilität eines Tests nicht exakt replizierbar. Bei der Regression zur Mitte kommt hinzu, dass die Extremität u. a. durch die ungewöhnliche Kombination vieler förderlicher bzw. hinderlicher Fehlerfaktoren bewirkt wird, die in dieser spezifischen Konstellation kaum wieder auftreten werden. Das Konfidenzintervall ist daher asymmetrisch, d. h., es ist zur Skalenmitte hin größer als zum Extrembereich hin. Je niedriger die Retest-Reliabilität eines Tests ist, desto stärker ist die Regression zur Mitte (► Abschn. 2.6.2.2). Im Grunde tragen alle Faktoren, die die Korrelation zwischen 2 Messungen reduzieren, zur Regression zur Mitte bei (Lohman und Korb 2006). Für die Schätzung des Messwertes bei einer Testwiederholung steht eine simple Formel zur Verfügung (vgl. Lohman und Korb 2006), die wir für z-Werte umgeformt haben (vgl. ► Abschn. 2.6.2.2):

Formel zur Schätzung der Regression zur Mitte

$$\hat{z}_2 = z_1 * r_{12}$$

dabei bedeuten:

z_1 = beobachteter z-Wert zum Testzeitpunkt 1

\hat{z}_2 = geschätzter z-Wert zum Testzeitpunkt 2

r_{12} = Retest-Reliabilität oder Korrelation zwischen Test 1 und Test 2

► Beispiel

Alle Kinder eines Schulbezirks werden vor der Einschulung mit einem Schuleingangstest ($r_{12}=.77$) untersucht. Ein Kind erzielt einen Standardwert von 75, der einem z-Wert von -2,5 entspricht. In die Formel $\hat{z}_2 = z_1 * r_{12}$ eingesetzt ergibt sich $\hat{z}_2 = -2,5 * .77 = -1,93$. Der 2. Messwert wird also voraussichtlich 1,93 Standardabweichungen unter dem Mittelwert liegen. Bei Standardwerten ($M=100$, $SD=10$) entspricht das einem Wert von 81. ◀

Wer nun vermutet, das wäre statistische Zauberei, muss sich eines Besseren belehren lassen. Lohman und Korb (2006) haben Daten aus einer bereits publizierten Längsschnittstudie analysiert, in der über 6000 Schülerinnen und Schüler von der 3. bis zur 8. Klasse jährlich den gleichen kognitiven Leistungstest bearbeiteten. Sie ermittelten für die 3. Klasse, welche Kinder zu den besten 3 % gehörten und verfolgten deren Position im folgenden Jahr und darüber hinaus. Obwohl der Test eine hohe Retest-Reliabilität von $r_{tt}=.91$ (von Klasse 3 zu Klasse 4) aufwies, gehörten in der 4. Klasse nur noch gut 60 % dieser Personen zu den besten 3 %. Lohman und Korb (2006, Tab. 1, S. 457) haben für verschiedene Cut-off-Werte und Retest-Reliabilitäten berechnet, wie viele Personen bei der 2. Messung noch in die gleiche Kategorie fallen sollten. Für die besten 3 % und $r_{tt}=.90$ geben sie 60 % an – das beobachtete Ergebnis deckt sich damit sehr gut. Für die weniger stabilen Subtests war der Regressionseffekt erwartungsgemäß noch stärker; in der 4. Klasse gehörten nur noch 50 % zu den besten 3 %.

Bei Testwiederholung gehen viele der Besten verloren

Wie findet man wirklich extreme Merkmalsauprägungen?

Lohman und Korb (2006) diskutieren 3 Strategien, die auch in ▶ Abschn. 5.1.3.2 als Verrechnungsregeln in Urteilsmodellen vorgestellt werden. Sie gehen exemplarisch von einer Korrelation zwischen zwei Tests von $r=.80$ aus und wollen die „wirklich“ besten 3 % der Testpersonen finden. Dazu wenden sie die Formel zur Schätzung der Regression zur Mitte (s. o.) an. Nach der Und-Regel (in Test 1 *und* Test 2 zu den besten 3 % gehören) verbleiben nur 1,35 %. Die Oder-Regel (in Test 1 *oder* Test 2 zu den besten 3 % gehören) ist mit 4,65 % Treffern zu „liberal“. Dem Ziel am nächsten kommt man mit der Kombinationsregel (Werte jeder Person über Test 1 und Test 2 gemittelt); sie führt zu einem Anteil von 2,4 % an Personen mit den besten Testwerten. Um die „wirklich“ 3 % besten zu finden, müsste hier der Cut-off-Wert für den Mittelwert noch leicht unter 3 % abgesenkt werden. Es sei daran erinnert, dass bei der Selektion der 3 % Personen mit den schlechtesten Testwerten das Gleiche gilt.

Kombination von 2 Testwerten zur Schätzung des wahren Werts

Eine Alternative wäre, auf Schuleingangstests ganz zu verzichten und das Ergebnis der schulischen Bewährung abzuwarten. Die mit einem schulischen Versagen verbundenen negativen Lernerfahrungen lassen es jedoch geboten erscheinen, einem noch nicht schulfähigen Kind möglichst bald die ständigen Überforderungserlebnisse zu ersparen. Die Nachteile einer Zurückstellung haben sich verringert, seit Vorschulklassen zurückgestellten Schulanfängerinnen bzw. -anfängern eine angemessene Lernumwelt bereitstellen.

Verzicht auf Schulreifetests

Schulreife ist ein mehrdimensionales Konstrukt (Snow 2006). Daher bietet es sich an, mehrere Bereiche zu betrachten, um spezifische Defizite zu erkennen. In ▶ Abschn. 7.4.1.1 werden Tests vorgestellt, die zur Schuleingangsdagnostik verwendet werden können. Ergänzend oder auch alternativ bietet sich der Einsatz von Entwicklungstests (▶ Abschn. 3.2.5) an. Die Psychologische Diagnostik kann detaillierte Ratschläge für eine gezielte Förderung begründen. Durch eine entsprechende Intervention so früh wie möglich – noch vor der Einschulung oder zu Beginn der Schulzeit – können eventuell die erkannten Rückstände aufgeholt werden. Kindertagesstätten und Schulen haben bereits geeignete Trainingsprogramme aufgenommen (Kamermeyer 2010).

Schulreife ist mehrdimensional

Kamermeyer (2010) zufolge sind 3 Merkmale oder Merkmalsbereiche besonders wichtig für den späteren Schulerfolg: die phonologische Bewusstheit, mengen- und zahlenbezogenes Vorwissen sowie bereichsübergreifende Fähigkeiten. Für den Schriftspracherwerb ist die *phonologische Bewusstheit*

Bestimmte Merkmale für späteren Schulerfolg relevant

Geringer Zusammenhang zwischen nichtkognitiven Fertigkeiten und Schulleistung

7

von großer Bedeutung. Darunter versteht man die Einsicht in die Lautstruktur der Sprache (reimen, Silben segmentieren, Anlaute erkennen etc.). *Mengen- und zahlenbezogenes Vorwissen* stellt eine bedeutsame Voraussetzung für das Mathematiklernen dar. Zu den *bereichsübergreifenden Fähigkeiten* gehören die Selbstdisziplin und die Fähigkeit, soziale Informationen angemessen zu verarbeiten. Die letztgenannten Fähigkeiten fallen in eine Kategorie, die man als nichtkognitive Fertigkeiten bezeichnen kann.

In einer Metaanalyse haben sich Smithers et al. (2018) mit der Relevanz eines breiten Spektrums nichtkognitiver Fertigkeiten (dazu zählten sie neben den oben genannten Beispielen u. a. auch exekutive Funktionen, Aufmerksamkeit und Gewissenhaftigkeit) für Schulleistungen (Lesen/Schreiben, Rechnen) befasst. Die in den eingeschlossenen Studien untersuchten Kinder waren normal entwickelt und bei der Messung der schulischen Leistungen bis zu 12 Jahre alt. In den einzelnen Studien wurden unterschiedliche Maße für die Effektstärke verwendet. Da kein einheitliches Effektstärkemaß zur Anwendung kam, geben Smithers et al. (2018) die Ergebnisse neutral in Standardabweichungen an. In experimentellen und quasi-experimentellen Interventionsstudien (wobei nur Studien mit hoher Qualität berücksichtigt wurden) betrug der Effekt bei sprachlichen Fertigkeiten 0,37 und bei rechnerischen Fertigkeiten durchschnittlich 0,38 Standardabweichungen. In Beobachtungsstudien waren die Effekte mit durchschnittlich 0,16 bzw. 0,17 Standardabweichungen deutlich kleiner. Zum Vergleich sei der Zusammenhang zwischen Intelligenz und Schulleistung in der Grundschule genannt, der (nur korrigiert für den Stichprobenfehler) $r = .44$ beträgt (Roth et al. 2015), was einer Effektstärke von $d = 0,98$ entspricht. Im Vergleich zum Effekt kognitiver Faktoren auf die Schulleistung ist der Effekt der von Smithers et al. (2018) untersuchten „nichtkognitiven“ Faktoren also relativ klein.

7.1.2 Diagnostik zur Feststellung von sonderpädagogischem Förderbedarf

Definition

„**Sonderpädagogischer Förderbedarf** ist bei Kindern und Jugendlichen anzunehmen, die in ihren Bildungs-, Entwicklungs- und Lernmöglichkeiten so beeinträchtigt sind, dass sie im Unterricht der allgemeinen Schule ohne sonderpädagogische Unterstützung nicht hinreichend gefördert werden können“ (Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland 1994, S. 5).

Diese Definition findet sich in den Ländergesetzen und Verordnungen wieder, so etwa in Brandenburg (Sonderpädagogik-Verordnung, § 1 Absatz 5; ▶ <https://bravors.brandenburg.de/verordnungen/sopv#1>).

10 Schulformen für die sonderpädagogische Förderung

Bevor wir uns der „Feststellung“ von sonderpädagogischem Förderbedarf (der Diagnostik) zuwenden, ist ein Blick auf die vorhandenen Förderschwerpunkte hilfreich (► Tab. 7.1).

Der ► Tab. 7.1 kann man entnehmen, dass 2017/2018 über eine halbe Million Kinder und Jugendliche in Deutschland eine sonderpädagogische Förderung erhalten haben. Das entspricht 7,2 % der Schülerinnen und Schüler in den Klassen 1 bis 10 an den allgemeinbildenden Schulen. Den höchsten Anteil hat der Förderschwerpunkt „Lernen“, gefolgt von den Schwerpunkten „geistige Entwicklung“ und „emotionale und soziale Entwicklung“ (früher:

Tab. 7.1 Förderschwerpunkte und Anzahl der Schülerinnen und Schüler im Schuljahr 2017/2018

Förderschwerpunkte	Schülerinnen und Schüler mit sonderpädagogischer Förderung		Förderquote (%)
	An Förderschulen	An allgemeinen Schulen	
Insgesamt	317.480	227.485	7,2
Lernen	86.200	100.425	2,5
Sehen	4615	4553	0,1
Hören	10.615	10.337	0,3
Sprache	28.843	27.271	0,8
Körperliche u. motorische Entwicklung	23.808	13.172	0,5
Geistige Entwicklung	79.373	11.832	1,2
Emotionale u. soziale Entwicklung	39.883	51.741	1,2
Lernen, Sprache, emotionale u. soziale Entwicklung	19.755	0	0,3
Förderschwerpunkt übergreifend	3246	1494	0,1
Noch keinem Förderbedarf zugeordnet	10.225	1241	0,2
Schulen für Kranke	10.917	419	–

Quelle: Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland (2020, S. 5, mit freundlicher Genehmigung). Die Förderquote bezieht sich auf den Anteil an den 7.361.885 Schülerinnen und Schüler, die im gleichen Zeitraum eine allgemeinbildende Schule der Klassen 1 bis 10 besuchten

„Erziehungsschwierige“). Die Förderung erfolgt sowohl an allgemeinen Schulen als auch an speziellen Förderschulen. Insgesamt 10 Schulformen stehen für verschiedene Förderschwerpunkte zur Verfügung. Sehr hoch ist der Anteil an Schülerinnen und Schülern, die eine Förderschule besuchen, bei den Förderschwerpunkten „geistige Entwicklung“ sowie „Lernen, Sprache, emotionale und soziale Entwicklung“. Die Schulen für Kranke sind zumeist Krankenhäusern angegliedert, sodass es nicht verwundert, dass die Schülerinnen und Schüler in der Regel dort unterrichtet werden und nicht in einer allgemeinen Schule.

Warum findet die Förderung nicht ganz überwiegend in speziellen Förderschulen statt? Dafür gibt es mehrere Gründe; einer davon ist die Idee der Inklusion, die auch in der UN-Behindertenrechtskonvention festgeschrieben ist. Das „Übereinkommen über die Rechte von Menschen mit Behinderungen“, das 2006 von der UN-Generalversammlung beschlossen wurde, ist 2009 auch in Deutschland in Kraft getreten. Artikel 24 betrifft die Bildung. Dort steht u. a. (Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderungen 2017):

UN-Behindertenrechtskonvention

» UN-Behindertenrechtskonvention, Artikel 24 Bildung

- (2) Bei der Verwirklichung dieses Rechts stellen die Vertragsstaaten sicher, dass
- Menschen mit Behinderungen nicht aufgrund von Behinderung vom allgemeinen Bildungssystem ausgeschlossen werden und dass Kinder mit Behinderungen nicht aufgrund von Behinderung vom unentgeltlichen und obligatorischen Grundschulunterricht oder vom Besuch weiterführender Schulen ausgeschlossen werden;
 - Menschen mit Behinderungen gleichberechtigt mit anderen in der Gemeinschaft, in der sie leben, Zugang zu einem integrativen, hochwertigen und unentgeltlichen Unterricht an Grundschulen und weiterführenden Schulen haben;
 - angemessene Vorkehrungen für die Bedürfnisse des Einzelnen getroffen werden;
 - Menschen mit Behinderungen innerhalb des allgemeinen Bildungssystems die notwendige Unterstützung geleistet wird, um ihre erfolgreiche Bildung zu erleichtern;
 - in Übereinstimmung mit dem Ziel der vollständigen Integration wirksame individuell angepasste Unterstützungsmaßnahmen in einem Umfeld, das die bestmögliche schulische und soziale Entwicklung gestattet, angeboten werden.

Menschen mit Behinderungen soll eine umfassende Partizipation in der Gesellschaft und speziell auch im Bildungssystem ermöglicht werden. Konkret bedeutet das für den schulischen Bereich, dass „Inklusion“ möglich sein muss. Während früher etwa Kinder und Jugendliche mit einer Intelligenzminderung ganz überwiegend Förderschulen besuchten, haben sie heute das Recht, auch eine allgemeinbildende Schule zu besuchen. Aus Sicht der Eltern bietet sich dadurch die Chance, dass ihre Kinder besser integriert werden. So schreibt etwa das Hessische Kultusministerium ([2020](#)) auf seiner Homepage:

- » Den Eltern der betroffenen Schülerinnen und Schülern steht es selbstverständlich nach wie vor frei, selbst zu entscheiden, ob ihr Kind in einer allgemeinen oder einer Förderschule unterrichtet wird. **Die Eltern sollen den Förderort wählen können, den sie für das Wohl ihres Kindes als am besten geeignet beurteilen.**

Inklusion ermöglichen

Und im hessischen Schulgesetz in der Fassung vom 30. Juni 2017 wird der Förderauftrag dem entsprechend in § 49 festgelegt:

» Hessisches Schulgesetz, § 49

- Kinder und Jugendliche, die zur Gewährleistung ihrer körperlichen, sozialen und emotionalen sowie kognitiven Entwicklung in der Schule sonderpädagogischer Hilfen bedürfen, haben einen Anspruch auf sonderpädagogische Förderung.
- Den Anspruch auf sonderpädagogische Förderung erfüllen die allgemein bildenden und beruflichen Schulen nach § 11 Abs. 3, die nicht Förderschulen sind (allgemeine Schulen), sowie die Förderschulen mit ihren verschiedenen Förder schwerpunkten nach § 50 Abs. 1.

Die Vermutung, bei einem Kind könne sonderpädagogischer Förderbedarf vorliegen, kommt oft schon vor der Einschulung auf. Liegt eine Sinnes- oder Körper-Behinderung oder eine Intelligenzminderung vor, so ist dies den Eltern zumeist schon lange vorher bekannt und sie haben daher in der Regel bereits Kontakt mit einer vorschulischen Frühförderereinrichtung. Dennoch kann der Wunsch bestehen, das Kind in die Regelschule zu schicken, weil dort eine bessere Förderung und/oder soziale Integration erwartet wird. Der Verdacht, dass eine Lern-, eine Sprachbehinderung oder eine Verhaltensstörung vorliegt, kann sich auch erst während der Schulzeit entwickeln.

Am Beispiel des Hessischen Schulgesetzes (in der Fassung vom 30. Juni 2017) wird das Vorgehen zur Feststellung von sonderpädagogischem Förderbedarf kurz skizziert. Entscheidend ist, ob ein „Anspruch auf sonderpädagogische Förderung“ bestehen kann. Wenn ja, können die Eltern beantragen, dass ihr Kind in eine Förderschule aufgenommen werden soll. Wenn sie das nicht tun, entscheidet nach § 54 Ziffer 1 die Schulleiterin oder der Schulleiter „nach Anhörung der Eltern und im Benehmen mit der Schulaufsichtsbehörde über Art, Umfang und Organisation der sonderpädagogischen Förderung“.

Damit ist geklärt, wer was beantragen kann. Oben lautete die Formulierung, dass ein Anspruch auf sonderpädagogische Förderung „bestehen kann“. Folglich ist nun zu überprüfen, ob ein solcher Anspruch auch tatsächlich besteht. In § 54 Ziffer 2 steht, was zu tun ist: Es wird ein „Förderausschuss“ einberufen, dem nach § 54 Ziffer 3 verschiedene Personen angehören:

1. Die Schulleiterin oder der Schulleiter
2. Eine Lehrkraft der allgemeinen Schule, die das Kind unterrichtet
3. Eine Lehrkraft des sonderpädagogischen Beratungs- und Förderzentrums oder der zuständigen Förderschule (im Auftrag der Schulaufsichtsbehörde)
4. Die Eltern des Kindes
5. Eine Vertreterin oder ein Vertreter des Schulträgers, wenn der Unterricht in der allgemeinen Schule besondere räumliche und sächliche Leistungen erfordert

Eventuell kommen weitere Personen mit beratender Stimme hinzu. Der Förderausschuss hat die Aufgabe, eine Empfehlung über Art, Umfang und Organisation der sonderpädagogischen Förderung vorzuschlagen, der eine Stellungnahme des sonderpädagogischen Beratungs- und Förderzentrums und, wenn erforderlich, ein schulärztliches sowie in Zweifelsfällen ein schulpsychologisches Gutachten zugrunde liegt. Die Empfehlung ist erst bindend, wenn sie auch von der Schulaufsichtsbehörde genehmigt wurde. Im Schulgesetz wird weiter geregelt, wie in Konfliktfällen zu verfahren ist.

Der Elternwunsch hat einen hohen Stellenwert. Auch wenn nicht die Eltern die Feststellung von sonderpädagogischem Förderbedarf beantragen, so sind sie als Mitglieder des Förderausschusses doch in den Entscheidungsprozess eingebunden. Die Entscheidung trifft am Ende das Staatliche Schulamt. Ein schulpsychologisches Gutachten kann dabei mitentscheidend sein; es wird aber nur bei Bedarf eingeholt.

■ „Lernbehinderung“

Wir gehen exemplarisch auf die größte Kategorie der Förderbedarfe, das Lernen, ein. Sie zeichnet sich durch eine begriffliche Unklarheit aus, zumindest, wenn man hier eine Behinderung („Lernbehinderung“) konstruieren will. Das diagnostische Vorgehen, das auch für andere Förderbedarfe gilt, wird kurz skizziert.

Sonderpädagogischer Förderbedarf
oft schon vor der Einschulung
absehbar

Förderausschuss

Elternwunsch, schulpsychologisches
Gutachten

„Lernbehinderung“ nicht definiert

Anstelle einer Definition

Sichtweisen ändern sich. Der Begriff „Lernbehinderung“ wurde in den 1960er-Jahren in die Fach- und Amtssprache in Deutschland eingeführt; im internationalen Sprachgebrauch existiert kein entsprechender Begriff (Grünke 2004). Wenn Parmar (2004) in einem Beitrag zu „Lernbehinderung“ auf eine offizielle amerikanische Definition von „specific learning disability“ verweist, so ist das irreführend. Die American Psychiatric Association (2019) stellt klar, dass sie darunter eine spezifische Beeinträchtigung schulischer Fertigkeiten wie das Lesen oder Rechnen versteht, die nicht mit mangelnder intellektueller Begabung, ökonomischer Benachteiligung etc. erklärt werden kann. Das wird in Deutschland als Teilleistungsstörung bezeichnet (► Abschn. 7.2.2). Früher gab es Schulen für Lernbehinderte. Aus Sicht dieser Institution und auch aus der des gesamten Schulsystems war es klar, dass es auch eine sog. „Lernbehinderung“ gibt. Die Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland (1994, S. 2) hat in ihren Empfehlungen zur sonderpädagogischen Förderung in den Schulen der Bundesrepublik Deutschland einleitend vorgeschlagen, sich von dieser Sichtweise der Institutionen zu lösen und statt von „Sonderschulbedürftigkeit“ lieber von „sonderpädagogischem Förderbedarf“ zu sprechen. Damit wird eine vom Individuum ausgehende Sichtweise bevorzugt. Zudem impliziert der Begriff, dass nicht allein eine Sonderschule für die Befriedigung des Förderbedarfs zuständig ist (heute leisten das zum Großteil die allgemeinen Schulen; vgl. □ Tab. 7.1). Im Bildungsbereich wird heute der Begriff „Lernbehinderung“ eher vermieden, obwohl in den Empfehlungen der Kultusministerkonferenz von 1994 der Begriff Behinderung sehr wohl gebraucht wurde, ebenso in der UN-Behindertenrechtskonvention (s. o.).

Keine Generalisierung auf alle Formen der Behinderungen

Angesichts der schon lange bestehenden begrifflichen Unklarheit ist es nicht verwunderlich, dass die Zahl der „Lernbehinderten“ variiert. Anfang der 1980er-Jahre galten in der BRD 3,75 % der Schülerinnen und Schüler als lernbehindert, in der DDR dagegen nur 2,0 %. Zu Beginn der 1990er-Jahre (nach der Wiedervereinigung) galten in den alten Bundesländern nur noch 2,1 % als lernbehindert, während die Quote in den neuen Bundesländern nun 3,1 % betrug (Langfeldt und Tent 1999). Der Vorschlag, auf den Begriff der Behinderung im Bildungsbereich zu verzichten, soll nicht generalisiert werden. Bei einer Seh- oder Hörbehinderung etwa ist es sinnvoll, von Behinderungen zu sprechen. Erstens liegen hier im Individuum verortete und zudem genau spezifizierbare Faktoren vor, während bei „Lernbehinderung“ interne (z. B. niedrige Intelligenz, niedrige Kompetenz in der deutschen Sprache, niedrige Lernmotivation) und externe Faktoren (z. B. wenig Unterstützung durch das Elternhaus, bildungsfeindliche Einstellung der sozialen Umgebung) in variabler Kombination zusammentreffen. Zweitens führt die Anerkennung eines bestimmten Grades der Behinderung zu einem Nachteilsausgleich.

Wie sieht das diagnostische Vorgehen aus, wenn es statt einer „Lernbehinderung“ um die Frage geht, ob bei einem Kind oder Jugendlichen Bedarf an sonderpädagogischer Förderung im Bereich des Lernens besteht? Die einfache Antwort lautet, dass genau das zu untersuchen ist. Ein Gutachtenauftrag könnte so lauten. Dabei ist allerdings zu beachten, dass im

Inhalte und Methoden des förderdiagnostischen Gutachtens vorgegeben

schulischen Bildungsbereich die Vorgehensweisen durch Gesetze, Verordnungen und zu verwendete Formblätter gesteuert wird. In Hessen regelt die „Verordnung über Unterricht, Erziehung und sonderpädagogische Förderung von Schülerinnen und Schülern mit Beeinträchtigungen oder Behinderungen (VOSB)“ vom 15. Mai 2012 (aktuellste verfügbare Fassung unter ► <https://www.rv.hessenrecht.hessen.de/bshc/document/hevr-SBUntErz-SoP%C3%A4dFVHEV2P2>) in § 28 den Inhalt des förderdiagnostischen Gutachtens gemäß § 54 Absatz 5 des Schulgesetzes. Darin wird nicht nur, wie bei einem Gutachten üblich (vgl. ► Abschn. 4.6), ein Auftrag formuliert, sondern es werden auch Inhalte und Methoden vorgeschrieben. Der Ansatz ist trotz der genannten Restriktionen brauchbar, weil er unter relativ standardisierten Bedingungen zu einer Antwort führt. Inhalte des förderdiagnostischen Gutachtens sind:

» § 28 VOSB Inhalt des förderdiagnostischen Gutachtens

(2) Das förderliche Gutachten enthält

1. ein auf die schulischen Anforderungen hin bezogenes Kompetenz- und Entwicklungsprofil mit Bezug auf das Lernumfeld,
2. Aussagen zur Wirkung eines angewandten Nachteilsausgleichs,
3. eine Darstellung gegebenenfalls erforderlicher geeigneter Lehr- und Lernmittel [...],
4. Empfehlungen über notwendige weitere Fördermaßnahmen unter anderem zur Weiterentwicklung des Lernens, der Sprache sowie der körperlichen, sozialen und emotionalen Entwicklung,
5. eindeutige Empfehlungen zu Art, Umfang und Organisation der zum Wohl des Kindes und seiner weiteren Entwicklung notwendigen sonderpädagogischen Förderung [...]

In methodischer Hinsicht wird vorgegeben:

» § 28 VOSB Inhalt des förderdiagnostischen Gutachtens

(1) Das förderdiagnostische Gutachten nach § 54 Abs. 5 des Schulgesetzes beruht auf

1. einer Darstellung des schulischen Lernstands anhand vorhandener individueller Förderpläne, Zeugnisse, der Anwendung des Nachteilsausgleichs und schulischer Stellungnahmen,
2. der Feststellung der Lernausgangslage und der Lernbedingungen anhand von Unterrichtshospitationen, Gesprächen mit den Eltern und mit Personen, die das Kind in schulischen und außerschulischen Einrichtungen fördern sowie der Auswertung diagnostischer Verfahren,
3. der Auswertung medizinischer Untersuchungsberichte und Stellungnahmen der Jugendhilfe oder anderer Maßnahmeträger,
4. dem Ausloten der Förderchancen aufgrund einer eingehenden Kind-Umfeld-Analyse unter Einbeziehung tatsächlicher oder einzurichtender schulischer und außerschulischer Fördermöglichkeiten.

Es existieren zum Teil sehr detaillierte Vorgaben, wie der Bedarf an sonderpädagogischer Förderung festzustellen ist. Speziell für die Begutachtung findet man sehr sinnvolle Vorgaben, wie die individuellen Stärken und Schwächen sowie das bisherige Lernumfeld und der aktuelle Lernstand eines be-

troffenen Kindes festgestellt werden soll. Die psychologische Diagnostik (im engeren Sinne) zur Feststellung der kognitiven Leistungsfähigkeit, der Motivation, der Aufmerksamkeit und Konzentration, der Merkfähigkeit – oder zu was auch immer sinnvolle psychologische Fragen (► Abschn. 4.3) formuliert werden können, findet in existierenden Vorgaben kaum Beachtung.

7.1.3 Diagnostik beim Übertritt in den tertiären Bildungsbereich

7.1.3.1 Auswahlverfahren

Bis 2019 zulässige Auswahlkriterien nach dem HRG

7

In der Bundesrepublik Deutschland übertrifft seit geraumer Zeit die Nachfrage zum mindesten nach bestimmten Studienplätzen das von den Universitäten bereitgehaltene Angebot. Es steht zu befürchten, dass auch in Zukunft eine Beschränkung von Zulassungen unausweichlich ist, was die Frage aufwirft, nach welchen Gesichtspunkten das vergleichsweise rare Gut „Studienplätze“ vergeben werden soll. Die Verteilung rarer Studienplätze wird stark durch politische Interessen (Stichworte: Bildung als Ländersache, dennoch auch Einflussnahme des Bundes; Autonomie der Hochschulen) und das Grundgesetz mit dem Recht auf freie Berufs- und Ausbildungswahl und dessen Auslegung durch das Bundesverfassungsgericht beeinflusst. Dabei finden auch wissenschaftliche Erkenntnisse Berücksichtigung. Hier sind vor allem Studien zur Ungleichheit der Abiturnoten in den Bundesländern und Erkenntnisse zur prognostischen Validität, Objektivität und Fairness von Auswahlkriterien zu nennen. Das Hochschulrahmengesetz (HRG) schrieb bis November 2019 vor, nach welchen Kriterien die Hochschulen ihre Studierenden auswählen dürfen. Dieses Rahmengesetz wurde mit minimalen Variationen in die Landesgesetze übertragen, die für die Hochschulen des jeweiligen Bundeslandes verbindlich sind. Auswahlkriterien nach HRG, § 32 Ziffer 3 Absatz 3 (Stand: August 2019) waren:

- Grad der Qualifikation in der Hochschulzulassungsberechtigung (Abiturnote, ggf. nach Länderrecht eine Berufsausbildung): Durchschnitt oder gewichtete Einzelnoten
- Ergebnis in einem fachspezifischen Studierfähigkeitstest
- Art der Berufsausbildung oder Berufstätigkeit
- Ergebnis in einem Auswahlgespräch, „das Aufschluss über die Motivation der Bewerberin oder des Bewerbers und über die Identifikation mit dem gewählten Studium und dem angestrebten Beruf geben sowie zur Vermeidung von Fehlvorstellungen über die Anforderungen des Studiums dienen soll“
- Eine Kombination der oben genannten Kriterien

Im Achten Gesetz zur Änderung des Hochschulrahmengesetzes vom 15. November 2019 wurde festgesetzt: „§ 32 wird aufgehoben“. Anlass war ein Urteil des Bundesverfassungsgerichts vom 19. Dezember 2017, in dem festgestellt wurde, dass bundes- und landesgesetzliche Vorschriften über das Verfahren zur Vergabe von Studienplätzen an staatlichen Hochschulen zur Zulassung zum Studium der Humanmedizin teilweise mit dem Grundgesetz unvereinbar sind. Zu dem Urteil wurden u. a. folgende Leitsätze formuliert:

» Leitsätze zum Urteil des Ersten Senats vom 19. Dezember 2017 (Auszug; Bundesverfassungsgericht 2017)

1. Regeln für die Verteilung knapper Studienplätze haben sich grundsätzlich am Kriterium der Eignung zu orientieren.[...]
2. Verfassungswidrig sind die gesetzlichen Vorschriften zum Auswahlverfahren der Hochschulen insofern,

- als der Gesetzgeber den Hochschulen ein eigenes Kriterienerfindungsrecht überlässt,
- als die Standardisierung und Strukturierung hochschuleigener Eignungsprüfungen nicht sichergestellt ist [...],
- als für einen hinreichenden Teil der Studienplätze neben der Abiturdurchschnittsnote keine weiteren Auswahlkriterien mit erheblichem Gewicht Berücksichtigung finden...
- als im Auswahlverfahren der Hochschulen die Abiturnoten berücksichtigt werden können, ohne einen Ausgleichsmechanismus für deren nur eingeschränkte länderübergreifende Vergleichbarkeit vorzusehen,
- [...]

Praktisch bedeutet das Urteil, dass eine Auswahl nach der Abiturnote zwar zulässig ist – aber nur, wenn die regionale Ungleichheit kompensiert wird. Neben der Abiturnote müssen auch andere Auswahlkriterien berücksichtigt werden. Infrage kommen Auswahlgespräche und Studierfähigkeitstests durch die Hochschulen, die als weiterhin grundsätzlich zulässig erklärt wurden. In der Begründung zum Urteil des Bundesverfassungsgerichts lautet es daher:

» Auch für die hochschuleigenen Eignungsprüfungsverfahren gilt, dass die Hochschulzulassung gleichheitsgerecht nach je einheitlichen Maßgaben grundsätzlich ausschließlich anhand der Eignung der Bewerberinnen und Bewerber erfolgen darf. Dabei genügt es, wenn die Hochschulen selbst die Standardisierung und Strukturierung ihrer Tests oder Auswahlgespräche transparent vornehmen. (Bundesverfassungsgericht 2017)

Das Urteil ist von grundsätzlicher Bedeutung und hat daher auch Auswirkungen auf andere Studiengänge, in denen eine Zulassungsbeschränkung besteht.

Urteil des
Bundesverfassungsgerichts vom 19.
Dezember 2017

Studierfähigkeitstests

Studierfähigkeitstests dienen dazu, die Eignung für ein Studium generell („allgemeiner Studierfähigkeitstest“) oder für einen bestimmten Studiengang, eventuell sogar einen bestimmten Studiengang an einer bestimmten Hochschule („fachspezifischer Studierfähigkeitstest“) festzustellen.

In den USA wird die Zulassung zum College oft vom Ergebnis eines allgemeinen Studierfähigkeitstests abhängig gemacht. Zwei Tests werden dort häufig eingesetzt: der *Scholastic Aptitude Test*, später auch *Scholastic Assessment Test*, kurz: *SAT* genannt (► <https://www.collegeboard.org/>), und der *American College Test – ACT* (► <https://www.act.org/>).

In Deutschland finden nur fachspezifische Studierfähigkeitstests Anwendung. Diese können wie der *Test für Medizinische Studiengänge – TMS* (► <https://cip.dmed.uni-heidelberg.de/tms-info/tms-info/index.php?id=tms-infostartsseite>) bundesweit durchgeführt werden oder auch nur von einer Hochschule. In Österreich wurde mit dem *MedAT – Aufnahmeverfahren Medizin* ein eigenständiger fachspezifischer Studierfähigkeitstest etabliert (► <https://www.mediinstudieren.at/>). In der Schweiz erfüllt der *Eignungstest für das Medizin-Studium – EMS* (► <https://www.doktortest.net/ems/aufnahmetest-medizinstudium-schweiz/>), der dem deutschen TMS sehr ähnlich ist, die gleiche Funktion. Während die oben genannten Test der Auswahl von Studierenden dienen, können Studierfähigkeitstests auch zu Beratungszwecken eingesetzt werden (► Abschn. 7.1.3.2).

Studierfähigkeitstests

Wie gut Studierfähigkeitstests, die Abiturnote und das Interview („Auswahlgespräche“) – alle zuvor als grundsätzlich zulässige Auswahlmethoden beschrieben – geeignet sind, den Studienerfolg vorherzusagen, wurde vielfach

Die Studiendauer wird hauptsächlich durch psychosoziale Faktoren vorhergesagt

Schulnoten und Studierfähigkeitstest sind gute Prädiktoren für den Studienerfolg

Fachspezifische Studierfähigkeitstests für Medizin und andere Fächer

untersucht. Die Ergebnisse sind in mehreren Metaanalysen zusammenfassend analysiert worden. Zu Beginn ist jedoch zu klären, was vorhergesagt werden soll: Studienleistung (Noten), Studiendauer, Studienabbruch, Studienzufriedenheit oder gar Berufserfolg. Berufserfolg ist, mit Ausnahme weniger Studienfächer (z. B. Lehramt), kein geeignetes Kriterium, weil sich nach dem Studium sehr verschiedene Beschäftigungsmöglichkeiten ergeben, die jeweils mit speziellen Anforderungen verbunden sind.

Robbins et al. (2004) haben die Prädiktoren von Studiennoten und Studiendauer analysiert und festgestellt, dass jeweils andere Prädiktoren bedeutsam sind: Studienleistungen (in Form von Noten) werden gut durch Studierfähigkeitstests und Schulnoten vorhergesagt, die Studiendauer dagegen eher durch psychosoziale Faktoren wie die Zufriedenheit mit der Institution und akademische Fertigkeiten (Kommunikationsfähigkeit, Zeitmanagement etc.). In einer deutschen Studie von Janke und Dickhäuser (2018) war es dementsprechend nicht möglich, die Studiendauer mit der Abitur- oder der Mathematiknote vorherzusagen ($r=.03$ bzw. $.11$, nicht signifikant).

Am gründlichsten wurde die Vorhersage der Studienleistungen erforscht. Die Ergebnisse sind in □ Tab. 7.2 zusammenfassend dargestellt.

Der □ Tab. 7.2 lässt sich entnehmen, dass Schulnoten und Studierfähigkeitstests die besten Prädiktoren des *Studienerfolgs* sind. Anzumerken ist, dass allgemeine Studierfähigkeitstests in Deutschland nicht zulässig sind. Durch eine fachspezifische Gewichtung einzelner Teile könnte ein allgemeiner Studierfähigkeitstest jedoch in mehrere fachspezifische umgewandelt werden. Fachspezifische Schulnoten (z. B. Mathematiknote beim Mathematikstudium) fallen deutlich gegenüber der Gesamtnote ab. Eine Ausnahme bildet die Mathematiknote bei der Vorhersage der Bachelornote im Fach Psychologie. Allerdings handelt es sich hierbei um eine Einzelstudie. Interviews sagen Studienleistungen kaum vorher.

In Deutschland war für das Fach Medizin ein *fachspezifischer Studierfähigkeits-test* (TMS) entwickelt worden. Allerdings wurde beschlossen, ihn

□ Tab. 7.2 Ergebnisse von Metaanalysen und einer Einzelstudie zur Vorhersage von Studienerfolg

Prädiktor	ρ	Anmerkungen (Quelle)
Schulnoten (Gesamtwert)	.52 ^{a, b, c} .41 ^b .41 ^{a, b} .41	Europäischer Raum (Trapmann et al. 2007) Überwiegend USA (Robbins et al. 2004) Überwiegend USA (Richardson et al. 2012) Deutschland; Bachelornote Psychologie (Janke und Dickhäuser 2018)
Studienfachbezogene Schulnoten	.36 ^a .63	Europäischer Raum (Trapmann et al. 2007) Deutschland; Mathematiknote – Bachelornote Psychologie (Janke und Dickhäuser 2018)
Allgemeine Studierfähigkeitstests	.37 ^b .33 ^{a, b} .40 ^{a, b}	Überwiegend USA (Robbins et al. 2004) USA: SAT (Richardson et al. 2012) USA: ACT (Richardson et al. 2012)
Fachspezifische Studierfähigkeits-tests	.48 ^{a, b, c}	Deutschsprachiger Raum (Hell et al. 2007a)
Strukturiertes Interview	.21 ^{b, c}	Weltweit (Hell et al. 2007b)
Unstrukturiertes Interview	.11 ^{b, c}	Weltweit (Hell et al. 2007b)

ρ = korrigierte Korrelation; korrigiert für:

^a Reliabilität des Prädiktors

^b Reliabilität des Kriteriums

^c Varianzeinschränkung

Bei Janke und Dickhäuser (2018) keine Korrektur vorgenommen

1997 letztmalig einzusetzen, weil durch die Beschränkung der Niederlassungsfreiheit für Ärzte damals weniger Studienbewerber/-innen in das Fach drängten, was den Aufwand nicht mehr gerechtfertigt hat. Inzwischen kommt ein Nachfolgetest an zahlreichen Universitäten in Deutschland und in der Schweiz (EMS) wieder zum Einsatz. Publizierte Validitätsbefunde sind in der Metaanalyse von Hell et al. (2007a) eingeslossen (► Tab. 7.2). Für den Zugang zum Studienfach Psychologie ist ab dem Wintersemester 2020/2021 in Baden-Württemberg ein fachspezifischer Studieneignungstest im Einsatz, den Studierende freiwillig absolvieren und somit ihre Chancen auf einen Studienplatz an teilnehmenden Universitäten erhöhen können. Zudem ist die Teilnahme an einem Online-Self-Assessment (► Abschn. 7.1.3.2 verpflichtend (Deutsche Gesellschaft für Psychologie 2020).

7.1.3.2 Online-Self-Assessments

Eine Alternative oder auch Ergänzung zu Studierfähigkeitstests sind sog. „Online-Self-Assessments“, die im deutschen Sprachraum enorm an Bedeutung gewonnen haben. Dafür gibt es einen einfachen Grund: Die Studienlandschaft ist unüberschaubar geworden. Der offizielle Studienführer für Deutschland „studienwahl.de“ (► <https://studienwahl.de/>) verzeichnet 18.660 Bachelor- und Masterstudiengänge an insgesamt 441 Hochschulen (Stand: Juni 2020). Nicht nur die schiere Zahl der Angebote macht die Suche nach einem passenden Studiengang schwer. Die Suche wird zusätzlich dadurch erschwert, dass ein Name kein Garant für den gesuchten Inhalt ist. Unter dem gleichen Namen, beispielsweise „Medienwissenschaft“, können sich unterschiedliche Studienangebote verbergen. Und ein Studium, das sich mit Medienwissenschaft beschäftigt, kann auch irgendwo unter einem anderen Namen geführt werden. Und selbst wenn jemand einen vermeintlich passenden Studiengang an einem bestimmten Ort gefunden hat, bleibt immer noch die Frage offen: „Wie gut passt dieser Studiengang zu mir?“

Studienlandschaft unüberschaubar

Definition

Online-Self-Assessments (kurz: OSA) sind diagnostische Verfahren (Assessments), die dazu dienen, einen passenden Studiengang zu finden. Sie werden von Studieninteressierten selbst („Self“) zu Hause durchgeführt, und zwar ausschließlich über das Internet („Online“). Sie unterscheiden sich von Auswahlverfahren der Hochschulen in mehrfacher Hinsicht: Die Teilnahme ist in der Regel freiwillig (einzelne Hochschulen verlangen jedoch vor der Bewerbung oder vor Aufnahme des Studiums den Nachweis, dass man an einem fachspezifischem Online-Self-Assessment der Hochschule teilgenommen hat), und die Hochschule erfährt das Ergebnis nicht. Ein Online-Self-Assessment kann auch Module enthalten, die der Information über den Studiengang dienen. Was mit einem Online-Self-Assessment gemessen wird, unterliegt nicht den Begrenzungen, die in Deutschland für Auswahlverfahren gelten (► Abschn. 7.1.3.1).

Breites Spektrum an Online-Self-Assessments

Die klassischen Beratungsangebote wie Studienberatung an den Hochschulen oder Internetauftritte von Hochschulen reichen nicht mehr aus. Seit über die ersten Online-Self-Assessments in Deutschland informiert wurde (Zimmerhofer et al. 2006), ist das Angebot stark gewachsen. Es wurde ein Online-Self-Assessment-Portal eingerichtet (► <https://www.osa-portal.de/>), das bei der Suche nach einem passenden Online-Self-Assessments für ein bestimmtes Studienfach hilft. Im August 2019 waren dort weit über 700 Angebote verzeichnet. Die Online-Self-Assessments unterscheiden sich darin, ob sie der Navigation zu dem passenden Studiengang dienen oder die Passung

zu einem bereits ins Auge gefassten Studiengang an einer bestimmten Hochschule prüfen. Ein zweites wichtiges Unterscheidungsmerkmal ist, ob sie rein diagnostisch arbeiten oder Informationen für einen Studiengang vermitteln. Beide Funktionen sind in einem Online-Self-Assessment oft kombiniert. Tatsächlich sind Online-Self-Assessments noch vielfältiger (s. Guttschick et al. 2019).

Selbstselektion anregen

7

Steuerungsfunktion für die Hochschulen

Ablauf eines Online-Self-Assessments

Anforderungsanalyse durch Prüfung der Validität ergänzen

Welchen Nutzen versprechen sich die Hochschulen als Anbieter von einem Online-Self-Assessment? Weil die Hochschule das Ergebnis nicht erfährt, kann sie nur darauf vertrauen, dass die Person, die sich selbst getestet hat, aus der Rückmeldung den richtigen Schluss zieht. Wird ihr eine gute Passung zum Studiengang rückgemeldet, sollte sie sich für den Studiengang entscheiden und bei einer schlechten Passung dagegen. So einfach funktioniert das natürlich nicht. Das Online-Self-Assessment-Ergebnis ist oft ein Element von mehreren, die zur Entscheidung für oder gegen einen Studiengang beitragen. Eltern und Freunde geben Ratschläge, und die Einschätzung der eigenen Fähigkeiten und Interessen spielt ebenfalls eine Rolle. Wie attraktiv der Studiengang und der Studienort erscheinen, ist auch relevant. Von einem sehr attraktiven Angebot wird man sich nicht so leicht durch eine negative Online-Self-Assessment-Rückmeldung abbringen lassen wie von einem wenig attraktiven Angebot. Daher können die Hochschulen nur darauf hoffen, dass die Selbstselektion durch das Online-Self-Assessment angeregt wird. Belastbare Ergebnisse, wie stark dieser Effekt ist, fehlen noch.

Aus Sicht der Hochschulen kommt noch eine andere Überlegung ins Spiel. Gibt es sehr viele Studieninteressierte – oder konkret sehr viele Bewerbungen auf einen Studienplatz –, kann es sich die Hochschule erlauben, in ihrem Online-Self-Assessment häufiger negative Rückmeldungen zu geben. Die Schwelle für eine positive Rückmeldung wird entsprechend hochgesetzt. Dadurch soll erreicht werden, dass möglichst viele „gute“ Bewerbungen eingehen. Umgekehrt liegt es nahe, bei Studiengängen mit schwacher Nachfrage die Schwelle für eine positive Rückmeldung abzusenken (natürlich nicht unter ein vertretbares Niveau – nicht vertretbare Empfehlungen sollten nicht ausgesprochen werden, auch wenn es aus Gründen des Bewerbendenmanagements sinnvoll erscheint). So kann ein Online-Self-Assessment auch als Rekrutierungsinstrument genutzt werden (Reiß 2019).

Wie läuft ein Online-Self-Assessment ab? Angesichts der großen Vielfalt an Online-Self-Assessments (s. o.) versuchen wir, ein typisches Online-Self-Assessment zu beschreiben. Dies könnte folgendermaßen aussehen: Wenn das passende Online-Self-Assessment gefunden wurde, muss man sich zumeist anmelden. Damit wird sichergestellt, dass man die Bearbeitung ohne Verlust der bereits gewonnenen Ergebnisse unterbrechen kann. In der Regel erhält man zunächst Informationen über den Ablauf, die Inhalte, den Zeitaufwand und über die Rückmeldung. Das Online-Self-Assessment soll ohne fremde Hilfe bearbeitet werden, weil es ausschließlich der Beratung dient. Die einzelnen Module können aus Leistungstests, Vorwissenstests zum Studiengang, Fragebögen zu Persönlichkeitsmerkmalen oder Interessen, die für den Studiengang relevant sind, sowie Informationen über den Studiengang bestehen. Der Zeitaufwand wird bei etwa 1 h liegen; es gibt aber auch deutlich kürzere Online-Self-Assessments. In der Regel erfolgt eine Rückmeldung erst am Ende. Es ist aber auch möglich, dass bei Fragen zur Informiertheit über den Studiengang sofort auf eine vom Nutzenden gewählte Antwort ein Feedback erfolgt. Die Rückmeldung wird Aussagen enthalten, wie gut man zu dem Studiengang passt. Sehr präzise Aussagen werden dabei vermieden, was u. a. der begrenzten Reliabilität der Skalen zuzuschreiben ist.

Bei den Unterschieden zwischen Auswahlverfahren der Hochschulen in Deutschland und Online-Self-Assessments (► Abschn. 7.1.3.1) wurde erwähnt, dass bei Online-Self-Assessments keine Beschränkung bei den Methoden und

dem Messgegenstand besteht. Dies ist ein unschätzbarer Vorteil. Ein Online-Self-Assessment basiert auf Anforderungsanalysen des Studiengangs, für das es entwickelt wurde – oder sollte es zumindest. Beispielsweise ergab eine Anforderungsanalyse bei 4 Studiengängen an der Universität Marburg, dass die Leistungsmotivation sehr bedeutsam ist. Deshalb wurde ein Fragebogenmodul „Leistungsmotivation“ entwickelt und evaluiert (Guttschick 2015). Der Gesamtwert des Fragebogens korrelierte sowohl mit der Studienleistung in Noten als auch mit der Studienzufriedenheit ($r = -.31$ bzw. $.25$). Je wichtiger die Leistungsmotivation der Anforderungsanalyse zufolge für einen Studiengang war, desto höher fielen diese Korrelationen aus.

Während Auswahlverfahren üblicherweise auf die Vorhersage der Studienleistungen und eventuell der Studiendauer abzielen, können Online-Self-Assessments auch andere Ziele verfolgen. Viele Studiengänge leiden unter einer hohen Abbruchquote, der oft schlechte Studienleistungen und Unzufriedenheit mit dem Studium vorausgehen. Mithilfe eines Online-Self-Assessments kann man versuchen, die Studienzufriedenheit und, als fernes Ziel, den Studienabbruch vorherzusagen (ein Studienabbruch kündigt sich oft durch eine niedrige Studienzufriedenheit an; Heublein et al. 2009). Im bereits erwähnten Marburger OSA-Projekt konnte beispielsweise aufgezeigt werden, dass für ein Studium der Wirtschaftswissenschaften ein zu geringes Vorwissen in Mathematik und eine zu geringe (studentische) Organisationsfähigkeit problematisch sind. Diese Merkmale wurden daher in einem Online-Self-Assessment durch einen Leistungstest bzw. durch einen Fragebogen erfasst. Je niedriger die beiden Merkmale ausgeprägt waren, desto geringer war später die Zufriedenheit mit den Studieninhalten und umso belastender wurde das Studium erlebt (Vorwissen Mathematik: $r = .33$ bzw. $.38$; studentische Organisationsfähigkeit: $r = .33$ bzw. $.37$). Die Abiturnote und auch die Intelligenz hatten bezüglich dieser Kriterien keine bedeutsame Vorhersagekraft (Hasenberg und Schmidt-Atzert 2014).

Nicht nur Anforderungsanalysen, sondern empirische Erkenntnisse über den Zusammenhang zwischen Merkmalen der Studierenden und Studienerfolg, Studienzufriedenheit und Studienabbruch können Hinweise auf relevante Variablen liefern. Beispielsweise haben Richardson et al. (2012) gezeigt, dass Prokrastination, d. h. die Neigung, Dinge aufzuschieben, negativ mit der Studienleistung korreliert ($r_{korr} = -.21$). Ähnlich hoch ($r_{korr} = .23$) ist der Zusammenhang zwischen dem Persönlichkeitsmerkmal Gewissenhaftigkeit und Studienerfolg (bei Trapmann et al. 2007: $r = .27$). Wenn nun auch eine Anforderungsanalyse für einen bestimmten Studiengang zeigt, dass ein solches Merkmal relevant ist, bestehen gute Chancen, damit einen nützlichen Baustein für ein Online-Self-Assessment gefunden zu haben.

An dieser Stelle ist leider noch eine Warnung nötig. Bietet eine Hochschule ein Online-Self-Assessment und damit eine Aussage über die Eignung für einen bestimmten Studiengang an, hat sie auch die Verantwortung dafür, dass dieses Instrument dem Anspruch gerecht wird. Man sieht es einem Online-Self-Assessment nicht an, ob es das misst, was es zu messen vorgibt, oder ob überhaupt die wirklich relevanten Anforderungsmerkmale ausgewählt worden sind. Die Zahl der Online-Self-Assessments, deren Validität überhaupt nachvollziehbar geprüft worden ist, dürfte für die über 700 im OSA-Portal aufgeführten Online-Self-Assessments bei etwa 3 % liegen (vgl. Guttschick et al. 2019). Die Gütekriterien für diagnostische Verfahren (► Kap. 2) können und sollten auf alle Bestandteile eines Online-Self-Assessments, von der Anforderungsanalyse bis zu den einzelnen Modulen (auch den Informationsmodulen) angewandt werden. Darüber hinaus gibt es einige Kriterien, die für Online-Self-Assessments spezifisch sind wie Nutzerfreundlichkeit, Qualität der Rückmeldung und Optimierung der Studienwahl (s. Schmitt und Schmidt-Atzert 2019).

Studienzufriedenheit kann durch Online-Self-Assessments vorhergesagt werden

Einschlägige empirische Befunde nutzen

Psychometrische Qualität der Online-Self-Assessments oft fraglich

Weiterführende Literatur

Zum Thema Schuleingangsdagnostik finden sich in dem von Schneider und Hasselhorn (2018) herausgegebenen Band zahlreiche Informationen.

Über Online-Self-Assessments an Hochschulen informieren das von Schmidt-Atzert et al. (2019) herausgegebene Buch sowie in kurzer Form ein Beitrag von Hasenberg et al. (2014).

7.2 Diagnostik bei Schulschwierigkeiten

7.2.1 Diagnostik bei Lernschwierigkeiten

Lernschwierigkeiten multifaktoriell bedingt

7

Persönlichkeitsmerkmale und biologische Faktoren

Schulisches und häusliches Umfeld betrachten

Die häufigsten Anlässe für eine diagnostische Untersuchung im Aufgabenbereich der Pädagogischen Psychologie sind *individuelle Lernschwierigkeiten*. Wir sprechen hier von schulischen Leistungsproblemen, die nicht auf umschriebene Defizite bei den Fertigkeiten Lesen, Schreiben oder Rechnen zurückzuführen sind (► Abschn. 7.2.2) oder die nicht so gravierend sind, dass sonderpädagogischer Förderbedarf im Bereich Lernen (früher „Lernbehinderung“) vorliegt (► Abschn. 7.1.2). „Lernschwierigkeiten“ sind kein fest definierter Begriff. Von Lernschwierigkeiten wird gesprochen, wenn die Zielerreichung beim Lernen erschwert oder verhindert ist (Heinecke-Müller 2019). Sie äußern sich in negativen Abweichungen der schulischen Leistung einzelner Schülerinnen und Schüler von klassenbezogenen Normen oder individuellen Erwartungen. Für die Beurteilung einer individuellen Leistung bildet in der Regel die Durchschnittsleistung der Klasse den Bezugsrahmen. Wird sie deutlich und nicht nur vorübergehend unterschritten, ist Anlass für diagnostische Maßnahmen gegeben, ohne die keine zielgerichteten Interventionen ergriffen werden können. Aber auch ein Nachlassen der Leistungen von bislang guten Schülerinnen oder Schülern kann diagnostische Maßnahmen initiieren, wenn die Leistungen längere Zeit dauerhaft hinter den individuellen Erwartungen zurückzubleiben drohen. Da Lernschwierigkeiten nach übereinstimmender Auffassung als multifaktoriell bedingt angesehen werden, stellt sich die Frage, welche Hypothesen durch diagnostische Maßnahmen sinnvollerweise geprüft werden sollen.

Orthmann Bless (2010) zufolge können die Ursachen für bestehende Lernschwierigkeiten bei der Person der Schülerin oder des Schülers und/oder der Lernsituation sowie in der Interaktion von Person und Situation gesucht werden. *Personenmerkmale* werden grob unterteilt in kognitive Persönlichkeitsmerkmale (z. B. Intelligenz, Lernstrategien), nichtkognitive Persönlichkeitsmerkmale (z. B. Motivation, Leistungsängstlichkeit) und organisch-biologische Voraussetzungen (z. B. Erkrankungen und Behinderungen). Innerhalb dieser Merkmalsgruppen kann noch einmal danach unterschieden werden, ob ein Merkmal übergreifend ist (z. B. allgemeine Schulunlust) oder fach- bzw. bereichsspezifisch (z. B. bereichsspezifisches Wissen). Auch die Dauer ist von Bedeutung. Eine Erkrankung kann chronisch sein oder vielleicht auch nur wenige Monate dauern, was bei einer Depression der Fall sein kann.

Situative oder Umweltfaktoren, die eventuell als Ursache für Lernschwierigkeiten infrage kommen, können im schulischen Umfeld, in der Familie und auch im Freundeskreis liegen. Schulische Faktoren sind etwa das Schul- oder Klassenklima, die Unterrichtsqualität, die Art der Leistungsbeurteilung sowie Mobbing. Im familiären Umkreis kommen fehlende Unterstützung beim Lernen, Konflikte zu Hause oder Scheidung der Eltern als ungünstige Faktoren infrage. Aus dem Freundeskreis kann eine negative Einstellung zur Schule oder die Ablehnung von schulischem Erfolg („Streber“) zu Lernproblemen beitragen.

Personen- und Situationsmerkmale können auch interagieren; eine Leistungsbeurteilung anhand vieler benoteter Prüfungen wird besonders bei leistungsschwachen Schülerinnen und Schülern mit hohem Anspruchsniveau oder mit ausgeprägter Prüfungsangst zu Problemen führen. Für eine interventionsvorbereitende Diagnostik ist neben der Identifikation von Defiziten oder generell ungünstigen Einflussfaktoren auch die Suche nach individuellen Stärken wie etwa eine hohe Belastbarkeit oder Gewissenhaftigkeit sowie nach fördernden Umweltfaktoren wie Hilfsangebote älterer Geschwister bei der Kontrolle der Hausaufgaben hilfreich.

Lernschwierigkeiten werden teilweise von Ärztinnen und Ärzten sowie Psychologinnen und Psychologen pathologisiert, also als psychische Störung betrachtet. Nach Internationaler statistischer Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, German Modification (ICD-10-GM) kommen dafür die Diagnosen F81.3 „Kombinierte Störungen schulischer Fertigkeiten“ (s. Lauth 2004) und F81.9 „Entwicklungsstörung schulischer Fertigkeiten, nicht näher bezeichnet“ (s. Castello et al. 2004) in Frage.

Auch Stärken und fördernde Faktoren beachten

Lernschwierigkeiten als psychische Störung

7.2.2 Diagnostik von Teilleistungsstörungen

Unter Teilleistungsstörungen versteht man Leistungsdefizite, die anders als eine Lernstörung oder eine „Lernbehinderung“ auf einen bestimmten Bereich schulischer Fertigkeiten beschränkt sind. Sie werden nach dem Diagnostic and Statistical Manual of Mental Disorders, 5. Auflage (DSM-5) als „spezifische Lernstörungen“ bezeichnet (Lindberg et al. 2018). Es können die Fertigkeiten zu Lesen, Schreiben oder Rechnen betroffen sein. Nach ICD-10-GM werden Teilleistungsstörungen als „umschriebene Entwicklungsstörungen schulischer Fertigkeiten“ eingeordnet, wobei zwischen folgenden Störungen differenziert wird (DIMDI 2019):

Teilleistungsstörungen sind auf einen Bereich schulischer Fertigkeiten beschränkt

F81 Umschriebene Entwicklungsstörungen schulischer Fertigkeiten

- F81.0 Lese- und Rechtschreibstörung
- F81.1 Isolierte Rechtschreibstörung
- F81.2 Rechenstörung
- F81.3 Kombinierte Störungen schulischer Fertigkeiten (als Restkategorie)
- F81.8 Sonstige Entwicklungsstörungen schulischer Fertigkeiten (inkl.: Entwicklungsbedingte expressive Schreibstörung)
- F81.9 Entwicklungsstörung schulischer Fertigkeiten, nicht näher bezeichnet (inkl.: Lernbehinderung, Lernstörung, Störung des Wissenserwerbs – jeweils ohne nähere Angaben)

Mit diesen Störungen befassen sich nicht nur Schulpsychologinnen und Schulpsychologen, sondern auch Psychologinnen und Psychologen, die in der Klinischen oder der Neuropsychologie tätig sind. Kinder mit entsprechenden Problemen fallen in der Regel meist erst in der Schule auf. Entwicklungsstörungen zeichnen sich nach ICD-10 durch den Beginn im Kleinkindalter oder in der Kindheit aus. Die Entwicklungseinschränkung oder -verzögerung betrifft Funktionen, die eng mit der biologischen Reifung des Zentralnervensystems verknüpft sind. Sie verlaufen stetig ohne Remissionen und Rezidive (DIMDI 2019). Nach ICD-10 liegt eine entsprechende Störung vor, „wenn die gezeigte Leistung im jeweiligen Teilbereich (Lesen, Schreiben, Rechnen) deutlich unter dem aufgrund der Intelligenz, der Klassenstufe oder des Alters zu erwartenden Niveau liegt. Probleme in der Lernentwicklung, die durch

körperliche Ursachen (z. B. Hör- oder Sehstörungen), äußerliche Faktoren (z. B. fehlende Schulung) oder allgemeine kognitive Beeinträchtigungen (z. B. Intelligenzminderung) begründet sind, werden demnach nicht als Lernstörung klassifiziert“ (Lindberg et al. 2018, S. 197).

Die Lese- und Rechtschreibstörung wird oft auch „Legasthenie“ genannt, und für die Rechenstörung hat sich auch der Begriff „Dyskalkulie“ etabliert. In der Fachliteratur finden sich außerhalb der ICD-Diagnostik verschiedene Definitionen für jede dieser Störungen, und es werden dementsprechend auch unterschiedliche diagnostische Kriterien genannt. Als Folge davon schwanken die Angaben zur Auftretenshäufigkeit erheblich.

Für die Diagnostik von spezifischen Lernstörungen ist es unerlässlich, mit gut normierten und am schulischen Kerncurriculum orientierten Testverfahren (Schulleistungstests) den altersgemäßen Lernstand, z. B. beim Rechnen, festzustellen. Überprüft wird, ob die Leistungen im Lesen, Schreiben bzw. Rechnen weit unterhalb der Alters- bzw. Klassennorm liegen. Da eine relativ niedrige Intelligenz als Ursache auszuschließen ist, muss die Intelligenz diagnostiziert werden. Dazu ist ein gut normierter Intelligenztest zu wählen, der die Intelligenz möglichst unabhängig von der vermutlich beeinträchtigten Fertigkeit (Lesen, Schreiben bzw. Rechnen) erfasst. Der Normwert der spezifische Fertigkeit (z. B. Lesen) muss weit unter dem Normwert der Intelligenz liegen (Lindberg et al. 2018). Ferner sind körperliche Ursachen wie Hör- oder Sehstörungen und Umweltfaktoren wie fehlende oder unregelmäßige Schulung (s. o.) durch ein diagnostisches Interview, einen Anamnesebogen oder eine Aktenanalyse auszuschließen.

Allerdings fehlt eine verbindliche Konvention, wie niedrig das Ergebnis im Schulleistungstest ausfallen muss und wie groß die Diskrepanz zwischen Schulleistung und Intelligenz sein muss (Fischbach et al. 2013). In einer deutschen Studie zur Prävalenz von spezifischen Lernstörungen (Fischbach et al. 2013) wurde festgelegt, dass das Ergebnis im Schulleistungstest über 1 Standardabweichung (ein verbreitetes Kriterium für „niedrig“) unter dem Altersdurchschnitt liegt ($T < 40$) und der Intelligenzquotient (IQ) mindestens 1,2 Standardabweichungen höher ist als dieser Wert (in der Forschung variiert dieser Diskrepanzwert zwischen 1,0 und 1,5 Standardabweichungen). Ein Kind mit einem Rechentestergebnis von $T = 39$ erfüllt das 1. Kriterium gerade. Im Intelligenztest war das 2. Kriterium dann gerade mit einem IQ von 102 erfüllt ($T = 39 + 12 = 51$). Kinder mit einem $IQ < 85$ wurden ganz ausgeschlossen.

Fischbach et al. (2013) untersuchten 2195 Kindern im Alter von durchschnittlich 8;8 Jahren in verschiedenen Grundschulen in Deutschland. Die ermittelten Prävalenzraten sind in □ Tab. 7.3 aufgeführt.

Demnach wiesen 24,6 % der untersuchten Grundschulkinder eine Lernstörung auf! Am häufigsten wurde mit 8,2 % eine isolierte Rechtsreibstörung diagnostiziert. Bemerkenswert sind die Geschlechterverhältnisse. Bei Lese- und/

□ Tab. 7.3 Prävalenz von Lernstörungen in der Grundschule

Lernstörung	Prävalenz (%)	Verhältnis Mädchen zu Jungen
Lese-Rechtschreibstörung	2,8	1 : 2,4
Isolierte Lesestörung	6,6	1 : 1,5
Isolierte Rechtschreibstörung	8,2	1 : 1,4
Isolierte Rechenstörung	5,0	3,3 : 1
Kombinierte Lernstörung	2,0	1,2 : 1

Quelle: Auszug aus Tab. 3 von Fischbach et al. (2013, © Hogrefe)

Legasthenie, Dyskalkulie

Niedrige Werte in Schulleistungstest und vergleichsweise hohe Intelligenz

7

Unklarheit bezüglich der Mindestwerte und der Diskrepanz

oder Rechtschreibstörung sind Jungen überrepräsentiert; eine Rechenstörung kommt bei Mädchen über $3 \times$ häufiger vor als bei Jungen.

Zur Epidemiologie der spezifischen Lernstörungen liegen weitere Studien vor (s. Moll et al. 2014, Tab. 1). Die Prävalenzraten variieren erheblich, was angesichts der unterschiedlichen Cut-off-Werte und zum Teil kleinen Stichproben nicht verwunderlich ist. Die Ergebnisse der Studie von Moll et al. (2014) berichten wir nicht, weil sie den Empfehlungen in DSM-5 folgend keine Diskrepanz zum IQ berücksichtigten und nur auf eine Standardabweichung (SD) von -1, -1,25 und -1,5 fokussierten. Bei Verwendung eines gut normierten Schulleistungstests ist zu erwarten, dass Werte von ≤ -1 SD bei 15,9 % der Kinder auftreten. Per Definition liegt die Prävalenzrate also bei 15,9 %. Abweichungen davon können auf eine „schlechte“ Normierung des Schulleistungstests oder eine „schlechte“ Stichprobenziehung in einer Prävalenzstudie zurückzuführen sein.

Für eine der Förderung dienende Diagnostik ist es naheliegend, die der problematischen Fertigkeit (z. B. Lesen) zugrunde liegende Basiskompetenzen zu untersuchen. Dann kann eine direkte Behandlung der identifizierten Defizite erfolgen, was sich in vielen Studien als eine wirkungsvolle Strategie erwiesen hat (Lindberg et al. 2018). Die spezifischen Lernstörungen gehen mit einer erhöhten Rate an anderen Störungen einher, wobei die Komorbidität mit ADHS am stärksten ausgeprägt ist. Visser et al. (2019) berichten in ihrer Literaturauswertung über eine Studie mit einer repräsentativen deutschen Stichprobe von Kindern mit Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS). Die Prävalenz für eine Lesestörung betrug 17,2 %, für eine Rechtschreibstörung 20,3 % und für eine kombinierte Lese- und Rechtschreibstörung (LRS) 22,2 %. Die Prävalenz für Rechenstörungen war bei ADHS nicht erhöht. Für eine Förderung kann auch das Vorliegen weiterer Störungen relevant sein. Die Forschung hat gezeigt, dass bei einer Lese- und/oder Rechtschreibstörung auch die Prävalenzen von Angststörungen und Störungen des Sozialverhaltens erhöht sind. Für die diesbezüglich weniger gut untersuchte Rechenstörung ist die Prävalenz von „Verhaltensproblemen“ offenbar erhöht (Visser et al. 2019). ADHS und Angststörungen können eventuell eine vorliegende Lernstörung verstärken und eine Intervention behindern.

Epidemiologie

Der Störung zugrunde liegende und komorbide Defizite erkennen

■ Rechenstörung (Dyskalkulie)

Exemplarisch wird nun die Diagnostik einer Rechenstörung beschrieben. Da Rechnen erst in der Schule verlangt wird, werden Rechenstörungen überwiegend in der Grundschule, manchmal auch erst später in der weiterführenden Schule erkannt. Für betroffene Kinder stellt die Erweiterung des Zahlenraums über 100 sowie der Verzicht auf Hilfsstrategien wie Abzählen mit den Fingern eine große Herausforderung dar. Einige Kinder können ihr Defizit anfangs durch Auswendiglernen, ein gutes Arbeitsgedächtnis etc. kompensieren, aber ihre Mathematikleistungen fallen dann irgendwann doch gegenüber dem Klassendurchschnitt ab.

Wird zumeist in der Grundschule erkannt

Aussagen zum diagnostischen Vorgehen macht eine am 25. Februar 2018 bei der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) hinterlegte evidenz- und konsensbasierte Leitlinie (AWMF 2018). Sie wurde federführend von der Deutschen Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e. V. unter der Mitwirkung von über 20 Berufsverbänden aus verschiedenen Bereichen sowie 2 Experten erarbeitet. Beteiligt waren u. a. der Berufsverband der Kinder- und Jugendärzte, der Bundesverband Legasthenie & Dyskalkulie e. V., die Deutsche Gesellschaft für Psychologie und der Deutsche Lehrerverband. Neben der eigentlichen Leitlinie liegen auch ein Leitlinienreport sowie weitere Materialien vor (s. ► <https://www.awmf.org/leitlinien/detail/ll/028-046>.

Evidenz- und konsensbasierte Leitlinie von 2018

[html](#)). „Evidenzbasiert“ bedeutet, dass die gesamte einschlägige Forschungsliteratur aufgearbeitet wurde, „konsensbasiert“ ist die Leitlinie, weil ein repräsentatives Gremium von Verbänden daran mitgearbeitet hat und bei einzelnen Statements auch der Grad der Übereinstimmung angegeben wird (z. B. „Empfehlungsgrad A, starke Empfehlung, starker Konsens: 100 % Zustimmung“). Wir beziehen uns im Folgenden auf die Langfassung der „S. 3-Leitlinie: Diagnostik und Behandlung der Rechenstörung“ (AWMF 2018) und zitieren sie als „Leitlinie“.

Auch eine Leitlinie zu Lese- und/oder Rechtschreibstörung

Bei der AWMF wurde 2015 übrigens auch eine ebenfalls evidenz- und konsensbasierte Leitlinie „Lese- und/oder Rechtschreibstörung bei Kindern und Jugendlichen, Diagnostik und Behandlung“ veröffentlicht (AWMF 2015).

Definition

„Die **Rechenstörung** ist [...] eine umschriebene Entwicklungsstörung schulischer Fertigkeiten. Sie ist in den einschlägigen internationalen Klassifikationssystemen (ICD, DSM) definiert. Wie bei den anderen umschriebenen Entwicklungsstörungen (Motorik, Sprache) handelt es sich bei den umschriebenen Entwicklungsstörungen schulischer Fertigkeiten um persistierende Störungen mit Krankheitswert“ (aus der Präambel der Leitlinie; AWMF 2018).

In der deutschen Ausgabe der ICD-10-GM (DIMDI 2019) wird die Diagnose F81.2 erläutert: „Diese Störung besteht in einer umschriebenen Beeinträchtigung von Rechenfertigkeiten, die nicht allein durch eine allgemeine Intelligenzminderung oder eine unangemessene Schulung erkläbar ist. Das Defizit betrifft vor allem die Beherrschung grundlegender Rechenfertigkeiten, wie Addition, Subtraktion, Multiplikation und Division, weniger die höheren mathematischen Fertigkeiten, die für Algebra, Trigonometrie, Geometrie oder Differential- und Integralrechnung benötigt werden“ (AWMF 2018, S. 5).

Begründung und Erläuterung der Kriterien

Die „Leitlinie“ verlangt bei der Diagnostik die Prüfung von 3 Arten von Kriterien (Abb. 7.2):

- Zur Prüfung der *psychometrischen Kriterien* kommen ausschließlich Tests in Frage. Für die „Leitlinien“ wurden die verfügbaren Tests evaluiert und in eine Rangreihe gebracht. Die besten 50 % werden empfohlen.
- Die *qualitativen Kriterien* sind für die Differentialdiagnostik wichtig. So sollen andere Ursachen für die Rechenprobleme (z. B. eine Hirnschädigung oder eine unzureichende Schulung) ausgeschlossen und ggf. Hinweise auf komorbide Störungen entdeckt werden.
- Die *klinischen Kriterien* sind ebenfalls für die Differentialdiagnostik wichtig. So sollen Hirnschädigungen und neurogenetische Störungen (z. B. Turner-Syndrom), bisher unentdeckte Seh- oder Hörstörungen sowie eine Intelligenzminderung ausgeschlossen werden. Nach ICD-10 ist ein IQ < 70 ein Ausschlussgrund für die Diagnose „Rechenstörung“. Bei der Auswahl des Intelligenztests ist ein figuraler Test zu bevorzugen, weil eine Minderleistung bei numerischen Aufgaben durch die Rechenschwierigkeiten zu stande kommen kann. Das gleiche gilt bedingt auch für verbale Aufgaben, weil möglicherweise eine komorbide Lesestörung vorliegt.

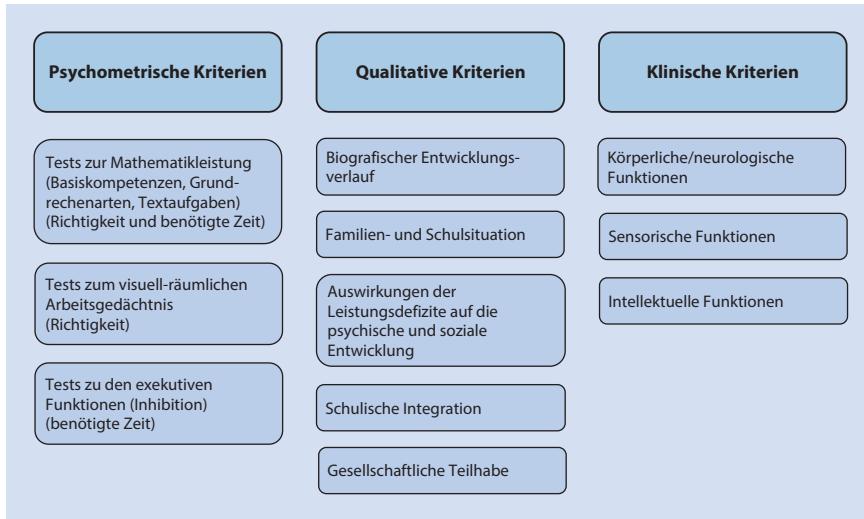


Abb. 7.2 Kriterien zur Diagnostik einer Rechenstörung nach Leitlinie der AWMF (2018, mit freundlicher Genehmigung der Deutschen Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie e. V.)

Empfehlungen zur Diagnostik einer Rechenstörung laut „Leitlinie“ (AWMF 2018, S. 20 f.)

„Die Diagnostik einer Rechenstörung soll beinhalten:

- psychometrische Tests zur Erfassung der Mathematikleistung (Basiskompetenzen, Grundrechenarten, Textaufgaben), der Leistung im Bereich des visuell-räumlichen Arbeitsgedächtnisses und der Leistung im Bereich der Exekutiven Funktionen (Inhibition),
(Empfehlungsgrad A, starke Empfehlung, starker Konsens: 100 % Zustimmung)
- die klinische Untersuchung einschließlich der körperlichen/neurologischen, sensorischen und intellektuellen Funktionen, (KKP)
- des biographischen Entwicklungsverlaufs, (KKP)
- der Familien- und der Schulsituation, (KKP)
- die Auswirkungen der Leistungsdefizite auf die psychische und soziale Entwicklung, (KKP)
- die schulische Integration, (KKP)
- die gesellschaftliche Teilhabe. (KKP)
(starker Konsens: 100 % Zustimmung)“

Anmerkung: KPP bedeutet „Klinischer Konsenspunkt“, die Zustimmung beträgt hier 100 %.

Weiter wird ausgeführt, dass für die Diagnose einer Rechenstörung unterdurchschnittliche Leistungen im Bereich der Mathematik vorliegen sollen, wobei „unterdurchschnittlich“ bedeutet, dass die Leistungen nach der Alters- oder der Klassennorm mindestens 1,5 Standardabweichungen unter dem Durchschnitt liegen. Wenn die klinischen und qualitativen Kriterien (Abb. 7.2) den Verdacht einer Rechenstörung unterstützen, soll es genügen, wenn die Diskrepanz nur 1 Standardabweichung beträgt. Zur Diagnostik der Mathematikleistung sollen Tests verwendet werden, die in 2 Tabellen (für 2 Altersbereiche) empfohlen werden. Dazu gehören auch die Serie Deutscher Mathematiktests DEMAT für die Klassen 1 bis 6 (Abschn. 7.4.1.2).

Eindeutige Diskrepanzkriterien für die Mathematikleistung	Die für die Diagnose verlangte Minderleistung im Bereich der Mathematik stellt ein einfaches Diskrepanzkriterium dar (nur die Mathematikleistung zählt). Zumindest implizit wird jedoch ein doppeltes Diskrepanzkriterium verwendet: Die beobachtete Minderleistung im Rechentest soll nicht mit der Intelligenz erklärt werden können. Der Leitlinie zufolge (AWMF 2018, S. 25) soll ein IQ unter 70 durch Einsatz eines figuralen („nonverbalen“) Intelligenztests ausgeschlossen werden.
Intelligenzminderung ausschließen	Dazu ist anzumerken, dass die Intelligenzdiagnostik in vielen Fällen zu schwer zu bewertenden Ergebnissen führen wird. Der für die Leitlinie durchgeföhrten Metaanalyse (AWMF 2018, S. 17) zufolge ist die Intelligenz (gemeint ist schlussfolgerndes Denken) bei Kindern mit einer Rechenstörung (im Vergleich zu denen ohne) deutlich vermindert (Effektstärke 0,85). Das Intelligenztestergebnis wird in vielen Fällen nahe an der Grenze eines IQ-Wertes von 70 liegen. Bei Berücksichtigung des Konfidenzintervalls für den IQ wird sich dann die Frage stellen, ob eine Intelligenzminderung sicher ausgeschlossen werden kann oder nicht.
Korrelate einer Rechenstörung	Aufmerksame Leserinnen und Leser haben vielleicht bemerkt, dass bei den psychometrischen Kriterien (Abb. 7.2) auch das Arbeitsgedächtnis und exekutive Funktionen genannt wurden. Defizite dieser Funktionen gehören zum „Profil“ einer Rechenstörung. Es handelt sich hierbei also „nur“ um Korrelate und nicht um Definitionselemente. Empirische Befunde sprechen dafür, dass Kinder mit Dyskalkulie in Bezug auf diese Funktionen häufiger Defizite aufweisen als Kinder ohne Dyskalkulie. Es bleibt allerdings unklar, welche Bedeutung auffälligen Befunden in diesen Bereichen zukommt. Für die Diagnose einer Rechenstörung sind sie jedenfalls nicht maßgeblich.
Screening auf Vorliegen komorbider Störungen	Eine weitere diagnostisch relevante Empfehlung der „Leitlinie“ betrifft die Suche nach möglicherweise vorliegenden komorbiden Störungen. Begründet wird dies mit einer erhöhten Prävalenz von bestimmten Störungen, wenn eine Rechenstörung vorliegt.
Förderrelevante Diagnostik	» Bei der Diagnostik einer Rechenstörung soll ein diagnostisches Screening auf das Vorhandensein komorbider Störungen stattfinden. Dabei sind besonders Symptome anderer schulischer Entwicklungsstörungen (LRS), Symptome aus dem ADHS-Spektrum sowie Symptome aus dem internalisierenden (insbesondere Mathematik-, Prüfungs- bzw. Schulangst) und externalisierenden Störungsspektrum zu berücksichtigen. (AWMF 2018, S. 38).
In der „Leitlinie“ werden Förderprogramme genannt, deren Wirksamkeit wissenschaftlich belegt ist. Ein Förderprogramm soll danach ausgewählt werden, dass es bei einem Kind oder Jugendlichen speziell geeignet ist. Das setzt eine förderrelevante Diagnostik voraus. Bei Rechenstörungen werden Subtypen unterschieden, die sich nach dem Fokus der Defizite richten (Lindberg et al. 2018). Beispiele für Fertigkeiten, die bei einer Rechenstörung defizitär und damit förderrelevant sein können, sind die Zahlen-Mengen-Zuordnung, die Zahlenbewusstheit und das Zahlengedächtnis.	In der „Leitlinie“ werden Förderprogramme genannt, deren Wirksamkeit wissenschaftlich belegt ist. Ein Förderprogramm soll danach ausgewählt werden, dass es bei einem Kind oder Jugendlichen speziell geeignet ist. Das setzt eine förderrelevante Diagnostik voraus. Bei Rechenstörungen werden Subtypen unterschieden, die sich nach dem Fokus der Defizite richten (Lindberg et al. 2018). Beispiele für Fertigkeiten, die bei einer Rechenstörung defizitär und damit förderrelevant sein können, sind die Zahlen-Mengen-Zuordnung, die Zahlenbewusstheit und das Zahlengedächtnis.
Weiterführende Literatur	Über die Diagnostik des Leseverständnisses informiert der von Lenhard und Schneider (2009) herausgegebene Band und über die Diagnostik mathematischer Kompetenzen der von Hasselhorn et al. (2013) herausgegebene Band. Für eine vertiefte Analyse zum Thema Rechenstörungen sei etwa auf Jacobs et al. (2013) oder Jacobs und Petermann (2007, 2012) verwiesen. Speziell für die Diagnostik von Entwicklungsdefiziten, die sich im schulischen Bereich auswirken können, kann der Band <i>Diagnostik im Vorschulalter</i> (Esser et al. 2015) empfohlen werden.

7.3 Hochbegabungsdiagnostik

Definitionsprobleme Ein wesentliches Problem bei der Diagnostik von *Hochbegabung* liegt bereits in der Begriffsbestimmung von Hochbegabung. Eine einheitliche Definition gibt es nicht:

- » Begabung bezeichnet allgemein das leistungsbezogene Potenzial eines Menschen.[...] Begabung oder Hochbegabung beziehen sich dabei immer auf ein bestimmtes Aktionsfeld (*begabt wofür?*) (Preckel und Vock 2013, S. 12).

Zumeist denkt man bei Hochbegabung an intellektuelle Hochbegabung; es gibt aber auch Menschen, die etwa musikalisch oder sportlich hochbegabt sind.

Zwei Dimensionen spielen in der Definition von Hochbegabung eine Rolle: die inhaltliche Breite des Konzepts und die Operationalisierung über eine in der Person liegende Kompetenz oder die Performanz (sichtbare Leistung) der Person (Tab. 7.4). Kompetenz und Leistung stehen in einer engen Beziehung zueinander. Kinder mit einer sehr hohen Intelligenz werden in der Regel in der Schule gute Noten erzielen. Die Beziehung wird durch andere Faktoren in der Person (z. B. Leistungsmotivation) und in der Umwelt (z. B. Förderung) moderiert. Mit den Begriffen „Under- und Overachiever“ (s. u.) wird zum Ausdruck gebracht, dass es erwartungswidrige, aber erklärbare Diskrepanzen zwischen dem intellektuellen Potenzial und der gezeigten Leistung in der Schule geben kann. Bei Musik oder Sport wird man zunächst an die beobachtbare Leistung denken und Menschen als musikalisch oder sportlich hochbegabt bezeichnen, die in dem jeweiligen Bereich durch besondere Leistungen aufgefallen sind. Dieser Leistung liegt vermutlich auch ein besonderes Potenzial zugrunde.

Begabung als Potenzial

Zweidimensionale Betrachtung des Konzepts der Hochbegabung

Wir betrachten in diesem Kapitel Hochbegabung als eine Kompetenz und damit als ein Potenzial für außergewöhnliche Leistungen. Ferner beschränken wir uns auf eine einzige Fähigkeit, nämlich die Intelligenz. In Tab. 7.4 ist diese Kombination in der Zelle links oben verortet. Gut begründbar und auch weitgehend konsensfähig ist die Festlegung auf ein Kriterium, und zwar das der Allgemeinen Intelligenz. Man setzt weiterhin fest, dass von Hochbegabung zu sprechen ist, wenn die Allgemeine Intelligenz 2 Standardabweichungen über dem Populationsmittelwert liegt (Holling und Kanning 1999; Rost et al. 2006). Diese Konzeption ist vor allem für die Praxis nützlich, da sie eine klare diagnostische Entscheidung ermöglicht.

Tab. 7.4 Definitionsansätze für Hochbegabung

Breite des Konzepts	Hochbegabung wird definiert über ...	
	Kompetenz	Performanz
Eindimensional	Sehr hohe Ausprägung einer Fähigkeit (z. B. Intelligenz)	Außergewöhnliche Leistung in einem Bereich (z. B. Schulleistung)
Mehrdimensional	Sehr hohe Ausprägung von mehr als einer Fähigkeit (z. B. Intelligenz und Kreativität)	Außergewöhnliche Leistungen in mehr als einem Bereich (z. B. Mathematik und Musik)

Quelle: In Anlehnung an Preckel und Vock (2013, © Hogrefe)

Hochbegabung als IQ ≥ 130 definiert

Definition

Unter **Hochbegabung** verstehen wir eine herausragend hohe allgemeine Intelligenz. Der IQ liegt mindestens bei 130.

Durch Kombination von Kriterien wird Hochbegabung wegdefiniert

Dieser Definition liegen 4 Entscheidungen zugrunde: Erstens wird Hochbegabung als intellektuelle (und nicht irgendeine andere, s. u.) Hochbegabung verstanden. Zweitens wird mit dem Rückgriff auf den IQ eine operationale Definition vorgenommen, weil Intelligenztests die beste Methode zur Messung der Intelligenz sind. Drittens wird die Intelligenz als allgemeine Intelligenz (► Abschn. 3.2.3.1) verstanden. Das impliziert, dass für die Messung Tests zu verwenden sind, die nicht ausschließlich verbale, numerische oder figurale Aufgaben verwenden und die nicht ausschließlich eine spezifische Intelligenzkomponente (mit Ausnahme von schlussfolgerndem Denken als Kernbereich der Intelligenz) erfassen. Viertens wird die Grenze exakt festgelegt, und zwar bei einem IQ ab 130. Damit gelten per Definition 2,3 % der Menschen als hochbegabt. Die Definition deckt sich mit der Festlegung auf hohe allgemeine Intelligenz und auf einen IQ ab 130 u. a. mit den Ausführungen von Rost (2009).

Einige Autorinnen und Autoren präferieren eine Hochbegabungsdiagnostik ausschließlich anhand der Allgemeinen Intelligenz, andere schließen auch andere Fähigkeitsbereiche wie soziale Intelligenz oder Kreativität mit ein. Dementsprechend wird je nach Begriffsverständnis das Urteil, ob eine Hochbegabung vorliegt, unterschiedlich ausfallen. Vor allem reduziert sich bei einem mehrdimensionalen Definitionsansatz die Zahl der Hochbegabten. Der Aufwand, um einen hochbegabten Menschen zu entdecken, nimmt mit jedem weiteren Kriterium immens zu. Hanses und Rost (1998) haben berechnet, wie viele Personen in Abhängigkeit von der Anzahl der geforderten Kriterien (bei einem festgelegten Cut-off-Wert) untersucht werden müssten, um 50 Hochbegabte zu finden. Daraus lässt sich ablesen, wie hoch der Anteil der Hochbegabten in der Population sein muss. Verlangt man, dass ein hochbegabter Mensch in Bezug auf 1 Kriterium (z. B. Intelligenz) zu den oberen 2 % der Verteilung gehört, sind 2500 Personen zu untersuchen. Schon bei 2 Kriterien (unter der Annahme, dass sie zu .30 korrelieren) erhöht sich die Zahl auf 30.048 Personen. Die Hinzunahme des 2. Kriteriums führt dazu, dass jetzt nicht mehr 2 % der Population als hochbegabt gelten, sondern nur noch 0,17 %. Bei 3 Kriterien (die ebenfalls zu .30 miteinander korrelieren) verringert sich der Anteil der Hochbegabten bereits auf 0,03 %. Diese Modellrechnung macht deutlich, dass die Vorannahmen über das Konzept der Hochbegabung enorme praktische Konsequenzen haben. Durch die Forderung nach immer weiteren Kriterien lässt sich die Hochbegabung schlüssig „wegdefinieren“. Außerdem wird mit jedem weiteren Kriterium konzeptuell unklarer, was die Kombination von Multitalenten inhaltlich bedeutet.

Begabung und Leistung Unter Intelligenz wird das *Potenzial* einer Person verstanden, kognitive Leistungen zu erbringen. Eine hoch intelligente Person kann gute Leistungen in der Schule oder etwa im Beruf zeigen, muss dies aber nicht tun. Motivationale Gründe oder ungünstige Lern- und Arbeitsbedingungen können dazu führen, dass die Person nicht die Leistungen zeigt, zu der sie eigentlich fähig wäre.

Die strikte Unterscheidung zwischen Fähigkeit (Potenzial) und Performance führt dazu, dass bei einer kategorialen Betrachtung folgende 3 Typen von Hochbegabten resultieren: *Underachiever* (die Leistungen sind niedriger als nach dem Potenzial zu erwarten wäre), *Achiever* (Hochbegabte, deren Leistungen ihren Fähigkeiten entsprechen) und *Overachiever* (Schülerinnen und Schüler, die herausragende Leistungen zeigen, die aber in dieser Höhe

Hochbegabte können auch „Underachiever“ sein

nicht durch ihre Intelligenz erklärbar sind). Wird in der Schule nur auf eine außergewöhnliche Performanz (z. B. sehr gute Leistungen) geachtet, fallen die hochbegabten Underachiever oft nicht auf. Dagegen werden Overachiever entdeckt und – zu Unrecht – als hochbegabt eingestuft.

Hochbegabungsdiagnostik durch Lehrkräfte Die Schule ist der Ort, an dem Hochbegabte vor allem auffallen sollten. Aber können Lehrerinnen und Lehrer eine herausragende intellektuelle Begabung ohne Zuhilfenahme von Tests erkennen? Wild (1993) hat in einer groß angelegten Studie an Drittklässlern die Übereinstimmung zwischen Lehrerurteilen und Intelligenztestergebnissen überprüft. Die Lehrerstichprobe umfasste 388 Lehrkräfte, die insgesamt über 7000 Schülerinnen und Schüler beurteilten. Die Intelligenz wurde mit 3 Tests gemessen: Grundintelligenztest – Skala 2 (CFT 20), Zahlen-Verbindungs-Test (ZVT) und Sprachliche Analogien 3/4 (SPA). Die Lehrerinnen und Lehrer stuften die Intelligenz ihrer Schülerinnen und Schüler auf 7-stufigen Ratingskalen von „extrem schwach“ bis „exzellent“ ein. Sie erhielten Informationen über den Inhalt (auch Itembeispiele) und den Aufbau der Intelligenztests, an denen ihr Urteil später überprüft werden sollte. Pro Schülerinnen und Schüler gaben sie für jeden der 3 Tests eine Prognose ab. Weiterhin nominierten sie einige ihrer Schülerinnen und Schüler als voraussichtlich hochbegabt. Dazu diente eine Liste von 15 begabungsrelevanten Merkmalen (z. B. formal-logisches Denken, Merkfähigkeit; solche Checklisten finden zum Teil in der Hochbegabungsdiagnostik Verwendung). Pro begabungsrelevantem Merkmal durften die Lehrkräfte maximal 3 Schülerinnen bzw. Schüler mit hoher Merkmalsausprägung benennen.

Die Korrelationen zwischen den Testleistungen und Ratings sowie den Nominierungen variierten sehr stark zwischen den Klassen. Einige Lehrerinnen und Lehrer schätzten die Intelligenz ihrer Schülerinnen und Schüler also recht gut ein, andere erwiesen sich als schlechte Diagnostikerinnen bzw. Diagnostiker. Über alle Klassen hinweg korrelierte die Intelligenztestleistung (aggregiert über die 3 Tests) mit den ebenfalls gemittelten Einschätzungen der Lehrerinnen und Lehrer immerhin zu $r = .59$. Für die Nomination als hochbegabt fiel die vergleichbare Korrelation mit $r = .47$ etwas niedriger aus.

Von großer praktischer Bedeutung sind die Trefferquoten, die mit einem Intelligenzrating und einem Nominierungsverfahren erzielt werden. Von den Schülerinnen und Schülern, die nach dem Lehrerinnen- bzw. Lehrerurteil „exzellent“ begabt sind, erwiesen sich – gemäß den Intelligenztestergebnissen – 35,1 % als tatsächlich hochbegabt. Der Rest hatte zu niedrige Intelligenztestergebnisse, um als hochbegabt zu gelten. Man kann auch umgekehrt fragen, wie viele der tatsächlich hochbegabten Schülerinnen und Schüler ($IQ \geq 130$) durch eine Lehrerinnen- bzw. Lehrerbeurteilung entdeckt wurden: Es sind lediglich 16,4 %! Insgesamt belegt diese Untersuchung eindrucksvoll, dass Urteile von Lehrenden wenig brauchbar sind, um Hochbegabte zu entdecken.

Lehrerurteile eignen sich aus pragmatischen Gründen zumindest für eine Vorselektion. Damit möglichst viele wirklich Hochbegabte gefunden werden, muss man den Ergebnissen dieser Studie zufolge all jene Schülerinnen und Schüler einer gründlichen Intelligenzdiagnostik unterziehen, die von den Lehrkräften mindestens als „gut“ begabt (3. Stufe auf der 7-stufigen Skala) beurteilt werden. Bei einer derart groben Vorauswahl würden lediglich 1,5 % der Hochbegabten nicht entdeckt.

Außer Lehrerinnen und Lehrern können auch andere Personen zur Entdeckung von Hochbegabten beitragen. Neben den Eltern sind hier Mitschülerinnen und Mitschüler sowie Freunde und auch die Hochbegabten selbst zu nennen. Die Güte dieser Quellen ist allerdings als kritisch zu beurteilen (Rost et al. 2006).

Lehrurteile über die Begabung ihrer Schülerinnen und Schüler

Relativ hohe Korrelation mit IQ

Im Extrembereich ($IQ \geq 130$)
Übereinstimmung gering

Lehrerurteil als Screening

Mindestens 1 Test zur Allgemeinen Intelligenz – besser aber 2

Anforderungen an den Intelligenztest Welche Anforderungen sind an einen Intelligenztest zur Feststellung von Hochbegabung zu stellen? Erstens sollte der Test ein breites g-Maß darstellen, die Intelligenz also über mehrere Teilbereiche prüfen. Werden beispielsweise nur numerische Testaufgaben verwendet, kann der Testwert (z. B. bedingt durch besondere schulische Förderung in Mathematik) im Vergleich zu anderen Begabungsbereichen erhöht sein und zu einer Überschätzung der Intelligenz führen. Umgekehrt können eine Teilleistungsschwäche oder eine mangelnde schulische Förderung dazu führen, dass die Allgemeine Intelligenz unterschätzt wird. Das Phänomen der Regression zur Mitte (► Abschn. 7.1.1) führt jedoch dazu, dass es bei einem Screening zu falsch positiven Urteilen kommt. Bedingt durch mangelnde Messgenauigkeit kann es auch zu einigen falsch negativen Urteilen kommen. Anstelle eines einzigen Tests ist es besser, 2 Intelligenztests zu verwenden. Der Mittelwert aus beiden Tests stellt die beste Schätzung dar (► Abschn. 7.1.1).

Aktuelle und repräsentative Normen, Differenzierung im oberen Bereich

Zweitens sind aktuelle Normen zu fordern, die für die infrage kommende Altersgruppe auf einer hinreichend großen und repräsentativen Stichprobe (also beispielsweise kein Übergewicht von Gymnasiastinnen und Gymnasiasten) basieren. Durch die beobachtete Zunahme der Intelligenztestleistungen im Laufe der Zeit (Flynn-Effekt) führt die Verwendung überalterter Normen zwangsläufig dazu, dass zu viele Personen als hochbegabt diagnostiziert werden. Drittens muss der Test im oberen Leistungsbereich gut differenzieren. Dazu sollten die Normen weit über einen IQ von 130 hinausgehen. Ein Test, der speziell zur Hochbegabungsdiagnostik entwickelt wurde und der diesen Anforderungen auch gerecht wird, ist der *Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB)* von Jäger et al. (2006; s. auch ► Abschn. 3.2.3.2).

Spezialbegabung ist nicht Hochbegabung

Spezialbegabungen Menschen können – neben der Intelligenz – auch in anderen Bereichen besonders begabt sein. Beispiele sind Mathematik, Kunst, Sport und Musik. Diese Begabungen oder Talente sollten nicht zum Begriff der Hochbegabung zählen, der für eine sehr hohe allgemeine Intelligenz reserviert bleiben soll. Im Einzelfall stellt sich aber die Frage, ob eine Spezialbegabung nicht Ausdruck einer außergewöhnlich hohen Intelligenz ist. Kognitive Fähigkeiten sind positiv korreliert. Herausragende Leistungen in der Mathematik werden zumeist mit einer sehr hohen Intelligenz einhergehen. Für eine explizit nicht intellektuelle Spitzenbegabung kann der Begriff „Talent“ verwendet werden (Rost 2001).

Bedeutung der Diagnose „Hochbegabung“ Nach der Diagnose „Hochbegabung“ fragen sich die Betroffenen, zumeist sind es die Eltern eines Kindes und dessen Lehrerinnen bzw. Lehrer, was nun zu tun ist. Zunächst einmal ist festzustellen, dass Hochbegabung nicht mit einer Benachteiligung in anderen Merkmalsbereichen „bezahlt“ wird. Es ist ein Mythos, der von einigen Elternvereinigungen mit Hinweis auf Einzelfälle genährt wird, dass hochbegabte Kinder im sozialen oder emotionalen Bereich als Folge ihrer Hochbegabung leiden. Im Gegenteil:

- » Hochbegabte Kinder gehen beispielsweise stärker aus sich heraus, sind warmherziger, emotional stabiler, ruhiger, fröhlicher, enthusiastischer, natürlicher als Schüler mittlerer oder unterer Intelligenz. (Rost 2001, S. 180)

Keine Benachteiligung zu erkennen

Bereits die berühmte Terman-Studie, in der der Lebensweg von 1528 hochbegabten Kinder und einer normal intelligenten Kontrollgruppe verfolgt wurde, erbrachte viele Belege für Vorteile der Hochbegabten (Oden 1968). Auch neue Querschnittsstudien belegen, dass hochbegabte Menschen im Leben nicht benachteiligt sind. In einer dieser Studien (Wirthwein et al. 2019) wurden Schülerinnen und Schüler von Gymnasien mit einem IQ über 130 mit durchschnittlich begabten Mitschülerinnen und Mitschülern (IQ 85 bis 115)

verglichen. Die hochbegabten Jugendlichen hatten in allen Fächern deutlich bessere Noten und beschrieben sich u. a. als offener und leistungsmotivierter. Eine Metaanalyse (Stricker et al. 2019) galt der Frage, ob sich normal- und hochbegabte Menschen bezüglich ihres Perfektionismus unterscheiden. Zwei Komponenten von Perfektionismus wurden unterschieden: Besorgtheit (Zweifel, alles richtig zu machen) und Streben nach Perfektionismus. Hochbegabte haben offenbar geringfügig höhere Ansprüche an sich selbst (Hedges $g=0,03$) – die aber nicht von Besorgtheit begleitet werden, diese Ansprüche nicht erfüllen zu können.

Da eine Begabung nicht automatisch zu entsprechenden Leistungen führt, liegt es nahe, über Fördermaßnahmen nachzudenken, die Hochbegabten hilft, ihr Potenzial zu entfalten. Im Einzelfall ist unter Berücksichtigung der individuellen Lern- und Lebensbedingungen zu erwägen, ob eine Fördermaßnahme (z. B. Überspringen einer Klasse) überhaupt angebracht ist – und wenn ja, welche (s. auch Interview mit Prof. Dr. Detlef H. Rost).

Helfen, Potenzial zu entfalten

Weiterführende Literatur

Zum Thema Hochbegabung können die Bände *Diagnostik von Hochbegabung* (Preckel et al. 2010) und *Begabungen und Talente* (Trautwein und Hasselhorn 2017) aus der Serie „Tests und Trends in der pädagogisch-psychologischen Diagnostik“ sowie das Lehrbuch *Hochbegabung* (Preckel und Vock 2013) empfohlen werden.

Interview mit Prof. Dr. Detlef H. Rost zum Thema „Hochbegabung: Begabungsdiagnostische Beratung“



Prof. Rost war bis zu seiner Pensionierung Professor für Pädagogische Psychologie und Entwicklungspsychologie am FB Psychologie der Philipps-Universität Marburg. Er leitet dort weiter die Begabungsdiagnostische Beratungsstelle (BRAIN)

Wenn man Hochbegabung als IQ von mindestens 130 definiert, müssen bei einer Normalverteilung der Intelligenz rund 2 % der Menschen hochbegabt sein. In Deutschland gibt es demnach etwa 1,6 Mio. Hochbegabte. Wie viele davon wissen Ihrer Einschätzung nach, dass sie zu dieser Gruppe gehören?

Viele ältere Hochbegabte haben im Laufe ihres Lebensvollzugs schon, wie es ein Betroffener einmal ausdrückte, gemerkt, dass sie „nicht dumm“ sind. Hochbegabte Schülerinnen und Schüler, insbesondere wenn sie noch die Grundschule besuchen, sind sich im Regelfall nicht bewusst,

dass sie zur Gruppe der Hochbegabten gehören – und das ist pädagogisch-psychologisch auch gut so, und es gibt keinerlei Änderungsbedarf.

Sie leiten seit 1999 die Begabungsdiagnostische Beratungsstelle *BRAIN* an der Philipps-Universität Marburg, die das Land Hessen eingerichtet hat. Warum finanziert das Land immerhin vier Teilzeitstellen (60 %), wo doch jede Psychologin und jeder Psychologe mit Hilfe eines Intelligenztests feststellen kann, ob ein Mensch hochbegabt ist?

Unsere Erfahrung zeigt leider, dass sich nicht wenige Psychologinnen und Psychologen mit einer

soliden (d. h. guten, psychodiagnostischen Standards genügenden) Diagnostik kognitiver Leistungsfähigkeit und einer differenzierten Gutachten erstellung ausgesprochen schwer tun – von den vielen Pädagoginnen bzw. Pädagogen und (Kinder-)Ärztinnen bzw. Ärzten, die trotz fehlender fachlicher Kompetenz Begabungsdiagnostik betreiben und sachlich falsche Befunde mitteilen, einmal ganz abgesehen. Nicht selten werden veraltete Tests eingesetzt, mit dem Resultat, dass wegen hochgradig veralteter Normen Kinder und Jugendliche als hochbegabt diagnostiziert werden, die deutlich von einer Hochbegabung entfernt sind („falsch positive“ Diagnose als Auswirkung des „Flynn-Effekts“).

Wer kommt zu Ihnen in die Beratungsstelle und warum?

Die Altersverteilung der fast 10.000 Kinder und Jugendlichen, deren Eltern bislang mit *BRAIN* Kontakt aufgenommen haben, sieht wie folgt aus: Der größte Teil, nämlich 51 %, war zwischen 6 und 9 Jahre alt, 18 % befanden sich im Vorschulalter mit der Fragestellung „vorzeitige Einschulung?“, 17 % waren im Alter von 10 bis 12 und 10 % von 13 bis 15 Jahren. Nur 4 % waren über 15 Jahre alt. Fast drei Viertel der vorgestellten Kinder und Jugendlichen waren Jungen.

Diagnostik ist kein Selbstzweck – das IQ-Feststellen reicht nicht aus. In fast allen Fällen liegen bei *BRAIN* über die eigentliche Begabungsdiagnostik hinausgehende spezielle Beratungsanliegen vor. Bei diesen (Mehrfachnennungen möglich) ergibt sich folgende Rangordnung: Schullaufbahnberatung, einschließlich Über-springen und vorzeitige Einschulung, ist zu 31 % gefragt, 29 % suchen (außerschulische) Fördermöglichkeiten; Verhaltensauffälligkeiten werden zu 29 % thematisiert. Schulische Langeweile in Verbindung mit Unterforderung wird von 28 % genannt. Schulische Leistungsprobleme (15 %) und Interaktionsschwierigkeiten mit den Peers (11 %) werden dagegen zu einem deutlich geringeren Anteil angesprochen. Rein „präventive“ Anfragen liegen zu 18 % vor. Andere

Gründe machen jeweils 5 % oder weniger aus (Probleme mit Lehrkräften: 5 %, Erziehungsprobleme: 4 %, Absicherung einer Vordiagnose und diskrepanter Entwicklung: je 3 %, Überforderung der Eltern: 2 %). In die Sammelkategorie „Rest“ fallen 6 %.

Zu einer Beratung gehört vermutlich mehr als nur die Intelligenzmessung.

Je nach Problemlage kommen ergänzend Persönlichkeits- und Interessenfragebögen, Tests zur Erfassung der Peer- und Familienbeziehungen, zum Lern- und Arbeitsverhalten, zum Klassenklima etc. hinzu. Eine ausführliche problembezogene Anamnese und Exploration sowie Verhaltensbeobachtung bei der Test- und Fragebogenbearbeitung sind integraler Diagnostikbestandteil.

Bei schwierigeren Fällen werden die Ergebnisse und Beratungsempfehlungen in den wöchentlichen Teamsitzungen vorbesprochen. Im ausführlichen Beratungsgespräch (meistens 2–3 h dauernd) erläutert der Berater/die Beraterin sehr verständlich die Resultate der Diagnostik und die Empfehlungen anhand des detaillierten Gutachtens, und gemeinsam mit den Betroffenen (zumeist Eltern) werden realistische Veränderungsmöglichkeiten erarbeitet. Niemand verlässt *BRAIN*, ohne das Gutachten voll verstanden zu haben.

Noch zwei für die Beratungspraxis nicht unwichtige Argumente: Eine mit staatlicher Autorität ausgestattete begabungsdiagnostische Beratungsstelle kann in manchen Fällen eher Veränderungen anstoßen als frei praktizierende Psychologinnen und Psychologen. Da *BRAIN* keine Rechnung stellt, können zudem auch weniger Betroffene eine kompetente Beratung bekommen.

Es gibt Elternvereinigungen, die hochbegabte Kinder als bedauernswerte Geschöpfe darstellen, die in der Schule unter der Mittelmäßigkeit ihrer Mitschülerinnen und Mitschüler sowie der mangelnden Förderung durch ihre Lehrerinnen bzw. Lehrer leiden müssen. Wie stehen Sie dazu?

Dies ist ein weitverbreitetes Vorurteil. Im Marburger Hochbegabtenprojekt, einem seit 1987 laufenden Längsschnittprojekt, konnten

wir anhand einer nicht durch Lehrkräfte etc. vorausgelesenen Gruppe von Hochbegabten zeigen, dass – verglichen mit durchschnittlich Begabten – Hochbegabte in der Regel gut mit sich selbst, ihren Klassenkameraden und -kameraden und ihren Lehrkräften auskommen; jedenfalls haben sie nicht mehr Probleme als ihre nicht hochbegabten Peers.

Der Eindruck, es gäbe viele Hochbegabte mit massiveren Problemen, entspricht dem gängigen „Genie-Wahnsinn“-Vorurteil und wird in zahlreichen Elternratgebern auch heute noch gern gepflegt. Solche Ratgeberbücher sind oft von psychologischen Laien (z. B. von betroffenen Eltern oder Lehrkräften) verfasst worden. Zudem werden in diesen Broschüren (und in den Medien) gern interessante – d. h. plakativ-abweichende – Einzelfälle vorgestellt, von denen unzulässig auf „die“ Hochbegabten geschlossen wird. Ähnliche Aussagen wurden und werden immer noch von einschlägigen

Elternvereinen in die Welt gesetzt. Warum? Die Antwort liegt auf der Hand: In solchen Vereinigungen sammeln sich mehrheitlich Problemfälle: Selbsthilfegruppen für das pflegeleichte Sonnenscheinkind sind mir nicht bekannt.

Sie sind seit 2013 Gastprofessor an der Southwest University Chongqing (V. R. China). Was ist in China bezüglich Hochbegabung anders als bei uns?

Hochbegabung und Hochleistung stehen in China sehr hoch im Kurs. Kinder und Jugendliche werden deshalb ab der Mittelschule, d. h. ab etwa 12 Jahren, von ihren Familien kontinuierlich einem ziemlich starken Leistungsdruck ausgesetzt. In der Beratung muss den Eltern oftmals klar gemacht werden, dass der andauernde enorme Leistungsstress für die Entwicklung einer harmonischen Persönlichkeit alles andere als förderlich ist: Auch Hochbegabte haben ein Anrecht darauf, nicht ständig gefördert zu werden – in China, bei uns, überall.

7.4 Tests im Bildungsbereich

Angesichts der vielen Tests, die für das Bildungssystem verfügbar sind, können wir hier nur exemplarisch auf einige Verfahren eingehen. Ziel ist es, die Besonderheiten der jeweiligen Verfahrensgruppe herauszustellen.

7.4.1 Schultests

Schultests unterscheiden sich nicht grundsätzlich von Entwicklungstests (► Abschn. 3.2.5), sondern vor allem hinsichtlich der Spezifität ihrer Anforderungen. Während Entwicklungstests eher allgemeine Fähigkeiten erfassen, dienen Schultests dazu, Fähigkeiten und Fertigkeiten zu erfassen, die eine Voraussetzung für das Erbringen schulischer Leistungen darstellen, oder sie erfassen direkt schulische Leistungen. Die erste Aufgabe wird von Schuleingangstests übernommen. Schulleistungstests messen den Leistungsstand von Schülern in einem bestimmten Bereich.

7.4.1.1 Schuleingangstests

Schuleingangstests wurden früher meist als Schulreifetests bezeichnet. Die implizite Annahme, dass sich die Eignung von Kindern für den Schulunterricht durch (biologische) Reifung quasi mit der Zeit von selbst einstellt, ist sicher naiv. Irreführend ist auch der Begriff Schulfähigkeitstest, weil neben Fähigkeiten natürlich auch (erworrene, trainierbare) Fertigkeiten relevant sind. Schuleingangstests sollen prüfen, ob ein schulpflichtiges Kind den Anforderungen der Schule gewachsen ist. Durch ihren Einsatz bereits vor Schuleintritt kann eventuell verhindert werden, dass noch nicht schulfähige Kinder

durch eine zu frühe Einschulung überfordert und dadurch psychisch geschädigt werden. Zusätzlich kann ggf. im Einzelfall auch ein Förderbedarf festgestellt werden.

Viele Aufgabentypen

7

Die Herausforderung besteht darin, dass ein Verhalten vorhergesagt werden soll, das in dieser Form noch nicht vorliegt, sondern erst später erworben wird: Kinder lernen in der Schule das Schreiben; ein Einschulungstest soll vorhersagen, ob dies einem Kind gelingen wird, das zum Untersuchungszeitpunkt noch nicht schreiben kann. Mit den Tests versucht man, in einfacher und kindgemäßer Form jene Grundfertigkeiten zu erfassen, die Kinder benötigen, um in der Schule Lesen, Schreiben und Rechnen zu erlernen. Die Fähigkeit zur Formerfassung wird dabei als Voraussetzung zum Erlernen der grafischen Symbole, die Auffassung von Mengen bis 5 als Basis für erfolgreiche Teilnahme am Mathematikunterricht angesehen. Zeichen-Aufgaben sollen grundlegende schreibmotorische Fertigkeiten diagnostizieren. Viele Schulreifetests verlangen das Nachzeichnen von Formen und das Zeichnen eines Menschen. Nach Langfeldt und Tent (1999, S. 140) finden folgende Aufgabentypen in den 9 von den Autoren analysierten „Schulreifetests“ am häufigsten Verwendung (in Klammern ist die Anzahl der damals verfügbaren Tests mit diesem Aufgabentyp angegeben):

- Nachmalen von Formen (Figuren, Ziffern, Buchstaben und Kombinationen) (8)
- Mann-Zeichnungen (5)
- Wiederholtes Zeichnen abstrakter Figuren (Zaun, Muster) (5)
- Malen bzw. Legen vorgegebener oder kurz exponierter Mengen (4)
- Heraussuchen und Markieren identischer Figuren aus ähnlichen (4)
- Markieren von Bildern nach Sprachverständnis für Einzelsituationen (4)

Es sind nur wenige Schuleingangstests verfügbar; die meisten davon sind älteren Datums. Vorgestellt werden 2 Tests neueren Datums.

Screening des Entwicklungsstandes bei Einschulungsuntersuchungen (S-ENS). Als eines der wenigen relativ neuen Verfahren ist das S-ENS von Döpfner et al. (2005) zu nennen. Es soll Hinweise auf mögliche Entwicklungsdefizite liefern und wurde damals an einer großen Stichprobe ($N=27.000$) normiert. Folgende Entwicklungsbereiche werden dazu untersucht (in Klammern sind die Namen der 8 Untertests aufgeführt):

- Körperkoordination (seitliches Hin- und Herspringen)
- Visuomotorik (Gestaltrekonstruktion, Gestaltreproduktion)
- Visuelle Wahrnehmung und Informationsverarbeitung (gleichnamiger Subtest; Aufgaben: Erkennen identischer figuraler Vorgaben, Auswahl einer Ergänzungsfürfigur nach bestimmten Regeln)
- Sprachkompetenz und auditive Informationsverarbeitung (Pseudowörter nachsprechen, Wörter ergänzen, Sätze nachsprechen)
- Artikulation (gleichnamiger Untertest)

Würzburger Vorschultest (WVT) Der WVT von Endlich et al. (2017) ist zur Feststellung der Eignung für die Schule verwendbar, weil er schriftsprachliche und mathematische Kompetenzen erfasst, die für das Lesen, Rechnen und Schreiben von Bedeutung sind. Der WVT besteht aus insgesamt 29 Untertests, die sich 3 Bereichen (als Module bezeichnet) zuordnen lassen: frühe schriftsprachliche Fähigkeiten, sprachliche Fähigkeiten sowie mathematische (Vorläufer-)Fertigkeiten. Der Test wurde an 417 Kindern im Kindergartenjahr vor der Einschulung aus 6 Bundesländern normiert. Er kann von Erzieherinnen und Erziehern oder Lehrkräften durchgeführt werden, und zwar als kompletter Test oder auch als einzelne Module.

7.4.1.2 Schulleistungstests

Lehrerinnen und Lehrer beurteilen die Leistungen ihrer Schülerinnen und Schüler zumeist durch Vergleiche innerhalb der Schulklasse. Schulleistungstests wurden konstruiert, um schulische Leistungen unter standardisierten Bedingungen zu erfassen und um sie anhand von überregionalen Normen (in der Regel bundesweiten) zu beurteilen. Es gibt sie vor allem für die Fächer Deutsch und Mathematik für verschiedene Klassenstufen. Sie werden meist so konstruiert, dass sie mit ihren Aufgaben die bundesweiten Bildungsstandards in einem Fach abbilden. Damit besteht auch die Gefahr, dass die Inhaltsvalidität verloren geht, wenn sich die Bildungsstandards verändern. Wenn Schulleistungstests im Rahmen einer Schullaufbahnberatung eingesetzt werden, sollte der Nachweis vorliegen, dass spätere Schulleistungen in dem Fach mit dem Test besser vorhergesagt werden als mit der aktuellen Fachnote oder einem Aggregat aus den letzten Fachnoten (inkrementelle Validität).

Leistungsstand in einem Schulfach nach bundesweiten Maßstäben bestimmen

Wozu braucht man Schulleistungstests?

Schulleistungstests können insbesondere für diese Fragestellungen verwendet werden:

- Wie ist der Leistungsstand der Schulklasse in z. B. Deutsch im bundesweiten Vergleich?
- Wie ist der Leistungsstand einer Schülerin in dem Schulfach – und zwar unabhängig von den (vielleicht zu strengen oder zu milden) Noten? Dies kann etwa bei einem geplanten Schulwechsel helfen, die passende Klassenstufe zu finden.
- Wie hoch oder niedrig ist die Rechenfertigkeit eines Schülers im Vergleich zu seiner Intelligenz? Der Vergleich ist für die Diagnose einer Teilleistungsstörung (Legasthenie, hier: Dyskalkulie) erforderlich.
- Wo genau hat die Schülerin Schwächen im Bereich der Rechtschreibung? Einige Tests helfen, zu erkennen, welche Fehler besonders oft vorkommen. Dies kann für die Förderung hilfreich sein.

Rechtschreibtests

Deutscher Rechtschreibtest (DERET) Schulleistungstests sind meistens für 1 oder allenfalls 2 aufeinanderfolgende Schuljahre entwickelt worden. Manchmal liegen mehrere Tests vor, die aufeinanderfolgende Altersbereiche abdecken; beispielsweise liegt eine Gruppe deutscher Rechtschreibtests vor: *DERET 1–2+* für das 1. und 2. Schuljahr (Stock und Schneider 2008a), *DERET 3–4+* für das 3. und 4. Schuljahr (Stock und Schneider 2008b) und *DERET 5–6+* für das 5. und 6. Schuljahr (Martinez Méndez et al. 2015). Die Tests verwenden 2 Aufgabenarten, nämlich ein Diktat und einen Lückentext, in dem nur an bestimmten Stellen Wörter einzutragen sind. Das etwas aufwendiger auszuwertende Diktat wurde mit dem Hinweis auf seine ökologische Validität für den Schulunterricht aufgenommen. Da Parallelformen vorliegen, können sie auch gut im Klassenverband eingesetzt werden, ohne dass ein Abschreiben zu befürchten ist. Beim DERET 1–2+ und dem DERET 3–4+ wurde bei den Diktaten die Inhaltsvalidität auf besondere Weise sichergestellt. Bei der Entwicklung der Diktattexte orientierten sich die Autorinnen und Autoren an den Grundschullehrplänen aller 16 Bundesländer (inklusive der darin vorhandenen Grundwortschätze) und den Wörterlisten der gängigsten Schulbücher für den Rechtschreibunterricht im Grundschulbereich. Die Reliabilität liegt in einem hohen Bereich (Cronbachs $\alpha = .89$ bis $.93$, r_{tt} nach 6 Wochen = $.81$ bis $.94$ – je nach Klassenstufe. Die Validität wird u. a.

Inhalts valide Rechtschreibtests für verschiedene Klassenstufen

durch hohe Korrelationen mit dem Lehrerurteil über die Rechtschreibleistung ($r=.58$ bis $.79$) und die Leseleistung ($r=.59$ bis $.73$) eindrucksvoll belegt. Die Normierungsstichprobe setzt sich für jede Klassenstufe aus über 2500 Kindern aus allen deutschen Bundesländern zusammen, wobei die Schülerinnen und Schüler aus den einzelnen Bundesländern anteilmäßig so vertreten sind, wie es dem Anteil schulpflichtiger Kinder im Bundesgebiet entspricht. In einer Testrezension (Gasteiger-Klicpera und Sticker 2011, S. 77) erfahren die DERET 1–2+ und 3–4+ eine sehr gute Bewertung: „Zusammenfassend betrachtet handelt es sich um solide konstruierte Tests, die zur Erfassung der Rechtschreibfähigkeiten nicht nur im Querschnitt, sondern auch im Längsschnitt sehr zu empfehlen sind.“ Der DERET 5–6+ enthält neben einem Fließtextdiktat auch ein Diktat einzelner Sätze sowie ein Lückensatzdiktat. Die Reliabilität liegt ebenfalls im hohen Bereich. Der wichtigste Beleg für die Validität ist auch hier die Inhaltsvalidität: Der Test wurde unter Beachtung der Lehrpläne der Bundesländer für das Fach Deutsch bzw. für die Rechtschreibkompetenz von Fünft- bis Siebtklässlern entwickelt. Die Normierungsstichprobe wurde in 5 großen Bundesländern rekrutiert und umfasst insgesamt 12.552 Schülerinnen und Schüler unterschiedlicher Schulformen. Die Tests DERET 1–2+ und DERET 3–4+ gehören zu den Verfahren, die in der evidenz- und konsensbasierten Leitlinie „Diagnostik und Behandlung von Kindern und Jugendlichen mit Lese- und/oder Rechtschreibstörung (► https://www.bvl-legasthenie.de/images/static/pdfs/Leitlinien/LF_Leitlinie.pdf) zur Messung der Lese- und oder Rechtschreibleistungen empfohlen werden“.

DRT 5 mit einem Aufgabentyp und Fehleranalyse

Diagnostischer Rechtschreibtest (DRT) Eine andere Serie von Rechtschreibtests wurde ebenfalls für die 1. bis 5. Klasse konzipiert (DRT 1, DRT 2, DRT 3, DRT 4 und DRT 5). Die Tests für die 4. und die 5. Klasse wurden 2017 in aktualisierter 3. Auflage mit neuen Normen vorgelegt. Die anderen Tests erschienen zuletzt 2003 (Stand: Mai 2020). Exemplarisch gehen wir kurz auf den DRT 5 (Grund et al. 2017) ein. Auch dieser Test beansprucht, die Rechtschreibleistung zu messen. Er verwendet nur einen Aufgabentyp, nämlich einen Lückentext. Insgesamt 51 Wörter müssen nach Diktat in vorgegebene Lückensätze eingetragen werden. Die Reliabilität der beiden Parallelformen ist mit $\alpha=.93$ und einer Retest-Reliabilität von .87 bzw. .95 hoch. Die Validität wurde über die Korrelation mit einer längeren Schreibaufgabe erfolgreich geprüft ($r=.71$ bis $.95$). Der DRT 5 wurde an einer Stichprobe von 3492 Schülern aus 10 Bundesländern neu normiert. Er bietet spezifische Normen für Haupt- und Werkrealschulen, Mittelschulen, Realschulen, Gemeinschaftsschulen, Oberschulen, Realschulen plus, Regelschulen, regionale Schulen, Sekundarschulen sowie für rechtschreibschwache Schüler/-innen im Gymnasium. Eine Besonderheit ist die für die 3. Auflage überarbeitete differenzierte Fehleranalyse (auch die DERET-Tests bieten eine qualitative Fehleranalyse). Sie soll Probleme mit der Lautunterscheidung und Lautfolge, der Buchstabenverbindungen (st/sp, pf, qu), der Dopplung/Dehnung, der Morphemkonstanz in verschiedenen Wortformen, der Ableitung des ä von a und des Endbuchstabens durch Verlängern, beim Präfix ver-/vor- sowie bei der Groß- und Kleinschreibung erfassen.

Mathematiktests

DEMAT für bestimmte Klassenstufen entwickelt

Deutscher Mathematiktest (DEMAT) Eine Serie von Mathematiktests beginnt mit dem DEMAT 1+ für 1. Klassen und reicht mit dem DEMAT 6+ bis zur 6. Klasse; zusätzlich liegt mit dem DEMAT 9 ein Test für 9. Klassen vor. Den Tests liegen die Lehrpläne aus allen Bundesländern zugrunde. Sie sind damit inhalts valide. Die Eichstichproben setzen sich aus Schülerinnen und Schü-

lern vieler oder sogar aller Bundesländer zusammen. Wir stellen kurz den derzeit (Stand: Mai 2020) neuesten Test vor, den DEMAT 3+ (Roick et al. 2018). Der Test umfasst 3 Subtests mit insgesamt 31 Items: Arithmetik (mit Aufgaben zu Zahlenstrahlen, Additionen, Subtraktionen und Multiplikationen), Sachrechnen (Sachrechnungen und Längen umrechnen) und Geometrie (Spiegelzeichnungen, Formen legen und Längen schätzen). Die mit Cronbachs α geschätzte Reliabilität ist mit .83 für einen Leistungstest allenfalls moderat; offenbar sind die Testteile nicht hinreichend homogen (s. Angaben zu den Aufgaben in den 3 Subtests oben) für eine Reliabilitätsschätzung über die interne Konsistenz. Der relativ hohe Zusammenhang mit der Mathematiknote ($r = -.63$) und mit einer landesweiten Vergleichsarbeit im Fach Mathematik ($r = -.66$) spricht einerseits für die Konstruktvalidität des Tests, lässt andererseits aber erkennen, dass mit dem Test partiell etwas anderes als mit der Mathematiknote erfasst wird. Normen liegen für die letzten 6 Wochen des 3. Schuljahres und die ersten 6 Wochen des 4. Schuljahres vor. Die Normierungsstichprobe umfasst 6185 Grundschulkinder aus allen deutschen Bundesländern. Sämtliche DEMAT-Tests werden in der Leitlinie zur Rechenstörung (► Abschn. 7.2.2) zur Diagnostik der Rechenstörung (ab einschließlich Ende 1. Klasse) empfohlen.

7.4.2 Tests zur Evaluierung des Bildungssystems

7.4.2.1 Programme for International Student Assessment (PISA)

PISA ist in aller Munde. Einprägsam ist auch der viel gebrauchte Begriff „PISA-Schock“ als Reaktion auf das unerwartet schlechte Abschneiden Deutschlands in der Studie 2000. Wir finden die PISA-Studie immerhin so wichtig, dass wir sie in ► Abschn. 1.6 bei den Meilensteinen in der Geschichte der psychologischen Diagnostik aufgeführt haben. PISA wird kontrovers diskutiert; die Meinungen über PISA gehen weit auseinander. Ein neutrales Statement über PISA lautet: „Ein Projekt, das die Welt der schulischen Bildung grundlegend verändert hat“ (Funke und Spinath 2014, S. 137). PISA ist wohl das bekannteste, aber nicht das einzige Projekt dieser Art.

Was ist PISA genau? Wir berichten zunächst ein paar Fakten über das Projekt (OECD 2018b) und gehen dann auf die Diagnostik ein, die sozusagen der Kern von PISA darstellt. PISA steht für „Programme for International Student Assessment“. Organisatorisch zuständig ist die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD). Diese verfolgt das Ziel, „eine Politik zu befördern, die das Leben der Menschen weltweit in wirtschaftlicher und sozialer Hinsicht verbessert“ (OECD 2018a). Ergebnisse der PISA-Studien dienen hauptsächlich der Politikberatung. Seit 2000 werden im 3-jährigen Turnus Schülerinnen und Schüler am Ende der Pflichtschulzeit (Alter: 15 Jahre) in vielen Ländern getestet. 2018 nahmen insgesamt 80 Staaten teil, darunter auch Deutschland. Getestet wurde über eine halbe Million Personen. Gemessen werden Lesekompetenz, mathematische Kompetenz und naturwissenschaftliche Grundbildung. Die Schwerpunkte wechseln jeweils. So lag 2015 der Schwerpunkt auf dem Bereich Naturwissenschaften; für diesen Bereich wurden also besonders viele Items eingesetzt. Ergänzend wird die Kompetenz in einem innovativen Bereich untersucht, 2015 war das kollaboratives Problemlösen. Zusätzlich werden weitere Daten mit Fragebögen erhoben. Kernstück ist jedoch die Leistung in den 3 Bereichen.

Weltweite Messung der Schulleistung in 3 Kernbereichen

Auch wenn für viele das Abschneiden im internationalen Vergleich wichtig ist, so liefern die PISA-Studien noch weitere Erkenntnisse wie auf die folgende Frage: Mit welchen Faktoren hängt die gemessene Schulleistung zu-

sammen? Beispiele für solche Faktoren sind Geschlecht, Bildung der Eltern, Bildungsausgaben in den Ländern.

Der Test Da sich der Test von Jahr zu Jahr leicht verändert, beziehen wir uns hier auf den von 2015 (OECD 2016). Soweit dies in den beteiligten Ländern möglich war, wurde der Test computerbasiert durchgeführt. Wir beschränken uns hier auf den Bereich Naturwissenschaften und beginnen mit der Frage, wie der Messgegenstand definiert wurde.

Naturwissenschaftliche
Grundbildung definiert

7

184 Items – nach verschiedenen
Kriterien klassifiziert

Definition naturwissenschaftlicher Kompetenzen (OECD 2016, S. 58)

„Gemäß der PISA-Definition ist eine Person, die über eine naturwissenschaftliche Grundbildung verfügt, dazu in der Lage und bereit, sich argumentativ mit Naturwissenschaften und Technologie auseinanderzusetzen. Dies erfordert die Kompetenzen, um:

- **Phänomene naturwissenschaftlich zu erklären** – Erklärungen für eine Reihe von natürlichen und technologischen Phänomenen erkennen, anbieten und bewerten,
- **naturwissenschaftliche Forschung zu bewerten und naturwissenschaftliche Untersuchungen zu planen** – naturwissenschaftliche Untersuchungen beschreiben und bewerten und Wege vorschlagen, um Fragen naturwissenschaftlich anzugehen,
- **Daten und Evidenz naturwissenschaftlich zu interpretieren** – Daten, Behauptungen und Argumente in verschiedenen Darstellungen analysieren und bewerten und angemessene naturwissenschaftliche Schlüsse ziehen.“

Die Items, auch die aus dem Bereich Mathematik und Lesen, können nach Inhalten, Kontexten und Prozessen klassifiziert werden. Wir betrachten die Systematik für die insgesamt 184 Items für den Bereich Naturwissenschaften nach Angabe in Schiepe-Tiska et al. (2016). Man beachte, dass die Items mehrfach klassifiziert werden und daher in allen 3 Bereichen die Summe 184 beträgt:

- *Inhalte:*
 - Physikalische Systeme (61 Items)
 - Lebende Systeme (74 Items)
 - Erd- und Weltraumsysteme (49 Items)
- *Kontexte* (die Menschen jeweils persönlich, die Gesellschaft oder das Leben weltweit betreffen können):
 - Gesundheit und Krankheit (22 Items)
 - Natürliche Ressourcen (46 Items)
 - Umwelt (34 Items)
 - Risiken/Gefahren (20 Items)
 - Grenzen von Naturwissenschaft und Technik (64 Items)
- *(Denk- oder Lösungs-)Prozesse:*
 - Phänomene naturwissenschaftlich erklären (89 Items)
 - Naturwissenschaftliche Forschung bewerten und Untersuchungen planen (39 Items)
 - Daten und Evidenz naturwissenschaftlich interpretieren (56 Items)

Die Items wurden zum Teil neu konstruiert, zum Teil stammen sie aus der Erhebung 2012. Sie sind unterschiedlich schwer; das kognitive Anforderungsni-veau wird als „niedrig“, „mittel“ oder „hoch“ eingestuft. Bezuglich des Ant-wortformats ist zu sagen, dass neben Multiple-Choice-Aufgaben auch offene Antworten vorkamen.

Zur Veranschaulichung ist unten ein Item ausgeführt (unter der angegebe-nen Quelle finden sich weitere Itembeispiele). Das aufgeführte Item kann nä-her klassifiziert werden (OECD 2016, S. 478):

- Aufgabentyp: Einfache Multiple-Choice-Aufgabe
- Kompetenz: Phänomene naturwissenschaftlich erklären
- Wissensbereich – System: Konzeptuelles Wissen – lebende Systeme
- Kontext: Global – Umweltqualität
- Schwierigkeitsgrad: 501 (Stufe 3)

Itembeispiel

Beziehe dich auf „Vogelzug“ auf der rech-ten Seite. Klicke eine Antwort an, um die Frage zu beantworten

Die meisten Zugvögel versammeln sich in einem Gebiet und ziehen nicht einzeln, sondern in großen Gruppen. Dieses Ver-halten ist eine Folge der Evolution. Welche der folgenden Aussagen ist die beste natur-wissenschaftliche Erklärung für die Evo-lution dieses Verhaltens bei den meisten Zugvögeln?

- Vögel, die einzeln oder in kleinen Grup-pen zogen, haben mit geringerer Wahr-scheinlichkeit überlebt und Nachkom-men bekommen
- Vögel, die einzeln oder in kleinen Grup-pen zogen, haben mit höherer Wahr-scheinlichkeit passendes Futter gefun-den
- Das Fliegen in großen Gruppen ermög-lichte es anderen Vogelarten, sich dem Zug anzuschließen
- Durch das Fliegen in großen Gruppen hatte jeder einzelne Vogel bessere Chan-cen, einen Nistplatz zu finden

Der Vogelzug ist eine jahreszeitenbedingte große Wanderung der Vögel zu und von ihren Brutstätten. Jedes Jahr zählen Frei-willige die Zugvögel an bestimmten Or-ten. Wissenschaftler fangen einige der Vö-gel ein und kennzeichnen ihre Beine mit ei-ner Kombination aus farbigen Ringen und Fähnchen. Die Wissenschaftler nutzen die Sichtungen gekennzeichneter Vögel zusam-men mit den Zählungen der Freiwilligen, um die Zugrouten von Vögeln zu bestim-men

[Foto mit beringtem Vogel]

Quelle: ► <https://www.oecd.org/pisa/PISA2015Questions/platform/index.html?user=&-domain=SCI&unit=S656-BirdMigration&lang=deu-DEU> (© OECD)

Ergebnisdarstellung Die Ergebnisse werden in einer Metrik mit dem Mittel-wert 500 und einer Standardabweichung von 100 Punkten angegeben (vgl. auch ► Abschn. 2.6.4). Dieser Metrik wurde anhand der Daten der ers-ten PISA-Erhebung erstellt und nachfolgend verwendet. 2015 lag der Mittel-wert über alle Länder im Bereich Naturwissenschaften bei 493 Punkten, also leicht unter dem Mittelwert der ersten PISA-Erhebung. Die Punkte werden nach einem feststehenden Schema in 7 Kompetenzstufen von 6 bis 1 transfor-miert, wobei 1 noch einmal in 1a und 1b (die niedrigste Stufe) unterteilt wird (s. □ Tab. 7.5 für Kurzbeschreibungen).

7 Kompetenzstufen

■ Tab. 7.5 Kompetenzstufen im PISA-Test 2015 für den Bereich Naturwissenschaften

Stufe (Punkte)	Erläuterung
6 (mindestens 708 Punkte)	Auf Stufe 6 können Schüler auf miteinander verknüpfte wissenschaftliche Ideen und Konzepte aus den Bereichen Physik, Lebenswissenschaften, Geologie und Astronomie zurückgreifen und inhaltliches, prozedurales und epistemisches Wissen nutzen, um Erklärungshypothesen neuer naturwissenschaftlicher Phänomene, Ereignisse und Prozesse anzubieten oder Vorhersagen zu treffen. Bei der Interpretation von Daten und Befunden sind sie in der Lage, zwischen relevanten und irrelevanten Informationen zu unterscheiden, und sie können auf Wissen zurückgreifen, das außerhalb des normalen Lehrplans erworben wurde. Sie können zwischen Argumenten unterscheiden, die auf naturwissenschaftlicher Evidenz und Theorie beruhen, und denjenigen, die auf anderen Erwägungen basieren. Schüler, deren Leistungen auf Stufe 6 liegen, können konkurrierende Gestaltungen komplexer Versuche, Feldstudien oder Simulationen evaluieren und ihre Entscheidungen begründen.
etc.	
1b (mindestens 261 Punkte)	Auf Stufe 1b können Schüler grundlegendes bzw. aus dem Alltag bekanntes naturwissenschaftliches Wissen einsetzen, um Aspekte vertrauter oder einfacher Phänomene zu erkennen. Sie sind in der Lage, einfache Datenstrukturen zu identifizieren, grundlegende naturwissenschaftliche Begriffe zu erkennen und expliziten Anweisungen zu folgen, um ein einfaches naturwissenschaftliches Verfahren durchzuführen.

Quelle: Auszug aus OECD (2016, Abb. I.2.6, S. 67 f., © OECD)

■ Tab. 7.6 Ausgewählte Ergebnisse aus der PISA-Studie 2015 zum Bereich Naturwissenschaften

Land	Punkte	% Kompetenzstufe <2
Singapur	556	10
Japan	538	10
Estland	534	09
...		
Deutschland	509	17
...		
Kosovo	378	68
Algerien	376	71
Dominikanische Republik	332	86

Quelle: Auszug aus OECD (2016, S. 225, © OECD)

Deutschland im oberen Mittelfeld

Ausgewählte Ergebnisse In ■ Tab. 7.6 ist das Ergebnis im Bereich Naturwissenschaften für Deutschland in der PISA-Studie 2015 im Vergleich zu den 3 besten und den 3 schlechtesten Länderergebnissen im PISA-Test aufgeführt. Deutschland liegt mit 509 Punkten leicht über dem internationalen Durchschnitt (das gilt übrigens auch für Lesekompetenz mit 509 und für Mathe-

matik mit 506 Punkten). 17 % der Schülerinnen und Schüler fielen in die sehr niedrigen Kompetenzstufen 1a und 1b. Man sieht, dass die Spannbreite international sehr groß ist. Das Schlusslicht ist die Dominikanische Republik, die mit 332 Punkten 168 Punkte unter dem theoretischen Mittelwert von 500 lag, was 1,68 Standardabweichungen entspricht. Außerdem erreichten dort 86 % der Schülerinnen und Schüler nicht einmal Kompetenzstufe 2. Singapur, Japan und Estland bilden die Spitzengruppe (für weitere Ausführungen hierzu s. auch ► Abschn. 2.6.4).

Deutsche Politikerinnen und Politiker wollen vermutlich wissen, warum Deutschland „nur“ im oberen Mittelfeld liegt. Der Abstand zu Singapur beträgt umgerechnet immerhin etwa 7 IQ-Punkte. Sie werden dazu zahlreiche Fakten finden, die aber nicht selbsterklärend sind. Beispielsweise ist in Deutschland die Varianz zwischen den Schulen deutlich größer als im OECD-Durchschnitt (OECD 2016, □ Abb. 1.6.11, S. 246). Die Varianz zwischen den Schulen wird in Deutschland in hohem Maße durch Unterschiede im sozioökonomischen Profil der Schülerschaft und Schulen „erklärt“ (OECD 2016, □ Abb. 1.6.13, S. 248). „Erklärt“ steht in Anführungszeichen, um dem Missverständnis vorzubeugen, dass es sich um eine kausale Erklärung handelt; vielmehr geht es hier um eine statistische Varianzaufklärung, also nur um eine korrelative Beziehung zwischen 2 Variablen. Der Unterschied zwischen den Schulen hängt zudem leicht damit zusammen, dass in Deutschland in den benachteiligten Schulen nach Angaben der Schulleitungen zu wenig Personal zur Verfügung stehe (OECD 2016, □ Abb. 1.6.14, S. 251). Sorgfältig und fachlich kompetent interpretiert lassen sich aus den Ergebnissen Maßnahmen zur Optimierung der Schulpädagogik ableiten.

Erklärungsansätze

7.4.2.2 Andere Bildungstests

Neben PISA sind im deutschen Bildungssystem auch andere systematische Testuntersuchungen zu schulrelevanten Kompetenzen von Bedeutung: Die *Trends in International Mathematics and Science Study (TIMSS)* und die *Internationale Grundschul-Lese-Untersuchung (IGLU)*, international als *Progress in International Reading Literacy Study (PIRLS)* bezeichnet, sind wohl die bekanntesten. Mit PISA gemeinsam haben sie den längs- und querschnittlichen Forschungsansatz. Sie werden also in einem bestimmten Turnus fortgeführt und sie erlauben für jeden Messzeitpunkt einen Vergleich mit anderen Ländern. Bei der Darstellung der Ergebnisse findet die gleiche Metrik (Mittelwert = 500, Standardabweichung = 100) wie bei PISA Verwendung. Auch in Bezug auf den Messgegenstand gibt es Gemeinsamkeiten. In □ Tab. 7.7 sind einige Informationen über die Studien aufgeführt.

Weitere längs- und querschnittliche Forschungsansätze

Weiterführende Literatur

Eine kurze, aber informative Abhandlung über das PISA-Projekt findet sich bei Funke und Spina (2014).

■ Tab. 7.7 Informationen zu den Studien TIMMS und IGLU (PIRLS)

Merkmale	Bildungsstudie	
	TIMMS	IGLU (PIRLS)
Messgegenstand	Mathematische und naturwissenschaftliche Kompetenzen	Lesekompetenz
Zielgruppe	In Deutschland: Schülerinnen und Schüler am Ende der 4. Jahrgangsstufe	Schülerinnen und Schüler am Ende der 4. Jahrgangsstufe
Turnus und Beginn (Deutschland)	Alle 4 Jahre (seit 2007)	Alle 5 Jahre (seit 2001)
Anzahl der Länder, die zuletzt teilgenommen haben (Stand: 2015)	49 Länder bei Mathematik und 47 bei Naturwissenschaften (Klasse 4)	50 Länder
Ergebnis (Punkte) für Deutschland	522 (Mathematik) 528 (Naturwissenschaften)	537
Top-Länder (Punkte für a Mathematik und b Naturwissenschaften)	a) Singapur (618), Hongkong (615), Korea (608) b) Singapur (590), Korea (589), Japan (569)	Russland (581), Singapur (576), Hongkong (569)
Leitung	International Association for the Evaluation of Educational Achievement (IEA)	International Association for the Evaluation of Educational Achievement (IEA)
Webseite	► https://timssandpirls.bc.edu/timss-landing.html	► https://timssandpirls.bc.edu/timss-landing.html
Literaturhinweis	Wendt et al. (2016)	Hußmann et al. (2017)

7.5 Zusammenfassung

Die Diagnostik in der Pädagogischen Psychologie ist ein weites Feld. Sie befasst sich mit Individuen – von Kindern im Vorschulalter bis zu Jugendlichen, die einen Studienplatz suchen – und auch mit der Evaluierung der Institution Schule.

Der individuumbezogene Ansatz richtet sich auch auf die Navigation im Schulsystem. Vor dem Schuleintritt wird bei Bedarf geprüft, ob ein Kind den Anforderungen der Schule schon gerecht wird und wo eventuell noch kompensierbare Defizite vorliegen. Geeignete Tests wurden vorgestellt. Wir haben uns hier wegen der großen Zahl von zu erwarteten Fehlurteilen gegen ein allgemeines Screening zur Feststellung der „Schulreife“ ausgesprochen. Eine anlassbezogene psychologische Diagnostik ist vorzuziehen. Gegen Screenings zur Entdeckung sowohl von Schwächen als auch von hoher Begabung spricht nicht nur die begrenzte Validität der Verfahren, sondern auch das Phänomen der Regression zur Mitte. Es besagt, dass (niedrige oder hohe) Extremwerte auch messfehlerbedingt sehr niedrig bzw. sehr hoch sein können und daher oft nicht replizierbar sind.

In manchen Fällen wird schon vor Schuleintritt festgestellt, dass ein Kind besonderen Förderbedarf hat, der dann entweder in einer speziellen Förderschule oder auch in der Regelschule angeboten wird. Manchmal wird die Frage, ob sonderpädagogischer Förderbedarf besteht, auch erst in der Grundschule aufgeworfen. Das diagnostische Vorgehen zur Feststellung von sonderpädagogischem Förderbedarf wird von Landesgesetzen und von den Schulbehörden stark reglementiert. Bei der Entscheidung für eine bestimmte Schulform kommt dem Elternwunsch eine große Bedeutung zu. Wenn Eltern-

wunsch und Einschätzung der Schulbehörde einander gegenüberstehen, kann psychologische Diagnostik manchmal zu einer Lösung beitragen. Grundsätzlich kann während der ganzen Schullaufbahn auch die Frage auftreten, ob ein Wechsel innerhalb der Regelschule zu empfehlen ist. Der Wechsel kann von einer Schulform (z. B. Gymnasium) auf eine andere (z. B. Realschule) oder auch innerhalb einer Schulform passieren. Speziell bei Kindern, die vielleicht hochbegabt sind, kann pädagogisch-psychologische Diagnostik sinnvoll sein, um die Chancen und Risiken besser abzuschätzen, die mit dem Über-springen einer Klasse einhergehen können. Beim Übertritt von der Schule auf eine Hochschule besteht oftmals Beratungsbedarf verbunden mit Diagnostik. Insbesondere die Frage, ob man für ein bestimmtes Studienfach geeignet ist, stellt sich häufig. Ein Grund dafür ist die schwer überschaubare Hochschullandschaft mit Tausenden von Studienangeboten. Zur Feststellung der Passung haben wir 2 Ansätze vorgestellt, und zwar die Auswahl von Studierenden durch die Hochschulen und die Eignungsfeststellung durch die Studieninteressierten selbst mithilfe von Online-Self-Assessments. Online-Self-Assessments haben lediglich eine Beratungsfunktion. Die Hochschule erfährt das Ergebnis nicht, hofft aber auf eine gewisse Selbstselektion zur Gewährleistung der Passung.

In der Schulzeit (teilweise auch schon im Vorschulalter) wird bei einem Kind manchmal der Verdacht geäußert, dass „Lernschwierigkeiten“ vorliegen. Diese können genereller Art oder auf eine Teilleistungsschwäche zurückzuführen sein. Die Ursache für allgemeine Lernprobleme können im Individuum (kognitive, motivationale, krankheitsbedingte Probleme), aber auch in seiner Umwelt (z. B. mangelnde Förderung, Mobbing) zu suchen sein. Lernschwierigkeiten werden manchmal über eine ICD-Diagnostik als psychische Störung pathologisiert. Bei der Diagnostik einer Lese- und Rechtschreib- sowie einer Rechenstörung kann auf fundierte Leitlinien zurückgegriffen werden. Exemplarisch wurde die Leitlinie zur Rechenstörung mit ihren Empfehlungen zum diagnostischen Vorgehen ausführlich vorgestellt.

Als eine besondere Fragestellung wurde die Diagnostik von Hochbegabung behandelt. Hochbegabung wird als herausragend hohe Intelligenz ($IQ \geq 130$) definiert und von Spezialbegabungen (wie etwa musikalischer Begabung) abgegrenzt. Die Diagnostik erfolgt mithilfe eines oder besser von 2 gut normierten Tests zur Allgemeinen Intelligenz. Für Beratungszwecke ist oft eine umfassende Diagnostik angebracht, die auch nichtkognitive Faktoren einbezieht. Hochbegabung geht nicht immer mit herausragenden Schulleistungen einher, und nicht immer ist eine herausragende Schulleistung mit Hochbegabung zu erklären. Hilfreich sind hier die Konzepte Under- und Overachiever. Sie helfen auch zu verstehen, warum Lehrerurteile über die Begabung von Schülerinnen und Schülern oftmals nicht mit deren Intelligenztestergebnissen übereinstimmen.

Im Bildungsbereich werden Schulleistungstests eingesetzt, um den Lernstand in einem Fach nach repräsentativen Normen für eine Klassenstufe zu beurteilen. Entsprechende Tests zu den schulischen Fertigkeiten Lesen, Schreiben und Rechnen benötigt man auch zur Diagnostik von Lese- und Rechtschreib- sowie Rechenstörungen. Etablierte Serien von Schulleistungstests wurden kurz vorgestellt. In länderübergreifenden Studien werden regelmäßig Tests eingesetzt, um schulische Kompetenzen zu messen. Ausführlicher vorgestellt wurde das PISA-Projekt und das damit verbundene diagnostische Vorgehen. Exemplarisch haben wir erläutert, wie naturwissenschaftliche Kompetenzen definiert und gemessen werden. Andere Bildungsstudien wie TIMMS und IGLU wurden nur kurz erwähnt. Alle zusammen dienen dazu, unser Bildungssystem sowohl im längsschnittlichen Vergleich als auch im Vergleich zu den Bildungssystemen anderer Länder zu evaluieren.

Weiterführende Literatur

Fallbeispiele zu schulpsychologischen Fragestellungen finden sich in dem von Kubinger und Ortner (2010) herausgegebenen Buch.

Einen immer noch lesenswerten Überblick über angewandte Fragen der pädagogisch-psychologischen Diagnostik geben Langfeldt und Tent (1999). Zu einzelnen Themen, die für bestimmte diagnostische Fragen relevant sein können finden sich in dem von Rost et al. 2018 herausgegebenen *Handwörterbuch Pädagogische Psychologie*, das in der 5. Auflage fast 1000 Seiten umfasst, kompetente Informationen.

7 ? Übungsfragen

— Abschn. 7.1:

- Welche 3 im Vorschulalter messbaren Fähigkeiten sind für den späteren Schulerfolg besonders relevant?
- Welches Problem besteht bei einem Screening zur Entdeckung mangelnder Schulreife, wenn der Test eine prognostische Validität von $r = .51$ hat und der Anteil nicht schulreifer Kinder mit 2 % angenommen wird?
- Was versteht man unter „Regression zur Mitte“ und welche Bedeutung hat sie für ein Screening von Menschen, die einer seltenen Gruppe (z. B. Hochbegabte, Minderbegabte) angehören?
- Nennen Sie wenigstens 3 Sonderschularten in Deutschland!
- Auf welche allgemeinen Faktoren werden Lernschwierigkeiten in der Schule zurückgeführt? Nennen Sie auch Beispiele!
- Studieninteressierte können an einem Online-Self-Assessment teilnehmen. Was ist das und wozu soll es dienen?

— Abschn. 7.2:

- Nennen Sie 2 umschriebene Entwicklungsstörungen schulischer Fertigkeiten!
- Welches sind nach der Leitlinie (AWMF 2018) die Hauptkriterien für die Diagnose einer Rechenstörung?

— Abschn. 7.3:

- Wie wird Hochbegabung definiert, und warum ist eine Definition über mehrere Begabungsmerkmale problematisch?
- Wie gut können Lehrer/-innen einer Studie von Wild zufolge die Begabung ihrer Schüler/-innen beurteilen und wie gut erkennen sie Hochbegabung?
- Was versteht man im Rahmen von Hochbegabung unter Under- und Overachiever?

— Abschn. 7.4:

- Für welche Fragestellungen kann man in Schulleistungstests verwenden?
- Was leisten die PISA-Tests?

Literatur

-
- American Psychiatric Association. (2019). What is specific learning disorder? ► <https://www.psychiatry.org/patients-families/specific-learning-disorder/what-is-specific-learning-disorder>. Zugriffen: 26. März 2020.
- Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF). (2015). S3-Leitlinie: Lese- und/oder Rechtschreibstörung bei Kindern und Jugendlichen, Diagnostik und Behandlung. Registernummer 028–044. ► <https://www.awmf.org/leitlinien/detail/ll/028-044.html>. Zugriffen: 26. Mai 2015.
- Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF). (2018). S3-Leitlinie: Diagnostik und Behandlung der Rechenstörung. Registernummer 028–046. Stand: 25.02.2018. ► https://www.awmf.org/uploads/tx_szleitlinien/028-046l_S3_Rechenst%C3%B6rung-2018-03_1.pdf. Zugriffen: 23. Mai 2020.
- Beauftragte der Bundesregierung für die Belange von Menschen mit Behinderungen (Hrsg.). (2017). Die UN-Behindertenrechtskonvention: Übereinkommen der Vereinten Nationen

- über die Rechte von Menschen mit Behinderung (amtliche, gemeinsame Übersetzung von Deutschland, Österreich, Schweiz und Lichtenstein). ► https://www.behindertenbeauftragter.de/DE/Koordinierungsstelle/UNKonvention/UNKonvention_node.html. Zugegriffen: 27. März 2020.
- Bundesverfassungsgericht. (2017). Leitsätze zum Urteil des Ersten Senats vom 19. Dezember 2017. ► https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2017/12/lsl20171219_lbv000314.html. Zugegriffen: 22. Mai 2020.
- Castello, A., Grünke, M., & Beelmann, A. (2004). Lernschwächen bei Entwicklungsretardierungen. In G. W. Lauth, M. Grünke, & J. C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (S. 78–89). Göttingen: Hogrefe.
- Deutsche Gesellschaft für Psychologie. (2020). Mitteilungsdetail: Neues Studierendenauswahlverfahren für Psychologie in Baden-Württemberg. Mitteilung vom 08.01.2020. ► https://www.dgps.de/index.php?id=143&tx_ttnews%5D=1951&cHash=a08c001c1963b76374648945006ab909. Zugegriffen: 22. Mai 2020.
- Deutsches Institut für Medizinische Dokumentation und Information (DIMDI). (2019). Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision – German Modification. ► <https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-who/kode-suche/htmlamlt2019/chapter-v.htm>. Zugegriffen: 22. Mai 2020.
- Döpfner, M., Dietmair, I., Mersmann, H., Simon, K. & Trost-Brinkhues, G. (2005). *S-ENS: Screening des Entwicklungsstandes bei Einschulungsuntersuchungen*. Göttingen: Hogrefe.
- Endlich, D., Berger, N., Küspert, P., Lenhard, W., Marx, P., Weber, J., et al. (2017). *WVT: Würzburger Vorschultest: Erfassung schriftsprachlicher und mathematischer (Vorläufer-) Fertigkeiten und sprachlicher Kompetenzen im letzten Kindergartenjahr*. Göttingen: Hogrefe.
- Esser, G., Hasselhorn, M., & Schneider, W. (Hrsg.). (2015). *Diagnostik im Vorschulalter*. Göttingen: Hogrefe.
- Fischbach, A., Schuchardt, K., Brandenburg, J., Kleczewski, J., Balke-Melcher, C., Schmidt, C., et al. (2013). Prävalenz von Lernschwächen und Lernstörungen: Zur Bedeutung der Diagnosekriterien. *Lernen und Lernstörungen* 2, 65–76.
- Funke, J., & Spinath, B. (2014). Die PISA-Studien: Ein transdisziplinäres Projekt verändert die Bildungswelt. In G. Jüttemann (Hrsg.), *Entwicklungen der Menschheit. Humanwissenschaften in der Perspektive der Integration* (S. 137–144). Lengerich: Pabst.
- Gasteiger-Klicpera, B., & Sticker, E. (2011). TBS-TK Rezension: „Deutscher Rechtschreibtest für das erste und zweite/dritte und vierte Schuljahr, DERET 1-2+3-4“. *Psychologische Rundschau* 63, 75–77.
- Götz, L., Lingel, K., & Schneider, W. (2013). *DEMAT 6+: Deutscher Mathematiktest für sechste Klassen*. Göttingen: Hogrefe.
- Grund, M., Leonhart, R., & Naumann, C. L. (2017). *DRT 5: Diagnostischer Rechtschreibtest für 5. Klassen* (3. Aufl.). Göttingen: Hogrefe.
- Grünke, M. (2004). Lernbehinderung. In G. W. Lauth, M. Grünke, & J. C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (S. 65–77). Göttingen: Hogrefe.
- Guttschick, K. (2015). Konstruktion und Validierung eines Leistungsmotivationstests für Online-Self-Assessments. Dissertation, Philipps-Universität, Marburg.
- Guttschick, K., Hasenberg, S., & Schmidt-Atzert, L. (2019). Status Quo in der deutschen OSA-Landschaft. In L. Schmidt-Atzert, M. Schütz, & G. Stemmler (Hrsg.), *Online-Self-Assessments an Hochschulen* (S. 25–38). Lengerich: Pabst Science Publishers.
- Hanses, P., & Rost, D. H. (1998). Das „Drama“ der hochbegabten Underachiever – „Gewöhnliche“ oder „außergewöhnliche“ Underachiever? *Zeitschrift für Pädagogische Psychologie* 21, 53–71.
- Hasenberg, S., & Schmidt-Atzert, L. (2014). Zur Vorhersage der Studienzufriedenheit durch internetbasierte Self-Assessments. *Empirische Pädagogik* 28, 19–35.
- Hasenberg, S., Guttschick, K., Schmidt-Atzert, L., Stemmler, G., Kohlhaas, G., Schütz, M., et al. (2014). Unterstützung beim Übergang von der Schule zur Hochschule durch präzise Studieninformationen und Online-Self-Assessments. *Zeitschrift für Hochschulentwicklung* 9, 116–129.
- Hasselhorn, M., Heinze, A., Schneider, W., & Trautwein, U. (Hrsg.). (2013). *Diagnostik mathematischer Kompetenzen*. Göttingen: Hogrefe.
- Heinecke-Müller, M. (2019). Lernschwierigkeiten. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (19. Aufl.). Göttingen: Hogrefe. ► <https://portal.hogrefe.com/dorsch/lernschwierigkeiten/>. Zugegriffen: 23. Mai 2020.
- Hell, B., Trapmann, S., & Schuler, H. (2007a). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik* 21, 251–270.
- Hell, B., Trapmann, S., Weigand, S., & Schuler, H. (2007b). Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung – eine Metaanalyse. *Psychologische Rundschau* 58, 93–102.
- Hessisches Kultusministerium. (2020). Teilhabe ermöglichen: Sonderpädagogische Förderung und inklusiver Unterricht in Hessen. ► <https://kultusministerium.hessen.de/foerderangebote/>

- sonderpaedagogische-foerderung/sonderpaedagogische-foerderung-und-inklusiver. Zugegriffen: 22. Mai 2020.
- Heublein, U., Hutzsch, C., Schreiber, J., Sommer, D., & Besuch, G. (2009). *Ursachen des Studienabbruchs in Bachelor- und in herkömmlichen Studiengängen: Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/08*. Hannover: Hochschul-Informations-System.
- Holling, H., & Kanning, U. P. (1999). *Hochbegabung: Forschungsergebnisse und Fördermöglichkeiten*. Göttingen: Hogrefe.
- Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., et al. (Hrsg.). (2017). *IGLU 2016: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.
- Jacobs, C., & Petermann, F. (2007). *Rechenstörungen* (Serie Leitfaden Kinder- und Jugendpsychotherapie, Bd. 9). Göttingen: Hogrefe.
- Jacobs, C., & Petermann, F. (2012). *Diagnostik von Rechenstörungen* (2. Aufl.). Göttingen: Hogrefe.
- Jacobs, C., Petermann, F., & Tischler, L. (2013). Rechenstörung. In F. Petermann (Hrsg.), *Lehrbuch der Klinischen Kinderpsychologie* (7. Aufl., S. 181–206). Göttingen: Hogrefe.
- Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., et al. (2006). *BIS-HB: Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik*. Göttingen: Hogrefe.
- Janke, S., & Dickhäuser, O. (2018). Zur prognostischen Güte von Zulassungskriterien im Psychologiestudium für Studienerfolgsindikatoren. *Psychologische Rundschau* 69, 160–168.
- Kammermeyer. (2010). Schulreife und Schulfähigkeit. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 718–728). Weinheim: Beltz.
- Kern, A. (1963). *Sitzenbleiberei und Schulreife: ein psychologisch-pädagogischer Beitrag zu einer inneren Reform der Grundschule* (4. Aufl.). Freiburg: Herder.
- Krajewski, K., Küspert, P., & Schneider, W. (2002). *DEMAT 1+: Deutscher Mathematiktest für erste Klassen*. Göttingen: Beltz.
- Kubinger, K. D., & Ortner, T. M. (2010). *Psychologische Diagnostik in Fallbeispielen*. Göttingen: Hogrefe.
- Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland. (1994). Empfehlungen zur sonderpädagogischen Förderung in den Schulen der Bundesrepublik Deutschland: Beschluss der Kultusministerkonferenz vom 06.05.1994. ► https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1994/1994_05_06-Empfehlung-sonderpaed-Foerderung.pdf. Zugegriffen: 22. Mai 2020.
- Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland. (2020). Datensammlung Sonderpädagogische Förderung in Förderschulen 2017/2018 (Korrekturfassung vom 21.02.2020 – geänderte Daten im Blatt Quoten für die allgemeinen Schulen). ► https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Aus_Sopae_2017.pdf. Zugegriffen: 22. Mai 2020.
- Langfeldt, H.-P., & Tent, L. (1999). *Pädagogisch-psychologische Diagnostik. Band 2: Anwendungsbereiche und Praxisfelder*. Göttingen: Hogrefe.
- Lauth, G. W. (2004). Allgemeine Lernschwäche (Kombinierte Schulleistungsstörung nach ICD-10). In G. W. Lauth, M. Grünke, & J. C. Brunstein (Hrsg.), *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (S. 55–64). Göttingen: Hogrefe.
- Lenhard, W., & Schneider, W. (Hrsg.). (2009). *Diagnostik und Förderung des Leseverständnisses*. Göttingen: Hogrefe.
- Lindberg, S., Hasselhorn, M., & Lonnemann, J. (2018). Förderrelevante Diagnostik bei Lernstörungen. *Lernen und Lernstörungen* 7, 197–201.
- Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted* 29, 451–484.
- Martinez Méndez, R., Schneider, W., & Hasselhorn, M. (2015). *DERET 5-6+: Deutscher Rechtschreibtest für fünfte und sechste Klassen*. Göttingen: Hogrefe.
- Moll, K., Kunze, S., Neuhoff, N., Bruder, J., & Schulte-Körne, G. (2014). Specific learning disorder: Prevalence and gender differences. *PLoS ONE* 9, e103537.
- Oden, M. H. (1968). The fulfillment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs* 77, 3–93.
- Organisation for Economic Co-operation and Development (OECD). (2016). *PISA 2015 Ergebnisse (Band I): Exzellenz und Chancengerechtigkeit in der Bildung*. Gütersloh: Bertelsmann.
- Organisation for Economic Co-operation and Development (OECD). (2018a). Die OECD. Unser Ziel: Bessere Politik für ein besseres Leben. ► <https://www.oecd.org/berlin/dieoecd/>. Zugegriffen: 23. Mai 2020.
- Organisation for Economic Co-operation and Development (OECD). (2018b). What is PISA? ► <https://www.oecd.org/pisa/>. Zugegriffen: 23. Mai 2020.
- Orthmann Bless, D. (2010). Lernschwierigkeiten. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 471–479). Weinheim: Beltz.

- Parmar, R. S. (2004). Lernbehinderung – Ein Forschungsüberblick zur Situation in den USA am Beispiel von Lesen und Mathematik. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete* 1, 43–56.
- Preckel, F., & Vock, M. (2013). *Hochbegabung: Ein Lehrbuch zu Grundlagen, Diagnostik und Fördermöglichkeiten*. Göttingen: Hogrefe.
- Preckel, F., Schneider, W., & Holling, H. (Hrsg.). (2010). *Diagnostik von Hochbegabung*. Göttingen: Hogrefe.
- Reiß, S. (2019). Steuerungsfunktion von Online-Self-Assessments. In L. Schmidt-Atzert, M. Schütz, & G. Stemmler (Hrsg.), *Online-Self-Assessments an Hochschulen* (S. 53–63). Lengerich: Pabst Science Publishers.
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin* 138, 353–387.
- Richter, T., Souvignier, E., Hertel, S., Heyder, A., & Kunina-Habenicht, O. (2019). Positionsstatement zur Lage der Pädagogischen Psychologie in Forschung und Lehre. *Psychologische Rundschau* 70, 109–118.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin* 130, 261–288.
- Roick, T., Göltz, D., & Hasselhorn, M. (2018). *DEMAT 3+: Deutscher Mathematiktest für dritte Klassen* (2. Aufl.). Göttingen: Hogrefe.
- Rost, D. H. (2001). Hochbegabung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 239–248). Weinheim: Beltz, PVU.
- Rost, D. H. (2009). *Hochbegabte und hochleistende Jugendliche. Befunde aus dem Marburger Hochbegabtenprojekt* (2. Aufl.). Münster: Waxmann.
- Rost, D. H., Sparfeldt, J. R., & Schilling, S. R. (2006). Hochbegabung. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 187–222). Berlin, Heidelberg: Springer.
- Rost, D. H., Sparfeldt, J. R., & Buch, S. (Hrsg.). (2018). *Handwörterbuch Pädagogische Psychologie* (5. Aufl.). Weinheim: Beltz.
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence* 53, 118–137.
- Schiepe-Tiska, A., Rönnebeck, S., Schöps, K., Neumann, K., Schmidtner, S., Parchmann, I., et al. (2016). Naturwissenschaftliche Kompetenz in PISA 2015 – Ergebnisse des internationalen Vergleichs mit einem modifizierten Testansatz. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Hrsg.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (S. 45–98). Münster: Waxmann.
- Schmidt, S., Ennemoser, M., & Krajewski, K. (2013). *DEMAT 9: Deutscher Mathematiktest für neunte Klassen*. Göttingen: Hogrefe.
- Schmitt, M., & Schmidt-Atzert, L. (2019). Gütekriterien für Online-Self-Assessments. In L. Schmidt-Atzert, M. Schütz, & G. Stemmler (Hrsg.), *Online-Self-Assessments an Hochschulen* (S. 99–114). Lengerich: Pabst Science Publishers.
- Schmidt-Atzert, L., Schütz, M., & Stemmler, G. (Hrsg.). (2019). *Online-Self-Assessments an Hochschulen*. Lengerich: Pabst Science Publishers.
- Schneider, W., & Hasselhorn, M. (Hrsg.). (2018). *Schuleingangsdiagnostik*. Göttingen: Hogrefe.
- Smithers, L. G., Sawyer, A. C. P., Chittleborough, C. R., Davies, N. M., Davey Smith, G., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour* 2, 867–880.
- Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development* 17, 7–41.
- Stock, C., & Schneider, W. (2008a). *DERET 1-2+: Deutscher Rechtschreibtest für das erste und zweite Schuljahr*. Göttingen: Hogrefe.
- Stock, C., & Schneider, W. (2008b). *DERET 3-4+: Deutscher Rechtschreibtest für das dritte und vierte Schuljahr*. Göttingen: Hogrefe.
- Stricker, J., Buecker, S., Schneider, M., & Preckel, F. (2019). Intellectual giftedness and multidimensional perfectionism: A meta-analytic review. *Educational Psychology Review*. doi: ▶ <https://doi.org/10.1007/s10648-019-09504-1>.
- Trapmann, S., Hell, B., Hirn, J.-O. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie* 215, 132–151.
- Trautwein, U., & Hasselhorn, M. (Hrsg.). (2017). *Begabungen und Talente*. Göttingen: Hogrefe.
- Visser, L., Büttner, G., & Hasselhorn, M. (2019). Komorbidität spezifischer Lernstörungen und psychischer Aufälligkeiten: Ein Literaturüberblick. *Lernen und Lernstörungen* 8, 7–20.
- Wendt, H., Bos, W., Selter, C., Köller, O., Schwippert, K., & Kasper, D. (Hrsg.). (2016). *TIMSS 2015: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

- Wild, K.-P. (1993). Hochbegabtendiagnostik durch Lehrer. In: D. H. Rost (Hrsg.), *Lebensumweltanalyse hochbegabter Kinder* (S. 236–261). Göttingen: Hogrefe.
- Wirthwein, L., Bergold, S., Preckel, F., & Steinmayr, R. (2019). Personality and school functioning of intellectually gifted and nongifted adolescents: Self-perceptions and parents' assessments. *Learning and Individual Differences* 73, 16–29.
- Zimmerhofer, A., Heukamp, V. M., & Hornke, L. F. (2006). Ein Schritt zur fundierteren Studienfachwahl – webbasierte Self-Assessments in der Praxis. *Report Psychologie* 31, 62–72.

Diagnostik in der Klinischen Psychologie und Psychotherapie

Thomas Fydrich

Inhaltsverzeichnis

- 8.1 Aufgaben der klinisch-psychologischen Diagnostik – 690**
 - 8.1.1 Rahmenbedingungen für klinisch-psychologische Diagnostik und Intervention – 692
 - 8.1.2 Das diagnostische Interview – 694
- 8.2 Problem-, Verhaltens- und Plananalyse als Ansatz der kognitiv-verhaltenstherapeutischen Diagnostik – 696**
- 8.3 Psychische Störungen und ihre Klassifikation – 700**
 - 8.3.1 Klassifikation psychischer Störungen – 701
 - 8.3.2 Diagnostische Verfahren zur Klassifikation psychischer Störungen – 706
- 8.4 Psychometrische Verfahren – 708**
 - 8.4.1 Verhaltenstheoretisch und kognitiv orientierte Fragebogenverfahren – 708
 - 8.4.2 Beobachtungsmethoden – 712
 - 8.4.3 Persönlichkeitstests in der Klinischen Psychologie und Psychotherapie – 713
 - 8.4.4 Verfahren und Ansätze auf klientenzentrierter, psychodynamischer, systemischer und interpersoneller Grundlage – 714
- 8.5 Verbindung von Diagnostik und Intervention:
Die Indikation – 717**
- 8.6 Erfolgskontrolle als Teil der Qualitätssicherung – 719**
 - 8.6.1 Zieldefinition, Therapieverlaufs- und Veränderungsdiagnostik – 720
 - 8.6.2 Kriterium der klinisch bedeutsamen Verbesserung – 721
- 8.7 Zusammenfassung – 724**
- Literatur – 726**

8.1 Aufgaben der klinisch-psychologischen Diagnostik

Definition

Gegenstand der **Klinischen Psychologie** ist die Erforschung von Entstehung und Aufrechterhaltung psychischer Störungen sowie auch körperlicher Störungen, bei denen psychische Faktoren eine bedeutsame Rolle spielen. Zur Klinischen Psychologie gehören die Diagnostik der entsprechenden Erkrankungen (psychischen Störungen) und Probleme sowie die Entwicklung und Überprüfung der Wirksamkeit psychologischer bzw. psychotherapeutischer Behandlungen. Dabei werden umfassend Erkenntnisse und Forschungsmethoden aus den Grundlagenfächern der Psychologie genutzt.

8

Störungswissen

Veränderungswissen

Ist- und Soll-Zustand feststellen

Aufgaben der Diagnostik

Zu den wichtigsten Aufgaben der Klinischen Psychologie gehört zum einen die Forschung über die Entstehung von psychischen Störungen, zum anderen die Überprüfung von Modellen darüber, wie psychische Störungen aufrechterhalten werden, wie sie „funktionieren“. Dieses „Störungswissen“ stellt die Voraussetzung für angemessene psychotherapeutische bzw. sonstige klinisch-psychologische Interventionen (z. B. Beratung) dar.

Ziel psychotherapeutischer Behandlung und sonstiger klinisch-psychologischer Interventionen ist es, die vorhandene Erkrankung oder Störung entweder zu beseitigen oder zumindest zu lindern und die mit der Problematik verbundenen Einschränkungen und das persönliche Leid zu reduzieren. Um dieses Ziel erreichen zu können, ist es notwendig, Kenntnisse darüber zu haben, wie eine Behandlung durchgeführt werden sollte, damit sich der gewünschte Erfolg mit hoher Wahrscheinlichkeit einstellt („Veränderungswissen“).

Notwendige Voraussetzung für eine erfolgreiche Anwendung des Störungs- und Veränderungswissens ist die Berücksichtigung *individueller Besonderheiten* der zu behandelnden Person (oder auch eines „Systems“; z. B. einer Partnerschaft). Die zentrale Zielstellung der Diagnostik besteht in der Beantwortung der Frage, welche Intervention für welche Person mit welchem Problem zu welcher Zeit durch welche Therapeutin bzw. welchen Therapeuten am ehesten zu den gewünschten Veränderungen führt (► Abschn. 8.3).

Um der Antwort auf diese Frage so nahe wie möglich zu kommen, müssen sowohl der *Ist-Zustand* (d. h. die Beschwerden bzw. die Symptomatik sowie die Bedingungen, unter denen sie auftreten) als auch der *Soll-Zustand* (das Ziel oder zumindest die Richtung der Veränderung) festgestellt werden. Dies ist die zentrale Aufgabe der klinisch-psychologischen Diagnostik. Wird diese Aufgabe nicht in angemessener Weise gelöst, kann keine ethisch vertretbare und professionell begründete Intervention stattfinden.

Zu den wichtigsten Aufgaben der klinisch-psychologischen Diagnostik gehören daher die in der folgenden Übersicht dargestellten 5 Punkte (vgl. auch Fydrich 2011; Wittchen und Hoyer 2011).

Die wichtigsten Aufgaben der klinisch-psychologischen Diagnostik

- *Qualitative und quantitative Beschreibung der vorliegenden Problematik:* Hierzu gehören die Beschreibung der Symptome bzw. des Problems sowie die Erhebung zentraler Aspekte der Problematik. Wichtig sind vor allem die Häufigkeit, die Intensität und die Dauer der Symptomatik sowie die Art der Einschränkungen, die sich für den bzw. die Betroffene dadurch ergeben. Darüber hinaus werden Bedingungen und Faktoren erhoben, unter denen die Probleme auftreten, sich verstärken oder verringern.
- *Klassifikation der psychischen Störung:* Die Klassifikation psychischer Störungen hat einerseits zum Ziel, die meist komplexe Information über die vorhandene Problematik so zu reduzieren bzw. zusammenzufassen, dass ein professioneller Informationsaustausch möglich wird. Klassifikationssysteme sind für die Feststellung notwendig, ob die vorliegenden Probleme bzw. Störungen krankheitswertig sind. Bei Weitem nicht alle psychischen Probleme erfüllen die Kriterien für eine psychische Störung bzw. Erkrankung. Weiterhin geben sie notwendige Hinweise für die Indikation und die Differentialindikation hinsichtlich der Behandlung und sie ermöglichen die Orientierung an aktuellen, empirisch fundierten Behandlungsleitlinien.
- *Exploration von besonderen lebensgeschichtlichen Bedingungen bei der Entstehung und dem bisherigen Verlauf der Störung:* Diese Informationen sind für die individuelle Planung der Behandlung von Bedeutung und spielen eine wichtige Rolle für das umfassende Verständnis der Problematik.
- *Beobachtung des Verlaufs der Intervention und der Veränderung der Symptomatik (adaptive Diagnostik, Verlaufsdiagnostik):* Nach der Feststellung des „Ist-Zustands“ zu Beginn der Therapie muss die Behandlung in der Regel an den jeweiligen Verlauf der Therapie angepasst werden. Es gehört zur Routine der klinisch-psychologischen Intervention, den Prozess der Veränderung zu beobachten und die Therapieschritte entsprechend anzupassen.
- *Überprüfung des Therapieerfolgs (Sicherung der Ergebnisqualität):* Das Ergebnis von Behandlungen sollte anhand zuverlässiger und valider Kriterien möglichst objektiv geprüft werden.

Beschreibung

Klassifikation

Lebensgeschichtliche Exploration

Veränderungen beobachten

Evaluation

Diagnostische Ansätze und Methoden

Um den genannten Aufgaben der klinisch-psychologischen Diagnostik gerecht zu werden, wurde eine Vielzahl diagnostischer Ansätze und Methoden entwickelt. Zur Orientierung lassen sich die Verfahren in zweierlei Weise gruppieren:

- Nach dem „System“, das sie beobachten: Hierbei ist es sinnvoll, sich an zentralen und allgemeinen Aspekten des menschlichen Erlebens und Verhaltens zu orientieren. Zu diesen Ebenen des Erlebens und Verhaltens gehören:
 - Körperliche Aspekte (z. B. physiologische Aktivierung wie die Erhöhung der Herzfrequenz bei Ängsten)
 - Gedanken und Gefühle (kognitiv-emotionale Ebene; z. B. Hoffnungslosigkeit, negative Erwartungen)
 - Verhalten (motorisch und sprachlich; z. B. Vermeiden von Situationen, die Angst machen)
 - Situative Faktoren (z. B. allein sein; Anwesenheit von Personen) und das Verhalten der Umwelt (z. B. der Familienmitglieder oder von Kolleginnen und Kollegen)

- Nach der *eingesetzten diagnostischen Methode*: Hierzu können folgende gezählt werden:
 - Das (offene) diagnostische Gespräch
 - Strukturierte und standardisierte klinisch-psychologische Interviews (z. B. zur objektiven Erhebung diagnostischer Informationen, die für eine Klassifikation der Störung notwendig sind)
 - Fragebogen- und Testverfahren (z. B. zur Erhebung personenbezogener Informationen; zum standardisierten Erfassen der Symptomatik und von Persönlichkeitsvariablen)
 - Beobachtungsmethoden (z. B. die Erhebung, wie oft ein bestimmtes, Leid bringendes Verhalten auftritt)
 - Psychophysiologische und biologische Verfahren (z. B. das Messen von Muskelspannungen bei einer Person, die unter Spannungskopfschmerzen leidet)

Unterschiedliche Störungsmodelle

8

Die im Folgenden dargestellten Verfahren gehen von unterschiedlichen theoretischen *Modellvorstellungen* über die Entstehung und die „Funktion“ psychischer Störungen aus. Im vorliegenden Text werden eher solche Verfahren dargestellt, die dem verhaltenstheoretischen (lerntheoretisch fundierten) Störungsparadigma nahestehen. Einige Abschnitte sind jedoch auch spezifischen diagnostischen Verfahren gewidmet, die sich an psychodynamischen, humanistisch-klientenzentrierten oder systemischen Störungsmodellen orientieren (zur Darstellung unterschiedlicher Störungs- und Interventionsmodelle s. Perrez und Baumann 2005). Dabei wird weitgehend auf die Diagnostik bei Erwachsenen eingegangen. Für den Bereich Kinder- und Jugendlichenpsychotherapie wird auf Esser (2008), Schneider und Margraf (2019) oder Steinhäusen und von Aster (1999) verwiesen.

8.1.1 Rahmenbedingungen für klinisch-psychologische Diagnostik und Intervention

Rahmenbedingungen

Klinisch-psychologische Diagnostik findet immer in einem professionellen Rahmen statt. Dieser Rahmen ist in den meisten Fällen gekennzeichnet durch

- institutionelle, berufs- und sozialrechtliche Bedingungen,
- die vorherrschenden Probleme bzw. die psychische Störung,
- die Motivation der Hilfesuchenden,
- Therapeutinnen- bzw. Therapeutenvariablen.

In Abhängigkeit verschiedener Rahmenbedingungen können sowohl der Inhalt als auch die Aufgabe und das Ziel der klinisch-psychologischen Diagnostik beträchtlich variieren.

Beispiele für institutionelle Rahmenbedingungen und damit zusammenhängende spezifische Aufgaben der klinisch-psychologischen Diagnostik

- *Private Praxis (niedergelassene Psychologische oder Ärztliche Psychotherapeutinnen und -therapeuten):* Diagnostik und klinisch-psychologische Intervention; primär Einzel- oder Gruppenpsychotherapie; klassifikatorische und dimensionale Diagnostik psychischer Störungen und Erkrankungen.
- *Psychiatrische Klinik:* meist Kombination von pharmakologischen, psychotherapeutischen und weiteren Therapiemaßnahmen für Menschen mit stark ausgeprägten psychischen Störungen, die oft eine Selbstgefährdung und/oder eine weitgehende Einschränkung der Fähigkeit, den Alltag zu bewältigen, mit sich bringen.
- *Psychosomatische Kliniken:* stationäre, zumeist psychotherapeutische Behandlung und Rehabilitation bei psychischen Störungen und körperlichen Erkrankungen, bei denen psychische Faktoren eine bedeutsame Rolle spielen. Hierzu zählt auch die psychologisch-psychotherapeutische Mitbehandlung von chronischen somatischen Erkrankungen (z. B. in der Psychoonkologie oder Psychokardiologie).
- *Berufliche Rehabilitation:* Diagnostik von Fähigkeiten und Fertigkeiten, die die Wiedereingliederung in die bisherige Berufstätigkeit oder die Umschaltung in einen neuen Arbeitsbereich zum Ziel haben; Einsatz dafür geeigneter therapeutischer Interventionen und Fördermaßnahmen.
- *Beratungsstellen:* Beratung, z. B. von Alkohol- und Drogenabhängigen; Beratungsstellen für Paare, Familien, Erziehungsfragen, Studierende. Indikationsstellung für umfassendere medizinische und/oder psychotherapeutische, ambulante oder stationäre Behandlung und entsprechende Vermittlung.
- *Allgemeinmedizinische Kliniken:* psychologische Konsiliar- und Liaisedienste primär bei körperlichen Erkrankungen, bei denen ein Einfluss psychologischer Faktoren auf die Entstehung, den Verlauf und möglicherweise auch auf die Genesung bekannt oder wahrscheinlich ist. Ziel der psychotherapeutischen, zumeist begleitenden Behandlung ist oft auch die Anpassung an die durch die somatische Erkrankung bedingte Beeinträchtigung des Alltags.

Beispiele für institutionelle Rahmenbedingungen

Die wichtigsten Rahmenbedingungen von Patientinnen- bzw. Patientenseite sind die vorherrschende *Problematik*, *Symptomatik* bzw. *Störung* und die damit im Zusammenhang stehende *Motivation*, die zum Aufsuchen professioneller Hilfe geführt hat. Grundsätzlich verschiedene Voraussetzungen bestehen beispielsweise dann, wenn ein Patient bzw. eine Patientin nach längerem Überlegen selbst zu dem Entschluss gekommen ist, eine psychotherapeutische Behandlung in Anspruch nehmen zu wollen, oder wenn – beispielsweise im Rahmen einer hausärztlichen Behandlung – eine Psychotherapie oder psychologischen Beratung empfohlen wurde, nachdem dort nach einer schon lange andauernden Schmerzerkrankung ein Missbrauch von Schmerzmitteln erkennbar geworden ist.

In Abhängigkeit von der Hauptproblematik und der Schwere der Erkrankung muss zu Beginn der Behandlung entschieden werden,

- a) ob überhaupt eine psychotherapeutische oder psychiatrische Behandlung indiziert ist und
- b) welche Behandlung in welchem Behandlungsrahmen bzw. welcher Behandlungseinrichtung und mit welcher Frequenz angemessen ist.

Patientenbezogene Rahmenbedingungen

Therapeutenbezogene Rahmenbedingungen

Für diese Indikationsentscheidung ist eine Beratung oder ggf. auch eine ausführliche Diagnostik durch niedergelassene Psychotherapeutinnen bzw. Psychotherapeuten oder fachärztlich durch einen Psychiater bzw. eine Psychiaterin notwendig.

Auf der Seite der Behandlerinnen und Behandler hat vor allem die akademische und klinische Ausbildung und damit auch die theoretische und praktische Orientierung weitreichende Konsequenzen für die Art der durchgeführten Diagnostik und die Schwerpunkte in der Methodik der therapeutischen Interventionen. So sehen beispielsweise Therapeutinnen und Therapeuten, die auf der Grundlage psychodynamischer Ansätze arbeiten, die klassifikatorische Diagnostik psychischer Störungen nach der Internationalen Klassifikation psychischer Störungen (ICD) oder dem Diagnostischen und Statistischen Manual Psychischer Störungen (DSM) als eher wenig hilfreich an, da diese kaum Berührungspunkte mit den psychodynamischen Theorievorstellungen aufweisen. In psychodynamisch orientierten Psychotherapien ist vielmehr die Diagnostik aktueller (interpersoneller oder intrapsychischer) Konflikte sowie die der psychischen Struktur von Bedeutung. Verhaltenstheoretisch arbeitende Psychotherapeutinnen und -therapeuten betonen im Rahmen ihrer Problem- und Verhaltensanalyse dagegen in besonderer Weise die Symptomatik der vorliegenden Störung sowie die funktionellen und situativen Bedingungen, die dazu beitragen, dass die Problematik weiter besteht bzw. immer wieder neu auftritt.

8.1.2 Das diagnostische Interview

Therapeutische Kompetenzen

Zur Grundlage der Gesprächsführung in der klinisch-psychologischen Praxis gehört neben dem theoretischen Wissen über psychische Störungen und Problembereiche vor allem auch eine Reihe wichtiger praktischer Fertigkeiten. Eine zentrale Voraussetzung für den Aufbau einer problem- und lösungsorientierten therapeutischen Arbeitsbeziehung ist, dass der Therapeut bzw. die Therapeutin sich von Beginn des Kontakts an in die Lage und in die Problematik der Ratsuchenden hineinversetzt (vgl. auch ▶ Abschn. 3.7).

Entscheidend bei der ersten Kontaktaufnahme – dies betrifft schon den telefonischen Erstkontakt – ist daher, dass eine Reihe von Bedingungen berücksichtigt werden muss, die auf der Seite des Patienten/der Patientin bzw. des Hilfesuchenden in der Regel gegeben sind. Hierzu gehören:

- Schwellenängste und Unsicherheiten bezüglich der Wahl des therapeutischen Ansatzes und des Therapeuten oder der Therapeutin
- Scham darüber, professionelle Hilfe in Anspruch nehmen zu müssen
- Hoffnung, die erwartete Hilfe zu bekommen
- Möglicherweise Frustration über eine bisher erfolglose Suche nach einem Therapieplatz und damit verbundenen Wartezeiten
- Gegebenenfalls Art, Umfang und Erfolg vorausgegangener Therapien

Empathische Haltung

Zu Beginn und im Verlauf jedes professionellen Kontakts ist es notwendig, dass die Therapeutin bzw. der Therapeut mit einer *empathischen Grundhaltung* diese Faktoren berücksichtigt. Neben dem Einfühlungsvermögen erwarten Patientinnen und Patienten mit Recht gleichzeitig auch ein hohes Maß an Professionalität. Das therapeutische Gespräch bzw. das therapeutische Interview von Psychotherapeutinnen bzw. -therapeuten und professionellen Beraterinnen bzw. Beratern unterscheidet sich in wesentlichen Punkten von unterstützenden Gesprächen im Freundes- oder Bekanntenkreis. Wichtigste formale Bestimmungsmerkmale des therapeutischen Gesprächs sind die Nichtreziprozität der Beziehung, ein institutioneller, rechtlich und

zeitlich geregelter Rahmen für die Gespräche, die Ausbildung der Therapeutinnen und Therapeuten sowie die finanzielle Honorierung der Thermen im sozialrechtlichen System der Kranken- und Rentenversicherungen bzw. der Institution, in der die Therapie stattfindet.

Zu den wichtigsten zentralen Inhalten beim Erwerb der notwendigen Kompetenzen in der Ausbildung in Psychotherapie und Klinischer Psychologie gehört insbesondere das Wissen über (psychische) Störungen inklusive der Kenntnisse und Fertigkeiten, um eine systematische Diagnostik durchführen zu können (*Störungswissen*). Weitere notwendige Kenntnisse und Fertigkeiten betreffen die Indikationsstellung und das Beherrschene konkreter Behandlungsmethoden und -techniken (*Veränderungswissen*). Außerdem ist empathisches Verhalten und das „Sich-einstellen-Können“ auf Personen mit unterschiedlichen Motiven und Persönlichkeitseigenschaften, verschiedenem Bildungshintergrund oder ethnischer Herkunft eine wichtige und notwendige Kompetenz von klinischen Psychologinnen und Psychotherapeuten (*Interaktionswissen/Interaktionskompetenz*).

Interaktionskompetenz

Im Rahmen der ersten diagnostischen Gespräche entsteht durch gezieltes, empathisches Erheben von Informationen zur aktuellen Problematik, ihrer Geschichte und möglicher ätiologischer Zusammenhänge eine professionelle, therapeutische und problem- bzw. lösungsorientierte Arbeitsbeziehung. Auf der therapeutischen Seite ist diese durch Empathie und Parteilichkeit für den Ratsuchenden sowie durch die Fähigkeit der Perspektivübernahme gekennzeichnet. Vertrauen, Offenheit sowie Veränderungsmotivation sind günstige Patientenbedingungen für den Aufbau einer zielorientierten Beziehung.

! Vorsicht

Das diagnostische Interview dient der Erhebung problemrelevanter Informationen mit dem Ziel,

- das Problem genau zu beschreiben,
- zu verstehen, wie das Problem gegenwärtig aufrechterhalten wird, und
- Hintergründe der Lebens-, Problem- und Behandlungsgeschichte zu erfassen.

Ziel des diagnostischen Interviews

Entsprechend können 2 Schwerpunkte bzw. Funktionen des klinisch-psychologischen Interviews unterschieden werden: erstens die Exploration der aktuellen Problematik und zweitens die Erhebung des lebensgeschichtlichen Hintergrunds.

8.1.2.1 Exploration der aktuellen Problematik

Da die *Beschwerden* eines Patienten bzw. einer Patientin in der Regel die Hauptmotivation für das Aufsuchen einer Behandlung oder Beratung darstellen, sollten diese zu Beginn eines Kontakts im Zentrum des Gesprächs stehen. Hauptthemen der problemzentrierten Exploration sind dabei die *aktuelle Symptomatik* bzw. die verschiedenen Facetten des *gegenwärtigen Problems*. Hierzu gehört in erster Linie die genaue Beschreibung des Problems und der damit zusammenhängenden Beschwerden und Einschränkungen. Dabei sind neben situativen Aspekten (in welchen Situationen tritt das Problem auf) vor allem die Häufigkeit und die Dauer des Auftretens, die Intensität sowie der Grad der Beeinträchtigung in verschiedenen Lebensbereichen von zentraler Wichtigkeit. Exploriert werden müssen neben problembezogenem Erleben die Gedanken und Gefühle des Patienten bzw. der Patientin, die dazu gehörenden körperlichen Reaktionen und das verbale und motorische Verhalten. Erhoben werden müssen die Situationen, in denen die Problematik auftritt, sowie Reaktionen des sozialen Umfelds auf das Problem bzw. das problematische Verhalten des Betroffenen.

Validierung

Die Exploration wird dabei wie folgt strukturiert. Erstens sollte die individuelle Problematik auf der Basis klinisch-psychologischer Modelle über die Entstehung und die Perpetuierung einer Störung sowohl für die Therapeutin bzw. den Therapeuten als auch für die Patientin bzw. den Patienten *verstehbar* werden. Zweitens muss die wissenschaftlich-fachliche Erklärung der Problematik und des (klinisch-psychologischen) Erklärungsmodells mit den subjektiven Einschätzungen der Problematik durch die Ratsuchenden *abgestimmt* werden (Validierung).

Problemstrukturierung

Diese Abstimmung ist notwendig, damit Patientinnen bzw. Patienten oder Ratsuchende das in der Therapie dargestellte wissenschaftlich fundierte Modell der Störung bzw. der Problematik und den darauf fußenden Behandlungsansatz verstehen und für sich annehmen können. Die Abstimmung wissenschaftlich fundierter Störungsmodelle mit dem Erleben und dem Verhalten von Patientinnen und Patienten stellt eine wichtige Voraussetzung für das Gelingen des therapeutischen Zusammenarbeits dar und wird zudem über ein transparentes Vorgehen der Therapeutin bzw. des Therapeuten erreicht. Patientinnen und Patienten sollte klar sein, was und warum die Therapeutin bzw. der Therapeut bestimmte Fragen stellt und wie sie/er den Prozess der Therapie steuert. Dabei helfen auch strukturierende Fragen sowie Zusammenfassungen der Therapeutin bzw. des Therapeuten. Damit wird eine kooperative therapeutische Arbeitsbeziehung unterstützt, in der Patientinnen und Patienten erleben, dass die Therapeutin bzw. der Therapeut sie versteht und fortlaufend sehen kann, und der therapeutisch eingeschlagene und realisierte Weg nachvollziehbar ist.

8.1.2.2 Problemvorgeschichte und biografische Anamnese

Bei der Exploration der *Problemvorgeschichte* werden Informationen gewonnen, die für die Entwicklung der Problematik bedeutsam waren. Um gezielt nach Entstehungsbedingungen zu fragen, die aus empirisch fundierten Störungsmodellen bekannt sind, müssen wissenschaftlich fundierte Störungskenntnisse vorhanden sein. Zu den Variablen, die in diesem Zusammenhang zu erheben sind, gehören, *wann* das Problem zum ersten Mal auftrat, *welche Bedingungen* aus der Sicht der Patientin bzw. des Patienten zum ersten Auftreten geführt haben oder damit in zeitlichem Zusammenhang standen sowie Informationen zu *anderen Problemen* bzw. *Symptomen* in der Vorgeschichte. Weiterhin müssen Vorbehandlungen und deren Verlauf sowie Art und Umfang bisheriger *eigener Lösungsversuche für das Problem* erfragt werden. Zu den wichtigen Daten der *Lebensgeschichte* zählen auch Informationen zu Kindheit und Adoleszenz, Eltern und Elternhaus, Geschwistern, Erziehung, Entwicklung sozialer Kontakte zu Gleichaltrigen, Partnerschaften, sexueller Entwicklung sowie schul- und berufsbezogene Informationen.

Auch bei der Exploration der Problemvorgeschichte gilt, dass die *subjektiven Vorstellungen* des bzw. der Betroffenen über die Entstehung und Aufrechterhaltung der eigenen Problematik sehr bedeutsam sind. Zwischen der individuellen Erklärung der Problematik in Hinblick auf die Entstehung und der Art, wie sie aufrechterhalten wird, besteht oft ein enger Zusammenhang.

Berücksichtigung subjektiver
Vorstellungen wichtig

8.2 Problem-, Verhaltens- und Plananalyse als Ansatz der kognitiv-verhaltenstherapeutischen Diagnostik

Diagnostik funktionaler
Zusammenhänge

Im Mittelpunkt der Analyse von Verhalten, Kognitionen, Emotionen und körperlichen Reaktionen innerhalb des kognitiv-verhaltenstherapeutischen Ansatzes steht die Diagnostik von *funktionalen Zusammenhängen* zwischen dem sog. „problematischen Verhalten“ einerseits sowie *antecedenten* und

konsequenteren Bedingungen und Ereignissen andererseits. Ziel dieser Form der Diagnostik ist in erster Linie, diejenigen Bedingungen festzustellen, die die Aufrechterhaltung der Probleme erklären können.

Auf diese Analyse bauen im Weiteren die Bestimmung von therapeutischen Zielen sowie die Indikationsstellung für bestimmte therapeutische Interventionen auf. Die psychische Störung wird als Problem betrachtet, das sowohl *beobachtbares Verhalten* als auch *persönliches Erleben* (Emotion, Denken) und *körperliche Zustände und Reaktionen* (z. B. Unwohlsein, Antriebslosigkeit, Schmerzen oder Herzrasen) umfasst. Antezedente und konsequente Bedingungen bzw. Ereignisse des „problematischen Verhaltens“ können sowohl externer als auch interner Art sein. Hierzu gehören z. B. bei Personen mit starken Ängsten vor engen Räumen bestimmte Räume und deren Beschaffenheit (z. B. kein Tageslicht, viele andere Menschen als externe Bedingungen), aber auch eigene Gedanken (z. B. der Gedanke daran, nicht schnell genug aus einem Raum herauszukommen) oder die Wahrnehmung von Körperempfindungen (z. B. Herzklopfen oder Schwindelgefühl) als interne auslösende Bedingungen.

Bei einer Problem- bzw. Verhaltensanalyse wird zunächst das *Verhalten in Situationen* möglichst genau erhoben und beschrieben (z. B. Bartling et al. 2016; Kanfer et al. 2012). Zentraler Schritt ist dabei zunächst die genaue Exploration und Beschreibung des Problems bzw. der Symptomatik. Dabei werden jeweils 4 zentralen Modalitäten des Erlebens und des Verhaltens berücksichtigt:

- Physiologie: Welche körperlichen Vorgänge treten auf?
- Kognition: Welche Gedanken und Gefühle gehören zur Symptomatik bzw. gehen mit ihr einher?
- Emotion: Welche Gefühle treten auf?
- Verhalten: Was tut die Person?

Elemente der Verhaltensanalyse

Die genannten 4 Modalitäten der Symptomatik folgen nicht unbedingt sukzessiv aufeinander, sondern sind oft simultan ablaufende und auftretende Anteile des Erlebens und des Verhaltens.

Mit dem Ziel, ein funktionales Bedingungsmodell als psychologische Erklärung des Problems bzw. der Symptomatik zu erstellen, werden weiterhin die inneren und äußereren situativen Komponenten analysiert, unter denen das Problem auftritt. Dabei wird darauf geachtet, welche internen und/oder externen Bedingungen mit einer Variation wichtiger Aspekte des Problems (Häufigkeit, Intensität, Dauer) einhergehen. Neben den vorausgehenden (antezedenten) situativen Bedingungen sind auch die nachfolgenden (konsequenteren) Bedingungen oder Ereignisse von großer Bedeutung.

Die durch eine Verhaltensanalyse explorierten Zusammenhänge werden als funktionales Bedingungsmodell nach den Paradigmen des *klassischen* und des *operanten Lernens* dargestellt. Überträgt man das Modell des klassischen Konditionierens (respondentes Lernen) auf den Bereich psychischer Störungen und Probleme, so können Situationen, die zunächst „neutral“ sind und in denen unangenehme (u. a. physiologische) Erlebnisse oder Reaktionen aufgetreten sind, zu Situationen werden, in denen diese Reaktionen dann häufig „automatisch“ stattfinden. Mit dem Modell des operanten Konditionierens wird erklärt, dass dem Verhalten nachfolgende Bedingungen oder Ereignisse eine Verstärkerfunktion für die betroffene Person und damit für das „problematische Verhalten“ haben können, sodass dieses mit größerer Wahrscheinlichkeit in den entsprechenden Situationen erneut auftritt.

4 Modalitäten

Interne und externe Bedingungen

Klassisches und operantes Lernen

► Beispiel

Ein klinisch-psychologisches Beispiel für eine respondente Reaktion ist das starke, plötzlich auftretende Herzklopfen von Frau A., das von Gefühlen der Unsicherheit und Ängstlichkeit begleitet wird. Es tritt immer dann auf, wenn bei ihr das Telefon klingelt. Hintergrund dieser Symptomatik ist bei ihr, dass in der Vergangenheit ihr früherer Partner häufig angerufen und ihr in aggressiver Weise wiederholt Vorwürfe über die Art der von ihr durchgesetzten Trennung gemacht hat.

Ein weiteres Beispiel für eine mit dem Modell der klassischen Konditionierung erklärbaren Reaktion ist die starke Angst von Herrn I., während er sich mit seinem Auto auf dem Weg zu seiner Arbeitsstelle einer großen Brücke nähert. Auf dieser Brücke hatte er vor Kurzem einen heftigen Panikanfall, bei dem er glaubte, die Kontrolle über sich und das Auto zu verlieren. Operant erklärbar ist, dass es Herr I. seit der erlebten Panikattacke vermeidet, über die Brücke zu fahren. Er fährt einen großen Umweg zur Arbeitsstelle. Dadurch treten die Ängste kaum noch auf. Das Vermeidungsverhalten kann durch negative Verstärkung (Reduktion der stark aversiven Angst) erklärt werden. ◀

8

Hypothesenprüfung

Verhaltensgleichung

Kognitive Komponenten

Horizontale und vertikale Verhaltensanalyse/Plananalyse

Die auf dieser Grundlage entwickelte Problem- und Verhaltensanalyse kann vor allem als ein Modell für die Erklärung von aktuell auftretenden Problemen (Symptomen) im Sinne der *Aufrechterhaltung* von Problemen verstanden werden. Die explorierten Zusammenhänge sind dabei in einem konkreten Fall als diagnostische und therapeutische *Hypothesen* zu verstehen, die im Verlauf der Therapie immer wieder an den aktuellen Informationsstand adaptiert werden müssen.

Kern des funktionalen Bedingungsmodells ist die sog. „Verhaltensgleichung“, in der das Verhalten selbst (R = Reaktion, Verhalten) sowie die auslösenden Bedingungen (S = Situation) und die nachfolgenden, meist verstärkenden Bedingungen (C = Konsequenz) in ihren funktionalen Zusammenhängen beschrieben werden. Die Verhaltensgleichung nach dem respondenten Modell enthält nur 2 Aspekte (S-R), die nach dem operanten Modell in der ursprünglichen Form 3 Komponenten (S-R-C).

In der Folge dieser „klassischen“ S-R-C-Darstellung wurden von verschiedenen Autoren zusätzliche Faktoren in die Gleichung eingeführt, um die Bedeutung anderer Variablen bei der Entstehung und der Aufrechterhaltung problematischer Verhaltensweisen und Symptome zu betonen. Bekannt wurde vor allem die Verhaltensgleichung von Kanfer und Saslow (1976), die zwischen die S- und R-Komponente zusätzlich eine O-Variable (O = Organismus) einfügten, um bei der funktionalen Erklärung der Problematik körperliche Aspekte (z. B. Alkohol, Müdigkeit, Behinderung, Krankheit) mit zu berücksichtigen. Weiterhin enthielt die Kanfer'sche Gleichung noch den Aspekt „K“ (Kontingenzen), womit die aus den Lerntheorien bekannte Bedeutsamkeit von Häufigkeit, Intensität, Dauer und Aufeinanderfolge der Konsequenzen betont wurde. Von anderen Autoren wurden zusätzlich kognitive Aspekte im Rahmen des funktionalen Bedingungsmodells betont und als „E“ (Erwartung) mit in die Gleichung aufgenommen. Ähnlich kann das Modell von Bartling et al. (2016) gesehen werden. Die Autoren differenzieren auf der Ebene des Verhaltens als kognitiven Aspekt den Wahrnehmungsprozess (WP) und die innere Verarbeitung (IV). Sowohl auf der Ebene der Situation als auch der Konsequenzen (bei Bartling et al. 2016, mit „K“ symbolisiert) werden externe und interne Aspekte unterschieden (S_e, S_i und K_e, K_i).

Mit der sog. „kognitiven Wende“ wurden in der Verhaltenstherapie wesentlich stärker als in der klassischen Verhaltensanalyse die Rolle von Kognitionen, Einstellungen, Erwartungen, persönlichen Zielen oder

Intentionen bei der Problemanalyse und der Diagnostik psychischer Störungen berücksichtigt. In unterschiedlichen Störungstheorien und nachfolgend auch im Bereich der kognitiv-verhaltenstherapeutischen Diagnostik spielen beispielsweise *kognitive Schemata* (z. B. Beck et al. 1996), „beliefs“ (Ellis und Grieger 1995) oder *Handlungspläne* und *persönliche Ziele* eine zentrale Rolle bei der Erklärung von klinisch relevantem Verhalten und Erleben. Im Unterschied zur Problem- und Verhaltensanalyse, bei der vor allem das Verhalten in Situationen beschrieben wird, dient die Exploration komplexer Handlungsziele und -pläne der Diagnostik von individuell bedeutsamen, das konkrete Handeln leitenden Regeln, Plänen und Kognitionen. Diese werden daher auch als „handlungsleitende Kognitionen“ bezeichnet. Den entsprechenden Teil der Exploration und Beschreibung nennt man „vertikale Verhaltensanalyse“ (oder *Plananalyse*) im Unterschied zur „horizontalen Verhaltensanalyse“, bei der das Verhalten in Situationen primär auf einer Zeitachse abgebildet wird (Bartling et al. 2016; Caspar 2008).

Zu den Verfahren und Methoden, die bei der Problem-, Verhaltens- und Plananalyse hilfreich sein können, gehören *Verhaltensbeobachtungen*, (diagnostische) *Rollenspiele* sowie *Beschreibungen von Verhalten* durch Dritte (► Abb. 8.1; ► Abschn. 8.4.2). Auch kann es zu den ersten Aufgaben und Übungen der Klienten gehören, ihr Verhalten und Denken zu beobachten und zu protokollieren (z. B. „Drei-Spalten-Technik“, bei der die Situation, das Verhalten und auftretende Gedanken notiert werden; vgl. z. B. Linden und Hautzinger 2015). Die in ► Abschn. 8.3.1 genannten störungsspezifischen Verfahren, beispielsweise das Mobilitätsinventar von Ehlers et al. (2001), sind ein anschauliches Beispiel dafür, dass sich die Erhebung von Intensität und Häufigkeit problematischen Verhaltens (hier die Vermeidung von angstauslösenden Situationen) und die Verhaltensanalyse gut ergänzen. In diesem Inventar werden spezifische Situationen (Stimuli) erfragt, die vom Patienten unter den Bedingungen „allein“ oder „mit jemandem zusammen“ gemieden werden. Die Liste angenehmer Ereignisse (s. Hautzinger 2013) ist ein Beispiel für das Erfassen von Konsequenzen und Aktivitäten, die im Sinne einer Verstärkung von Bedeutung sein können.



► Abb. 8.1 Im Rahmen der Verhaltensanalyse werden individuell bedeutsame Situationen (hier eine Partysituation) genau erfasst, zu denen z. B. ein gemeinsames Essen gehört

Störungen mit Krankheitswert

8.3 Psychische Störungen und ihre Klassifikation

Die Definition davon, was eigentlich eine psychische Störung (Erkrankung) ausmacht, ist sowohl für die Forschung als auch für sozialrechtliche Aspekte der psychotherapeutischen oder psychiatrischen Versorgung und damit auch für das Versicherungswesen (vor allem Krankenversicherung und Rentenversicherung) von großer Wichtigkeit. Für die sozialrechtlich bedeutsame Frage, ob die Kosten für Behandlungen von Personen mit einer psychischen Störung übernommen werden, muss das Vorliegen einer Störung „mit Krankheitswert“ festgestellt werden. Diese Bedingung ist sogar Teil der Definition von Psychotherapie, wie sie im Psychotherapeutengesetz (PsychThG, in der Fassung vom 15. November 2019) formuliert wird. Dort heißt es in § 1 Absatz 2:

§ 1 PsychThG Berufsbezeichnung, Berufsausübung

(2) Ausübung der Psychotherapie im Sinne dieses Gesetzes ist jede mittels wissenschaftlich geprüfter und anerkannter psychotherapeutischer Verfahren oder Methoden berufs- oder geschäftsmäßig vorgenommene Tätigkeit zur Feststellung, Heilung oder Linderung von Störungen mit Krankheitswert, bei denen Psychotherapie indiziert ist.

Zentrale Komponenten für die Definition psychischer Störungen (vgl. DSM-5; American Psychiatric Association 2015, S. 26).

- Das Erleben und Verhalten lässt sich charakterisieren durch eine klinisch bedeutsame Störung in den Kognitionen, der Emotionsregulation oder des Verhaltens.
- Die Störung ist Ausdruck von dysfunktionalen psychologischen, biologischen oder entwicklungsbezogenen Prozessen, denen psychische und seelische Funktionen zugrunde liegen.
- Die Störung ist typischerweise verbunden mit bedeutsamen Leid oder Beeinträchtigungen hinsichtlich sozialer, berufs- oder ausbildungsbezogener und anderer Aktivitäten.
- Normativ oder kulturell anerkannte Reaktionen auf übliche Stressoren (z. B. Tod eines Angehörigen) oder sozial abweichende Verhaltensweisen und Konflikte sind nicht als psychische Störungen zu betrachten; es sei denn ihnen liegt eine der oben genannten Dysfunktionen zugrunde.

Abgrenzung klinischer Relevanz

In der Regel müssen mehrere, jedoch nicht alle genannten Punkte zutreffen. Zusätzlich müssen die Kriterien für die Diagnose mindestens einer psychischen Störung nach DSM oder ICD erfüllt sein. Es ist jedoch wichtig, dass das Zutreffen einzelner der oben benannten Komponenten allein nicht unbedingt indikativ für psychische Störungen sein muss. So können beispielsweise Normabweichungen durchaus erwünscht und Ausdruck besonderer Fähigkeiten sein. Normen sind abhängig von der ethnischen und sozialen Herkunft. Weiterhin ist zu beachten, dass persönliches Empfinden von Leid dann *nicht* auftritt, wenn die Störung „ich-synton“ ist, d. h., wenn die damit verbundenen Erlebens- und Verhaltensweisen von der betroffenen Person als angemessen, richtig und „zu sich gehörig“ erlebt werden. Dies kann beispielsweise bei einer Person mit einer paranoiden Persönlichkeitsstörung der Fall sein, die die eigenen Familienmitglieder „bespitzelt“, um Beweise für deren feindselige Haltung zu erhalten. Schlaflosigkeit und Konzentrationsprobleme, die z. B. für einige Zeit nach einem Autounfall auftreten, sind

ebenso nicht zwingend Indikatoren für eine psychische Erkrankung. Gleichsam erfüllt eine starke Angst vor Fliegen, die mit der Vermeidung jeglicher Flugreisen einhergeht, nicht die Kriterien für eine psychische Erkrankung (in diesem Fall einer spezifischen Phobie), wenn sie nicht oder nur in geringem Ausmaß mit Beeinträchtigungen einhergeht, z. B. weil die Person nicht auf Flugreisen angewiesen ist.

8.3.1 Klassifikation psychischer Störungen

Die aktuell gültigen und in Wissenschaft und klinischer Praxis eingesetzten Klassifikationssysteme für psychische Störungen sind

- die *Internationale Klassifikation psychischer Störungen* in ihrer 10. Revision (ICD-10; Kapitel V), die von der Weltgesundheitsorganisation (WHO) herausgegeben wird (Dilling et al. 2010), und
- das Klassifikationssystem der Amerikanischen Psychiatrischen Vereinigung (American Psychiatric Association), das *Diagnostische und Statistische Manual Psychischer Störungen* in der 5. Revision (DSM-5; American Psychiatric Association 2015, 2018).

ICD-10, ICD-11 und DSM-5

Für den europäischen Sprachraum und für das Gesundheitswesen in Deutschland ist aktuell die ICD-10 maßgeblich, die voraussichtlich ab dem Jahr 2022 von der 11. Revision (ICD-11) abgelöst wird. Das DSM-5 wird häufig (teilweise zusätzlich) in forschungsorientierten Einrichtungen eingesetzt. Beide Systeme basieren auf dem Prinzip der *operational und deskriptiv definierten Diagnostik*. Es wird weitgehend auf die früher üblichen ätiologisch und nosologisch (d. h. auf der Basis von Krankheitslehrern) orientierten Ordnungskriterien verzichtet. Ätiologische Faktoren sind in beiden Systemen im Wesentlichen nur bei den organisch bedingten psychischen Störungen und bei den Anpassungsstörungen als Definitionskriterien enthalten. Durch die weitgehend deskriptive Charakteristik sind die Klassifikationssysteme für die unterschiedlichen klinisch-psychologischen und ätiologietheoretischen Ansätze und Psychotherapieverfahren akzeptabler geworden.

Operationale und deskriptive Diagnostik

8.3.1.1 Internationale Klassifikation psychischer Störungen

Die Abkürzung ICD steht für „International Classification of Diseases“. Dieses Klassifikationssystem gliedert Störungen und Erkrankungen aller Art in 21 Kapitel. Für den Bereich der psychischen Störungen ist das Kapitel V relevant, das seit 1991 als 10. revidierte Fassung (ICD-10) verfügbar ist. Aktuell ist in Deutschland die 10. Revision der ICD in der Version GM (= German Modification) von 2019 gültig. Rechtsgrundlage für die Verpflichtung zur Klassifikation nach ICD-10 in der jeweils gültigen Fassung sind § 301 und § 295 des Sozialgesetzbuchs Fünftes Buch (SBG V), die sich auf die stationäre und ambulante Versorgung beziehen. Hierzu heißt es in den entsprechenden Paragrafen:

ICD-10 und ICD-10-GM

§ 301 Absatz 2 SGB V Krankenhäuser und § 295 Absatz 1 SGB V Abrechnung ärztlicher Leistungen (Auszug)

Die Diagnosen [...] sind nach der Internationalen Klassifikation der Krankheiten in der jeweiligen vom Deutschen Institut für medizinische Dokumentation und Information im Auftrag des Bundesministeriums für Gesundheit und Soziale Sicherung herausgegebenen deutschen Fassung zu verschlüsseln.[...].

Die ICD-11 wurde im Mai 2019 verabschiedet und soll ab 2022 in Deutschland in Kraft treten. Aktuell gültige Versionen sowie Informationen über beide Revisionen sowie aktuelle Anpassungen sind einsehbar über die Internetseite des Deutschen Instituts für Medizinische Dokumentation und Information (kurz: DIMDI; ► <https://www.dimdi.de/>).

10 Hauptgruppen der ICD-10

8

Die 10 Hauptgruppen zur Klassifikation psychischer Störungen nach ICD-10

Die ICD-10-Klassifikation für psychische Störungen (Kapitel V; F-Kodierungen) umfasst 10 Hauptgruppen:

- F0: Organische, einschließlich symptomatische psychische Störungen
- F1: Psychische und Verhaltensstörungen durch psychotrope Substanzen
- F2: Schizophrenie, schizotyp und wahnhafte Störungen
- F3: Affektive Störungen
- F4: Neurotische, Belastungs- und somatoforme Störungen
- F5: Verhaltensauffälligkeiten in Verbindung mit körperlichen Störungen und Faktoren
- F6: Persönlichkeits- und Verhaltensstörungen
- F7: Intelligenzminderung
- F8: Entwicklungsstörungen
- F9: Verhaltens- und emotionale Störungen mit Beginn in der Kindheit und Jugend

Jede dieser in der Übersicht genannten Hauptgruppen trägt den Kennbuchstaben „F“ für das relevante Kapitel „psychische Störungen“ sowie eine fortlaufende Nummerierung der Hauptgruppen. Die jeweils nächste Stelle der Verschlüsselung stellt die nächste, spezifischere Ebene der klassifikatorischen Einteilung dar. So gehört die „depressive Episode“ (F32) zur Hauptkategorie der „Affektiven Störungen“ (F3). Weitere Spezifizierungen der depressiven Episode sind in der nachfolgenden Übersicht dargestellt (Dilling et al. 2010).

Spezifizierungen der depressiven Episode

F32 depressive Episode

- F32.0: Leichte depressive Episode
- F32.00: ohne somatische Symptome
- F32.01: mit somatischen Symptomen
- F32.1: Mittelgradige depressive Episode
- F32.10: ohne somatische Symptome
- F32.11: mit somatischen Symptomen
- F32.2: Schwere depressive Episode ohne psychotische Symptome
- F32.3: Schwere depressive Episode mit psychotischen Symptomen
- F32.30: synthyme psychotische Symptome
- F32.31: parathyme psychotische Symptome
- F32.8: Sonstige depressive Episode
- F32.9: Nicht näher bezeichnete depressive Episode

ICD-11

Es gibt im Vergleich zur ICD-10 eine Reihe bedeutsamer Änderungen in der ICD-11. Das Kapitel psychische Störungen ist nun das 6. Kapitel mit dem Titel „Psychische Störungen sowie Störungen des Verhaltens und der neurologischen Entwicklung“ (Übersetzung durch den Autor). In der grundsätzlichen Beschreibung von ► Kap. 6 formuliert die WHO Folgendes:

- » Psychische, verhaltensbezogene und neurologische Entwicklungsstörungen umfassen Syndrome, die durch klinisch bedeutsame Störungen der individuellen Kognitionen, Emotionsregulation oder Entwicklungsprozesse gekennzeichnet sind, die den psychischen oder Verhaltensfunktionen zugrunde liegen. Diese Störungen gehen in der Regel mit Belastungen oder Beeinträchtigungen in persönlichen, familiären, sozialen, bildungsbezogenen, beruflichen oder anderen bedeutsamen Lebensbereichen einher. (WHO 2019; Übersetzung durch den Autor).

Eine wesentliche Änderung (ähnlich wie in DSM-5) im ICD-11 ist für die Diagnostik von Persönlichkeitsstörungen vorgesehen. Mit Ausnahme der Borderline-Persönlichkeitsstörung wird dort eine durchgehend dimensionale Klassifikation verfolgt. Sind die allgemeinen Kriterien für das Vorliegen einer Persönlichkeitsstörung erfüllt, wird der Schweregrad der Störung (leicht, mäßig, schwer) auf der Basis der festgestellten Funktionsbeeinträchtigung skaliert. Zusätzlich werden die Domänen „negative Affektivität“, „Dissozialität“, „Enthemmung“, „Zwanghaftigkeit“ und „Distanziertheit“ beurteilt.

8.3.1.2 Diagnostisches und Statistisches Manual Psychischer Störungen

Wesentlicher Unterschied zwischen dem DSM und der ICD ist, dass das von der WHO herausgegebene ICD den Anspruch verfolgt, die diagnostischen Kriterien für sämtliche Erkrankungen zu spezifizieren. Das Kapitel F (psychische Störungen) ist nur eines von 21 Kapiteln bzw. Erkrankungsgruppen. Dies führt dazu, dass sowohl die Beschreibung der einzelnen Störungen und ihrer Symptome als auch die theoretischen, epidemiologischen und wissenschaftlichen Begründungen der einzelnen Diagnosen wesentlich kürzer ausfallen als im DSM. Dies zeigt sich nicht zuletzt darin, dass das DSM mit 1300 Seiten etwa 6× so umfangreich ist wie das Kapitel F der ICD-10. Das DSM (und die darauf fußenden diagnostischen Verfahren) beschreiben die Prozedur, mit der einzelne Diagnosen abgeleitet und überprüft werden können, umfassender und sehr viel eindeutiger. Hierdurch ergibt sich auch eine höhere Zuverlässigkeit (Reliabilität) des DSM-Systems im Vergleich zur ICD-10. Daher hat das DSM als Kompendium vor allem in der Forschung zu psychischen Störungen und in der Epidemiologie einen besonders hohen Stellenwert. Darüber hinaus können die nach DSM gestellten Diagnosen über die jeweiligen diagnostischen Referenzziffern nahezu ohne Ausnahme in ICD-Diagnosen überführt werden.

Von dem bis zur 4. Revision (DSM-IV) vorhandenen sog. „axialen System“ mit 5 Achsen (klinische Störungen, Persönlichkeitsstörungen und geistige Behinderungen, medizinische Krankheitsfaktoren, psychosoziale und umweltbedingte Probleme sowie Beurteilung des globalen Funktionsniveaus) wurde in der 5. Revision des DSM (DSM-5) aus methodischen Gründen (vor allem geringe oder nicht vorhandene diskriminative Validität oder geringe Reliabilität einzelner Faktoren) Abstand genommen.

Die Autorinnen und Autoren der amerikanischen Originalversion des DSM-5 weisen in ihrem Vorwort auf eine Reihe von Punkten hin, die bei der Erarbeitung des DSM berücksichtigt wurden (American Psychiatric Association 2013). Hierzu gehören die Berücksichtigung entwicklungsbezogener Aspekte bei der Beschreibung der Erkrankungen, die Integration neuer genetischer Befunde sowie von Ergebnissen der Bildgebungsforschung, Zusammenfassungen ähnlicher Krankheitsgruppen (z. B. Autismus-Spektrum-Störung), die Vereinfachung der Klassifikation von bipolaren und depressiven Störungen, eine konsistentere Strukturierung der substanzbezogenen Störungen, eine Erhöhung der Spezifität für leichte und schwere kognitive Störungen sowie eine neue (dimensionale) Konzeptualisierung der

Grundlegende Unterschiede zwischen ICD und DSM

Vergleich DSM-IV/DSM-5

DSM-5

Persönlichkeitsstörungen, die parallel zur herkömmlichen kategorialen Ein- teilung dieser Störungsgruppe vorgestellt wird.

Im DSM-5 werden dimensionale Aspekte psychischer Erkrankungen bzw. Symptome deutlich stärker als in früheren Ausgaben oder im ICD-10 berücksichtigt (s. u.).

Die 22 Hauptkategorien bzw. Störungsgruppen nach DSM-5

1. Störungen der neuronalen und mentalen Entwicklung
2. Schizophrene Spektrum und andere psychotische Störungen
3. Bipolare und verwandte Störungen
4. Depressive Störungen
5. Angststörungen
6. Zwangsstörungen und verwandte Störungen
7. Trauma- und belastungsbezogene Störungen
8. Dissoziative Störungen
9. Somatische Belastungsstörung und verwandte Störungen
10. Fütter- und Essstörungen
11. Ausscheidungsstörungen
12. Schlaf-Wach-Störungen
13. Sexuelle Funktionsstörungen
14. Geschlechtsdysphorie
15. Disruptive Impulskontroll- und Sozialverhaltensstörungen
16. Störungen im Zusammenhang mit psychotropen Substanzen und abhängigen Verhaltensweisen
17. Neurokognitive Störungen
18. Persönlichkeitsstörungen
19. Paraphile Störungen
20. Andere psychische Störungen
21. Medikamenteninduzierte Bewegungsstörungen und andere unerwünschte Medikamentenwirkungen
22. Andere klinisch relevante Probleme

8

Persönlichkeitsstörungen im DSM-5

Seit den 1990er-Jahren gibt es eine umfassende Diskussion über die Validität von Diagnosen dieser Kategorien. Dem grundsätzlich kategorialen Ansatz der Diagnostik wird auf der Basis zahlreicher Forschungsarbeiten ein dimensionaler Ansatz gegenübergestellt. Im Ergebnis wird im DSM-5 sowohl eine überarbeitete Version der (kategorial definierten) Persönlichkeitsstörungen nach DSM-IV als auch ein dimensionales Konzept der Diagnostik von Persönlichkeitsstörungen dargestellt.

Kategoriale definierte Diagnosen

- Paranoide Persönlichkeitsstörung
- Schizoide Persönlichkeitsstörung
- Schizotypische Persönlichkeitsstörung
- Antisoziale Persönlichkeitsstörung
- Borderline-Persönlichkeitsstörung
- Histrionische Persönlichkeitsstörung
- Narzisstische Persönlichkeitsstörung
- Vermeidend-selbstunsichere Persönlichkeitsstörung
- Dependente Persönlichkeitsstörung
- Zwanghafte Persönlichkeitsstörung

Die dimensionale Perspektive fasst Persönlichkeitsstörungen als „unangepasste Varianten von Persönlichkeitszügen mit fließenden Übergängen sowohl zur Normalität als auch zueinander“ auf (American Psychiatric Association 2015, S. 885). Grundsätzlich gilt, dass bei der Diagnostik von Persönlichkeitsstörungen immer die allgemeinen Kriterien für Persönlichkeitsstörungen sorgfältig geprüft werden müssen (American Psychiatric Association 2015, S. 1045 f.). Weiterhin werden das Funktionsniveau der Persönlichkeit, problematische Persönlichkeitsmerkmale sowie die sog. „Durchgängigkeit“ und Stabilität der Merkmale festgestellt. Es werden nur noch 6 spezifische Störungen gelistet – die antisoziale, vermeidend-selbstunsichere, Borderline-, narzisstische, zwanghafte und schizotypale Persönlichkeitsstörung. Im alternativen Modell nach DSM-5 wird also eine Kombination von charakteristischen Persönlichkeitsmerkmalen mit dimensionalen Aussagen zum Funktionsniveau sowie zu Domänen und Facetten von Persönlichkeitsmerkmalen kombiniert.

Persönlichkeitsstörungen sind im ICD-10 als Kodierungen unter der Hauptgruppe F6 zu finden. Dort werden jedoch lediglich 7 spezifische Persönlichkeitsstörungen genannt, die für die genannten weitgehend denen im DSM-5 entsprechen; 2 weitere Störungen (die schizotypische und narzisstische Persönlichkeitsstörung) sind in der ICD-10 entweder in einer anderen diagnostischen Hauptgruppe oder unter „andere Persönlichkeitsstörungen“ zu finden.

8.3.1.3 Bewertung und Vergleich der Klassifikationssysteme

Ein großer Vorteil der operationalen Definition psychischer Störungen besteht darin, dass die Zuverlässigkeit (*Reliabilität*) der Diagnosen und damit auch die *Validität* der klinischen Diagnostik psychischer Störungen im Vergleich zu früheren Klassifikationssystemen deutlich verbessert werden konnten. Die Abkehr vom Krankheitsbegriff zugunsten des Begriffs der psychischen Störung dokumentiert zudem eine tendenzielle Abwendung vom organisch orientierten Krankheitsmodell bei psychischen Störungen.

Erhöhung von Reliabilität und Validität

Die WHO und die American Psychiatric Association bemühten sich um eine möglichst weitgehende Abstimmung der Klassifikationssysteme ICD-10/ICD-11 bzw. DSM-5. Wenngleich die Anzahl der Hauptgruppen in beiden Klassifikationssystemen unterschiedlich ist, so sind die einzelnen Diagnosen doch in wesentlichen Punkten *gut vergleichbar*. Querverweise über ICD-Diagnoseschlüssel im DSM-5 helfen, die korrespondierende Diagnose nach ICD-10 festzustellen. Eine gewisse Ausnahme stellt die Diagnostik von Persönlichkeitsstörungen dar, bei denen es sowohl hinsichtlich der Diagnosen selbst als auch der konkreten Operationalisierungen teilweise deutliche Unterschiede zwischen der ICD und dem DSM gibt.

Diagnosen gut vergleichbar

Sowohl für die ICD-10 als auch für das DSM 5 gilt das sog. „Komorbiditätsprinzip“. Dies bedeutet, dass es kein hierarchisches Vorgehen bei der Feststellung von Diagnosen gibt. Das heißt, wann immer die Kriterien für mehrere psychische Störungen erfüllt sind, sollen diese auch aufgeführt werden; es sei denn, dass bei einzelnen Diagnosen explizit darauf hingewiesen wird, wenn bestimmte Ko-Diagnosen nicht gestellt werden dürfen. Der Begriff der Komorbidität kommt aus dem Bereich der Organmedizin und bedeutet ursprünglich, dass bei einer Person zu einem gegebenen Zeitpunkt mehrere Krankheiten vorliegen. Bei der Verwendung des Begriffs im Bereich psychischer Störungen ist jedoch zu bedenken, dass zum einen psychische Störungen meist keine Krankheitseinheiten entsprechend des organmedizinischen Modells sind und zum anderen viele der Symptome in ähnlicher Form bei unterschiedlichen Diagnosen auftreten (Symptomüberlappung) und dadurch die Komorbiditätsrate artifiziell erhöht sein kann.

Komorbiditätsprinzip

Weiterführende Literatur und Internetressourcen

Für einen umfassenderen Überblick über die Kodiersysteme ICD-10 und DSM-5 sei auf Wittchen und Hoyer (2011) sowie das DSM-5 (American Psychiatric Association 2013, 2015, 2018) selbst verwiesen. Die genauen Kriterien sowie die aktuelle Form der ICD-10-GM finden sich online unter: ► <https://www.icd-code.de>.

8.3.2 Diagnostische Verfahren zur Klassifikation psychischer Störungen

Reliabilität strukturierter Diagnostik

Untersuchungen zur Reliabilität der klassifikatorischen Diagnostik zeigen, dass psychische Störungen mit strukturierten, halbstrukturierten oder standardisierten Interviews mit mindestens zufriedenstellender Zuverlässigkeit erhoben werden können (► Abb. 8.2). Zuverlässigkeitskennwerte gibt es jedoch nur für standardisierte oder strukturierte diagnostische Verfahren. Für die meisten der diagnostizierbaren Störungen werden Kappa-Werte für die Interrater- und/oder Test-Retest-Reliabilität zwischen ,50 und ,95 festgestellt (z. B. Fydrich et al. 1996; In-Albon et al. 2008). Dabei gibt es teilweise deutliche Unterschiede in der Zuverlässigkeit für die Diagnostik verschiedener Störungen. So ist – selbst beim Einsatz strukturierter oder standardisierter Diagnostik – die Reliabilität für die Diagnostik somatoformer Störungen geringer als für die meisten anderen Störungen und liegt in verschiedenen Studien oft unter einem Zuverlässigkeitswert (Kappa) von ,50. Zu den Faktoren, die die Zuverlässigkeit der Interviews beeinflussen, gehören Güte und Umfang der Ausbildung und des Trainings der Interviewenden sowie die Methodik der Interviewverfahren selbst. Voll standardisierte Interviews (z. B. das *Composite International Diagnostic Interview, CIDI*; ein computerisiertes Verfahren, bei dem die Fragen und deren Reihenfolge vom Computer vorgegeben werden) haben eine höhere Reliabilität als Interviewverfahren, in denen seitens der Interviewenden auch klinische Einschätzungen notwendig sind (wie z. B. im *Strukturierten Klinischen Interview für DSM-Diagnosen, SKID*).

Am häufigsten eingesetzte Verfahren

Der Einsatz standardisierter und strukturierter Diagnoseverfahren setzt in der Regel ein systematisches Training, meist auch klinische Erfahrungen voraus. Im deutschsprachigen Raum steht eine Reihe strukturierter und standardisierter Interviews zur klassifikatorischen Diagnostik psychischer Störungen zur Verfügung. Am häufigsten werden folgende Interviews eingesetzt:



► Abb. 8.2 Strukturierte klinische Interviews, auch solche, die sich an den Klassifikationskriterien von DSM oder ICD orientieren, finden im Einzelgespräch statt

Strukturierte und standardisierte Interviews zur klassifikatorischen Diagnostik psychischer Störungen (Auswahl)

— Für die Diagnostik von *Symptomstörungen*:

- IDCL: Internationale Diagnose Checklisten für ICD-10 (Hiller et al. 1997)
- SKID-5-CV: Strukturiertes Klinisches Interview für DSM-5-Störungen – Klinische Version (Beesdo-Baum et al. 2019)
- DIPS Open Access: Diagnostisches Interview bei psychischen Störungen (Margraf et al. 2017)
- Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen (Margraf und Cwik 2017)
- Kinder-DIPS Open Access: Diagnostisches Interview bei psychischen Störungen im Kindes- und Jugendalter (Schneider et al. 2017)
- CIDI und DIA-X-CIDI: Composite International Diagnostic Interview (Wittchen et al. 1997); DIA-X bezeichnet die voll strukturierte und computerisierte Version dieses Interviews.

— Für die Diagnostik von *Persönlichkeitsstörungen*:

- IDCL-P: Internationale Diagnose Checklisten für Persönlichkeitsstörungen (Bronisch et al. 1995)
- SCID-5-PD: Strukturiertes Klinisches Interview für DSM-5 – Persönlichkeitsstörungen (Beesdo-Baum et al. 2019)
- IPDE: International Personality Disorder Examination (Deutsch: ICD-10-Modul; Mombour et al. 1996)

Der Vergleich dieser Verfahren zeigt, dass Checklistenverfahren (IDCL und IDCL-P) den geringsten Durchführungsaufwand erfordern und daher im klinischen Alltag vergleichsweise ökonomisch eingesetzt werden können. Allerdings können damit meist nur konfirmatorische Diagnosen gestellt werden. Das bedeutet, dass Verdachtsdiagnosen zwar bestätigt, aber mögliche andere oder zusätzliche („komorbide“) Störungen übersehen werden können.

Die strukturierten klinischen Interviews (SKID-5-CV und -5-PD; DIPS Open Access) sowie das CIDI sind die vergleichsweise zuverlässigsten Verfahren zur umfassenden kategorialen Diagnostik psychischer Störungen. Sie werden im klinischen Alltag wegen des relativ hohen Zeitaufwands von 2 oder mehr Stunden oft nicht eingesetzt. Zur Vereinfachung wird daher im klinischen Kontext den strukturierten Verfahren oft ein Screening-Interview oder ein Screening-Fragebogen vorgeschaltet, sodass im Anschluss nur relevante Teile des Interviews durchgeführt werden müssen.

Das CIDI als voll computerisierte Interviewversion kann nach Empfehlung der Autoren auch von trainierten Interviewerinnen und Interviewern ohne umfassendes klinisches Training durchgeführt werden, da die Antworten der Patientinnen bzw. Patienten nicht klinisch beurteilt werden müssen. Dies führt zur Besonderheit, dass das CIDI eine besonders hohe Sensitivität hat (Störungen werden schon bei geringer Symptomausprägung als vorhanden eingestuft); leider geht jedoch damit für einige diagnostische Bereiche eine geringe Spezifität einher, was zu einer Verringerung der diskriminanten Validität führt. Das CIDI wurde in den für die Epidemiologie psychischer Störungen sehr wichtigen Studien des Robert-Koch-Instituts (Bundesgesundheitssurvey 1999 und 2009) eingesetzt, die Befunde über recht hohe (Jahres-)Prävalenzen psychischer Erkrankungen in Höhe von rund 30 % der Bevölkerung aufweisen.

Ökonomische Checklistenverfahren

Zuverlässige standardisierte und strukturierte Interviews

Hohe Sensitivität und geringe Spezifität beim CIDI

Zusätzlich zur strukturierten oder standardisierten Klassifikation psychischer Störungen wird zur systematischen Erhebung des psychopathologischen Befunds, vor allem im psychiatrischen Bereich, ein weiteres Beurteilungssystem, erarbeitet von der „Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie“, eingesetzt, das AMDP-System (AMDP 2018). Damit werden durch den/die Diagnostiker/-in bzw. Kliniker/-in (als Fremdeinschätzung) 100 ggf. vorliegende psychische und 40 somatische Symptome beurteilt, aus denen Informationen über 9 Syndrome sowie 3 übergeordnete Syndromkomplexe (paranoid-halluzinatorische, depressive und psychoorganische Symptomatik) gewonnen werden. Informationsquellen sind hierbei entweder die Patientin bzw. der Patient selbst und/oder Dritte (z. B. Verwandte, Pflegepersonal).

8.4 Psychometrische Verfahren

8.4.1 Verhaltenstheoretisch und kognitiv orientierte Fragebogenverfahren

8

Kategoriale und dimensionale Diagnostik

Vorteile von Fragebogenverfahren

Standardisierte und strukturierte diagnostische Verfahren entsprechend der Kriterien der ICD oder des DSM führen zu kategorialen Urteilen über das Vorliegen psychischer Störungen. Demgegenüber werden mit Fragebögen in der Regel dimensionale Befunde über den Ausprägungsgrad der Problematik bzw. Symptomatik erfasst. Beispielsweise kann die Prüfung vorliegender Symptomatik auf Basis der IDC-10-Kriterien zu einer Diagnose einer „Depressiven Episode“ führen, und der Einsatz des Beck Depressions-Inventars (BDI; Testverfahren zur Erfassung depressiver Symptome) führt zu Informationen über die Stärke der Depressivität (Abb. 8.3). Wichtig ist, dass in der Regel mit einem psychometrischen Testverfahren keine Aussage über das Vorliegen einer psychischen Störung gemacht werden kann, da mit Testverfahren meist nicht systematisch die in den Klassifikationssystemen für psychische Störungen festgelegten diagnostischen Kriterien überprüft werden.

Ein Vorteil der Verwendung von *Fragebogenverfahren* (nicht nur) in der Klinischen Psychologie und der Psychotherapie liegt darin, dass sie vergleichsweise kostengünstig und zeitsparend eingesetzt werden können. Weiterhin werden sie zumeist den methodischen Anforderungen nach Objektivität, Reliabilität und Validität weitgehend gerecht und sie sind dazu geeignet, eine dimensionale Einschätzung des Schweregrads einer Symptomatik vorzunehmen.

Im Folgenden werden einige wichtige Verfahren genannt. Nähere Details sind den Empfehlungen von Fydrich (2011) zur Diagnostik im Bereich der Psychotherapie im Erwachsenen- sowie im Kinder- und Jugendlichenbereich zu entnehmen.

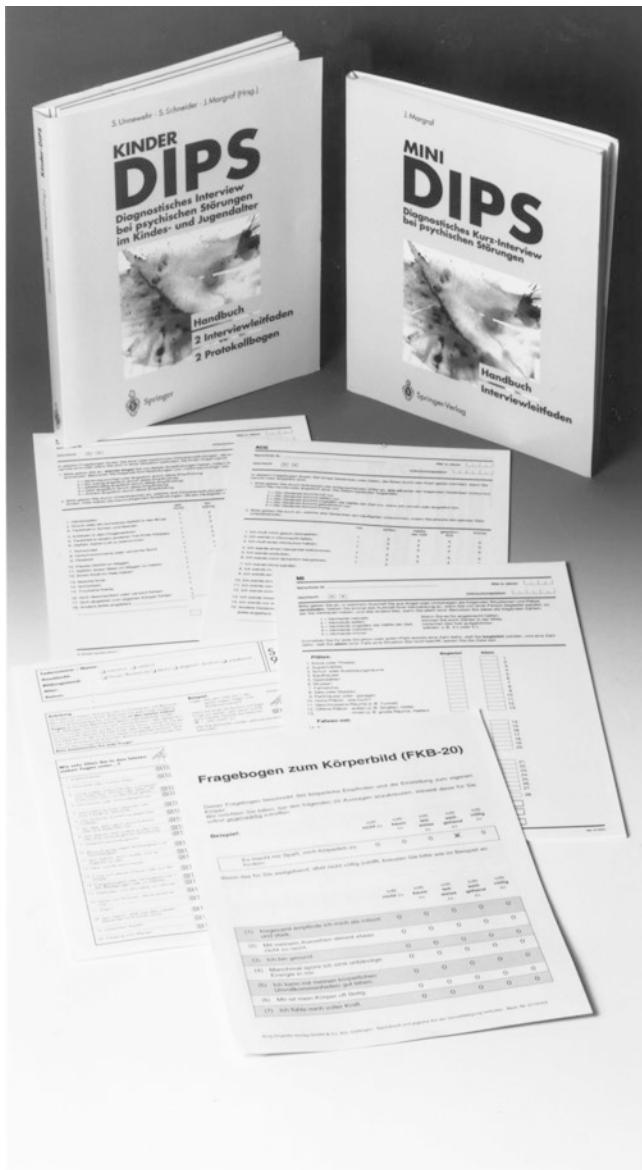


Abb. 8.3 Strukturierte Interviews und Fragebogenverfahren werden zur klassifikatorischen Diagnostik sowie zur Diagnostik verschiedener Aspekte psychischer Störungen eingesetzt

Psychodiagnostische Verfahren für die Psychotherapie bei Erwachsenen (Auswahl)	
Störungsumfassende Diagnostik; Lebensqualität	
BSCL und Mini-SCL: Brief-Symptom-Checklist und Mini-Symptom-Checklist zur Erfassung der subjektiven Beeinträchtigung durch körperliche und psychische Symptome (Franke 2017a, b; Kurzformen des SCL-90-S) CGI: Clinical Global Impression (Busner und Targum 2007) ISR: ICD-10-Symptom-Rating (Tritt et al. 2015) OQ-45.2: Outcome Questionnaire-45.2 (Lambert et al. 2004) SCL-90-S: Symptom-Checklist-90-Standard von Derogatis, deutsche Fassung (Franke 2014) SF-36 (SF-12): Fragebogen zum Gesundheitszustand (Morfeld et al. 2011)	
Interaktions- und Beziehungsdiagnostik; Persönlichkeitsdiagnostik	
HAQ: Helping Alliance Questionnaire, deutsche Fassung (Bassler et al. 1995) WAI-SR: Working Alliance Questionnaire – short revised, deutsche Fassung (Wilmers et al. 2008) IIP-D: Inventar zur Erfassung interpersonaler Probleme – Deutsche Version (Horowitz et al. 2016) PSSI: Persönlichkeits-Stil- und Störungs-Inventar (Kuhl und Kazén 2009)	
Störungsspezifische Diagnostik	
Alkoholabhängigkeit/-missbrauch	LAST: Lübecker Alkoholabhängigkeits- und -missbrauchs-Screening (Rumpf et al. 2001) SESA: Skala zur Erfassung der Schwere der Alkoholabhängigkeit (John et al. 2001) FFT: Fragebogen zum funktionalen Trinken (Belitz-Weihmann und Metzler 1997) WHO ASSIST: Alcohol, Smoking and Substance Involvement Screening, deutsche Übersetzung (Schütz et al. 2005)
Ängste und Phobien	AKV: Fragebogen zu körperbezogenen Ängsten, Kognitionen und Vermeidung (Ehlers et al. 2001) BAI: Beck Angst-Inventar (Margraf und Ehlers 2007) PAS: Panik- und Agoraphobie-Skala (Bandelow 2016) SPA: Soziale Phobie und Angst Inventar (Fydrich 2002) SIAS: Soziale-Interaktions-Angst-Skala (von Consbruch et al. 2016) SPS: Soziale-Phobie-Skala (von Consbruch et al. 2016) LSAS: Liebowitz-Soziale-Angst-Skala (von Consbruch et al. 2016) PSWQ: Penn State Worry Questionnaire (Stöber und Bittencourt 1998) IES-R: Impact of Event Scale – revidierte Form (Maercker und Schützwohl 1998) PSS: PTSD ^a Symptom Scale (Foa et al. 1993) MPSS: Modified PTSD ^a Symptom Scale, deutsche Version (Spitzer et al. 2001) HAMA: Hamilton Anxiety Scale (Hamilton 1959)
Depressivität	ADS: Allgemeine Depressionsskala (Hautzinger et al. 2012) BDI-II: Beck Depressions-Inventar, Revision (Hautzinger et al. 2009) HAM-D: Hamilton Depressionsskala (Hamilton 1986) HADS-D: Hospital Anxiety and Depression Scale – Deutsche Version (Herrmann-Lingen et al. 2010)
Essstörungen	EDI-2: Eating Disorder Inventory-2, deutsche Version (Paul und Thiel 2004)
Psychotische Störungen Somatoforme Störungen/Schmerz	BPRS: Brief Psychiatric Rating Scale (CIPS 2015) PANSS: Positive and Negative Syndrome Scale (Kay et al. 1989) SOMS: Screening für Somatoforme Störungen (Rief und Hiller 2007) KSI: Kieler Schmerz-Inventar (Hasenbring 1994) SES: Die Schmerzempfindungs-Skala (Geissner 1996) MASK-P: Multiaxiale Schmerzklassifikation. Psychosoziale Dimension (Klinger et al. 2016) WI-IAS: Internationale Skalen für Hypochondrie (Hiller et al. 2004)
Zwangsstörungen	Y-BOCS: Yale-Brown Obsessive Compulsive Scale (Goodman et al. 1989) OCI-R: Obsessive-Compulsive Inventory-Revised, deutsche Version (Gönner et al. 2007)
Borderline-Störung und Dissoziation	BSL: Borderline-Symptom-Liste (Bohus et al. 2001) FGG: Fragebogen zu Gedanken und Gefühlen (Renneberg und Seehausen 2010) FDS: Fragebogen zu Dissoziativen Symptomen (Spitzer et al. 2005) SKID-D: Strukturiertes Klinisches Interview für Dissoziative Störungen (Gast und Rodewald 2004)

^aPosttraumatic Stress Disorder (Posttraumatische Belastungsstörung)

Psychodiagnostische Verfahren für die Psychotherapie bei Kindern und Jugendlichen (Auswahl)	
Breitband- und Screeningverfahren	
DISYPS-III: Diagnostik-System für psychische Störungen nach ICD-10 und DSM-5 für Kinder und Jugendliche – III (Döpfner und Götz-Dorten 2017)	
CBCL: Child Behavior Checklist in verschiedenen Versionen: – CBCL 1½–5: Versionen für Eltern von Kindern im Alter von 18 Monaten bis 5 Jahren (Achenbach & Arbeitsgruppe Deutsche Child Behavior Checklist 2000) – CBCL/6–18R, TRF/6–18R, YSR/11–18R: Versionen für 4- bis 18-Jährige und als Version YSR für 11- bis 18-Jährige als Fragebogen sowie als TRF = Teacher's Report Form für Lehrer (Döpfner et al. 2014)	
MEI: Mannheimer Elterninventar (Interviewverfahren; Esser et al. 1989)	
VBV 3–6: Verhaltensbeurteilungsbogen für Vorschulkinder (Döpfner et al. 2018)	
ILK: Inventar zur Erfassung der Lebensqualität bei Kindern und Jugendlichen (Mattejat und Remschmidt 2003)	
SDQ: Fragebogen zu Stärken und Schwächen (The Strengths and Difficulties Questionnaire; Goodman 1997; dt. ► https://psydix.org/psychologische-testverfahren/sdq-d/)	
Störungsspezifische Verfahren	
Aufmerksamkeitsdefizit-/Hyperaktivitätsstörungen und oppositionelle Verhaltensstörungen	DISYPS-III: Spezielle Skalen aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (Döpfner und Götz-Dorten 2017) EL-PF: Elterninterview über Problemsituationen in der Familie (Döpfner et al. 2005) DAT: Dortmunder Aufmerksamkeitstest (Lauth 2003)
Angststörungen	SPAIK: Sozialephobie- und -angstinventar für Kinder (Melfsen et al. 2001) KAT-III: Kinder-Angst-Test-III (Tewes und Naumann 2016) DAI: Differentielles Leistungsangst Inventar (Rost und Schermer 2007) AFS: Angstfragebogen für Schüler (Wieczorkowski et al. 2016) DISYPS-III: Spezielle Skalen aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (Döpfner und Götz-Dorten 2017)
Zwangsstörungen	CY-BOCS: Children's Yale-Brown Obsessive Compulsive Scale, deutsche Fassung (Döpfner 1999) HZI-K: Hamburger Zwangsinventar – Kurzversion (ab 16 Jahre; Klepsch et al. 1993) LOI-K: Leyton Obsessive Compulsive Inventory, deutsche Fassung (Döpfner 1999) DISYPS-III: Spezielle Skalen aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (Döpfner und Götz-Dorten 2017)
Störungen des Sozialverhaltens	STAXI-2: State-Trait-Ärgerausdrucks-Inventar – 2 (ab 16 Jahre; Rohrmann et al. 2013) STAXI-2 kJ: State-Trait-Ärgerausdrucks-Inventar – 2 für Kinder und Jugendliche (von 9 bis 16 Jahren; Kupper und Rohrmann 2016) EAS: Erfassungsbogen für aggressives Verhalten in konkreten Situationen (Petermann und Petermann 2015) DISYPS-III: Spezielle Skalen aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (Döpfner und Götz-Dorten 2017) BAV: Beobachtungsbogen für 3- bis 6-jährige Kinder (Frey et al. 2008)
Depressive Störungen	DTK-II: Depressionstest für Kinder – II (Rossmann 2004) DIKJ: Depressionsinventar für Kinder und Jugendliche (Stiensmeier-Pelster et al. 2014) BDI-II: Beck Depressions-Inventar, Revision (ab 16 Jahre; Hautzinger et al. 2009) ADS: Allgemeine Depressionsskala (ab 16 Jahre; Hautzinger et al. 2012) DISYPS-III: Spezielle Skalen aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (Döpfner und Götz-Dorten 2017)
Essstörungen	ANIS: Anorexia-Nervosa-Inventar zur Selbstbeurteilung (Fichter und Keeser 1980) EDI-2: Eating Disorder Inventory-2 (Paul und Thiel 2004) SIAB: Strukturiertes Inventar für Anorektische und Bulimische Essstörungen (Fichter und Quadflieg 1999) EAT: Eating Attitude Test, deutsche Fassung (Steinhausen und von Asten 1999)
Somatisierungsstörungen	GBB-KJ: Gießener Beschwerdebogen für Kinder und Jugendliche (Barkmann und Brähler 2009)
Tic-Störungen	Yale-Tourette-Syndrom-Symptomliste (Steinhausen und von Asten 1999)

Weiterführende Literatur

Weitere Informationen sind den Übersichten von Ahle et al. (2006), Fydrich (2011), Horn (2006) sowie Schneider und Margraf (2019) zu entnehmen.

8.4.2 Beobachtungsmethoden

Systematische Beobachtungen sind vor allem im Bereich der Diagnostik von Verhaltensweisen bei Kindern und Jugendlichen von Bedeutung. Sie finden aber auch systematische Anwendung in der Forschung sowie bei der Selbst- oder Fremdbeobachtung von Personen in störungsrelevanten Situationen (z. B. in angstauslösenden Situationen). Folgende Aspekte der Kategorisierung von Beobachtungsmethoden können unterschieden werden:

- In-vivo-Beobachtung (in der natürlichen Umgebung)
- Strukturierte Beobachtung (meist im Labor oder in einer „künstlichen“ Umgebung)
- Selbstbeobachtung
- Verhaltenstests

In-vivo-Beobachtungen finden überwiegend in natürlichen Umgebungen statt. Anwendungsbeispiele sind etwa die Beobachtung aggressiven Verhaltens von Kindern im Kindergarten oder in der Schule oder – bei Erziehungsproblemen – die Beobachtung der Interaktionen von Eltern und Kindern in ihrer Wohnung.

Bei *strukturierten Beobachtungen* kann den Klientinnen und Klienten gezielt eine Aufgabe gestellt werden. Beispielsweise werden Paare instruiert, sich über ein konfliktbehaftetes Thema auseinanderzusetzen, Personen mit sozialen Ängsten können im Rahmen diagnostischer Rollenspiele aufgefordert werden, mit einer ihnen unbekannten Person ein Gespräch zu beginnen und weiterzuführen, und Personen mit phobischen Ängsten können beobachtet werden, wie sie sich einer gefürchteten Situation nähern und welche Reaktionen dabei auftreten. Die Verhaltensbeobachtungen folgen dabei zu meist konkreten Beobachtungsrichtlinien und Kriterien. Als abhängige Variablen können beispielsweise die Häufigkeit, die Dauer und die Intensität einer Aktivität erhoben werden. Möglich ist auch eine Auswertung und Beurteilung, beispielsweise der beobachtbaren sozialen Kompetenz (Fydrich und Bürgener 2005).

Verfahren zur *Selbstbeobachtung* eignen sich besonders gut, um im klinischen Kontext problematische Verhaltensweisen, Kognitionen, Gefühle und körperliche Reaktionen in der alltäglichen Umwelt systematisch zu beobachten und zu protokollieren. So werden beispielsweise in gesundheitspsychologischen Programmen zum Beenden des Rauchens die Teilnehmenden instruiert, jeweils die Situation, in der geraucht wurde, sowie die vorausgegangenen und begleitenden Gedanken zu notieren. Bei der Behandlung von Patientinnen und Patienten mit (chronischen) Schmerzen können sog. „Schmerztagebücher“ dabei helfen, einen Zusammenhang zwischen körperlichen oder psychischen Belastungen und Schmerzhäufigkeit und -intensität herauszufinden. Mit Schmerztagebüchern wird die Stärke und Frequenz von Schmerzen sowie der Zusammenhang mit Aktivitäten, Gedanken und Situationen über den Tag hinweg von den Patientinnen und Patienten protokolliert. Ein weiteres Beispiel für Selbstbeobachtung ist, dass betroffene Personen im Rahmen einer kognitiven Therapie bei Depressionen „dysfunktionale“ Gedanken und die Situationen, in denen sie auftreten, beobachten und protokollieren.

Eine Kombination von (durch den Therapeuten bzw. die Therapeutin) strukturierter Beobachtungssituation und Selbstbeobachtung im natürlichen Umfeld stellen sog. „Verhaltenstests“ dar. So kann beispielsweise eine Person, die unter einer Agoraphobie leidet, aufgefordert werden, den gefürchteten Supermarkt nahe der eigenen Wohnung aufzusuchen. Dabei soll sie die Intensität der erlebten Angst auf einer Skala von 0 bis 100 einzustufen. Wenn derartige Verhaltenstests zu Beginn, am Ende und eventuell auch bei einem Nachuntersuchungstermin eingesetzt werden, eignen sie sich auch

Strukturierte Beobachtungen

Selbstbeobachtung: eigenes Verhalten beobachten und protokollieren

Verhaltenstests zur Selbstbeobachtung und zur Überprüfung von Befürchtungen

gut zur verhaltensorientierten Erfassung des Therapieerfolgs. Verhaltens-
tests werden zudem in der Verhaltenstherapie oft als therapeutische Maß-
nahmen eingesetzt. Dabei werden Patientinnen oder Patienten instruiert,
in (meist lange Zeit vermiedene) Situationen zu gehen, um individuelle Be-
fürchtungen zu überprüfen. Beispielsweise kann die Befürchtung einer Per-
son mit sozialen Ängsten, von vielen anderen Menschen in einer Gaststätte
beobachtet zu werden, dadurch überprüft werden, dass sie sich im Rahmen
der Behandlung in diese Situation begibt und dabei darauf achtet, von wem
und von wie vielen Personen sie tatsächlich beobachtet wird.

Nachteile der Beobachtungsverfahren sind eine häufig unzureichende
Reliabilität (z. B. durch Verzerrung der Einschätzung durch Vorinformation
oder Halo-Effekte) und die nicht bekannte oder geringe Validität (Reprä-
sentativität der Beobachtungssituation und der beobachteten Verhaltensan-
teile). Dies liegt u. a. an der Reaktivität des Beobachtungsprozesses: Das
zu beobachtende Verhalten, die Gedanken oder andere Reaktionen werden
allein schon durch die Beobachtung (d. h. ohne weitere Intervention) häu-
fig verändert. Dieses Problem kann dadurch etwas entschärft werden, dass
statt partizipierender Beobachtung Videoaufzeichnungen vorgenommen
werden, die später von Beobachterinnen und Beobachtern beurteilt werden.
Leichter sind solche Vorehrungen in entsprechend eingerichteten Räumen
(Beratungsstelle, Labor) zu realisieren, in denen Beobachtungen ggf. auch
durch Einwegscheiben möglich sind.

Nachteile

Während die Reaktivität von Verhaltensbeobachtungen einerseits im
Hinblick auf die Validität der Beobachtung problematisch sein kann, ist sie
andererseits in manchen anderen Bereichen klinisch wünschenswert. So hat
die Anleitung zur Selbstbeobachtung in einem Programm zur Raucherin-
nen- und Raucherentwöhnung durchaus zum Ziel, schon dadurch das Rau-
chen zu reduzieren.

Spezifische Techniken und Verfahren zur Diagnostik störungs- und the-
rapierelevanter Kognitionen werden ausführlicher von Bastine und Tuschen
(1996) dargestellt und diskutiert. Hierzu gehören z. B. die Strategie der „Als-
ob-Methode“, bei der Patienten bzw. Patientinnen die entsprechende prob-
lematische Situation nachspielen oder sich diese innerlich vorstellen und die
dabei auftretenden Gedanken exploriert werden; außerdem die Technik des
Gedankenauflistens (thought listing), das stichprobenmäßige Erfassen von
Gedanken (thought sampling), die Methode des lauten Denkens sowie der
oben schon erwähnte Einsatz von Tagebüchern bei der Selbstbeobachtung.

8.4.3 Persönlichkeitstests in der Klinischen Psychologie und Psychotherapie

Persönlichkeitstests liegt in der Regel das Konzept zugrunde, dass die da-
mit erfassten Merkmale relativ zeitstabile Erlebens- und Verhaltenswei-
sen sind (Trait-Konzept). Daher werden in der Klinischen Psychologie und
Psychotherapie bei einigen Fragestellungen solche Persönlichkeitsmerkmale
erhoben, die sich empirisch als Prädiktoren für den Verlauf psychothera-
peutischer Behandlungen als bedeutsam erwiesen haben. Weiterhin kann
die Kenntnis hoher Ausprägung einzelner Persönlichkeitsmerkmale (z. B.
geringe Offenheit) hilfreich sein, um die therapeutische Beziehung zu Pati-
entinnen und Patienten zu gestalten. Teilweise wird auch davon ausgegan-
gen, dass im Rahmen von Therapien ungünstige Ausprägungen von Per-
sonlichkeitsmerkmalen (z. B. „Neurotizismus“) verändert werden können.
Daher werden Persönlichkeitstests gelegentlich auch im Bereich der *Psycho-
therapieevaluation* angewendet, indem sie vor und am Ende der Behandlung
oder zu katamnestischen Zeitpunkten eingesetzt werden.

Trait-Konzept und
Psychotherapieevaluation

Mangels spezifischer Testverfahren wurden früher auch im klinischen Kontext Persönlichkeitstests zur Eingangs-, teilweise auch zur Verlaufsdiagnostik eingesetzt. Dabei wurden zum Teil nur einzelne Skalen der oft umfangreichen Testverfahren genutzt. Persönlichkeitstests (► Abschn. 3.3.3), die vergleichsweise häufig im Rahmen der klinisch-psychologischen Diagnostik eingesetzt werden, sind folgende:

In der klinisch-psychologischen Diagnostik eingesetzte Persönlichkeitstests (Auswahl)

- MMPI-2: Minnesota Multiphasic Personality Inventory-2 (Hathaway et al. 2000; ► Abschn. 3.3.3.1)
- FPI-R: Freiburger Persönlichkeitsinventar – revidierte Fassung (Fahrenberg et al. 2020; ► Abschn. 3.3.3.3)
- TIPI: Trierer Integriertes Persönlichkeitsinventar (Becker 2003; ► Abschn. 3.3.3.3)
- NEO-FFI: NEO-Fünf-Faktoren-Inventar (Borkenau und Ostendorf 1993; ► Abschn. 3.3.3.5)

8

Therapierelevante Persönlichkeitsmerkmale

Relevante *Skalen* aus den genannten Fragebogen sind beispielsweise „Depression“ und „Psychasthenie“ (MMPI), „Lebenszufriedenheit“, „Aggressivität“, „körperliche Beschwerden“ und „Emotionalität“ (FPI-R), „Neurotizismus“ (NEO-FFI) oder Faktoren der seelischen Gesundheit (z. B. „Fröhlichkeit“, „Tatendrang“, „Selbstvertrauen“), wie sie mit dem TIPI erfasst werden.

Da Persönlichkeitstests ihrem Ansatz nach jedoch so konzipiert sind, dass sie psychologische Konstrukte möglichst stabil erfassen, werden sie – wenn überhaupt – nur im Rahmen der Eingangsdagnostik verwendet. Für die Beurteilung von klinisch bedeutsamen Veränderungen sind sie methodisch nicht angemessen und werden wegen der geringen oder nicht vorhandenen Relevanz für klinische bedeutsame Merkmale heute kaum noch verwendet.

Fazit Zur Bewertung mehrdimensionaler Persönlichkeitstests in der klinisch-psychologischen Diagnostik soll betont werden, dass sie im Sinne von Screeninginstrumenten relevante Informationen für die Eingangsdagnostik sowie die Therapieplanung liefern können. Mit solchen Verfahren können einzelne klinisch relevante Persönlichkeitsaspekte erfasst werden, die im Kontext der Behandlung für manche Diagnostiker von Interesse sein können. Zum Einsatz bei der Kontrolle von Therapieverlauf und -erfolg sind Persönlichkeitstests jedoch wegen des zugrunde liegenden Trait-Konzepts nicht geeignet.

8.4.4 Verfahren und Ansätze auf klientenzentrierter, psychodynamischer, systemischer und interpersoneller Grundlage

8.4.4.1 Verfahren auf der Grundlage der klientenzentrierten Gesprächspsychotherapie

Leitsatz der klientenzentrierten (oder personzentrierten) Gesprächspsychotherapie ist die Überzeugung, dass die Klientin bzw. der Klient selbst am besten über sich und sein Problem im Bilde ist und daher die Richtung und den Verlauf des therapeutischen Gesprächs bestimmen sollte. Der Einsatz formalisierter diagnostischer Verfahren spielt aus diesem Grund eine eher untergeordnete Rolle und dient allenfalls zur Ergänzung der Gesprächsmethode.

Einige Verfahren wurden jedoch speziell zur Unterstützung gesprächsdiagnostischer Maßnahmen konstruiert. Ihr Ziel liegt vor allem darin, die Perspektive der Klientin bzw. des Klienten abzubilden. So werden eher ideografische Methoden genutzt, die sich nicht an einer allgemeinen statistischen Norm orientieren (vgl. Eckert et al. 2006).

Ein Beispiel hierfür ist die sog. „Q-Sort-Technik“. Dabei sortiert die Klientin bzw. der Klient vorgegebene Aussagen nach dem Grad, mit dem sie auf sie/ihn zutreffen (meist abgestuft zwischen „trifft auf mich voll und ganz zu“ und „trifft gar nicht zu“). Neben dem auf diese Art erfasssten (realen) Selbstkonzept können die Karten auch entsprechend eines idealen Selbstkonzepts sortiert werden. Da in der klientenzentrierten Gesprächspsychotherapie eine Annäherung von realem und idealem Selbstbild angestrebt wird, kann dieses Verfahren zur Kontrolle des Therapieverlaufs und -erfolgs im Rahmen der Gesprächspsychotherapie (nach Rogers) eingesetzt werden.

Q-Sort-Technik

Fragebogenverfahren, die im Kontext des Gesprächspsychotherapeutischen Ansatzes entwickelt wurden, beziehen sich oft auf die Diagnostik des Selbst- und des Fremdbilds sowie des Erlebens und des Verhaltens. Beispiele für Verfahren, die in der Forschung und im therapeutischen Kontext eingesetzt werden, sind folgende:

Fragebogenverfahren in der Gesprächspsychotherapie

Fragebogenverfahren in der Gesprächspsychotherapie (Auswahl)

- Berger-Skala zur Erfassung der Selbstakzeptanz (Bergeman und Johann 1993).
- VEV: Veränderungsfragebogen des Erlebens und Verhaltens (Zielke und Kopf-Mehnert 1978): In diesem Verfahren sollen Klienten bzw. Klientinnen nach Abschluss der Behandlung 42 Aussagen der Art „Ich habe mehr Selbstvertrauen“ im Vergleich zum Beginn der Therapie beurteilen.
- VLB: Veränderungsfragebogen für Lebensbereiche (Itten und Grawe 2002).
- KASSL: Kieler Änderungssensitive Symptomliste (Zielke 1979): Sie erfasst mit 50 Fragebogen-Items eine Reihe von Beschwerden. Auf faktorenanalytischer Basis wurden die Skalen „sozialer Kontakt“, „Stimmung“, „Beruf“ sowie „Leistung und Konzentration“ etabliert. Zusätzlich kann ein Gesamtwert zur Symptombelastung gebildet werden (vgl. auch ▶ Abschn. 8.3).

8.4.4.2 Psychodynamisch orientierte Verfahren

Neben der Entwicklung psychodynamisch orientierter Interviewkonzepte, die klassischerweise zu den wichtigsten diagnostischen Methoden der Psychoanalyse gehören, gibt es auch eine Reihe von strukturierten oder standardisierten Interview- und Testverfahren, denen explizit ein psychodynamisches bzw. psychoanalytisches Konzept zugrunde liegt. Exemplarisch erwähnt sei hier der Ansatz des *Zentralen Beziehungskonfliktthemas* (*ZBKT*) von Luborsky et al. (1992), eine formalisierte psychodynamische Diagnostik, die in Form eines standardisierten Interviews subjektiv bedeutsame Beziehungsepisoden erhebt. Ausgewertet werden die Interaktionsmuster von Patientinnen und Patienten anhand der eigenen Wünsche, der Reaktionen anderer und der anschließenden Reaktionen des/der Interviewten.

Zentrales Beziehungskonfliktthema (*ZBKT*)

Umfassende Aktivitäten gibt es seit über 10 Jahren im Rahmen der *Operationalisierten Psychodynamischen Diagnostik* (*OPD-2*) des Arbeitskreises OPD (2014). Ziel dieses Ansatzes ist es, diagnostische Konzepte der psychodynamischen Theorie und Psychotherapie mit möglichst großer Zuverlässigkeit (Reliabilität) zu erfassen. Über nichtformalisierte oder standardisierte diagnostische Gespräche werden theoretisch relevante Bereiche auf 5 Achsen abgebildet. Operationalisiert werden dabei die folgenden Ebenen:

Operationalisierte
Psychodynamische Diagnostik
(*OPD-2*)

- Krankheitserleben und Behandlungsvoraussetzungen
- Beziehungsebene
- Zeitlich überdauernde Konflikte
- Psychische Struktur
- Symptom- und Syndromebene, die auch die klassifikatorische Diagnostik nach ICD-10 berücksichtigt

Psychischer und Sozial-Kommunikativer Befund (PSKB)

Gießen-Test – II (GT-II)

8

Mit dem *Psychischen und Sozial-Kommunikativen Befund (PSKB)* von Rudolf (1993) stufen Therapeutinnen und Therapeuten auf der Basis des anamnestischen Gesprächs einerseits psychische Störungen und berichtete Konflikte und Ereignisse hinsichtlich ihrer Symptomatik (psychischer Befund) ein, andererseits das Ich-Erleben, die soziale Bewältigung sowie Reaktionen auf belastende Lebensereignisse. Das Verfahren kann auch als Fragebogen eingesetzt werden.

In früheren Studien, auch außerhalb psychodynamischer Einrichtungen und Fragestellungen, wurde der *Gießen-Test (GT)*, häufig eingesetzt; die derzeit aktuelle Version ist der *GT-II* von Beckmann et al. (2012; ▶ Abschn. 3.3.3.3). Hierbei handelt es sich um ein Fragebogenverfahren mit 40 Items, die insgesamt 6 Skalen zugeordnet werden. Im Gegensatz zu projektiven Verfahren zeichnet sich der GT-II durch eine vergleichsweise bessere Objektivität, Reliabilität und Ökonomie aus. Anders als bei vielen Persönlichkeitsskalen werden in besonderer Weise soziale Einstellungen und soziales Verhalten berücksichtigt. Es soll erfasst werden, wie Personen sich selbst in psychodynamisch relevanten Kategorien in Beziehung zu anderen darstellen. Theoretischer Hintergrund des Inventars sind psychoanalytische, rollentheoretische und interaktionistische Gesichtspunkte. Im GT-II werden Selbst- und Fremdbild auf den Dimensionen „soziale Resonanz“, „Dominanz“, „Kontrolle“ (unter- vs. überkontrolliert), „Grundstimmung“ (hypomanisch vs. depressiv), „Durchlässigkeit“ (durchlässig vs. retentiv) und „soziale Potenz“ erfasst (▶ Abschn. 3.3.3.3).

Pathogene Muster erkennen

Zirkuläres Fragen

Genogramm und Organigramm

8.4.4.3 Systemische Therapie und interpersonale Diagnostik

Bei den *systemischen Ansätzen* liegt der Fokus der Diagnostik und Behandlung nicht in erster Linie auf der „gestörten“ Person, die professionelle Hilfe sucht; vielmehr werden psychische Probleme in funktionalem Zusammenhang mit den aktuellen Lebensumständen gesehen. Dabei spielen nahe Bezugspersonen und die Familie eine zentrale Rolle. Die Person, die in Behandlung kommt, wird als „Indexpatient“ betrachtet. Als „gestört“ bzw. dysfunktional gilt entsprechend das soziale Bezugssystem, das den „Problemträger“ braucht, um in einer eingespielten Homöostase bleiben zu können. In der systemischen Therapie ist es aus diesen Gründen weitgehend unwichtig, welcher Kategorie psychischer Störungen die Symptomatik des „Indexpatienten“ zuzuordnen ist. Ziel der Diagnostik ist vielmehr das *Erkennen der pathogenen Muster* in den Beziehungen zu den Familienmitgliedern und zu anderen wichtigen Personen.

Hieraus resultieren einige Besonderheiten des diagnostischen Vorgehens bei der Problemexploration und der Therapie. Mit der *Methode des zirkulären Fragens* werden Familienmitglieder (bzw. Mitglieder des untersuchten „Systems“) nacheinander über persönliche Sichtweisen und Mutmaßungen über die jeweils anderen Beteiligten befragt. Dadurch erhalten alle Beteiligten gleichzeitig Informationen zu den unterschiedlichen Sichtweisen der Einzelnen auf die jeweiligen Beziehungen und zu den interaktionellen Motiven.

Informationen über Familien oder soziale Systeme werden im Rahmen der systemischen Therapie auch häufig mittels eines *Genogramms* erhoben. Hierbei werden Aussagen, beispielsweise zu Mitgliedern einer Familie, meist

einschließlich der Berücksichtigung zurückliegender Generationen und weiterer wichtiger dazugehöriger Personen hinsichtlich ihrer Beziehungen und biografischen Besonderheiten erhoben und grafisch dargestellt. Mit dem *Organigramm* können Organisationen und Verbände in ihrer hierarchischen Struktur und den wechselseitigen formellen und informellen Abhängigkeiten von Stellen bzw. Personen grafisch dargestellt werden (ausführlich dazu von Schlippe und Schweitzer 2016).

In der interpersonalen Diagnostik ist die *Strukturelle Analyse Sozialer Beziehungen (SASB)* bzw. Structural Analysis of Social Behavior von Benjamin (1974; deutsche Fassung: Tress 2003) ein bekanntes und – vor allem in der Forschung – vielfach genutztes System. Mit der SASB können 3 unterschiedliche Aspekte des interpersonellen Verhaltens und Erlebens erfasst werden: der Fokus auf andere, der Fokus auf das Selbst und das Umgehen mit sich selbst (Introjekt). Das sehr komplexe System kann als Beobachtungsverfahren nur nach einem ausführlichen Training mit ausreichen- der Zuverlässigkeit eingesetzt werden.

Mit dem *Inventar zur Erfassung interpersonaler Probleme – Deutsche Version (IIP-D)* von Horowitz et al. (2016) werden Schwierigkeiten in sozialen Beziehungen auf mehreren Ebenen erfasst: Intimität, Aggressivität, Assertivität (Selbstsicherheit), Unabhängigkeit und Geselligkeit.

Im Sinne der Diagnostik von interpersonalen Kompetenzen und Resourcen spielt die Erfassung von sozialer Unterstützung im sozialen Netz eine wichtige Rolle innerhalb der klinisch-psychologischen Diagnostik. Fydrich und Sommer (2003) gehen von 3 zentralen Aspekten sozialer Unterstützung aus: praktische Unterstützung, emotionale Unterstützung und soziale Integration. Der *Fragebogen zur sozialen Unterstützung (F-SozU)* von Fydrich et al. (2007) erfasst in seiner ausführlichen Form (54 Items) neben diesen 3 Merkmalen zusätzlich auch mögliche Belastungen durch das soziale Netz und durch konkrete Personen, die als sozial unterstützend oder belastend erlebt werden. Kurzformen mit 22 und 14 Items ergeben reliable und valide Maße für allgemeine soziale Unterstützung.

Strukturelle Analyse Sozialer Beziehungen (SASB)

Inventar zur Erfassung interpersonaler Probleme (IIP-D)

Fragebogen zur sozialen Unterstützung (F-SozU)

8.5 Verbindung von Diagnostik und Intervention: Die Indikation

Der *praktische Wert* diagnostischer Prozeduren und Ergebnisse misst sich daran, in welchem Umfang diagnostische Ergebnisse Handlungsanweisungen für bestimmte therapeutische Entscheidungen und den Einsatz von spezifischen Interventionsverfahren geben. Es geht darum, wie gut bestimmte Patienten mit konkreten Problemen bzw. psychischen Störungen den verfügbaren therapeutischen Einrichtungen oder Settings (Einzel-, Paar-, Gruppentherapie) und Therapeuten zugeordnet werden können. Entsprechend der unterschiedlichen Zielrichtung der Diagnostik kann auch zwischen der *selektiven Indikation* (Zuordnungsproblem, Selektionsstrategie) und der *adaptiven Indikation* (Anpassung der therapeutischen Intervention an den Einzelfall und an den therapeutischen Prozess, prozessuale Indikation, Modifikationsstrategie) unterschieden werden (vgl. Stieglitz et al. 2001).

Selektive und adaptive Indikation

Die Indikationsfrage wird häufig als eines der wichtigsten Probleme der psychotherapeutischen Praxis und damit auch der klinisch-psychologischen Forschung erachtet. Die *Aufgaben der Indikationsstellung* in der Psychotherapie bestehen in einer hierarchischen Entscheidung über folgende 3 Fragen:

Aufgaben der Indikationsstellung und Therapieplanung

Störungsspezifische Behandlungsansätze

Bedeutung der klassifikatorischen Diagnostik

Evidenzbasierte Behandlungsleitlinien

- *Psychotherapie-Indikation:* Ist im konkreten Fall überhaupt eine Psychotherapie angezeigt?
- *Behandlungsbezogene Indikation:* Welche psychotherapeutischen Maßnahmen sind angebracht? Bezieht sich diese Frage auf die Entscheidung, welches Therapieverfahren für einen Patienten am ehesten geeignet ist, handelt es sich um eine Fragestellung der differentiellen Indikation.
- *Adaptive oder prozessuale Indikation:* Wie können die Maßnahmen an den Einzelfall bzw. den Verlauf der Behandlung angepasst werden?

Differentielle Indikation

Die zentrale Frage der *differentiellen Indikation* lautet nach Paul (1967, S. 111): „Welches ist für dieses Individuum mit diesem spezifischen Problem die effektivste Behandlung, durch wen und unter welchen Umständen?“ Um auf diese sehr komplexe Frage eine empirisch begründete Antwort zu finden, wäre die Untersuchung einer Vielzahl von Faktoren in einem multifaktoriellen Versuchsplan nötig. Dies übersteigt jedoch die Möglichkeiten der empirischen Psychotherapieforschung. Damit dennoch eine angemessene Indikation gestellt werden kann, sollten die bekannten Ergebnisse aus Therapiestudien mit Blick auf den konkreten Fall genutzt werden.

Schwerpunkt der Forschung in der Klinischen Psychologie und Psychotherapie waren und sind sowohl die Weiterentwicklung störungsbezogenen Grundlagenwissens als auch die Entwicklung und Überprüfung von psychotherapeutischen Interventionen, die aus dieser Grundlagenforschung abgeleitet wurden. So konnten mittlerweile zahlreiche *spezifische Behandlungsverfahren* entwickelt werden, die auf bestimmte Störungsbilder zugeschnitten sind und sich in der Anwendung als besonders wirksam erwiesen haben. Auch zeigen zahlreiche Metaanalysen zur Psychotherapie, in denen die Ergebnisse einer Vielzahl von *Psychotherapiestudien* zusammengefasst wurden, dass im Bereich der Psychotherapie keinesfalls von einer Uniformität ausgegangen werden kann, nach der alle Psychotherapien und alle Psychotherapeuten und -therapeutinnen unabhängig von ihrer theoretischen Orientierung mit ihrer Arbeit etwa gleich gute Erfolge erzielen.

Mit dem selektiv-indikativen Entscheidungsprozess, mit dem bestimmte Behandlungsverfahren spezifischen Störungen zugeordnet werden, wird deutlich, dass die klassifikatorische Diagnostik einen besonders hohen Stellenwert für *therapierelevante Entscheidungen* hat. Die Diagnostik auf der Grundlage der verfügbaren Klassifikationskompendien (ICD-10 oder DSM-5) kann – in Kombination mit den verfügbaren Befunden aus der Grundlagen- und Interventionsforschung – zuverlässige Hinweise auf eine differentielle Therapieindikation geben.

So sind sowohl das Erscheinungsbild als auch die Modelle zur Entstehung und Aufrechterhaltung beispielsweise von Depressionen von denen bei Phobien oder Essstörungen deutlich verschieden. Entsprechend wurden spezifische Behandlungsmethoden entwickelt, die die Spezifität der Störung berücksichtigen. In den letzten Jahrzehnten wurden die Befunde zur Effektivität von Behandlungsmethoden in Metaanalysen zusammengefasst. Eine wichtige Rolle spielten dabei die fortlaufenden Zusammenfassungen der „Empirically Supported Therapies“ (Chambless und Hollon 1998; Chambless und Ollendick 2001). Weiterhin finden auch in der Psychotherapie die störungsspezifische Indikation und Behandlung ihren Niederschlag in der seit etwa Mitte der 1990er-Jahre zunehmenden Entwicklung von *Behandlungsleitlinien*. Dies sind systematisch entwickelte, wissenschaftlich begründete und praxisorientierte Hilfen zur Entscheidungsfindung über die angemessene therapeutische Vorgehensweise bei speziellen gesundheitlichen Problemen. Hauptzweck ist die Darstellung des fachlichen Entwicklungsstands zu einer Erkrankung oder zu einem Problemkreis, mit denen den Angehörigen des Berufs eine evidenzbasierte Orientierung im Sinne von

Entscheidungshilfen und Handlungsempfehlungen gegeben wird (► <https://www.awmf.org/>). Auch die Überblicksarbeiten der Cochrane Collaboration sind störungsspezifisch organisiert (► <https://www.cochrane.de/>).

Neben der Zuordnung zentraler Probleme bzw. psychischer Störungen zu spezifischen Behandlungsansätzen ist auch die Frage nach einer möglichst optimalen Zuordnung von Therapeut/-in und Patient/-in von Bedeutung. Hierzu liegen jedoch deutlich weniger und zudem sehr widersprüchliche Ergebnisse vor (s. u.). Im Ergebnis ist daher bei der Behandlungsplanung eine Orientierung an empirisch fundierten störungsspezifischen Behandlungsleitlinien klar zu empfehlen.

Kompetenzorientierte und transdiagnostische Ansätze

Neben der leitlinienorientierten Behandlungsplanung werden in den letzten Jahren bei der Behandlungsplanung auch störungsübergreifende Faktoren berücksichtigt. So stellen Stenzel et al. (2010) das Verfahren der *Operationalisierten Fertigkeitsdiagnostik (OFD)* vor. Es werden – zusätzlich zur störungsspezifischen Diagnostik – verschiedene Fertigkeiten in 4 Lebensbereichen erfasst wie „Problemlösen“, „soziale Kompetenz“, „Stressbewältigung“, „Emotionsregulation“ und „Entspannungsfähigkeit“, die bei der Therapieplanung berücksichtigt werden.

Die Arbeitsgruppe um David Barlow (Barlow et al. 2019) stellt – ebenfalls basierend auf dem Konzept einer Kompetenzorientierung bei Therapeutinnen und Therapeuten – ein Konzept der *transdiagnostischen Behandlung emotionaler Störungen* vor. Die Autorinnen und Autoren betonen dabei die Notwendigkeit, die häufig vorhandene Komorbidität psychischer Störungen bei der Therapieplanung zu berücksichtigen, und entwickeln einen auf diesem Konzept aufbauenden Therapieansatz.

Die *Passung* von Therapeut/-in und Patient/-in wird vergleichsweise oft thematisiert. Die empirischen Befunde dazu sind jedoch eher gering. Beutler et al. (2004) fassen zusammen, dass beispielsweise gleichgeschlechtliche Therapeut-Klient-Dyaden etwas erfolgreicher sind als gegengeschlechtliche. Weiterhin gehen offensichtlich auch eine größere Ähnlichkeit von Interessen und eine ähnliche Schichtzugehörigkeit mit etwas besseren Behandlungsergebnissen einher. Andere Befunde weisen darauf hin, dass Patienten, die sich als sehr hilfebedürftig schildern, eher von einem direktiven, strukturierten Umgang profitieren, während solche mit hoher internaler Kontrolle und hohem Autonomiebedürfnis eher in Therapien Erfolg haben, die einem nondirektiven Ansatz folgen. In der klinisch-psychologischen Praxis ist zudem aus organisatorischen Gründen die Zuordnung von Therapeutinnen bzw. Therapeuten mit bestimmten persönlichen Merkmalen zu Klientinnen bzw. Klienten mit ihren Persönlichkeitsmerkmalen zumeist nicht möglich und – wenn überhaupt – deutlich schwieriger zu realisieren als eine Zuordnung von Störungsbildern zu spezifischen therapeutischen Ansätzen.

Passung von Therapeut/-in und Patient/-in

8.6 Erfolgskontrolle als Teil der Qualitätssicherung

Zur Sicherung der Ergebnisqualität klinisch-psychologischer Interventionen ist es notwendig, den Verlauf und den Erfolg der Behandlung zu evaluieren. Zur möglichst objektiven Beurteilung des Therapieerfolgs bieten sich verschiedene Strategien und Methoden an. Bei der *indirekten Veränderungsmessung* werden die relevanten (meist symptomorientierten) Verfahren, die

Veränderungsmessung

zu Beginn der Therapie eingesetzt wurden, am Ende der Behandlung erneut vorgegeben. Eine *direkte Veränderungsmessung* erfolgt in Form einer Einmalurhebung, z. B. mit dem Veränderungsfragebogen des Erlebens und Verhaltens (VEV) oder dem Veränderungsfragebogen für Lebensbereiche (VLB; ▶ Abschn. 8.4.4.1). Mit der Zielerreichungsbeurteilung (Goal Attainment Scaling, GAS) kann eingestuft werden, inwieweit die zu Beginn oder im Verlauf der Therapie gesetzten Ziele tatsächlich realisiert werden konnten. Zur Überprüfung der Stabilität des Therapieerfolgs sind katamnestische Untersuchungen unerlässlich, wobei zumindest eine Nachuntersuchung 12 Monate nach Abschluss der Behandlung zu empfehlen ist.

8.6.1 Zieldefinition, Therapieverlaufs- und Veränderungsdiagnostik

Exemplarisch werden zwei Instrumente dargestellt, in dem für die Psychotherapie spezifische Konstruktions- und Evaluationsschritte umgesetzt werden.

8.6.1.1 Goal Attainment Scaling (GAS)

Das GAS von Kiresuk und Sherman (1968) gehört zum Standard einer Psychotherapie und wird dazu verwendet, zu Beginn konkrete, für den Patienten bzw. die Patientin relevante Therapieziele zu formulieren. Bei der Formulierung der Ziele ist darauf zu achten, dass sie als Ergebnis der diagnostischen Gespräche in der oder den ersten Sitzungen erfolgt. Die Ziele sollen spezifisch und individuell formuliert werden. Zentral dabei ist, genaue Indikatoren zu bestimmen, an denen man erkennen kann, ob das jeweilige Ziel erreicht wurde oder in welchem Ausmaß es erreicht wurde. Beispiel für ein zunächst wenig spezifische Zielformulierung wäre: „Ich möchte weniger depressiv sein.“ Die dann notwendige therapeutische Frage zur Konkretisierung wäre: „Woran wird man erkennen können, dass Sie weniger depressiv sind als jetzt zu Beginn der Therapie?“ Hieraus könnten sich konkrete Indikatoren ergeben wie „Ich könnte wieder zur Arbeit gehen“, „Ich würde mehr mit meiner Familie unternehmen“, „Ich würde als Freizeitaktivität wieder zu den Chorproben gehen“. Im Verlauf der Behandlung können sich die Ziele verändern, neue dazukommen oder sie können neu operationalisiert werden. Das GAS kann auch im Verlauf der Therapie im Sinne einer Zwischenevaluation eingesetzt werden. Charakteristischerweise wird das Erreichen der Ziele anhand einer 5-stufigen Skala sowohl von dem Patienten bzw. der Patientin als auch von dem Therapeuten bzw. der Therapeutin beurteilt (genauere Beschreibung bei Geue et al. 2016).

8.6.1.2 Bochumer Veränderungsbogen-2000

Der Fragebogen wurde zunächst im Kontext der Psychotherapieforschung im Bereich der klientenzentrierten Gesprächspsychotherapie von Zielke und Kopf-Mehnert (1978) entwickelt und wurde von der Bochumer Arbeitsgruppe zum Bochumer Veränderungsbogen-2000 (Willutzki et al. 2013) weiterentwickelt. Es handelt sich um ein Verfahren mit 26 Items, bei dem am Ende der Therapie die von den Klienten/Klientinnen berichteten Veränderungen retrospektiv eingeschätzt werden. Die Patienten bzw. Patientinnen vergleichen den Zustand zu Beginn der Therapie mit dem aktuellen Zeitpunkt. Die Items liegen in Aussageform vor, die sowohl positive als auch negative Äußerungen über Veränderungen in einem bestimmten Zeitraum beinhalten. Dabei erfolgt die Einschätzung auf einer 7-stufigen Skala. Es folgt ein Beispiel für ein Item:

► Beispiel

„Im Vergleich zum Zeitpunkt vor der Therapie ...“

<input type="checkbox"/> bin ich mit mir unzufriedener.“	<input type="checkbox"/> bin ich mit mir zufriedener.“
--	--

(Kodierung: 1)	(Kodierung: 7)
----------------	----------------

Der Mittelwert der Antworten vermittelt als Ergebnis eine globale Einschätzung der durch die Therapie erreichten Veränderung (ein Wert >4 entspricht einer Verbesserung; Werte <4 weisen auf eine Verschlechterung hin). Im Weiteren kann der Fragebogen bzw. können einzelne Items des Fragebogens auch als Grundlage für eine direkte Rückmeldung an den Patienten bzw. die Patienten genutzt werden.

Probleme mit dieser Form der Diagnostik bestehen vor allem darin, dass bei einer einmaligen Erhebung am Ende einer Behandlung z. B. soziale Erwünschtheit das Antwortverhalten beeinflussen kann sowie die subjektive Bewertung des (therapeutischen) Aufwands eine bedeutende Rolle spielt. Zudem können – besonders bei längeren Therapien – Erinnerungseinflüsse die Angaben verfälschen und damit zu einer Verzerrung und Verfälschung der Erfolgseinschätzung führen. Untersuchungen von Willutzki et al. (2013) zeigen, dass – entgegen der Erwartung – Verbesserungen und Verschlechterungen in unspezifischer Weise sowohl bei behandelten Klienten/Klientinnen als auch in nicht behandelten klinischen (Warte-)Kontrollgruppen auftreten können.

Grenzen der direkten Veränderungsmessung

Insgesamt ist daher der Einsatz des Bochumer Veränderungsbogens-2000 zur Überprüfung des Therapieverlaufs und -erfolgs nur in Kombination mit anderen Verfahren zu empfehlen oder als „kleine Lösung“, wenn keine symptom- oder prozessbezogenen Daten zu Beginn oder im Verlauf der Therapie erhoben wurden.

Mehrfachmessung mit reliablen und validen Messverfahren Standard

Wegen der methodischen Nachteile gilt jedoch das Erfassen von Beschwerden und Symptomen in Form einer Mehrfachmessung mit möglichst reliablen und validen Messverfahren als Standard bei der Evaluation von klinisch-psychologischen bzw. psychotherapeutischen Behandlungen. Dabei werden die relevanten Maße mindestens vor und nach der Behandlung erfasst. Dabei kommen vor allem die in ► Abschn. 8.4.1 dargestellten psychometrischen Verfahren zum Einsatz.

8.6.2 Kriterium der klinisch bedeutsamen Verbesserung

Bei der Beurteilung der Wirksamkeit klinisch-psychologischer Interventionen gibt es grundsätzlich 2 unterschiedliche Ansätze bzw. Aufgaben. Im Bereich der *nomothetischen Evaluationsforschung* geht es darum, festzustellen, ob der Einsatz einer definierten Behandlung oder Intervention im Vergleich zu einer oder mehreren Kontroll- bzw. Vergleichsgruppen zu den erwarteten behandlungsbezogenen Veränderungen führt. Solche Untersuchungsdesigns folgen einem *gruppenstatistischen Paradigma*. Die wissenschaftlich größte Bedeutung haben dabei randomisierte kontrollierte Studien oder Untersuchungen mit parallelisierten Gruppen, da nur solche den Schluss auf einen kausalen Zusammenhang von psychologischer Intervention und Therapieverlauf bzw. -erfolg erlauben.

Randomisierte kontrollierte Studien

Die Ergebnisse werden in der Regel mittels statistischer Vergleiche hinsichtlich der Veränderungen innerhalb der Gruppen bzw. zwischen den beobachteten Gruppen ausgewertet und entsprechend interpretiert. Häufig werden die Befunde auch als Effektstärken dargestellt. Derartige gruppenstatistische

Auswertungen und Aussagen darüber, ob Effekte für die gesamte Gruppe statistisch bedeutsam, d. h. signifikant sind, hängen jedoch unter Umständen von Faktoren ab, die klinisch-psychologisch keine oder eine nur geringe Relevanz haben. So wird beispielsweise die statistische Signifikanz stark durch die Stichprobengröße und die Streuung der Messwerte beeinflusst. Mit steigender Anzahl von untersuchten Personen können sehr kleine Unterschiede in den Messvariablen zwar statistisch signifikant sein, jedoch ist dies nicht gleichbedeutend damit, dass die Unterschiede auch im klinischen Sinne bedeutsam sind.

Klinische Relevanz einer Veränderung

8

Klinisch bedeutsame Symptomreduktion wichtig

Für die Kontrolle der Effekte durchgeführter Behandlungen ist es sowohl bei der Untersuchung von Gruppen als auch im Einzelfall notwendig, eine Aussage über die Bedeutsamkeit einer beobachteten Veränderung treffen zu können. Die *klinische Relevanz einer Veränderung* bezieht sich somit auf die Beurteilung von Standards und Zielen, die neben der statistischen Beurteilung von Veränderungen auch von gruppenbezogenen Normen und darüber hinaus auch von den Betroffenen selbst und von der Einschätzung der Therapeuten bzw. Therapeutinnen oder von sozialrechtlich begründeten Normen des Gesundheitssystems (z. B. Begründungen von Krankschreibungen und Berentung) abhängen kann.

Eine Arbeitsgruppe um Jacobson hat zum Thema der Einschätzung der klinischen Bedeutsamkeit von Veränderungen durch klinisch-psychologische Interventionen bzw. Psychotherapie einen methodischen Vorschlag ausgearbeitet, der sowohl in der Psychotherapieforschung als auch bei der Einschätzung von Veränderungen im Einzelfall ein Standard für die Erfolgsbeurteilung von Therapien ist (Jacobson und Revenstorf 1988; Jacobson und Truax 1991; Jacobson et al. 1984). Wichtigster Kerngedanke dieses Ansatzes ist, dass die klinisch bedeutsame Verbesserung für die betroffene Person bestenfalls Symptomfreiheit („Heilung“) oder Problemlösung bedeutet; zumindest aber eine merkliche *Symptom- bzw. Problemreduktion* darstellen muss. Während Patienten bzw. Patientinnen vor der Behandlung zur Population von Personen gehören, bei denen eine Problemausprägung „mit Krankheitswert“ vorliegt, sollten die Personen nach der Behandlung mit hoher Wahrscheinlichkeit der Population angehören, die diese Problematik nicht aufweist. Diese Thematik wurde etwas allgemeiner und mit einem eher methodischen Fokus bereits in ▶ Abschn. 5.1.1 behandelt.

Jacobson und Truax (1991) schlagen 3 Möglichkeiten vor, wie diese Veränderungen operationalisiert werden können:

- a) Das Ausmaß der Symptomatik sollte nach der Behandlung mindestens 2 Standardabweichungen unter dem Mittelwert der Population liegen, bei der die entsprechende Störung vorliegt;
- b) die Symptomatik einer behandelten Person sollte nach der Therapie in ihrem Ausmaß innerhalb von 2 Standardabweichungen einer nicht gestörten Population liegen;
- c) die Ausprägung der Symptomatik muss nach einer Behandlung näher am Mittelwert der nicht gestörten Population als am Mittelwert der gestörten Population liegen.

Um eine Beurteilung nach dem Kriterium a oder c vorzunehmen, braucht man neben dem Wert für die Symptomatik des Patienten nach der Behandlung auch den Mittelwert und die Streuung für die entsprechend beeinträchtigte Population. Solche Normen sind zumeist vorhanden. Seltener verfügbar sind Informationen über den Mittelwert und die Streuung für nicht klinische („gesunde“) Stichproben, die für die Beurteilung der klinischen Bedeutsamkeit nach Kriterium b oder c benötigt werden. Für die Einschätzung nach Kriterium c müssen die Mittelwerte für beide Stichproben bekannt sein.

Operationalisierung von Therapieerfolg

Wie das folgende Beispiel (verändert nach Jacobson und Truax 1991) zeigt, kann die Einschätzung der klinischen Bedeutsamkeit sehr unterschiedlich ausfallen, je nachdem, welches der 3 Kriterien zugrunde gelegt wird.

Modellrechnung zur Einschätzung klinischer Kriterien

Die Untersuchung einer Gruppe von Personen mit Depressionen und einer nichtklinischen Vergleichsgruppe zeigen hypothetisch folgende Werte für die Ausprägung der Depressivität:

\bar{x}_0 Mittelwert der nichtdepressiven Gruppe: 40

s_0 Standardabweichung der nichtdepressiven Gruppe: 7,5

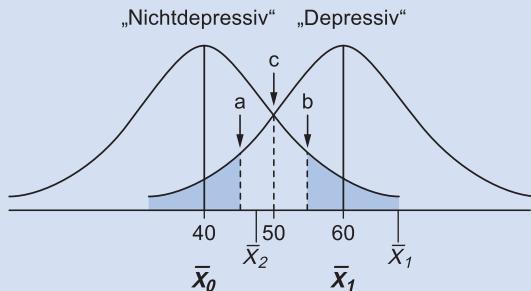
\bar{X}_1 Mittelwert der depressiven Gruppe: 60

s_1 Standardabweichung der depressiven Gruppe: 7,5

r_{tt} Test-Retest-Reliabilität des Depressionstests: ,80

\bar{x}_1 Depressionsmaß einer Person X zu Beginn der Therapie: 67,5

\bar{x}_2 Depressionsmaß der Person nach Abschluss der Therapie: 47,5



Beispiel der Verteilungen von Messwerten einer „depressiven“ und einer „nicht depressiven“ Stichprobe mit den Mittelwerten \bar{x}_1 und \bar{x}_0 sowie den Messwerten einer Person vor und nach der Behandlung (x_1, x_2) und den 3 Kriterien für Therapieerfolg (a, b, c).

In der obigen Darstellung der Messwertverteilungen wird davon ausgegangen, dass die Streuungen für die klinisch beeinträchtigte und die nichtklinische Stichprobe gleich sind. Die Ausführungen gelten nur unter der Annahme, dass das untersuchte Merkmal (Depressivität) in beiden Stichproben normalverteilt ist. Die Verteilungen des Merkmals in beiden Gruppen zeigen in dem Beispiel eine deutliche Überlappung. Die Abbildung macht deutlich, dass es nach den 3 Berechnungsmodi auch 3 unterschiedliche Kriterien (Schwellenwerte) für die Einschätzung der klinisch bedeutsamen Verbesserung gibt. Am „strengsten“ ist in dem Beispiel das *Kriterium a* (2 Standardabweichungen der dysfunktionalen Gruppe in Richtung der nichtgestörten Gruppe = 45). Nach diesem Kriterium hätte sich der gesundheitliche Zustand von Patient X noch nicht klinisch bedeutsam verbessert. Weniger streng ist *Kriterium c* (Mitte zwischen den Mittelwerten der funktionalen und der depressiven Gruppe = 50), und *Kriterium b* (innerhalb von 2 Standardabweichungen um den Mittelwert der nicht gestörten Gruppe = 55) ist in diesem Falle vergleichsweise „weich“. Nach Umrechnung in einen Prozentrang entspricht schon ein vergleichsweise großer Teil der dysfunktionalen Gruppe, nämlich 25 % ($Z = (55 - 60) / 7,5 = -0,67$) diesem Wert auch ohne Behandlung.

Diese Modellrechnung zeigt, dass es bei der Einschätzung der Kriterien für eine klinisch bedeutsame Verbesserung sehr darauf ankommt, welches Kriterium als Schwellenwert eingesetzt wird. In Anlehnung an Jacobson und

Auswahl des Kriteriums

Truax (1991) wird für die Entscheidung über die Auswahl des Schwellenwertes Folgendes vorgeschlagen:

- Wenn Normwerte für eine nichtklinische Population nicht verfügbar sind, so ist als einziges Kriterium a berechenbar.
- Wenn Normwerte für die nichtklinische (funktionale) Population und die klinische Population vorhanden sind und beide Verteilungen nur so weit überlappen, dass Kriterium b näher am Mittelwert der funktionalen Gruppe liegt als Kriterium c, so ist b das bessere (weil strengere) Kriterium.
- Bei größerer Überlappung der beiden Verteilungen (Kriterium b liegt weiter vom Mittelwert der funktionalen Stichprobe entfernt als c) sollte auch hier das strengere Kriterium (in diesem Fall c) gewählt werden.

Diagnostik weiterer relevanter klinisch-psychologischer Variablen

8

Bisher haben wir das Kriterium der klinisch bedeutsamen Veränderung betrachtet. Eine weitere bedeutsame Frage ist, ob sich der gesundheitliche Zustand einzelner (oder in einer Gruppenstudie: wie vieler) Patienten bzw. Patientinnen nach einer Intervention in dem relevanten Merkmal tatsächlich verbessert hat und – wenn ja – wie groß diese Veränderung ausfällt. Diese Frage ist umso bedeutsamer, je mehr sich die beiden Gruppen (die klinische und die nichtklinische) überlappen. Jacobson et al. (1984) sowie Christensen und Mendoza (1986) schlagen dazu die Berechnung eines *Veränderungsindex* (Reliable Change Index, RCI) vor (zur Berechnung s. ▶ Abschn. 2.6.2.2).

Neben der Veränderungsmessung, die gleichzeitig auch die zentrale Maßnahme der Überprüfung der Ergebnisqualität von Behandlungen darstellt, gibt es eine Reihe von diagnostischen Instrumenten, die sich insbesondere auf interpersonelle und Prozessvariablen der Intervention beziehen. Hierzu gehören Verfahren zum Erfassen von Therapiezielen, Therapiemotivation, Ressourcen und Fertigkeiten der Patientinnen und Patienten, Selbstwert, sozialen Netzwerken und sozialer Unterstützung sowie kognitiver Denkstile. Eine Auswahl hilfreicher Interviews bzw. psychometrischer Verfahren veranschaulicht folgende Liste:

Diagnostische Instrumente zur Erfassung von interpersonellen und Prozessvariablen der Intervention (Auswahl)

- BIT: Berner Inventar für Therapieziele (Grosse Holtforth 2001)
- BTSTB/BPSTB: Berner Therapeuten- und Patientenstundenbogen 2000 (Flückiger et al. 2010)
- BRI: Berner Ressourceninventar (Tröskens und Grawe 2004)
- FMP: Fragebogen zur Psychotherapiemotivation (Schneider et al. 1989)
- F-SozU: Fragebogen zur sozialen Unterstützung (Fydrich et al. 2007)
- B-IKS: Beck Inventar für Kognitive Schemata (Fydrich 2016)
- RSES: Rosenberg Skala zum Selbstwertgefühl (von Collani und Herzberg 2003)

8.7 Zusammenfassung

Für eine professionelle und auf wissenschaftlicher Grundlage beruhende Intervention ist eine auf theoretischen Modellen aufbauende und methodisch geprüfte Diagnostik unbedingte Voraussetzung. Hierfür müssen Diagnostikerinnen und Diagnostiker über gute Kompetenzen verfügen. Diese betreffen vor allem.

- a) Kenntnisse über die statistischen Grundlagen diagnostischer Prozesse,
- b) Kenntnisse über mögliche Fehler und Fehlerwahrscheinlichkeiten im diagnostischen Prozess und Kompetenzen des Umgangs mit möglichen Fehlern,

- c) Fertigkeiten, angemessene diagnostische Verfahren auszuwählen und fachgerecht einzusetzen und auszuwerten und
- d) auf den Befunden aufbauend korrekte handlungsorientierte Schlüsse zu ziehen, Empfehlungen auszusprechen oder Behandlungsindikationen zu stellen.

Bei allem ist eine hohe interpersonelle Kompetenz wichtig, um sich auf eine gegebene Person bzw. Situation mit Empathie einzustellen.

Eine zuverlässige und valide Diagnostik in der Klinischen Psychologie hat nicht nur eine hohe Relevanz bei der Beurteilung von Einzelfällen, sondern ist eine wesentliche Grundlage für die Entscheidung der Wissenschaftlichkeit von psychotherapeutischen Behandlungsansätzen, wie sie berufsrechtlich vom Wissenschaftlichen Beirat Psychotherapie ([► https://www.wbpsychotherapie.de/](https://www.wbpsychotherapie.de/)) oder sozialrechtlich vom Gemeinsamen Bundesausschuss ([► https://www.g-ba.de/](https://www.g-ba.de/)) getroffen werden. Die Orientierung an Diagnosen ist darüber hinaus maßgeblich für Behandlungsentscheidungen, die ihren Niederschlag in den empirisch begründeten Behandlungsleitlinien finden.

Daher hat auch in dem am 1. September 2020 in Kraft getretenen novelierten Psychotherapeutengesetz und in der Approbationsordnung für Psychotherapeutinnen und Psychotherapeuten, die am 4. März 2020 veröffentlicht wurde, der Erwerb diagnostischer Kompetenz berechtigterweise einen sehr hohen Stellenwert.

Weiterführende Literatur

- Geue, K., Strauß, B., & Brähler, E. (2016). *Diagnostische Verfahren in der Psychotherapie* (3. Aufl.). Göttingen: Hogrefe.
- Wittchen, U., & Hoyer, J. (2011). *Klinische Psychologie und Psychotherapie* (2. Aufl.). Berlin, Heidelberg: Springer.
- Schneider, S., & Margraf, J. (2019). *Lehrbuch der Verhaltenstherapie. Bd. 3: Störungen im Kindes- und Jugendalter* (2. Aufl.). Berlin, Heidelberg: Springer.
- Collegium Internationale Psychiatriae Scalarum. (CIPS). (Hrsg.) (2015). *Internationale Skalen für Psychiatrie* (6. Aufl.). Göttingen: Beltz Test.

?

Übungsfragen

- ► Abschn. 8.1–8.6:
 - Welches sind die 5 zentralen Aufgaben der klinisch-psychologischen Diagnostik?
 - Welche 3 Bereiche des individuellen Erlebens und Verhaltens stehen im Fokus der klinisch-psychologischen Diagnostik?
 - Welche 5 Gruppen diagnostischer Methoden werden in der klinisch-psychologischen Diagnostik unterschieden?
 - Nennen Sie einige Versorgungseinrichtungen, in denen klinisch-psychologische Diagnostik ein wichtiger Teil der professionellen Arbeit darstellt!
 - Welche 3 wesentlichen Kompetenzbereiche sind zur Ausübung einer klinisch-psychologischen bzw. psychotherapeutischen Tätigkeit notwendig?
 - Nennen Sie zentrale Themen der biografischen Anamnese!
 - Welches sind die zentralen Faktoren für die Definition einer psychischen Störung?
 - Was ist der Unterschied zwischen kategorialer und dimensionaler Diagnostik in der Klinischen Psychologie?
 - Wann ist der Einsatz von Verhaltensbeobachtungen angemessen?
 - Was bedeuten die Buchstaben S, R und C in der klassischen Verhaltensgleichung?

- Welches sind die 5 Achsen der Operationalisierten Psychodynamischen Diagnostik (OPD)?
- Was ist der Unterschied zwischen einer statistisch und einer klinisch bedeutsamen Veränderung der klinisch relevanten Symptomatik?

Literatur

- Achenbach, T. M., & Arbeitsgruppe Deutsche Child Behavior Checklist (2000). *CBCL 1½–5: Child Behavior Checklist 1½–5 – Deutsche Fassung. Elternfragebogen für Klein- und Vorschulkinder*. Köln: Arbeitsgruppe Kinder-, Jugend- und Familiendiagnostik (KJFD).
- Ahle, M. E., Döpfner, M., Könning, J., Mattejat, F., Müller, U., Walter, D., & Zumpf, H. (2006). Qualitätssicherung bei Therapien mit Kindern und Jugendlichen. In F. Mattejat (Hrsg.), *Lehrbuch der Psychotherapie* (Bd. 4, S. 197–206). München: CIP Medien.
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®), Fifth Edition*. Arlington, VA, American Psychiatric Association.
- American Psychiatric Association. (2015). *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5: Deutsche Ausgabe herausgegeben von Peter Falkai und Hans-Ulrich Wittchen*. Göttingen: Hogrefe.
- American Psychiatric Association. (2018). *Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5: Deutsche Ausgabe herausgegeben von Peter Falkai und Hans-Ulrich Wittchen* (2. Aufl.). Göttingen: Hogrefe.
- Arbeitsgemeinschaft für Methodik und Dokumentation (AMDP). (Hrsg.). (2018). *Das AMDP-System. Manual zur Dokumentation psychiatrischer Befunde* (10. Aufl.). Göttingen: Hogrefe.
- Arbeitskreis OPD. (Hrsg.). (2014). *OPD-2: Operationalisierte Psychodynamische Diagnostik. Das Manual für Diagnostik und Therapieplanung* (3. Aufl.). Göttingen: Hogrefe.
- Bandelow, B. (2016). *PAS: Panik- und Agoraphobie-Skala* (2. Aufl.). Göttingen: Hogrefe.
- Barkmann, C., & Brähler, E. (2009). *GBB-KJ: Gießener Beschwerdebogen für Kinder und Jugendliche* (2. Aufl.). Bern: Huber.
- Barlow, D. H., Farchione, T. J., Sauer-Zavala, S., Murray Latin H., Ellard, K. K., Bullis, J. R., Bentley, K. H., et al. (2019). *Transdiagnostische Behandlung emotionaler Störungen*. Göttingen: Hogrefe.
- Bartling, G., Echelmeyer, L., Engberding, M. (2016). *Problemanalyse im therapeutischen Prozess* (6. Aufl.). Stuttgart: Kohlhammer.
- Bassler, M., Potratz, B., & Krauthäuser, H. (1995). Der Helping Alliance Questionnaire (HAQ) von Luborsky. Möglichkeiten zur Evaluation des therapeutischen Prozesses von stationärer Psychotherapie. *Psychotherapeut* 40, 23–32.
- Bastine, R., & Tuschen, B. (1996). Klinisch-psychologische Diagnostik. In A. Ehlers & K. Hahlweg (Hrsg.), *Psychologische und biologische Grundlagen der Klinischen Psychologie* (Enzyklopädie der Psychologie, Serie Klinische Psychologie, Bd. 1, S. 195–268). Göttingen: Hogrefe.
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1996). *Kognitive Therapie der Depression* (5. Aufl.). Weinheim: Psychologie Verlags Union.
- Becker, P. (2003). *TIPI: Trierer Integriertes Persönlichkeitsinventar*. Göttingen: Hogrefe.
- Beckmann, D., Brähler, E., & Richter, H.-E. (2012). *GT-II: Der Gießen-Test – II*. Bern: Huber.
- Beesdo-Baum, K., Zaudig, M., & Wittchen, H.-U. (Hrsg.). (2019). *SCID-5-PD: Strukturiertes Klinisches Interview für DSM-5® – Persönlichkeitsstörungen*. Göttingen: Hogrefe.
- Belitz-Weihmann, E., & Metzler, P. (1997). *FFT: Fragebogen zum Funktionalen Trinken*. Frankfurt am Main: Sweets Test Services.
- Benjamin, L. S. (1974). Structural Analysis of Social Behavior. *Psychological Review* 81, 392–425.
- Bergeman, N., & Johann, G. K. (1993). *Berger-Skala zur Erfassung der Selbstakzeptanz*. Göttingen: Hogrefe.
- Beutler, L., Malik, M., Alinobamed, S., Harwood, T. M., Talebi, H., Noble, S., & Wong, E. (2004). Therapist variables. In M. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (4th ed., pp. 227–306). New York, NY: Wiley.
- Bohus, M., Limberger, M. F., Frank, U., Sender, I., Gratwohl, T., & Stieglitz, R. (2001). Entwicklung der Borderline-Symptom-Liste. *Psychotherapie, Psychosomatik, Medizinische Psychologie* 51, 201–211.
- Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa & McCrae*. Göttingen: Hogrefe.
- Bronisch, T., Hiller, W., Zaudig, M., & Mombour, W. (1995). *IDCL-P: Internationale Diagnose Checklisten für Persönlichkeitsstörungen nach ICD-10 und DMS-IV*. Bern: Huber.
- Busner, J., & Targum, S. D. (2007). The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)* 4, 28–37.
- Caspar, F. (2008). Plananalyse. In B. Röhrle, F. Caspar, & P. Schlottke (Hrsg.), *Lehrbuch der klinischpsychologischen Diagnostik* (S. 149–166). Stuttgart: Kohlhammer.

- Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology* 66, 7–18.
- Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology* 52, 685–716.
- Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC Index. *Behavior Therapy* 17, 305–308.
- von Collani, G., & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie* 24, 3–7.
- Collegium Internationale Psychiatriae Scalarum. (CIPS). (Hrsg.) (2015). *Internationale Skalen für Psychiatrie* (6. Aufl.). Göttingen: Beltz Test.
- von Consbruch, K., Stangier, U., & Heidenreich, T. (2016). *SOZAS: Skalen zur Sozialen Angststörung. Soziale-Phobie-Inventar (SPIN), Soziale-Interaktions-Angst-Skala (SIAS), Soziale-Phobie-Skala (SPS), Liebowitz-Soziale-Angst-Skala (LSAS). Manual*. Göttingen: Hogrefe.
- Dilling, H., Freyberger, H. J., & Cooper, J. E. (Hrsg.) (2010). *Taschenführer zur ICD-10-Klassifikation psychischer Störungen* (5. Aufl.). Bern: Huber.
- Döpfner, M. (1999). Zwangsstörungen. In H.-C. Steinhäusen, & M. von Aster (Hrsg.), *Verhaltenstherapie und Verhaltensmedizin bei Kindern und Jugendlichen* (2. Aufl., S. 271–326). Weinheim: Beltz, Psychologie Verlags Union.
- Döpfner, M., & Görtz-Dorten, A. (2017). *DISYPS-III: Diagnostik-System für psychische Störungen nach ICD-10 und DSM-5 für Kinder und Jugendliche – III*. Göttingen: Hogrefe.
- Döpfner, M., Dietmair, I., Mersmann, H., Simon, K., & Trost-Brinkhues, G. (2005). *S-ENS: Screening des Entwicklungsstandes bei Einschulungsuntersuchungen*. Göttingen: Hogrefe.
- Döpfner, M., Plück, J., Kinnen, C., & Arbeitsgruppe Deutsche Child Behavior Checklist (2014). *CBCL/6–18R, TRF/6–18R, YSR/11–18R: Deutsche Schulalter-Formen der Child Behavior Checklist von Thomas M. Achenbach. Elternfragebogen über das Verhalten von Kindern und Jugendlichen (CBCL/6–18R), Lehrerfragebogen über das Verhalten von Kindern und Jugendlichen (TRF/6–18R), Fragebogen für Jugendliche (YSR/11–18R)*. Göttingen: Hogrefe.
- Döpfner, M., Berner, W., Fleischmann, T., & Schmidt, M. H. (2018). *VBV 3-6: Verhaltensbeurteilungsbogen für Vorschulkinder* (2. Aufl.). Weinheim: Beltz Test.
- Eckert, J., Biermann-Ratjen, E., & Höger, D. (Hrsg.) (2006). *Gesprächspsychotherapie: Lehrbuch für die Praxis*. Berlin, Heidelberg: Springer.
- Ehlers, A., Margraf, J., & Chambless, D. (2001). *AKV: Fragebogen zu körperbezogenen Ängsten, Kognitionen und Vermeidung* (2. Aufl.). Weinheim: Beltz.
- Ellis, A., & Grieger, R. (1995). *Praxis der rational-emotiven Therapie* (2. Aufl.). Weinheim: Psychologie Verlags Union.
- Esser, G. (Hrsg.) (2008). *Lehrbuch der Klinischen Psychologie und Psychotherapie bei Kindern und Jugendlichen: Ein Lehrbuch* (3. Aufl.). Stuttgart: Thieme.
- Esser, G., Blanz, B., Geisel, B., & Laucht, M. (1989). *MEI: Mannheimer Elterninventar*. Weinheim: Beltz.
- Fahrenberg, J., Hampel, R., & Selg, H. (2020). *FPI-R: Freiburger Persönlichkeitsinventar* (9. Aufl.). Göttingen: Hogrefe.
- Fichter, M., & Keeser, W. (1980). Das Anorexia nervosa-Inventar zur Selbstbeurteilung (ANIS). *Archiv für Psychiatrie und Nervenkrankheiten* 288, 67–89.
- Fichter, M., & Quadflieg, N. (1999). *SIAB: Strukturiertes Inventar für Anorektische und Bulimische Essstörungen nach DSM-IV und ICD-10. Fragebogen (SIAB-S) und Interview (SIAB-EX)*. Handanweisung. Göttingen: Hogrefe.
- Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000. Ein Instrument zur Erfassung von Therapieprozessen. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 39, 71–79.
- Foa, E. B., Riggs, D. S., Dancu, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress* 6, 459–473.
- Franke, G. H. (2014). *SCL-90®-S: Symptom-Checklist-90®-Standard*. Göttingen: Hogrefe.
- Franke, G. H. (2017a). *BSCL: Brief-Symptom-Checklist*. Göttingen: Hogrefe.
- Franke, G. H. (2017b). *Mini-SCL: Mini-Symptom-Checklist*. Göttingen: Hogrefe.
- Frey, A., Duhm, E., & Althaus, D. (2008). *BBK 3-6: Beobachtungsbogen für 3- bis 6-jährige Kinder*. Göttingen: Hogrefe.
- Fydrich, T. (2002). SPAI – Soziale Phobie und Angst Inventar. In E. Brähler, J. Schumacher, & B. Strauß (Hrsg.), *Diagnostische Verfahren in der Psychotherapie* (S. 335–338). Göttingen: Hogrefe.
- Fydrich, T. (2011). Klinisch-psychologische Diagnostik. In L. F. Hornke, M. Amelang, & M. Kersting (Hrsg.) *Grundfragen und Anwendungsfelder psychologischer Diagnostik* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 1, S. 343–382). Göttingen: Hogrefe.
- Fydrich, T. (2016). Beck Inventar für kognitive Schemata. In K. Geue, B. Strauß, & E. Brähler (Hrsg.), *Diagnostische Verfahren in der Psychotherapie* (3. Aufl., S. 50–60). Göttingen: Hogrefe.

- Fydrich, T., & Bürgener, F. (2005). Ratingskalen für soziale Kompetenz. In N. Vriendt, & J. Margraf (Hrsg.), *Soziale Kompetenz, Soziale Unsicherheit, Soziale Phobie* (3. Aufl., S. 81–96). Baltmannsweiler: Schneider-Verlag Hohengehren.
- Fydrich, T., & Sommer, G. (2003). Diagnostik sozialer Unterstützung. In M. Jerusalem, & H. Weber (Hrsg.), *Psychologische Gesundheitsförderung* (S. 79–104). Göttingen: Hogrefe.
- Fydrich, T., Laireiter, A. R., Saile, H., & Engberding, M. (1996). Diagnostik und Evaluation in der Psychotherapie. *Zeitschrift für Klinische Psychologie* 25, 161–168.
- Fydrich, T., Sommer, G., & Brähler, E. (2007). *F-SozU: Fragebogen zur sozialen Unterstützung*. Göttingen: Hogrefe.
- Gast U., & Rodewald, F. (2004) Das Strukturierte Klinische Interview für Dissoziative Störungen (SKIDD). In A. Eckhardt-Henn, & S. O. Hoffmann (Hrsg.), *Diagnostik und Behandlung Dissoziativer Störungen* (S. 321–327). Stuttgart: Schattauer.
- Geissner, E. (1996). *SES: Die Schmerzempfindungs-Skala*. Göttingen: Hogrefe.
- Geue, K., Strauß, B., & Brähler, E. (2016). *Diagnostische Verfahren in der Psychotherapie* (3. Aufl.). Göttingen: Hogrefe.
- Gönner, S., Leonhart, R., & Ecker, W (2007). Das Zwangsinventar OCI-R – die deutsche Version des Obsessive Compulsive Inventory-Revised. Ein kurzes Selbstbeurteilungsinstrument zur mehrdimensionalen Messung von Zwangssymptomen. *Psychotherapie, Psychosomatik, Medizinische Psychologie* 57, 395–404.
- Goodman, R. (1997). SDQ – The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry* 38, 581–586.
- Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., et al. (1989). The Yale-Brown Obsessive Compulsive Scale: I. Development, use, and reliability. *Archives of General Psychiatry* 46, 1006–1011.
- Grosse Holtforth, M. (2001). Was möchten Patienten in ihrer Therapie erreichen? – Die Erfassung und Kategorisierung von Therapiezielen mit dem Berner Inventar für Therapieziele (BIT). *Verhaltenstherapie und psychosoziale Praxis* 33, 241–258.
- Hamilton M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology* 32, 50–55.
- Hamilton, M. (1986). The Hamilton rating scale for depression. In N. Sartorius & T. A. Ban (Eds.), *Assessment of depression* (pp. 278–296). Berlin: Springer.
- Hasenbring, M. (1994). *KSI: Kieler Schmerz-Inventar*. Bern: Huber.
- Hautzinger, M. (2013). *Kognitive Verhaltenstherapie bei Depressionen: Behandlungsleitungen und Materialien* (7. Aufl.). Weinheim: Beltz.
- Hautzinger, M., Keller, F., & Kühner, C. (2009). *BDI-II: Beck Depressions-Inventar II – Revision. Deutsche Ausgabe des Manuals. Original von Aaron T. Beck, Robert A. Steer, Gregory K. Brown* (2. Aufl.). Frankfurt am Main: Pearson.
- Hautzinger, M., Bailer, M., Hofmeister, D., & Keller, F. (2012). *ADS: Allgemeine Depressionsskala* (2. Aufl.). Göttingen: Hogrefe.
- Herrmann-Lingen, C., Buss, U., & Snaith, R. P. (2010). *HADS-D: Hospital Anxiety and Depression Scale – Deutsche Version. Deutsche Adaptation der Hospital Anxiety and Depression Scale (HADS) von R. P. Snaith und A. S. Zigmond* (3. Aufl.). Göttingen: Hogrefe.
- Hiller, W., Zaudig, M., & Mombour, W. (1997). *IDCL Internationale Diagnosen Checklisten für DSM-IV und ICD-10*. Göttingen: Hogrefe.
- Hiller, W., Rief, W., & Pilowsky, I. (2004). *WI-IAS: Internationale Skalen für Hypochondrie. Deutschsprachige Adaptation des Whiteley-Index (WI) und der Illness Attitude Scales (IAS)*. Göttingen: Hogrefe.
- Horn, H. (2006). Dokumentation und Evaluation in der Psychotherapie von Kindern und Jugendlichen. In H. Hopf, & E. Windaus (Hrsg.), *Lehrbuch der Psychotherapie* (Bd. 5, S. 83–96). München: CIP Medien.
- Horowitz, L. M., Thomas, A., Strauß, B., & Kordy, H. (2016). *IIP-D: Inventar zur Erfassung interpersonaler Probleme – Deutsche Version* (3. Aufl.). Weinheim: Beltz.
- In-Albon, T., Suppiger, A., Schlup, B., Wendler, S., Margraf, J., & Schneider, S. (2008). Validität des Diagnostischen Interviews bei psychischen Störungen (DIPS für DSM-IV-TR). *Zeitschrift für Klinische Psychologie und Psychotherapie* 37, 33–42.
- Itten, S., & Grawe, K. (2002). VLB – Veränderungsfragebogen für Lebensbereiche. In E. Brähler, J. Schumacher, & B. Strauss (Hrsg.), *Diagnostische Verfahren in der Psychotherapie* (2. Aufl., S. 382–384). Göttingen: Hogrefe.
- Jacobson, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment* 10, 133–145.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59, 12–19.
- Jacobson, N. S., Folette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy* 15, 336–352.

- John, U., Hapke, U., & Rumpf, H.-J. (2001). *SESA: Die Skala zur Erfassung der Schwere der Alkoholabhängigkeit*. Göttingen: Hogrefe.
- Kanfer, F. H., & Saslow, G. (1976). Verhaltenstheoretische Diagnostik. In D. Schulte (Hrsg.), *Diagnostik in der Verhaltenstherapie* (2. Aufl., S. 24–59). München: Urban & Schwarzenberg.
- Kanfer, F. H., Reinecker, H., & Schmelzer, D. (2012). *Selbstmanagement-Therapie* (5. Aufl.). Berlin: Springer.
- Kay, S. R., Opler, L. A., & Lindenmayer, J. P. (1989). The Positive and Negative Syndrome Scale (PANSS): rationale and standardization. *British Journal of Psychiatry* 1989, 155(Suppl 7): 59–65.
- Kiresuk, T. J., & Sherman, R. R. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal* 4, 443–453.
- Klepsch, R., Zaworka, W., Hand, I., Lünenschloß, K., & Jauernig, G. (1993). *HZI-K: Hamburger Zwangsinventar – Kurzform*. Weinheim: Beltz.
- Klinger, R., Hasenbring, M., & Pfingsten, M. (2016). *Multiaxiale Schmerzklassifikation. Psychosoziale Dimension – MASK-P*. Berlin, Heidelberg: Springer.
- Kuhl, J., & Kazén, M. (2009). *PSSI: Persönlichkeits-Stil- und Störungs-Inventar* (2. Aufl.). Göttingen: Hogrefe.
- Kupper, K., & Rohrmann, S. (2016). *STAXI-2 KJ: Das State-Trait-Ärgerausdrucks-Inventar – 2 für Kinder und Jugendliche. Deutschsprachige Adaptation und Erweiterung des State-Trait Anger Expression Inventory-2 Child and Adolescent (STAXI-2 C/A) von Thomas M. Brunner und Charles D. Spielberger*. Göttingen: Hogrefe.
- Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults* (pp. 191–234). Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.
- Lauth, G. W. (2003). DAT: Dortmunder Aufmerksamkeitstest. Göttingen: Hogrefe.
- Linden, M., & Hautzinger, M. (2015). *Verhaltenstherapiemanual* (6. Aufl.). Berlin: Springer.
- Luborsky, L., Albani, C., Eckert, R. (1992). Manual zur ZBKT-Methode (deutsche Übersetzung mit Ergänzungen). *Psychotherapie, Psychosomatik, Medizinische Psychologie Disc-Journal* 5.
- Maercker, A., & Schützwohl, M. (1998). Erfassung von psychischen Belastungsfolgen: Die Impact of Event Skala – revidierte Version. *Diagnostica* 44, 130–141.
- Margraf, J., & Ehlers, A. (2007). *BAI: Beck Angst-Inventar. Manual von Aaron T. Beck, Robert A. Steer – deutsche Bearbeitung*. Frankfurt am Main: Harcourt Test Services.
- Margraf, J., & Cwik, J. C. (2017). Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen. Bochum: Forschungs- und Behandlungszentrum für psychische Gesundheit, Ruhr-Universität Bochum. Doi: ► <https://doi.org/10.13154/rub.102.91>.
- Margraf, J., Cwik, J. C., Suppiger, A., & Schneider, S. (2017). *DIPS Open Access: Diagnostisches Interview bei psychischen Störungen*. Bochum: Forschungs- und Behandlungszentrum für psychische Gesundheit, Ruhr-Universität Bochum. doi: ► <https://doi.org/10.13154/rub.100.89>.
- Mattejat, F., & Remschmidt, H. (2003). Inventar zur Erfassung der Lebensqualität bei Kindern und Jugendlichen. In J. Schumacher, A. Klaiberg, & E. Brähler (Hrsg.), *Diagnostische Verfahren zu Lebensqualität und Wohlbefinden* (S. 176–179). Göttingen: Hogrefe.
- Hathaway, S. R., McKinley, J. C., & Engel, R. R. (2000). *MMPI-2: Minnesota Multiphasic Personality Inventory-2*. Bern: Huber.
- Melfsen, S., Florin, I., & Warnke, A. (2001). *SPAIK: Das Sozialphobie und -angstinventar für Kinder*. Göttingen: Hogrefe.
- Mombour, W., Zaudig, M., Berger, P., Gutierrez, K., Berner, W., Berger, K., von Cranach, M., Giglhuber, O., & von Bose, M. (1996). *International Personality Disorder Examination (IPDE), ICD-10-Modul von A. W. Loranger*. Bern: Huber.
- Morfeld, M., Kirchberger, I., & Bullinger, M. (2011). *SF-36 – Fragebogen zum Gesundheitszustand* (2. Aufl.). Göttingen: Hogrefe.
- Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology* 31, 109–118.
- Paul, T., & Thiel, A. (2004). *EDI-2: Eating Disorder Inventory-2. Deutsche Version*. Göttingen: Hogrefe.
- Perrez, M., & Baumann, U. (2005). *Lehrbuch Klinische Psychologie: Psychotherapie*. Bern: Huber.
- Petermann, F., & Petermann, U. (2015). *EAS: Erfassungsbogen für aggressives Verhalten in konkreten Situationen* (5. Aufl.). Göttingen: Hogrefe.
- Renneberg, B., & Seehausen, A. (2010). Fragebogen zu Gedanken und Gefühlen (FGG): Ein Screening Instrument für Borderline-spezifisches Denken. *Zeitschrift für Klinische Psychologie und Psychotherapie* 39, 170–178.
- Rief, W., & Hiller, W. (2007). *SOMS: Screening für Somatoforme Störungen* (2. Aufl.). Bern: Huber.
- Rohrmann, S., Hodapp, V., Schnell, K., Tibubos, A., Schwenkmezger, P., Spielberger, C. D. (2013). *STAXI-2: Das State-Trait-Ärgerausdrucks-Inventar – 2. Deutschsprachige Adaptation des State-Trait Anger Expression Inventory-2 (STAXI-2) von Charles D. Spielberger*. Bern: Huber.

- Rossmann, P. (2004). *DTK-II: Depressionstest für Kinder – II*. Göttingen: Hogrefe.
- Rost, D. H., & Schermer, F. J. (2007). *DAI: Differentielles Leistungsangst-Inventar* (2. Aufl.). Frankfurt am Main: Pearson.
- Rudolf, G. (1993). *Psychischer und Sozial-Kommunikativer Befund (PSKB). Ein Instrument zur standardisierten Erfassung neurotischer Befunde*. Göttingen: Hogrefe.
- Rumpf, H. J., Hapke, U., & John, U. (2001). *LAST: Lübecker Alkoholabhängigkeit- und -missbrauchs-Screening-Test*. Göttingen: Hogrefe.
- von Schlippe, A., & Schweitzer, J. (2016). *Lehrbuch der systemischen Therapie und Beratung* (3. Aufl.). Göttingen: Vandenhoeck & Ruprecht.
- Schneider, S., & Margraf, J. (2019). *Lehrbuch der Verhaltenstherapie. Bd. 3: Störungen im Kindes- und Jugendalter* (2. Aufl.). Heidelberg: Springer.
- Schneider, S., Pflug, V., In-Albon, T., & Margraf, J. (2017). *Kinder-DIPS Open Access: Diagnostisches Interview bei psychischen Störungen im Kindes- und Jugendalter*. Bochum: Forschungs- und Behandlungszentrum für psychische Gesundheit, Ruhr-Universität Bochum. doi: ▶ <https://doi.org/10.13154/rub.101.90>.
- Schneider, W., Basler, H.-D., & Beisenherz, B. (1989). *FMP: Fragebogen zur Psychotherapiemotivation*. Göttingen: Beltz Test.
- Schütz, C. G., Daamen, M., & van Niekerk, C. (2005). Deutsche Übersetzung des WHO ASSIST Screening-Fragebogens. *Sucht* 51, 265–271.
- Spitzer, C., Abraham, G., Reschke, K., & Freyberger, H. J. (2001). Die deutsche Version der Modified PTSD Symptom Scale (MPSS): Erste psychometrische Befunde zu einem Screeningverfahren für posttraumatische Symptomatik. *Zeitschrift für Klinische Psychologie und Psychotherapie* 30, 159–163.
- Spitzer, C., Stieglitz, R.-D., & Freyberger, H. J. (2005). FDS: Fragebogen zu Dissoziativen Symptomen. Ein Selbstbeurteilungsverfahren zur syndromalen Diagnostik dissoziativer Phänomene – Deutsche Adaptation der Dissociative Experiences Scale (DES) von E. Bernstein-Carlson und F. W. Putnam (3. Aufl.). Bern: Huber.
- Steinhausen, H. C., & von Aster, M. (Hrsg.). (2000). *Verhaltenstherapie und Verhaltensmedizin bei Kindern und Jugendlichen* (2. Aufl.). Weinheim: Psychologie Verlags Union.
- Stenzel, N., Krumm, S., & Rief, W. (2010). Therapieplanung mithilfe des Interviews zur operationalisierte Fertigkeitsdiagnostik (OFD). *Verhaltenstherapie* 20, 109–117.
- Stieglitz, R.-D., Baumann, U., & Freyberger, H. J. (Hrsg.). (2001). *Psychodiagnostik in Klinischer Psychologie, Psychiatrie, Psychotherapie* (2. Aufl.). Stuttgart: Thieme.
- Stiensmeier-Pelster, J., Braune-Krickau, M., Schürmann, M., & Duda, K. (2014). *DIKJ: Depressionsinventar für Kinder und Jugendliche* (3. Aufl.). Göttingen: Hogrefe.
- Stöber, J., & Bittencourt, J. (1998). Weekly assessment of worry: An adaptation of the Penn State Worry Questionnaire for monitoring changes during treatment. *Behaviour Research and Therapy* 36, 645–656.
- Tewes, A., & Naumann, A. (2016). *KAT-III: Kinder-Angst-Test-III. Drei Fragebögen zur Erfassung der Ängstlichkeit und von Zustandsängsten bei Kindern und Jugendlichen*. Göttingen: Hogrefe.
- Tress, W. (Hrsg.). (2003). *SASB: Die Strukturelle Analyse sozialen Verhaltens*. München: Psycho-sozial-Verlag.
- Tritt, K., von Heymann, F., Zaudig, M., Probst, T., Loew, T., Klapp, B., Söllner, W., Fydrich, T., & Bühner, M. (2015). *ICD-10-Symptom-Rating (ISR): Das Manual*. Ebook. Neobooks, München.
- Tröskens, A., & Grawe, K. (2004). Inkongruenzerleben aufgrund brachliegender und fehlender Ressourcen: Die Rolle von Ressourcenpotentialen und Ressourcenrealisierung für die Psychologische Therapie. *Verhaltenstherapie und psychosoziale Praxis* 36, 51–62.
- Wieczorkowski, W., Nickel, H., Janowski, A., Fittkau, B., Rauer, W., & Petermann, F. (2016). *AFS: Angstfragebogen für Schüler* (7. Aufl.). Göttingen: Hogrefe.
- Willutzki, U., Ülsmann, D., Schulte, D., & Veith, A. (2013). Direkte Veränderungsmessung in der Psychotherapie: Der Bochumer Veränderungsbogen-2000 (BVB-2000). *Zeitschrift für Klinische Psychologie und Psychotherapie* 42, 256–268.
- Wilmers, F., Munder, T., Leonhart, R., & Herzog, T. (2008). Die deutschsprachige Version des Working Alliance Inventory – short revised (WAI-SR) – Ein schulenübergreifendes, ökonomisches und empirisch validiertes Instrument zur Erfassung der therapeutischen Allianz. *Klinische Diagnostik und Evaluation* 1, 343–358.
- Wittchen, U., & Hoyer, J. (2011). *Klinische Psychologie und Psychotherapie* (2. Aufl.). Heidelberg: Springer.
- Wittchen, H.-U., Zaudig, M., & Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen/Achse II: Persönlichkeitsstörungen*. Göttingen: Hogrefe.
- World Health Organization (WHO). (2019). ICD-11 for Mortality and Morbidity Statistics (Version: 04/2019). ▶ <https://icd.who.int/browse11/l-m/en>. Zugegriffen: 04. Juni 2020.
- Zielke, M. (1979). *KASSL: Kieler Änderungssensitive Symptomliste*. Weinheim: Beltz.
- Zielke, M., & Kopf-Mehnert, C. (1978). *VEV: Veränderungsfragebogen des Erlebens und Verhaltens*. Weinheim: Beltz.



Diagnostik in weiteren Anwendungsfeldern

Lothar Schmidt-Atzert, Stefan Krumm und Manfred Amelang

Inhaltsverzeichnis

- 9.1 Neuropsychologische Diagnostik – 732**
 - 9.1.1 Neuropsychologische Untersuchung – 735
 - 9.1.2 Spezielle Probleme der neuropsychologischen Diagnostik – 738
- 9.2 Rechtspsychologische Diagnostik – 746**
 - 9.2.1 Glaubhaftigkeit von Zeuginnen und Zeugen – 749
 - 9.2.2 Schuldunfähigkeit und verminderte Schuldfähigkeit – 750
 - 9.2.3 Kriminalprognose – 751
 - 9.2.4 Familiengericht: Sorgerechtsentscheidungen – 758
- 9.3 Verkehrspychologische Diagnostik – 762**
 - 9.3.1 Begutachtung der Fahreignung für den Straßenverkehr – 763
 - 9.3.2 Spezielle Probleme der verkehrspychologischen Diagnostik – 771
- 9.4 Zusammenfassung – 774**
- Literatur – 777**

9.1 Neuropsychologische Diagnostik

Störungen aufgrund von Defekten im Zentralnervensystem

Organische Ursache manchmal unklar

Arbeitsplatz Klinik, Reha-Einrichtung, freie Praxis

Defizite nach Art und Schwere beschreiben

Gegenstand neuropsychologischer Diagnostik Die neuropsychologische Diagnostik befasst sich mit Störungen, die auf (angeborene oder erworbene) Defekte im Zentralnervensystem zurückzuführen sind.

- » Ziel der neuropsychologischen Diagnostik ist die Erfassung und Objektivierung von kognitiven und affektiven Funktionsstörungen nach einer Hirnfunktionsstörung oder Hirnschädigung und ggf. der emotionalen Reaktionen des Patienten auf diese Störungen. (Gesellschaft für Neuropsychologie et al. 2005, S. 185).

Das menschliche Gehirn kann durch vielfältige Einflüsse geschädigt werden. Die Einwirkung auf das Gehirn kann spezifisch sein und nur eine kleine Region betreffen, z. B. bei einem kleinen Hirntumor oder einem Schlaganfall. In anderen Fällen kann die Schädigung eher unspezifisch sein, wenn große Teile des Gehirns betroffen sind. Dazu gehören degenerative Erkrankungen wie Alzheimer, Gewalteinwirkung auf das gesamte Gehirn durch einen Unfall oder eine Sauerstoffunterversorgung bei einem schweren Herzinfarkt. Die Schädigung kann durch mechanische Einwirkungen (z. B. Schädel-Hirn-Verletzung), durch chemische Stoffe (Aufnahme neurotoxischer Stoffe durch die Atemluft oder die Nahrung), durch Bakterien (z. B. Neuroborreliose, ausgelöst durch Zeckenbiss), durch Viren (z. B. viral bedingte Hirnhautentzündung) oder etwa durch Sauerstoffmangel hervorgerufen werden. Bei manchen Erkrankungen des Gehirns kennt man die genaue Ursache (noch) nicht, obwohl sicher ist, dass das Zentralnervensystem betroffen ist (z. B. Alzheimer-Erkrankung). Neuropsychologinnen und Neuropsychologen befassen sich zum Teil aber auch mit Störungen, bei denen eine organische Ursache im Gehirn lediglich angenommen wird, die neuroanatomischen und neurophysiologischen Grundlagen jedoch noch nicht hinreichend geklärt sind. Dazu gehören Aufmerksamkeitsstörungen (ADHS), Legasthenie, Dyskalkulie (Rechenschwäche) und Entwicklungsstörungen, etwa im Bereich der Sprache. Die Frage, ob diese Störungen (partiell) genetisch bedingt oder erworben sind, ist für die Diagnostik von untergeordneter Bedeutung. In □ Tab. 9.1 sind einige relativ häufige neurologische Erkrankungen aufgeführt. Allerdings ist zu beachten, dass die hier aufgeführten Störungen nicht ausschließlich in den Arbeitsbereich von Neuropsychologinnen und Neuropsychologen fallen; viele dieser Störungen werden auch in anderen Disziplinen der Psychologie (z. B. Klinische Psychologie, Pädagogische Psychologie) diagnostiziert und ggf. behandelt.

Arbeitsgebiete Neuropsychologinnen und Neuropsychologen arbeiten überwiegend in neurologischen Kliniken, in rehabilitativen Einrichtungen, in psychiatrischen Kliniken und zunehmend auch in freien psychologischen Praxen. Die Gesellschaft für Neuropsychologie (► <https://www.gnp.de/>), die die fachlichen und beruflichen Interessen von Psychologinnen und Psychologen im Bereich der Neuropsychologie vertritt, verzeichnet etwa 1600 Mitglieder (Stand: Mai 2020). Die Zahl der praktizierenden Neuropsychologinnen und Neuropsychologen ist größer, da nicht alle Mitglied in dieser Gesellschaft sind.

Die neuropsychologische Diagnostik dient dazu, Defizite in diesen Funktionsbereichen nach ihrer Art und Schwere zu beschreiben. Innerhalb

Tab. 9.1 Ausgewählte neurologische Erkrankungen

Erkrankung	Neurologische Symptome	(Mögliche) Ursache	Häufigkeit ^a
Akute zerebrale Zirkulationsstörung („Schlaganfall“)	Ausfall der Funktion der Gehirnteile, die nicht mehr (hinreichend) mit Blut versorgt werden (z. B. <i>Aphasie, Neglect</i>)	Trombus (Blutgerinnsel)	200.000 (Neuerkrankungen pro Jahr)
Spontane intrakranielle Blutung	Bewusstseinstrübung	Gefäßmissbildung	Keine Angabe
Schädel-Hirn-Trauma	Leichte Bewusstseinsstörung bis Koma, u. a. auch Aufmerksamkeitsstörung	Verkehrsunfall mit Kopfverletzung	200.000 (Neuerkrankungen pro Jahr)
Gehirntumore (Krebs)	Zum Beispiel Kopfschmerz, epileptische Anfälle	Wie bei anderen Tumoren	Keine Angabe
Bakterielle und virale Entzündungen des Gehirns oder der Gehirnhäute	Zum Beispiel Kopfschmerz, Überempfindlichkeit gegen Licht und Geräusche, Bewusstseinsstörung	Bakterien (z. B. Staphylokokken) dringen über den Blutkreislauf ein, HIV-Infektion	Keine Angabe
Epilepsien (zerebral bedingte Krampfanfälle)	Verschiedene Formen von Anfällen, auch solche mit neuropsychologischen Symptomen	Tumor, Narbe	400.000 Erkrankte
Morbus Parkinson	Ruhetremor, der unter mentaler Belastung zunimmt, <i>Rigor, Akinese</i>	Degeneration dopaminerger Neurone in der Substantia nigra	200.000 Erkrankte
Multiple Sklerose	Ataxie	Autoimmunerkrankung des Zentralnervensystems	Jeder 1000. Mensch
Demenzen, u. a. vom Alzheimer-Typ	Gedächtnisstörungen, Wortfindungsstörungen	Alzheimer: Degeneration von Nervenzellen	Über 500.000 Erkrankte

Quelle: Auswahl, zusammengestellt nach Wallesch und Herrmann (2000). Die neurologischen Symptome sind unvollständig aufgeführt.

Akinese = verminderte Spontanmotorik (aber keine Lähmung); *Aphasia* = erworbene Sprachstörung, Ataxie = Koordinationsstörung mit unter- und überschießenden Bewegungen; *Neglect* = Wahrnehmungs- oder Aufmerksamkeitsstörung, bei der Objekte auf einer Körperseite nicht erkannt werden; *Rigor* = anhaltende Erhöhung der Muskelspannung

^aUngefähr Anzahl der Erkrankten bzw. der Neuerkrankungen pro Jahr in Deutschland

der Funktionsbereiche sind weitere *Differenzierungen* vorzunehmen. Beispielsweise wird bei Störungen im Bereich der Sprache zwischen verschiedenen Formen der Aphasie unterschieden: Schwierigkeiten beim Verstehen und bei der Produktion von gesprochener vs. geschriebener Sprache. Möglicherweise ist auch die Motorik der Sprechwerkzeuge beeinträchtigt mit Auswirkungen auf die Artikulation, den Klang der Stimme und die Melodie der Sprache. Manchmal wird ein Defizit auch erst im Rahmen einer diagnostischen Untersuchung entdeckt.

Betroffene Funktionsbereiche

Von Hirnschädigung betroffene Funktionsbereiche

Als Folge einer Hirnschädigung können in verschiedenen Funktionsbereichen Einschränken auftreten; hier eine Auflistung wichtiger Funktionsbereiche (Gesellschaft für Neuropsychologie et al. 2005, S. 185):

- Basale und höhere Wahrnehmungsleistungen
- Aufmerksamkeitsleistungen
- Gedächtnisfunktionen
- Planungs- und Kontrollfunktionen („exekutive Funktionen“)
- Sprache
- Sensomotorische Leistungen und motorische Planung
- Räumlich-perzeptive, räumlich-kognitive und räumlich-konstruktive Leistungen
- Zahlenverarbeitung und Rechenleistungen
- Intellektuelles Niveau und Leistungsprofil (aggregierte Kompetenz)
- Berufsabhängige Fertigkeiten und domänen spezifisches Wissen
- Affektivität und Persönlichkeit

Erklärung und Prognose

9

Medizinische Diagnostik im Vorfeld

Hypothesengeleitetes Vorgehen

Zweck der Quantifizierung von Funktionsbeeinträchtigungen

Auch wenn die Beschreibung und Klassifikation einer Störung die häufigste Aufgabe ist, kann in manchen Fällen auch eine Erklärung für eine Schädigung gesucht werden. So kann nach einem Unfall die Frage zu beantworten sein, ob die nun beobachteten Defizite ganz oder vielleicht nur teilweise unfallbedingt sind. Oftmals sind zudem Prognosen über den weiteren Verlauf erforderlich; Patientinnen und Patienten oder deren Angehörige wollen wissen, ob eine Wiederaufnahme des alten Berufs möglich ist, ob eine Umschaltung mit Erfolg in einen anderen Beruf führen kann oder ob und wie gut es der betroffenen Person gelingen wird, im Alltag zurechtzukommen.

In den meisten Fällen, besonders in neurologischen Kliniken und Rehabilitationseinrichtungen, geht der neuropsychologischen Diagnostik eine umfangreiche medizinische Diagnostik voraus. Mithilfe von bildgebenden Verfahren wie Positionen-Emissions-Tomografie (PET), funktioneller Kernspinntomografie (fMRT) und Magnetresonanztomografie (MRT) können viele pathologische Veränderungen des Gehirns erkannt und genau lokalisiert werden. Solche medizinischen Befunde sind für die Planung einer neuropsychologischen Untersuchung wichtig.

Wenn eine Schädigung durch die medizinische Diagnostik lokalisiert ist, kann der Bereich möglicher funktioneller Ausfälle oft gut eingegrenzt werden, da bekannt ist, zu welchen funktionalen Ausfällen bestimmte Schädigungen des Gehirns führen können. Die psychologische Diagnostik erfolgt nun hypothesengeleitet. Aber auch Berichte von Logopädinnen und Logopäden etwa, die bereits mit der Patientin oder dem Patienten gearbeitet haben, liefern weitere Hinweise auf die Art und den Umfang bestimmter Beeinträchtigungen.

Die neuropsychologische Diagnostik dient unter diesen Voraussetzungen nicht dazu, eine medizinische Diagnose zu überprüfen, sondern soll vielmehr die bereits bekannten Funktionsbeeinträchtigungen objektivieren, d. h. quantitativ genau bestimmen. Diese Quantifizierung erfüllt, je nach Fragestellung, unterschiedliche Aufgaben.

Beispiele für Fragestellungen in der neuropsychologischen Diagnostik

- Therapieindikation
- Dokumentation des Krankheitsverlaufs (Evaluation von Therapiemaßnahmen, Notwendigkeit weiterer Therapiemaßnahmen erkennen, „natürlichen“ Krankheitsverlauf verfolgen)
- Abschätzung der Auswirkung der Erkrankung auf die berufliche Wiedereingliederung und die private Lebensgestaltung
- Klärung von Versicherungsfragen

Im Rahmen von Therapien oder Reha-Maßnahmen reicht eine einmalige diagnostische Untersuchung in der Regel nicht aus. Verlaufsuntersuchungen dienen dazu, die Genesung genau zu verfolgen. So lässt sich zuverlässig erkennen, ob und wie gut die Patientin oder der Patient von den Therapiemaßnahmen profitiert. Manche Erkrankungen zeichnen sich durch einen charakteristischen Eigenverlauf aus, der durch therapeutische Maßnahmen nur wenig beeinflusst werden kann. Schubweise Verschlechterungen können sich mit stabilen Phasen abwechseln, oder es tritt eine kontinuierliche Verbesserung oder Verschlimmerung des Zustands ein. In diesem Fall wird der „natürliche“ Krankheitsverlauf abgebildet. Nach Abschluss der Rehabilitationsmaßnahmen bleiben manchmal noch Beeinträchtigungen bestehen, mit denen sich die Patientin oder der Patient arrangieren muss. Die Diagnostik hilft in diesem Fall, die Schädigung sowie ihre Auswirkungen auf eine berufliche Tätigkeit und die private Lebensgestaltung genau zu beschreiben. Es gilt, Kompensationsmöglichkeiten zu entdecken. Die Diagnostik wird dementsprechend breit angelegt sein und auch intakte Funktionsbereiche einbeziehen: Stärken wie auch weitere Schwächen, die vielleicht schon vor der Erkrankung bzw. Verletzung vorgelegen haben, sollen erkannt werden. Die Gesellschaft für Neuropsychologie rät in ihren Leitlinien zur Diagnostik und Therapie explizit von einer reinen Defizitorientierung ab und empfiehlt, auch die Ressourcen der Patientin bzw. des Patienten zu berücksichtigen (Gesellschaft für Neuropsychologie et al. 2005).

Für viele neurologische Patientinnen und Patienten führt ihre Erkrankung zu gravierenden Veränderungen in der Lebensführung. Deshalb werden im Rahmen der neuropsychologischen Diagnostik oft auch die psychischen Folgen der Hirnschädigung, die Krankheitsbewältigung und die emotionale Belastbarkeit der Betroffenen untersucht.

Schließlich dient neuropsychologische Diagnostik auch dazu, für Behörden und Versicherungen diverse Fragen zu klären. In □ Tab. 9.2 sind die wichtigsten Begutachtungsanlässe aufgeführt (ausführliche Informationen bei Neumann-Zielke et al. 2009).

Verlaufsdiagnostik

Auswirkungen auf Lebensführung beschreiben

Versicherungsfragen klären

Diagnostisches Interview

Neuropsychologische und andere Tests

9.1.1 Neuropsychologische Untersuchung

In einer neuropsychologischen Untersuchung hat das diagnostische Interview (Anamnese und, je nach Fragestellung, Exploration der selbst wahrgenommenen Funktionseinschränkungen, der sozialen und beruflichen Situation, der früheren Leistungsfähigkeit etc.) einen hohen Stellenwert.

Zur Quantifizierung der Funktionsbeeinträchtigungen finden zahlreiche Leistungstests Verwendung, die zum Teil speziell für die Neuropsychologie konstruiert wurden. Einige Tests, etwa Intelligenztests, wurden primär für andere Zwecke entwickelt, finden aber auch bei neuropsychologischen Fragestellungen Verwendung.

Tab. 9.2 Anlässe für neuropsychologische Begutachtungen

Auftraggeber	Mögliche Fragestellungen
<i>Gesetzliche Sozialversicherungen</i>	
Unfall (z. B. Berufsgenossenschaft)	Was ist in Folge der Berufskrankheit/des Berufsunfalls an erwerbsbezogener Leistungsfähigkeit verloren gegangen?
Rente (Deutsche Rentenversicherung)	Kann die Person ihren Beruf weniger als 6 h täglich ausüben?
<i>Private Versicherungen</i>	
Unfallversicherung	Was ist in Folge des Unfalls an (körperlicher und) geistiger Leistungsfähigkeit verloren gegangen?
Haftpflichtversicherung	Wie wirkt sich die Schädigung auf die berufliche und private Lebensführung aus?
Berufsunfähigkeitsversicherung	Was kann die Person noch leisten? Welche Fähigkeiten zur Berufsausführung sind verloren gegangen?
<i>Behörden und Gerichte</i>	
Sozialgericht	Welche Auswirkungen hat eine Behinderung auf alle Lebensbereiche?
Vormundschaftsgericht	Ist die Einsichtsfähigkeit oder die freie Willensbestimmung eingeschränkt oder aufgehoben? (Bei Fragen der Geschäfts-, Testierfähigkeit, Betreuung)
Straf- und Zivilgerichte	Ist die Einsichtsfähigkeit oder die freie Willensbestimmung eingeschränkt oder aufgehoben? (Bei Fragen der Schuldfähigkeit, Verhandlungsfähigkeit)

Quelle: Gesellschaft für Neuropsychologie et al. (2009, Tab. 3, © Hogrefe)

Benton-Test als Klassiker

Früher wurden neuropsychologische Tests häufig eingesetzt, um den Verdacht auf eine *hirnorganische Störung* zu verifizieren oder um eine hirnorganische Störung zu entdecken. Heute übernimmt überwiegend die medizinische Diagnostik diese Aufgabe. Das klassische Beispiel für einen Test zur Feststellung einer hirnorganischen Störung ist der Benton-Test (Benton-Sivan und Spreen 2009), dessen Erstveröffentlichung bereits 1946 erfolgte. Mit dem Test sollen Probleme der visuellen Merkfähigkeit festgestellt werden, die auf eine hirnorganische Schädigung hinweisen können. Die Aufgabe besteht (in der am häufigsten verwendeten Form A) darin, 1–3 einfache geometrische Figuren 10 s lang anzuschauen und sie dann aus dem Gedächtnis nachzuzeichnen. Wenn die Patientin oder der Patient motorisch beeinträchtigt ist, kann die Wahlform verwendet werden, bei der lediglich ein Wiedererkennen der gezeigten Vorlage verlangt wird.

Zur Intelligenzdiagnostik finden die gleichen Verfahren wie in anderen Praxisbereichen Verwendung (► Abschn. 3.2.3). Auch die meisten Aufmerksamkeits- und Konzentrationstests (► Abschn. 3.2.2) wurden nicht speziell für die Neuropsychologie entwickelt. Lediglich die Testbatterie zur Aufmerksamkeitsprüfung (TAP) von Zimmermann und Fimm (2017) wurde eigens für die neuropsychologische Diagnostik konstruiert. Es handelt sich dabei um eine computerbasierte Testbatterie mit vielen Untertests, die seit mindestens 1993 auf dem Markt ist und mehrfach aktualisiert wurde. Zudem liegt mit der KiTAP eine ähnlich aufgebaute Version für Kinder vor (Zimmermann et al. 2002).

Für Gedächtnistests besteht speziell in der Neuropsychologie ein großer Bedarf, weil viele Patientinnen und Patienten unter Gedächtnissstörungen leiden. Zu diesem Bereich liegen zahlreiche Testentwicklungen aus der Neuropsychologie vor. Das Lernmaterial kann visuell oder (bei Zahlen und sprachlichem Material) auch akustisch vorgegeben werden. Während beim Benton-Test (s. o.) lediglich einfache geometrische Figuren verwendet werden, kommen bei einem anderen sehr bekannten Gedächtnistest, der 4. Auflage der Wechsler Memory Scale (Wechsler 2012), figurale und verbale Stimuli vor. Die Wechsler Memory Scale wurde erstmals 1945 publiziert und

Intelligenz, Aufmerksamkeit und Konzentration

Gedächtnis

seitdem mehrmals überarbeitet. Die deutsche Ausgabe ist für den Altersbereich von 16 bis 90 Jahren normiert. Zu dem Test liegt eine umfangreiche Forschung vor.

Aphasien sind eine weitere Domäne der Neuropsychologie mit entsprechenden (aber wenigen) Testentwicklungen. Planung und Kontrolle von Verhalten (exeutive Funktionen) werden ebenfalls besonders in der Neuropsychologie thematisiert. Allerdings liegen dazu nur wenige Tests vor. Im deutschen Sprachraum ist der Aachener Aphasie Test (Huber et al. 1983) bekannt. Zur Messung der Reaktionsfähigkeit sowie verschiedener Aspekte der (Psycho-)Motorik, der Affektivität und der Persönlichkeit kann auf bewährte Tests zurückgegriffen werden, die nicht speziell für neuropsychologische Fragestellungen entwickelt wurden.

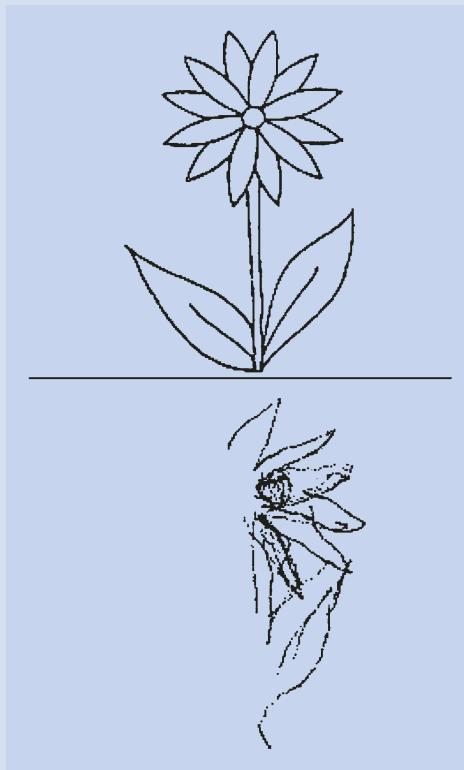
Ein Phänomen, das vielen Laien rätselhaft erscheint, ist der Neglect. Patientinnen und Patienten mit dieser Störung vernachlässigen Reize, die sich auf einer Seite des Wahrnehmungsfelds befinden. Sie reagieren nicht auf entsprechende Reize. Die Störung kann neben dem visuellen Bereich auch andere Sinnesmodalitäten betreffen. Als Erklärung für dieses Defizit wird eine Störung der Aufmerksamkeit, der mentalen Repräsentation der Umwelt und des neuronalen Raumkoordinatensystems diskutiert (Karnath 2009).

Psychomotorik, Affektivität und Persönlichkeit

Neglect – „Vernachlässigung“ einer Seite der Umwelt

Patientinnen und Patienten stoßen beispielsweise mit der Schulter am Türrahmen an, übersehen einen Stuhl oder essen nur eine Seite des Tellers leer. Obwohl die Augen intakt sind, nehmen sie eine Seite ihrer Umwelt nicht wahr. Folgende Abbildung zeigt, wie ein Neglect-Patient eine Blume wahrnimmt und abzeichnet.

Neglect: Eine Seite des Wahrnehmungsfelds ist „ausgeschaltet“



Aufgabe aus dem Neglect-Test (nach Fels und Geissner 1997, © Hogrefe) und Nachzeichnung eines Neglect-Patienten (mit freundlicher Genehmigung von Dipl.-Psych. R. Momtazi).

9.1.2 Spezielle Probleme der neuropsychologischen Diagnostik

Mensch als komplexes System

Bei einer neuropsychologischen Untersuchung kann sich die Diagnostikerin oder der Diagnostiker mit einer Reihe von Problemen konfrontiert sehen (s. auch Hartje 2004). Der Mensch ist als ein komplexes System zu verstehen; eine hirnorganische Störung macht sich daher unter Umständen im Verhalten nicht deutlich bemerkbar, weil das Defizit kompensiert werden kann. So kann eine Patientin oder ein Patient beispielsweise eine ausgeprägte Störung der Merkfähigkeit durch den Einsatz von Notizzetteln kompensieren.

Komorbidität

Umgekehrt lässt sich ein beobachtetes Leistungsdefizit eventuell nicht eindeutig einer Störung zuordnen, weil sich eine psychische Störung und/oder Medikamente ebenfalls auf die momentane Leistung auswirken. Leistungseinbußen, z. B. bei der Aufmerksamkeit, der Konzentration oder dem Gedächtnis, sind manchmal auf eine Depression zurückzuführen, die sich auch als Folge der mit der Hirnschädigung verbundenen veränderten Lebensumstände eingestellt haben kann. Depressive Menschen können Gedächtnisstörungen entwickeln, die an eine Demenz erinnern („Pseudodemenz“). Der Nachweis einer Depression reicht jedoch nicht aus, um eine Demenz auszuschließen, denn Demenz und Depression können durchaus gemeinsam auftreten (Komorbidität). Deshalb ist eine sehr sorgfältige Diagnostik angemessen, bei der manchmal nicht nur die naheliegenden Hypothesen geprüft, sondern auch alternative Erklärungsmöglichkeiten in Erwägung gezogen werden.

Zwei Probleme werden nun ausführlicher erörtert: die Einschätzung der prämorbid Leistungsfähigkeit und die Verfälschung.

Prämorbid Leistungsfähigkeit schätzen

Prämorbid Leistungsfähigkeit Die Beurteilung der kognitiven Leistungsfähigkeit vor einer Erkrankung oder einem Unfall kann beispielsweise bei Versicherungsfragen von großer Bedeutung sein, um den Verlust an geistiger Leistungsfähigkeit zu quantifizieren. Aber auch wenn man einschätzen will, wie weit eine degenerative Erkrankung fortgeschritten oder wie schwer sie ist, muss man wissen, wie groß die Leistungsfähigkeit vor der Erkrankung war.

Alle untersuchten Funktionen weisen auch ohne Schädigung eine große Variation auf. Eine schlechte Aufmerksamkeitsleistung kann bereits vor der vermeintlichen Schädigung vorgelegen haben, und eine heute durchschnittliche Gedächtnisleistung kann das Relikt einer früher exzellenten Merkfähigkeit sein. In ähnlicher Weise stellt sich das Problem der früheren Leistungsfähigkeit bei der Unterscheidung von normalen und pathologischen Veränderungen im Alter dar. Ist eine reduzierte Gedächtnisleistung Ausdruck eines pathologischen Prozesses oder liegt sie angesichts der früheren Fähigkeiten im Bereich des normalen Altersabbaus? Wenn keine früheren Testbefunde vorliegen – was in der Regel so sein wird –, kann der Ausgangszustand nur indirekt beurteilt werden. Grundsätzlich gibt es dazu 3 Möglichkeiten:

- Die Diagnostikerin oder der Diagnostiker kann Informationen über den früheren Zustand der Patientin oder des Patienten erheben und über eine „klinische Urteilsbildung“ zu einer Einschätzung kommen. Wichtig können Angaben über den ausgeübten Beruf, frühere Schul- und Studienleistungen, Hobbys sowie Leistungseinschätzungen durch Angehörige oder die Patientin bzw. den Patienten selbst sein.
- Die Diagnostikerin oder der Diagnostiker kann auf die „statistische Urteilsbildung“ vertrauen und eine Schätzformel wie die *Barona demographic regression equation* oder die *Oklahoma premorbid intelligence estimate-3* verwenden, die sich ebenfalls auf biografische Daten stützen (s. dazu McCaffrey und Vanderslice-Barr 2010).

Klinische Urteilsbildung auf Basis biografischer Daten

Statistische Urteilsbildung auf Basis biografischer Daten

- Die Diagnostikerin oder der Diagnostiker kann spezielle Tests zur Schätzung der prämorbiden Intelligenz einsetzen. Diesen Tests liegt die Überlegung zugrunde, dass sich bestimmte Intelligenzkomponenten wenig ändern, wenn das Gehirn geschädigt wird. Es handelt sich meist um Lesetests, von denen man weiß, dass sie hoch mit Intelligenztests und niedrig mit Tests zu anderen Merkmalen korrelieren. Der National Adult Reading Test (NART) besteht aus 50 Wörtern, die ungewöhnlich ausgesprochen werden (irreguläre Graphem-Phonem-Übereinstimmung). Die Testpersonen lesen die Wörter vor, und jedes richtig ausgesprochene Wort ergibt einen Punkt. Beim Cambridge Contextual Reading Test (CCRT) sind die Wörter in einen Satz eingebunden (McCaffrey und Vanderslice-Barr 2010). In Deutschland wird gerne der Mehrfachwahl-Wortschatz-Test in der Form B (MWT-B) von Lehrl (2005) zur Schätzung der prämorbiden Intelligenz verwendet. Die Aufgabe besteht darin, bei jedem Item unter Pseudowörtern ein „richtiges“ Wort zu finden (z. B. Sukiff – Fasek – Siuke – Fiskus – Fuske).

Lese- und Wortschatztests zur Schätzung der prämorbiden Intelligenz

Wir stellen 2 Studien vor, die sich mit der Brauchbarkeit von 2 im deutschen Sprachraum verwendbarer Instrumente befasst haben. In der 1. Studie kam der oben genannte MWT-B zum Einsatz (Hessler et al. 2013). An der Studie nahmen Mitglieder eines katholischen Frauenordens teil. Bei insgesamt 142 von 442 Schwestern, deren Daten verwertbar waren, wurden kognitive Beeinträchtigungen unterschiedlichen Grades festgestellt. Insgesamt 27 Schwestern hatten demnach eine moderate und 40 eine schwere kognitive Beeinträchtigung. Der MWT-B war jedoch bei einem Teil (26 % bzw. 75 %) dieser Personen gar nicht durchführbar. Bei den wenigen testbaren Schwestern lag der mit dem MWT-B gemessene IQ durchschnittlich bei 84 bzw. 83. Demnach hätten die Schwestern mit einer moderaten und schweren kognitiven Beeinträchtigung bereits früher eine deutlich verminderte (kristalline) Intelligenz gehabt. Hinzu kommt ein weiteres Problem: Der MWT-B Wert hing deutlich mit dem Bildungsniveau und der beruflichen Ausbildung der Schwestern zusammen. Die Autoren kommen zu dem Schluss, dass der MWT-B bei kognitiver Beeinträchtigung und Demenz nicht zur Schätzung der prämorbiden Intelligenz geeignet ist.

Schätzung anhand des MWT-B

In der 2. Studie (Jahn et al. 2013) wurde eine „Sozialformel“ zur Schätzung der prämorbiden Intelligenz entwickelt und mit gutem Erfolg auch kreuzvalidiert. Dazu machte eine Stichprobe von 612 Gesunden, die bezüglich Alter, Geschlecht und Schulabschluss weitgehend repräsentativ für die Gesamtbevölkerung war, eine Reihe von biografischen Angaben und bearbeitete u. a. einen Wechsler-Intelligenztest (HAWIE-R; ▶ Abschn. 3.2.3.2). Der Gesamt-IQ konnte in einer multiplen Regression mit den biografischen Merkmalen als Prädiktoren mit $R = .75$ (was einer Varianzaufklärung von 55 % entspricht) gut vorhergesagt werden. Starke Prädiktoren waren Internetnutzung, Schultyp, Mathematiknote, Musikspiel, berufliche Stellung und Qualität der gelesenen Bücher. Der Vorteil der Sozialformel gegenüber dem MWT-B liegt darin, dass die notwendigen Daten ggf. mit Unterstützung der Angehörigen auch bei schwer oder nicht testbaren Personen erhoben werden können. Ein direkter Vergleich von MWT-B und Sozialformel zur Schätzung der Intelligenz von Gesunden steht noch aus.

Schätzung anhand der Sozialformel

Anwendung der Sozialformel

Die Anwendung der Sozialformel soll an einem Beispiel von Jahn et al. (2013, S. 15 f.) demonstriert werden. Dazu werden die in Tab. 3 der Publikation aufgeführten Beta-Gewichte verwendet. Ein 46-jähriger Spediteur absolvierte nach der mittleren Reife (Mathematiknoten „meist so um 3 bis 4“) eine Ausbildung zum Kfz-Mechaniker (ohne Meisterbrief) und machte sich nach einem Umzug nach München vor 22 Jahren selbstständig. Als Kind erhielt er 2 Jahre Blockflötenunterricht und 1 Jahr Klavierunterricht (Mindestwert von 5 Jahren damit nicht erreicht). Bis zum Beginn einer schweren Depression vor 2 Jahren nutzte er das Internet, aber nur beruflich. Das politische und wirtschaftliche Geschehen verfolgt er über Fernsehen und Tageszeitung (Abonnement der *Süddeutschen Zeitung*). An Büchern hatte er allenfalls Biografien („über Adenauer und Churchill und so“) gelesen.

Sein Gesamt-IQ wurde mit der Formel für den HAWIE-R auf 112 geschätzt.

$$\text{IQ} = \text{Konstante } 81,78 + 2,174 \text{ (Geschlecht)} + 3,187 \text{ (Wohnort)} + 0 \text{ (Internetnutzung)} + 6,903 \text{ (Schulabschluss)} + 0 \text{ (Mathematiknote)} + 0 \text{ (Musizieren)} + 6,342 \text{ (Berufsstellung)} + 3,196 \text{ (Zeitungslektüre)} + 8,801 \text{ (Buchlektüre)}$$

Der HAWIE-R überschätzt die Intelligenz aufgrund zu alter Normen. Deshalb wurde der IQ anhand einer Tabelle in Erzberger und Engel (2010) in einen Wert für den Wechsler-Intelligenztest für Erwachsene (WIE; ► Abschn. 3.2.3.2) umgerechnet. Die Schätzung des Gesamt-IQ ergibt damit 101.

Häufig Simulation bei Hoffnung auf Entschädigung

Verfälschung Die Befragungen von praktizierenden Neuropsychologinnen und -psychologen und systematische Untersuchungen zeigen, dass häufig mit einer Simulation zu rechnen ist. Slick et al. (2011) schätzen, dass etwa 40 % der Patientinnen und Patienten, die auf eine Entschädigung hoffen, simulieren. Ihren Studienergebnissen zufolge setzen in Nordamerika etwa 50 % der Neuropsychologinnen und -psychologen spezielle Tests ein, mit denen das Vortäuschen einer niedrigen Leistung erkannt werden soll. In vielen Fällen sind mit dem Ergebnis einer neuropsychologischen Begutachtung erhebliche finanzielle Konsequenzen für die Patientin oder den Patienten verbunden. Beispielsweise erstattet die Unfall- oder die Haftpflichtversicherung des Schadenverursachers einen Großteil der Kosten oder leistet Entschädigungszahlungen, oder die Feststellung einer vollen Erwerbsminderung führt zu einer vorzeitigen Berentung. Deshalb ist damit zu rechnen, dass Betroffene eine Beeinträchtigung simulieren bzw. tatsächlich vorhandene Symptome übertreiben.

Definition

Mit dem Begriff der **Simulation** wird das absichtliche Vortäuschen einer Störung oder Beeinträchtigung bezeichnet; unter **Dissimulation** versteht man das absichtliche Verbergen einer Störung oder Beeinträchtigung. Bei einer Überreibung von Symptomen spricht man auch von **Aggravation**.

Simulation, Dissimulation und Aggravation

Die Begriffe suggerieren, dass hier ein Alles-oder-Nichts-Prinzip zugrunde liegt – jemand täuscht oder ist ehrlich. Es ist aber angebracht, von einem Kontinuum auszugehen (Iverson 2010): Patientinnen oder den Patienten, die ihre Symptome unter- oder übertreiben, tun dies mehr oder weniger stark.

Die amerikanische National Academy of Neuropsychology hat 2005 erklärt, dass es unerlässlich sei, die Gültigkeit der Angaben von Patientinnen

und Patienten zu überprüfen, damit man Vertrauen in die Ergebnisse von kognitiven- und Persönlichkeitsmaßen und die sich daraus abgeleiteten Diagnosen und Empfehlungen haben kann.

- » Clinical neuropsychologists are responsible for making determinations about the validity of the information and test data obtained during neuropsychological evaluations. (Bush et al. 2005, S. 419)

Auch die American Academy of Clinical Neuropsychology betont diese Notwendigkeit. In einer Konsensuskonferenz hat sie eine Reihe von Empfehlungen erarbeitet (Heilbronner et al. 2009). Dazu zählen etwa der Rat, grundsätzlich diagnostische Informationen aus mehreren unabhängigen Quellen zu verwenden, Tests zur Entdeckung von Verfälschung und auch Verfahren mit eingebauten Validitätsskalen (etwa das Minnesota Multiphasic Personality Inventory, MMPI-2; ▶ Abschn. 3.3.3.1) einzusetzen und dies auch in den Untersuchungsberichten genau zu dokumentieren. Eine sicherlich angemessene Empfehlung ist dabei, aus der Verfälschung bei einem Instrument auf eine Verfälschung bei anderen Verfahren zu schließen.

Wenn damit zu rechnen ist, dass durch Vortäuschen oder Übertreibung von Symptomen ein Vorteil entsteht, oder wenn der Verdacht besteht, dass sich eine Patientin oder ein Patient bei Tests nicht genügend anstrengt bzw. unrichtige oder unvollständige Angaben macht, kann und muss man – laut der National Academy of Neuropsychology – Symptomvalidierungstests (s. u.) und andere Verfahren einsetzen, um die Gültigkeit der Angaben und Testergebnisse zu beurteilen. Die Empfehlung der American Academy of Clinical Neuropsychology geht noch einen Schritt weiter, indem grundsätzlich zum Einsatz solcher Verfahren geraten wird.

Gültigkeit der Angaben unbedingt überprüfen

Mangelnde Anstrengung bei Leistungstests erkennen

► Beispiel

Ein 9-jähriges Kind und dessen Eltern waren nach einem Autounfall in einen Rechtsstreit verwickelt. Das Kind zeigte in unterschiedlichen Situationen (Schule, Rehabilitationszentrum, diagnostische Untersuchung) diskrepante Leistungen, das Muster der Ergebnisse in kognitiven Leistungstests war untypisch für die Verletzung, und es fiel bei mehreren Symptomvalidierungstests (s. u.) mit kritischen Werten auf. Der Verdacht auf Simulation wurde weiterhin durch zweifelhafte Angaben der Eltern zu den Veränderungen des Kindes seit dem Unfall genährt. Es bestand der Verdacht, dass seine Eltern und die Anwälte das Kind dazu gebracht hatten, Symptome einer Kopfverletzung zu simulieren (Slick et al. 2011, S. 461).◀

Diskrepanzen zwischen Datenquellen

Wie kann man Verfälschung erkennen? Anhaltspunkte können Diskrepanzen zwischen dem Selbstbericht einerseits und Fremdbericht bzw. Dokumenten andererseits sein. Ebenso können Diskrepanzen zwischen berichteten Beschwerden und beobachtbarem Verhalten Hinweise auf eine Verfälschung liefern, genauso wie Beschwerden oder Testergebnisse, die nicht zum neurologischen Status passen (Hartje 2004; Iverson 2010; Sturm 2000). Bei verschiedenen Tests finden Patientinnen und Patienten, die eine kognitive Beeinträchtigung vortäuschen wollen, nicht das richtige Maß und produzieren daher unglaublich niedrige Leistungen.

Kontrollskalen für Verfälschung

Besteht der Verdacht auf Simulation (oder Dissimulation), können zur Überprüfung 2 Arten von Tests eingesetzt werden: Erstens gibt es Tests und Fragebögen, die zusätzlich zu den Informationen über den eigentlichen Messgegenstand auch Kennwerte zur Verfälschung liefern. Ein bekanntes Beispiel ist der MMPI-2 (▶ Abschn. 3.3.3.1), der mehrere Kontrollskalen

hat, die zudem sehr gut erforscht sind (Iverson 2010, S. 102 ff.). Wie mithilfe des Tests d2 Simulation zu erkennen ist, wurde in ▶ Abschn. 3.2.2.2 bereits erläutert.

Symptom validity tests

Daneben stehen einige spezielle Verfahren zur Verfügung (s. auch Littmann 2000). Es handelt sich um Leistungstests, bei denen Simulanten die Testleistungen tatsächlich hirnorganisch gestörter Patientinnen bzw. Patienten nicht richtig einschätzen können und daher auffällig schlechte Leistungen erzielen. Diese Tests werden im Englischen „symptom validity tests“ genannt; einige sehr bekannte Verfahren sind in □ Tab. 9.3 aufgeführt.

Für die Bewertung der Tests zur Entdeckung von Simulation muss man wissen, wie gut sie funktionieren. Blaskewitz et al. (2008) untersuchten in Deutschland, wie gut mit 3 der in □ Tab. 9.3 genannten Tests (MVST, TOMM und FIT) Simulantinnen und Simulanten entdeckt werden können. In dieser Studie wurden Kinder (Durchschnittsalter: 9 Jahre) entweder instruiert, sich bei den Tests anzustrengen oder sich als schlecht darzustellen. Letztere sollten sich vorstellen, dass ein Zauberer kommt, um Kinder für seine Zauberschule zu suchen. Da er selbst nicht sehr gescheit sei, würde er keine Kinder mögen, die ihm überlegen sind. Allerdings sollten die Kinder nicht zu viele Fehler machen, weil der Zauberer auch keine dummen Kinder brauchen könnte. Die „Simulierenden“ zeigten im MVST, TOMM und FIT schlechtere Leistungen als die Kontrollgruppe. Kein Kind in der Kontrollgruppe wurde mit einem der 3 Verfahren als Simulant bzw. Simulant eingestuft. Mit dem MVST gelang

□ Tab. 9.3 Tests zum Erkennen von Simulation (symptom validity tests)

Test	Kurzbeschreibung
Word Memory Test (WMT)	20 Wortpaare (semantisch verwandte Begriffe wie Katze – Hund) werden in 2 Durchgängen nacheinander auf dem Bildschirm dargeboten. Dann folgt sofort ein Wiedererkennungstest mit 40 neuen Wortpaaren, die ein gelerntes Wort mit einem neuen kombinieren (z. B. Hund – Hase). Der Proband bzw. die Probandin soll angeben, wenn ein Wort (hier: Hund) in der Lernliste vorkam. Nach 30 min wird der 2. Wiedererkennungstest mit 40 neuen Wortpaaren (z. B. Hund – Ratte) durchgeführt. Kennwerte: Leistung in Test 1 und in Test 2, Konsistenz der Antworten in Test 1 und 2.
Medical Symptom Validity Test (MSVT) ^a	Vereinfachte Version des WMT mit 10 Wortpaaren. Der 2. Wiedererkennungstest folgt nach 10 min; hier wird jeweils ein Wort vorgegeben, und das fehlende muss ergänzt werden. Zusätzlich erfolgt ein Durchgang mit freier Erinnerung.
Test of Memory Malingering (TOMM)	50 Bilder bekannter Objekte werden nacheinander auf dem Bildschirm gezeigt. Sofort danach folgt ein Wiedererkennungstest mit jeweils 2 Antwortmöglichkeiten: Welches der 2 Bilder wurde zuvor gezeigt? Wiederholung von Lern- und Testdurchgang. Optional wird nach 30 min noch ein Wiedererkennungstest durchgeführt.
Computerized Assessment of Response Bias (CARB)	Jeweils eine 5-stellige Zahl wird auf dem Bildschirm gezeigt, danach folgt ein Wiedererkennungstest mit 2 Antwortmöglichkeiten; scheinbare Variation der Schwierigkeit durch Zeitintervall von bis zu 15 s zwischen Darbietung und Test.
Fifteen Item Test nach Rey (FIT) ^b	Auf einer Karte sind in 5 Zeilen je 3 Items (Buchstaben, Zahlen, Figuren) aufgeführt. Diese sind nach 10 s Darbietung nachzuzeichnen. Die Aufgabe ist trotz der vielen Items tatsächlich sehr leicht (bei den Buchstaben handelt es sich z. B. um A, B, C und a, b, c).

Quelle: Testbeschreibungen nach Slick et al. (2011, S. 465 ff.), mit freundlicher Genehmigung von Springer Publishing Company

^aFür eine Testbesprechung s. Carone (2009)

^bFür den Test sind auch andere Bezeichnungen gebräuchlich: Rey Memory Test, Rey 15-Item Memory Test, Rey's 15-Item Visual Memory Test, Memory for Fifteen Items Test (MFIT)

es, 90 % der „Simulierenden“ zu entdecken, mit dem TOMM wurden 68 % gefunden. Der FIT erwies sich als wenig geeignet; nur 10 % der „Simulierenden“ fielen auf. Die untersuchten Tests zeichnen sich demnach durch eine exzellente Spezifität und eine unterschiedlich gute Sensitivität (► Abschn. 5.1.3.1) aus.

In einer Metaanalyse zur Validität von Tests zur Entdeckung von Simulation befassten sich Sollman und Berry (2011) nur mit Studien, an denen Erwachsene teilnahmen. In einer Experimentalgruppe wurden die Testpersonen jeweils aufgefordert, sich als „gestört“ darzustellen. Für die 5 analysierten Tests, den Victoria Symptom Validity Test (VSVT), TOMM, Letter Memory Test (LMT), MVST und WMT, betrug die mittlere Effektstärke $d=1,55$, was als sehr großer Effekt anzusehen ist. Für die Praxis ist eine Frage wichtig: Wie hoch sind die Sensitivität und die Spezifität der einzelnen Tests? Die Sensitivität variierte zwischen 65,4 (TOMM) und 81,5 (VSVT). Mit den Tests wurden also 65,4 bis 81,5 % der Personen, die simulieren sollten, auch als Simulierende klassifiziert. Die Unterschiede sind jedoch nicht signifikant. Bei der Spezifität fanden sich jedoch signifikante Unterschiede. Sie war bei 4 Tests sehr hoch (VSVT: 95,5, TOMM: 93,8, LMT: 93,0, MSVT: 91,3) und für den WMT mit 69,4 ungewöhnlich niedrig (die Tests sind mit Ausnahme des VSVT und des LMT in □ Tab. 9.3 kurz beschrieben). Demnach besteht bei Anwendung des WMT die Gefahr, dass Menschen, die nicht simulieren, fälschlicherweise der Simulation bezichtigt werden.

Eine sehr praxisnahe Prüfung der Spezifität des MSVT haben Suesse et al. (2015) vorgelegt. Sie setzten den Test bei neurologischen Patientinnen und Patienten ein, die in einer neuropsychologischen Abteilung eines Oxfordner Krankenhauses diagnostisch untersucht wurden. Man kann davon ausgehen, dass kein Grund für eine Simulation vorlag; bei einem Großteil stand die Untersuchung wegen eines bevorstehenden neurologischen Eingriffs (Tiefe Hirnstimulation) an. Von den 124 präoperativen Patientinnen und Patienten wurden 12 (10 %) als Simulantinnen und Simulanten eingestuft. Bei anderen Untersuchungsanlässen betrug das Verhältnis 64 von 281 (23 %). Über beide Gruppen hinweg lag die Spezifität bei 81,2 % (im Vergleich zu 91,3 % in der Metaanalyse von Sollman und Berry 2011; s. o.). Bei einer sorgfältigen Betrachtung der Einzelfälle fanden sich möglicherweise passende Erklärungen für das auffällige Testergebnis, etwa eine schwere kognitive Beeinträchtigung, Schmerzen oder auch tatsächlich Simulation ($n=21$). Nach Ausschluss dieser Fälle blieben aber immer noch 21 Personen (5,2 %), bei denen sehr wahrscheinlich eine falsch positive Diagnose erstellt worden war. Als Fazit lassen sich 2 Erkenntnisse festhalten: Ein kritischer Wert in einem etablierten Symptom-Validitäts-Test ist noch kein Garant dafür, dass tatsächlich eine Simulation vorliegt. Auch bei stärker Nachkontrolle blieben im vorliegenden Fall 5 % (falsch positive) Fehldiagnosen übrig. Und es lohnt sich offenbar, bei einem positiven Testergebnis nach möglichen alternativen Erklärungen zu suchen (s. auch Interview mit Dipl.-Psych. Edmund Grieshaber).

Untersuchungen zum Erkennen von Simulation

Falsch positive Diagnose im Klinikalltag

Interview mit Dipl.-Psych. Edmund Grieshaber zum Thema „psychologische Diagnostik in der Neurologie“



9

Dipl.-Psych. Edmund Grieshaber, seit 1988 Leiter der Abteilung für Neuropsychologie der Neurologischen Klinik GmbH Bad Neustadt/Saale, Zertifizierung zum klinischen Neuropsychologen GNP, Psychologischer Psychotherapeut, Weiterbildungsermächtigung für Klinische Neuropsychologie und Supervisor GNP

(Folgendes Interview wurde mit kleinen redaktionellen Änderungen aus der vorigen Auflage des Buches übernommen).

Aus welchen Gründen kommen Patientinnen und Patienten in die neurologische Klinik, in der Sie tätig sind?

Die meisten Patienten kommen wegen eines apoplektischen Insults (Schlaganfall), eines raumfordern den Prozesses (Tumor) oder nach einem Schädel-Hirn-Trauma.

Welche Bedeutung hat die Diagnostik für Ihre Arbeit mit Patientinnen und Patienten?

Nur durch eine hypothesen geleitete Diagnostik wird es möglich, die Einschränkungen höherer Hirnleistungsfunktionen, die erkrankungs- oder verletzungsbedingt entstanden sind, zu objektivieren und sie qualitativ und quantitativ zu erfassen. Dieser Prozess ist für die Festlegung

der therapeutischen Interventionen von Bedeutung. Nur über die Erfassung der Störungsfelder in ihrer Intensität wird es möglich, die therapeutischen Anforderungen gezielt am momentanen Vermögen der Patientin bzw. des Patienten zu orientieren. Dadurch ist Förderung möglich, und es wird sichergestellt, dass die Patientin bzw. der Patient in den einzelnen Therapiesitzungen auch Erfolge realisieren kann.

Spätestens gegen Ende des Aufenthalts wird erneut eine neuropsychologische Diagnostik durchgeführt. Es sollen Veränderungen (hoffentlich Verbesserungen) erfasst und dokumentiert werden; damit wird festgestellt, ob und wo noch weiter Therapie erforderlich ist. Es muss geprüft werden, wie die berufliche Wiedereingliederung durchgeführt werden soll und was dabei zu beachten ist. Auch die Frage, ob eine Umschulung, innerbetriebliche Umsetzung oder Berentung erforderlich wird, kann Gegenstand der abschließenden Diagnostik sein. Circa 60 % der täglichen Arbeit dienen der Diagnostik (Aktenstudium, Anamnesegespräch, Planung der Untersuchung, Untersuchung, Auswertung, Interpretation, Befunderstellung, Empfehlung).

Machen die modernen bildgebenden Verfahren die psychologische Diagnostik nicht überflüssig? Wozu braucht man in einer apparativ gut ausgestatteten neurologischen Klinik noch psychologische Diagnostik?

Bildgebende Verfahren sind eine Hilfe in der hypothesen geleiteten Diagnostik. Die Kenntnis über Interaktion und Kommunikation der Hirnstrukturen ist aber bei Weitem noch nicht so präzise, um immer exakt beurteilen zu können, ob und in welcher Intensität Einschränkungen entstanden sind. Qualifizierung und Quantifizierung der Störungsfelder und natürlich auch der Nachweis der Wirksamkeit neuropsychologischer Therapie

erfordert gezielte neuropsychologische Diagnostik. Die apparativen medizinischen Verfahren wie CT, MRT u. Ä. leisten diese Aufgabe nicht. Teilweise sind auch neuropsychologische Einschränkungen zu objektivieren, obwohl bildgebende Verfahren keine Auffälligkeiten zeigen. In der Forschung wird das PET eingesetzt, um das Gehirn in Aktion sowie jene Hirnareale abzubilden, die bei spezifischen kognitiven Aktivitäten besonders angesprochen sind. Auch dieses bildgebende Verfahren wird aus Kostengründen für die neuropsychologische Praxis momentan noch selten eingesetzt.

Welche diagnostischen Verfahren setzen Sie besonders häufig ein und zu welchem Zweck?

Die Testbatterie zur Aufmerksamkeitsprüfung ist Standard, obwohl zu Recht auch viel Kritik an diesem Verfahren geübt wird. Im Bereich der Diagnostik von Gedächtnisstörungen kommen vorwiegend der Visuelle und Verbale Merkfähigkeitstest (VVM) und der Verbale Lern- und Merkfähigkeitstest (VLMT) zum Einsatz. Da wir auch viele ältere Patientinnen und Patienten haben, setzen wir auch oft das Nürnberger-Alters-Inventar (NAI) ein. Es hat den großen Vorteil, dass die Normen bis in den Altersbereich über 80 Jahre reichen. Wir verwenden es zur Messung von Gedächtnis- und konzentrativen Leistungen.

Haben Sie einen Wunsch an die Testentwicklerinnen bzw. Testentwickler und Testverlage, was die Verbesserung und Neuentwicklung von diagnostischen Verfahren angeht?

Ein erheblicher Teil unserer Patientinnen und Patienten sind ältere und

alte Menschen. Daher sind Normierungen an dieser Gruppe erforderlich; die Normen der jüngeren sind nicht einfach linear interpolierbar. Die Entwicklung im Gesundheitswesen erfordert auch bei uns, möglichst ökonomisch zu Resultaten zu gelangen. Der Untersuchung in Kleingruppen wird zunehmend mehr Bedeutung zukommen. Ein in der Gruppe einsetzbarer Test zur Prüfung mnestischer (Gedächtnis-)Leistungen wäre z. B. sehr hilfreich. Auch die Schriftgröße, insbesondere bei Konzentrationstests, sollte so gestaltet sein, dass die Zeichen problemlos erkannt werden.

Wenn Sie eine Psychologen- bzw. Psychologinnenstelle in Ihrem Team zu besetzen haben, welche Qualifikationsmerkmale sind aus Ihrer Sicht wichtig?

Aus der Wahl der Studienschwerpunkte sollte Interesse an diagnostischen Fragestellungen deutlich werden. Einschlägige, mehrmonatige Praktika in einer Einrichtung, die sich mit der Diagnostik und Therapie neuropsychologischer Fragen beschäftigt, sollten das Interesse dokumentieren. Auch Bereitschaft zu psychotherapeutischer Arbeit und zur Hilfestellung bei der Krankheitsverarbeitung und Entwicklung tragfähiger Zukunftsperspektiven für Betroffene und Familienangehörigen sind wichtig. Dem Mitfühlen, ohne durch eine Überidentifikation handlungsunfähig zu werden, kommt eine wichtige Bedeutung zu, ebenso eigenständigem und strukturiertem Arbeiten im interdisziplinären Team. Das Arbeitsfeld muss deutlich sein: schwerkranke Menschen, denen dieses Leiden häufig sehr anzusehen ist.

Weiterführende Literatur

Über die neuropsychologische Diagnostik sowie über Störungen, mit denen sich die psychologische Diagnostik zu befassen hat, informieren mehrere Beiträge in dem von Sturm et al. (2009) herausgegebenen Lehrbuch. Zu Fragen der neuropsychologischen Begutachtung sei auf die „Leitlinie der Gesellschaft für Neuropsychologie“ (Neumann-Zielke et al. 2015) verwiesen. Über neuropsychologische Tests zu bestimmten Störungsgruppen finden sich in den 3 Bänden des *Handbuchs neuropsychologischer Testverfahren* (Schellig et al. 2009; Schellig et al. 2018, 2019) umfangreiche Informationen. Fallbeispiele präsentieren Kubinger und Ortner (2010). Speziell über die Forschung zu Symptom-Validitäts-Tests informiert eine Literaturübersicht von Sweet und Breiting (2013).

9.2 Rechtspsychologische Diagnostik

Aufgaben und Fragestellungen

9

Klassische diagnostische Aufgaben in der Rechtspsychologie sind die Erstellung von Gerichtsgutachten zu Fragen der Schuldfähigkeit von Straftätern und Straftäterinnen, der Glaubhaftigkeit von Zeuginnen- bzw. Zeugenaussagen oder etwa des Sorgerechts in Scheidungsverfahren sowie die Untersuchung von Straftätern bzw. Straftäterinnen, die bereits überführt sind und sich nun in einer Haftanstalt befinden. Damit soll eine fundierte Auswahl von Behandlungsmaßnahmen gewährleistet und ggf. gegen Ende des Strafvollzugs eine Rückfallprognose erstellt werden.

Diagnostische Fragestellungen in der Rechtspsychologie

- In Strafverfahren:
 - Glaubhaftigkeit einer Zeuginnen- bzw. Zeugenaussage
 - Schuldfähigkeit eines Täters/einer Täterin
 - Strafrechtliche Verantwortlichkeit bei jugendlichen Tätern/Täterinnen
- Im Strafvollzug:
 - Erstellen eines Vollzugsplans
 - Vollzugslockerung
 - Kriminalprognose (vorzeitige Entlassung, Sicherheitsverwahrung)
- In Zivilverfahren:
 - Entzug der Geschäftsfähigkeit
 - Umgangs- und Sorgerecht für ein Kind nach Scheidung der Eltern
 - Entzug der elterlichen Sorge
- In Sozialgerichtsverfahren:
 - Arbeits- und Erwerbsfähigkeit
 - Voraussetzungen für eine Umschulungsmaßnahme
 - Berufsunfähigkeit

Glaubhaftigkeit der Aussagen von Zeuginnen und Zeugen

Die Fragestellungen sind so heterogen, dass ein einheitliches Vorgehen völlig abwegig wäre. Deshalb gehen wir exemplarisch auf 4 Bereiche ein, die in der Praxis eine große Bedeutung haben: auf die Beurteilung der Glaubhaftigkeit von Zeuginnen- bzw. Zeugenaussagen, auf die Schuldfähigkeit von Straftätern und Straftäterinnen, auf die Kriminalprognose und auf Sorgerechtsentscheidungen. Zur Beantwortung der oben genannten Fragestellungen in Sozialgerichtsverfahren kann auf eignungsdiagnostische Ansätze und Methoden zurückgegriffen werden (► Abschn. 5.2); wenn eine Hirnschädigung vorliegt, kommen auch neuropsychologische Methoden zum Einsatz (► Abschn. 9.1).

Aussageglaubhaftigkeit vs. Aussagetüchtigkeit

In Strafverfahren kommt der Aussage von Zeuginnen und Zeugen oft eine große Bedeutung zu. Wenn es stimmt, was sie berichten, wird die beschuldigte Person vielleicht verurteilt – oder auch nicht, wenn es sich um eine entlastende Aussage handelt. Manchmal sind jedoch Zweifel an den Aussagen von Zeuginnen und Zeugen angebracht. In diesem Fall kann das

Gericht eine Begutachtung veranlassen. Grundsätzlich ist zwischen der Glaubhaftigkeit der Aussage und der „Aussagetüchtigkeit“ der Zeugin bzw. des Zeugen zu unterscheiden (Volbert und Dahle 2010). Es geht also darum, was jemand gesagt und wer diese Aussage gemacht hat.

Auch wenn sich eine Zeugin oder ein Zeuge um eine korrekte Aussage bemüht, können Fehler passieren. Beobachtungen (und Aussagen darüber) decken sich oft nicht mit den Fakten, die beobachtet werden. Vielen Menschen unterlaufen teilweise gravierende Beobachtungsfehler, wie in empirischen Untersuchungen mit gestellten oder im Film gezeigten Ereignissen eindrucksvoll demonstriert wurde (z. B. Loftus 1979). Inzwischen konnten zahlreiche Faktoren, die sich auf die Identifikation durch Augenzeuginnen bzw. -zeugen auswirken, durch experimentelle Untersuchungen identifiziert werden (für eine Übersicht s. Wells und Olsen 2003).

Insgesamt belegen diese Untersuchungen, wie trügerisch es sein kann, sich auf menschliche Beobachtungen zu verlassen. Bei der Begutachtung von Zeugenaussagen werden aus diesen Gründen die situativen Bedingungen, unter denen die Aussage zustande gekommen ist, analysiert. Dabei spielen die Wahrnehmungsbedingungen (z. B. Beobachtungsdauer, mögliche Ablenkung der Aufmerksamkeit, sensorische Wahrnehmungsbedingungen), die Komplexität des Ereignisses und der zeitliche Abstand zwischen Ereignis und Bericht ebenso eine Rolle wie die Bedingungen, unter denen die Aussage aufgenommen wurde. Besonders bei Kindern kann sich eine (ungewollt) suggestive Befragung auf die Aussagen auswirken (Volbert 2000). Günstig sind Aufforderungen, zu einem Thema zu erzählen („Beschreibe doch einmal, wie die Person aussah“) und offene Fragen („Was hat die Person zu dir gesagt?“).

Menschen sind extrem anfällig für Fragen, in denen eine falsche Information unterstellt wird. Beispielsweise sehen Versuchspersonen, wie eine Person einem Mädchen die Geldbörse stiehlt. Danach werden sie gefragt, ob sie gesehen haben, wie das Mädchen dabei am Arm verletzt wurde. Tatsächlich wurde das Mädchen am Nacken verletzt. Dennoch bestätigen viele, dass sie gesehen haben, dass das Mädchen am Arm verletzt wurde. Der Missinformation-Effekt ist intensiv erforscht und vielfach bestätigt worden (Loftus 2005). Wir wissen z. B., dass jüngere Kinder und ältere Erwachsene anfälliger dafür sind. Falsche Informationen können nicht nur durch Suggestivfragen unterschoben werden, sondern können auch durch Gespräche zwischen Zeuginnen bzw. Zeugen durch öffentliche Berichterstattung ins Gedächtnis gelangen (Loftus 2005).

Beobachtungen decken sich oft nicht mit den Fakten

Situative Bedingungen der Aussage beachten

Beispiele für suggestive Befragung

- „Hat der Mann gesagt, du sollst mitkommen?“ (der Sachverhalt, dass der Mann das gesagt hat, wird unterstellt)
- „Könnte es sein, dass das Messer schon am Tatort lag?“ (Aufforderung zu einer Spekulation, implizite Erwartung)
- Verstärkung von Antworten (Nicken, „ah ja“, „gut beobachtet“), die ins Konzept des Interviewers/der Interviewerin passen (die Zeugin/der Zeuge erfährt, welche Antworten erwünscht sind)
- Wiederholung einer Frage im Verlauf des Interviews (erzeugt Druck, nun endlich die „richtige“ Antwort zu geben)

Die vorliegende Aussage wird danach analysiert, ob sie wahr ist. Im Grunde gibt es 3 Möglichkeiten, die auch als Hypothesen formuliert werden können:

- Die Aussage ist wahr.
- Es handelt sich um eine absichtliche Falschaussage (Lüge).
- Die Aussage ist subjektiv wahr, objektiv jedoch unwahr (Pseudoerinnerungen aufgrund von Suggestion).

Hypothesengeleitetes Vorgehen

Motive für eine Falschaussage

Diese Hypothesen werden geprüft; die Leitfrage lautet dabei: Kann „die Aussage anders als durch einen tatsächlichen Erlebnishintergrund zustande gekommen sein“? (Volbert und Dahle 2010, S. 31).

Für eine absichtliche *Falschaussage* muss die Zeugin oder der Zeuge eine entsprechende Motivation haben. Mögliche Motive sind u. a. Rache (der vermeintliche Täter soll bestraft werden), Angst vor negativen Konsequenzen (die Zeugin oder der Zeuge wird bedroht), Mitleid (das Kind will nicht, dass sein Vater ins Gefängnis kommt), Vorteile durch eine Verurteilung erlangen (das Sorgerecht für das Kind erhalten) oder auch Verdecken von eigenem Fehlverhalten (die Zeugin oder der Zeuge hätte die Tat verhindern können). Zur Beurteilung der Motivation für eine mögliche Falschaussage wird die Beziehung zwischen der beschuldigten Person und der Zeugin oder dem Zeugen analysiert. Auch wenn bei einer Zeugin bzw. einem Zeugen ein Motiv für eine Falschaussage erkennbar wird, bedeutet dies nicht zwangsläufig, dass eine Falschaussage gemacht wurde. Nicht jeder Mensch, der ein Motiv hat, handelt gemäß diesem Motiv. Auch könnte die Aussage durch Suggestion zustande gekommen sein. Deshalb müssen die Umstände und die Art der Befragung eruiert werden.

Die Aussage selbst kann inhaltsanalytisch untersucht werden. Die Gutachterin oder der Gutachter sucht dabei nach sog. „Realkennzeichen“ (vgl. Steller und Volbert 1997; Volbert und Steller 2014). Die Realkennzeichen stellen Kriterien dar, anhand derer erfundene Aussagen erkannt werden sollen.

9

„Kriteriumsbasierte Inhaltsanalyse“ – „reality monitoring“

Begriffsklärung

Bei der Suche nach Realkennzeichen handelt es sich um eine Inhaltsanalyse von Aussagen. Anhand bestimmter Kriterien sollen erlebnisbasierte von erfundenen Aussagen unterscheidbar sein. Der Ansatz ist deshalb auch als „kriteriumsbasierte Inhaltsanalyse“ bekannt. Eine eng verwandte inhaltsanalytische Methode ist unter dem Begriff „reality monitoring“ (Realitätsüberwachung) eingeführt worden (s. Masip et al. 2005). Berichte, die von selbst wahrgenommenen und erlebten Ereignissen herrühren, sollten u. a. mehr sensorische Informationen (Farben, Gerüche etc.), affektive Informationen und Kontextinformationen enthalten als erfundene Berichte. Dieser Ansatz hat also einen anderen theoretischen Hintergrund und war zumindest anfangs nicht zur Aufdeckung von Täuschung und strategischer Selbstdarstellung gedacht. Inhaltlich überlappen sich die Kriterien teilweise mit den Realkennzeichen (s. Volbert und Steller 2014).

Aussagen inhaltsanalytisch untersuchen

Das Vorliegen vieler Realkennzeichen in einer Aussage spricht dafür, dass die Zeugin bzw. der Zeuge das Ereignis selbst erlebt hat und keine „Erfindungen“ präsentiert. Dazu muss eine verschriftlichte Aussage sehr sorgfältig auf das Vorliegen von Realkennzeichen durchgesehen werden.

Realkennzeichen

Beispiele für Realkennzeichen

Das Vorliegen dieser Merkmale spricht für eine glaubwürdige Aussage:

- Logische Konsistenz
- Schilderungen von Komplikationen im Handlungsverlauf
- Schilderung ausgefallener Einzelheiten
- Schilderung eigener psychischer Vorgänge
- Eingeständnis von Erinnerungslücken

Wie viele Hinweise man bei der Analyse einer Aussage findet, hängt von verschiedenen Faktoren ab. Grundsätzlich gilt: Je älter Kinder sind, desto mehr Aussagen treffen sie, und zwar sowohl erfundene als auch erlebnisbasierte. Faktoren, die sich möglicherweise auf die Anzahl gefundener Realkennzeichen auswirken, sind außerdem folgende (Volbert und Steller 2014):

- Bereitschaft von Zeuginnen und Zeugen, eine Aussage zu machen
- Zeitlicher Abstand zum Ereignis
- Komplexität des Ereignisses
- Interviewtechnik

Ergebnis von vielen Faktoren abhängig

Ein anderes Problem besteht darin, dass es keine verbindlichen Standards (Normen) dafür gibt, wie viele Realkennzeichen vorliegen müssen, damit eine Aussage als sehr wahrscheinlich authentisch eingestuft werden kann. Es bleibt somit der Gutachterin oder dem Gutachter überlassen, die Zahl der vorgefundenen Realkennzeichen richtig zu bewerten.

Bewertung von Realkennzeichen durch Gutachter/-innen

Wie bei jedem diagnostischen Instrument ist zu fragen, wie valide es ist. Eine Validierungsstrategie besteht darin, die Anzahl der Realkennzeichen von erlebnisbasierten und erfundenen Berichten zu vergleichen. Dazu liegen viele Laboruntersuchungen vor, in denen man sicher sein kann, wie die Berichte zustande gekommen sind. Feldstudien haben den Nachteil, dass grundsätzlich schwer festzustellen ist, ob eine Aussage tatsächlich erlebnisbasiert oder erfunden ist. Zudem werden Zeuginnen und Zeugen, die absichtlich falsche Angaben machen, nicht durchgängig lügen. Vielmehr liegt es nahe, dass sie sich auf Details beschränken, die aus ihrer Sicht entscheidend und zudem nicht aus anderen Quellen verifizierbar sind.

Validierung

Für eine Metaanalyse haben Oberlader et al. (2016) insgesamt 56 Studien ausgewertet, in denen eine inhaltsanalytische Technik (nach dem Ansatz Realkennzeichen oder reality monitoring – s. o.) zur Anwendung kam. Über alle Studien hinweg ermittelten sie eine Effektstärke von Hedges $g = 1,03$. Setzt man die Sensitivität und Spezifität rechnerisch als gleich hoch an, entspricht dies einer Trefferquote von 70 % bei einer Rate von 30 % falschen Alarmen (falsch positive Urteile). Rechnet man kann die Effektstärke in eine Korrelation um, erhält man einen Validitätskoeffizienten von $r = .46$ (laut Normwertrechner; s. ► <https://www.psychometrica.de/effektstaerke.html>). Eine Kriteriumsvalidität in dieser Höhe ist beachtlich. Dennoch muss man die damit immer noch verbundene Rate von Fehlklassifikationen in der Praxis beachten. Man könnte den kritischen Wert so absenken, dass z. B. 95 % Treffer zu erwarten sind. Der Preis dafür wären aber 73 % falsche Alarme (Oberlader et al. 2016, Tab. 3).

Validität metaanalytisch bestätigt

Zwischen den beiden oben beschriebenen inhaltsanalytischen Ansätzen fand sich ein kleiner Unterschied zugunsten von reality monitoring (Hedges $g = 1,26$ vs. $0,97$). Was könnte man tun, um die Validität der Methode generell zu erhöhen? Die Moderatoranalysen liefern dazu Hinweise. Es ist offensichtlich besser, mit Lang- und nicht mit Kurzversionen der Kriterien zu arbeiten und nicht einfach Summenwerte zu berechnen, sondern die Kriterien optimal zu gewichten (jedenfalls führen Diskrimiananzanalysen zu einer höheren Trefferquote).

9.2.1 Glaubhaftigkeit von Zeuginnen und Zeugen

Nun richtet sich die Aufmerksamkeit auf die Fähigkeit von Zeuginnen und Zeugen, grundsätzlich zu dem Sachverhalt eine gültige Aussage zu machen. Eine niedrige Intelligenz, fehlender Erfahrungshintergrund, bestimmte psychische Störungen oder Alkohol- bzw. Drogenkonsum können Zweifel an dem Urteilsvermögen begründen.

Fähigkeitsmerkmale der Zeuginnen und Zeugen

► Beispiel

Antje F. (20 Jahre) sagt in einem Strafverfahren gegen Herrn S. aus. Der Zuhälter habe sie in der Wohnung eingeschlossen, ihr Rauschgift verkauft und sie zur Prostitution gezwungen. Nach 10 bis 11 Monaten wurde sie „seelisch und körperlich heruntergekommen“ vom Zuhälter in die Klinik gebracht. Ihre Glaubwürdigkeit als Zeugin wird dadurch erschüttert, dass ihr Erinnerungsvermögen möglicherweise durch mehrjährigen Heroingebrauch beeinträchtigt ist. Es liegt langjähriger Alkohol- und Drogenkonsum vor; bereits mit etwa 17 Jahren hat sie auch Heroin (ca. 3 g pro Tag) konsumiert. Drogengebrauch über längere Zeit kann die Persönlichkeit, die Konzentrationsfähigkeit, die Gedächtnisleistung etc. negativ verändern. Die Untersuchung ergibt, dass ihre Aussagen als glaubhaft erscheinen, da Kriterien wie Detailfülle, folgerichtige Handlungsverkettungen, teils widersprüchliche, aber geklärte Aussagen und Selbstbezeichnung erfüllt sind. In Tests zur Messung der Intelligenz, der Konzentrationsfähigkeit und der Merkfähigkeit erreicht sie durchschnittliche bis leicht überdurchschnittliche Werte. Der Gutachter kommt zu dem Schluss, dass keine Verschlechterung der intellektuellen Leistungsfähigkeit (hierzu stellt er einen Vergleich mit den früheren Schulleistungen an) erkennbar ist, keine Hinweise auf zerebral-pathologische Abbauprozesse vorliegen und die Erinnerungsfähigkeit nicht eingeschränkt ist. Er bejaht die Glaubwürdigkeit der Zeugin und die Glaubhaftigkeit ihrer Aussagen. Das Gericht hatte keine Zweifel an ihren Aussagen. Angesichts dieser Situation ließ sich der Angeklagte auf die wesentlichen Punkte der Anklageschrift ein (Reichert 1997).◀

Informationsquellen für Begutachtung

Für die Begutachtung von Zeuginnen und Zeugen als Person finden Akteninformationen (Gerichtsakten) Verwendung, weitere Informationen werden im diagnostischen Interview, eventuell auch mit dritten Personen (z. B. Eltern, Erzieher) gewonnen. Wenn sich die Frage nach einer psychischen Störung stellt, bieten sich dazu strukturierte klinische Interviews sowie bestimmte klinische Fragebögen an. Je nach Sachlage kommen auch Testverfahren (Intelligenztests, Konzentrationstests, Gedächtnistests etc.) zum Einsatz.

9.2.2 Schuldunfähigkeit und verminderte Schuldfähigkeit

§ 20 und § 21 StGB

In § 20 des deutschen Strafgesetzbuchs (StGB) wird festgelegt:

§ 20 StGB Schuldunfähigkeit wegen seelischer Störungen

Ohne Schuld handelt, wer bei Begehung der Tat wegen einer krankhaften seelischen Störung, wegen einer tiefgreifenden Bewusstseinsstörung oder wegen Schwachsins oder einer schweren anderen seelischen Abartigkeit unfähig ist, das Unrecht der Tat einzusehen oder nach dieser Einsicht zu handeln.

Eine verminderte Schuldfähigkeit liegt nach § 21 StGB vor:

§ 21 StGB Verminderte Schuldfähigkeit

Ist die Fähigkeit des Täters, das Unrecht der Tat einzusehen oder nach dieser Einsicht zu handeln, aus einem der in § 20 bezeichneten Gründen bei Begehung der Tat erheblich vermindert, so kann die Strafe [...] gemildert werden.

Einschränkung des Unrechtsbewusstseins bzw. der Steuerungsfähigkeit

Die Begutachtung der Schuldfähigkeit bzw. einer verminderten Schuldfähigkeit setzt an den in § 20 StGB genannten möglichen Bedingungen an. Es genügt jedoch nicht, dass eines dieser Kriterien vorliegt. Vielmehr muss daraus eine Aufhebung oder Einschränkung des Unrechtsbewusstseins oder der Steuerungsfähigkeit bei der Tat herrühren.

Die juristischen Begriffe können nicht eins zu eins in psychologische oder auch psychiatrische Kategorien übersetzt werden. Am einfachsten ist aus psychologischer Sicht mit dem Begriff des „Schwachsinn“ umzugehen, der als intellektuelle Minderbegabung bzw. geistige Behinderung aufzufassen ist. Zur Feststellung sind Intelligenztests geeignet. Eine wichtige Informationsquelle stellen auch biografische Daten wie Art und Dauer des Schulbesuchs, die Beschäftigung oder die Unterbringung in einem Heim für geistig behinderte Menschen dar. Bezugspersonen können im Interview nach Kompetenzen der Täterin bzw. des Täters zur Bewältigung alltäglicher Probleme befragt werden.

Die Feststellung einer krankhaften seelischen Störung oder einer schweren anderen seelischen Abartigkeit fällt in den Bereich der klinischen Diagnostik (► Kap. 8), denn hier geht es um psychiatrische Diagnosen. Zu den krankhaften seelischen Störungen zählen „organische, psychische und Verhaltensstörungen durch psychotrope Substanzen, Schizophrenien, wahnhaftes und psychotische Störungen, verschiedene affektive Störungen und zu den genannten Diagnosegruppen analoge Symptomatiken sowie Anfalls-erkrankungen“ (Scholz und Schmidt 2008, S. 403). Die anderen schweren seelischen Abartigkeiten sind überwiegend im Bereich gravierender Persönlichkeitsstörungen und bei den abnormen Gewohnheiten und Störungen der Impulskontrolle zu suchen. Aus der letztgenannten Kategorie sind insbesondere die pathologische Brandstiftung (Pyromanie, F63.1 nach Internationaler statistischer Klassifikation der Krankheiten, 10. Revision – ICD-10) und das pathologische Stehlen (Kleptomanie, F63.2) in der Praxis relevant.

Eine tiefgreifende Bewusstseinsstörung muss nicht krankhaft sein; auch ein psychisch gesunder Mensch kann sich bei der Tat in einer hochgradigen affektiven Erregung befunden haben (weitere Informationen dazu bei Scholz und Schmidt 2008). Bei der Begutachtung kommen daher der Analyse des Tathergangs und der Vorgeschichte, die zur Tat geführt hat, eine besondere Bedeutung zu. Weitere Faktoren sind etwa die Persönlichkeit der Täterin oder des Täters, deren Beziehung zum Opfer und Erinnerungsstörungen nach der Tat. Die nötigen Informationen erhält die Gutachterin oder der Gutachter durch Auswertung der gerichtlichen Akten, durch ein Interview mit der zu begutachtenden Person, aber auch anhand von Persönlichkeitsfragebögen oder projektiven Verfahren.

Geistige Behinderung

Krankhafte seelische Störungen

Tiefgreifende Bewusstseinsstörung
auch bei psychisch Gesunden

9.2.3 Kriminalprognose

Die Prognose des zukünftigen Verhaltens einer Straftäterin oder eines Straftäters hat eine erhebliche Bedeutung für die Auswahl und Bemessung der Strafe bzw. anderer Maßnahmen (z. B. Sicherungsverwahrung), die Ausgestaltung des Strafvollzugs und für dessen Beendigung (Dahle 2000, 2009). Eine Rückfallprognose wird benötigt, wenn eine straffällig gewordene Person einen Großteil ihrer Strafe verbüßt hat und nun eine Entscheidung über eine vorzeitige Haftentlassung auf Bewährung ansteht. In bestimmten Fällen wird geprüft, ob die Notwendigkeit einer Sicherungsverwahrung besteht.

Rückfallprognose

- Prognosen über künftiges delinquentes Verhalten sind schwer zu stellen, da gleich mehrere ungünstige Bedingungen zusammentreffen (vgl. Dahle 2000):
- Das vorherzusagende Verhalten tritt (zumindest bei zahlreichen Delikten) selten auf.
 - Viele Täterinnen und Täter bleiben Jahre oder gar Jahrzehnte lang unauffällig, um dann plötzlich wieder einschlägige Straftaten zu begehen.

Schwierige Randbedingungen für
Prognose

- Verhalten wird immer auch durch die Situation determiniert. In welche Situationen eine Straftäterin oder ein Straftäter einmal kommen wird, die sie oder ihn wieder in Versuchung bringen, ist ungewiss. Auch ungünstige Lebensumstände wie Arbeitslosigkeit oder das Zerbrechen einer Beziehung oder Ehe sind schwer vorherzusehen. Günstige Randbedingungen, die anfangs vorhanden sind und bei der Begutachtung berücksichtigt wurden (z. B. Alkoholabstinenz), können später wegfallen.
- Der Geltungszeitraum der Prognose ist gerade bei jungen Straftäterinnen und Straftätern sehr lang.

Nur Wahrscheinlichkeitsaussagen

Deshalb sind grundsätzlich nur Wahrscheinlichkeitsaussagen möglich („es ist zu erwarten, dass...“). Die Folgen einer Fehleinschätzung des Rückfallrisikos sind schwerwiegend, und zwar einerseits für die Opfer, andererseits für Straftäterinnen und Straftätern, wenn ihnen unbegründet die Freiheit vorenthalten wird.

Rückfallrisiko mit beobachteten „Gesetzmäßigkeiten“ abschätzen

Nomothetische Prognose Bei der Begutachtung können statistische Erkenntnisse über Rückfallrisiken in vergleichbaren Fällen genutzt werden. In diesem Fall spricht man von einer nomothetischen Prognose (Volbert und Dahle 2010, S. 71 f.). Die statistische Vorhersage muss sich auf Merkmale der Tat und der Delinquentin bzw. des Delinquents konzentrieren, die allgemein verfügbar sind. Durch gründliche Forschung kann manchmal aufgedeckt werden, dass andere, hinter dem leicht erfassbaren Merkmal stehende Faktoren für die Kriminalprognose entscheidend sind. So verliert der Faktor Hautfarbe bei der Prognose von künftigen Gewalttaten seine Vorhersagekraft, wenn die Kriminalität in der Nachbarschaft, in die sich eine früher straffällig gewordene Person nach ihrer Entlassung begibt, berücksichtigt wird (Monahan 2003). Die statistische Prognose vernachlässigt zudem zwangsläufig die seltenen, nicht bei allen Personen vorhandenen Risikofaktoren.

■ Tab. 9.4 zeigt ein Beispiel für eine Prognosetafel. Zu deren Erstellung hatte Gretenkord (2002) 188 Straftäter, die im Durchschnitt 8 Jahre lang in einer Klinik für gerichtliche Psychiatrie untergebracht waren, nach Variablen klassifiziert, die der internationalen Forschung zufolge (46 Studien) zur Vorhersage des Rückfallrisikos gut geeignet sind: Vorliegen einer Persönlichkeitsstörung (meist Psychopathie), früheres Gewaltdelikt, Gewalttätigkeit während der Unterbringung und Alter. Er überprüfte, ob die Patienten in einem Zeitraum von durchschnittlich 8 Jahren nach ihrer Entlassung einen Eintrag im Bundeszentralregister erhalten hatten, also wieder straffällig geworden waren.

■ Tab. 9.4 lässt sich entnehmen, dass das Rückfallrisiko deutlich mit dem Alter abnimmt. Prognostisch ungünstig sind eine Persönlichkeitsstörung, Vorstrafen wegen eines Gewaltdelikts und Gewalttätigkeit während des Maßregelvollzugs. Das höchste Risiko (Rückfallwahrscheinlichkeit von 65 %) haben Personen, die jung sind (Altersgruppe 20 Jahre) und 3 weitere Risikomerkmale aufweisen. Am unwahrscheinlichsten ist ein Rückfall bei älteren Personen (Altersgruppe: 60 Jahre), die weder eine Persönlichkeitsstörung noch eine Vorstrafe wegen eines Gewaltdelikts haben oder trotz einschlägiger Vorstrafe während ihres Klinikaufenthalts nicht gewalttätig geworden sind.

Beispiel für Prognosetafel

Rückfallrisiko aus Tabelle ablesen

Tab. 9.4 Beispiel für eine Prognosetafel zur Rückfallwahrscheinlichkeit (in Prozent) von männlichen Straftätern nach Entlassung aus dem Maßregelvollzug. (Nach Gretenkord 2002, mit freundlicher Genehmigung des LIT-Verlags)

Risikofaktor			Entlassungsalter		
Persönlichkeitsstörung	Vorstrafe mit Gewaltdelikt	Mindestens 2× gewalttätig	20 Jahre	40 Jahre	60 Jahre
Nein	Nein	Nein	6	2	1
		Ja	17	6	2
	Ja	Nein	15	6	2
		Ja	37	16	6
Ja	Nein	Nein	16	6	2
		Ja	39	18	7
	Ja	Nein	36	16	6
		Ja	65	38	17

oder eine Persönlichkeitsstörung haben, aber weder aufgrund ihrer Vorstrafen noch durch ihr Verhalten während des Maßregelvollzugs als gewalttätig gelten. Die Rückfallwahrscheinlichkeit liegt hier nur bei 1 oder 2 %.

Eine Alternative zu Prognosetafeln ist die Vorhersage eines Rückfalls mittels Regressionsgleichung. Aus Großbritannien stammt die Offender Group Reoffending Scale – Version 3 (s. Volbert und Dahle 2010, S. 77 f.), in die damals die Daten von etwa 79.000 ehemaligen Strafgefangenen eingingen. Die Gleichung basiert auf 6 Prädiktoren: Geschlecht, Alter bei der letzten einschlägigen Tat, aktuelles Alter, Anzahl früherer Verurteilungen, Alter bei der ersten Verurteilung und Art des Anlassdelikts (dazu liegt eine gesonderte Tabelle vor). In Deutschland konnten damit rechtskräftige Neuverurteilungen mit einer Korrelation von $r = .32$ bis $.37$ vorhergesagt werden (Volbert und Dahle 2010).

Regressionsgleichungen für Rückfallrisiko

Zur *Vorhersage bestimmter Delikte* (z. B. Gewaltdelikte, Sexualdelikte) sowie für bestimmte Delinquentengruppen liegen mehrere Instrumente vor. Die Verfahren fallen in die folgenden beiden Kategorien:

1. Selbstbeurteilungsinstrumente (insbesondere Fragebögen, beispielsweise zur Psychopathie)
2. Verfahren, in denen Expertinnen und Experten vorhandene oder auch eigens erhobene Informationen eintragen

Instrumente zur Rückfallprognose

Ein prominentes Beispiel für die 2. Kategorie ist das Historical-Clinical-Risk Management-20 (HCR-20), das weiter unten beschrieben wird. In einer Metaanalyse versuchte Walters (2006) zu klären, ob und wie sich die beiden Verfahrensgruppen in Bezug auf ihre Kriteriumsvalidität unterscheiden. Das Ergebnis war, dass die Verfahren zur Fremdbeurteilung von Risikomerkmalen (wie das HCR-20) eine erneute Delinquenz besser vorhersagen als Selbstbeurteilungsverfahren ($r = .31$ vs. $.20$ bzw. speziell bei Gewalttäten $r = .24$ vs. $.17$). Die Schlussfolgerung, deshalb auf Selbstbeurteilungsinstrumente zu verzichten, wäre falsch. Walters (2006) konnte nämlich zeigen, dass die Verfahren eine inkrementelle Validität haben und sich daher gut

ergänzen. Wir stellen mit dem HCR-20 ein sehr bekanntes Verfahren vor, dessen Ergebnisse auch in deutschen Studien als valide bestätigt werden konnten (Dahle et al. 2007).

Beim HCR-20 (Webster et al. 1997) handelt es sich um eine Checkliste, zu der auch eine deutsche Fassung vorliegt (Müller-Isberner et al. 1998). Sie soll zukünftiges gewalttägliches Verhalten vorhersagen. Eine Beurteilerin oder ein Beurteiler stuft 20 Risikofaktoren (= Items) auf einer 3-stufigen Skala (0 = „trifft definitiv nicht zu“; 1 = „trifft möglicherweise/teilweise zu“; 2 = „trifft sicher zu“) ein. Dabei soll auf alle denkbaren Informationsquellen wie Berichte von Behörden, Ämtern, Polizei und Staatsanwaltschaft oder auf Stellungnahmen von Psychologinnen bzw. Psychologen, Psychiaterinnen bzw. Psychiatern, Sozialarbeiterinnen bzw. -arbeitern und Krankenpflegepersonal zurückgegriffen werden; auch ein Interview ist möglich. Die Beurteilungen zu jedem Risiko werden zu bereichsspezifischen Skalenwerten sowie zu einem Gesamtwert addiert, der bei 20 Items und maximal 3 Punkten pro Item einen Wert von 60 erreichen kann. Seit 2013 liegt mit dem HCR-20 V3 die 3. Version des Verfahrens vor (Douglas et al. 2014). Eine deutschsprachige Ausgabe wurde von Müller-Isberner et al. (2014) herausgegeben. Über die Änderungen gegenüber der Vorgängerversion berichten Douglas et al. (2014), dass das HCR-20 V3 weiter 10 historische und je 5 klinische und Items zum Risikomanagement enthält und nur wenige Items ausgetauscht wurden. Allerdings kommen Subitems vor, z. B. gibt es zu H1 Gewaltanwendung die Subitems „als Kind“, „als Jugendliche(r)“, „als Erwachsene(r)“ (Kötter et al. 2014).

Aufbau des HCR-20 (nach Müller-Isberner et al. 1998)

— (H) Historische Items (Risikofaktoren aus der Vorgeschichte):

- H1 Frühere Gewaltanwendung
- H2 Geringes Alter bei erster Gewalttat
- H2a Geringes Alter bei Erstdelinquenz
- H3 Instabile Beziehungen
- H4 Probleme im Arbeitsbereich
- H5 Substanzmissbrauch
- H6 (Gravierende) seelische Störung (Erhebung mit DSM-5 oder ICD-10)
- H7 Psychopathie (Erhebung mit Hare Psychopathy Checklist-Revised [PCL])
- H8 Frühe Fehlanpassung
- H8a Inadäquater Erziehungsstil
- H8b Fehlverhalten in Kindheit und Jugend
- H9 Persönlichkeitsstörung (Erhebung mit DSM-5 oder ICD-10)
- H10 Frühere Verstöße gegen Auflagen

— (C) Klinische Items (gegenwärtige Risikofaktoren):

- C1 Mangel an Einsicht
- C2 Negative Einstellungen
- C3 Aktive Symptome (Erhebung mit DSM-5 oder ICD-10)
- C4 Impulsivität (z. B. Erhebung mit einer Impulsivitäts-Checkliste)
- C5 Fehlender Behandlungserfolg

— **(R) Risiko-Management (Vorhersage des zukünftigen Verhaltens unter den zu erwartenden äußeren Umständen):**

- R1 Fehlen realisierbarer Pläne
- R2 Destabilisierende Einflüsse
- R3 Mangel an Unterstützung
- R4 Fehlende Compliance
- R5 Stressoren

Anhand eines Itembeispiels kann das Messprinzip veranschaulicht werden:
Bei H1 (frühere Gewaltanwendung) gibt es folgende Antwortmöglichkeiten:

Itembeispiel aus dem HCR-20

- Keine frühere Gewalttätigkeit
- Mögliche oder weniger gravierende frühere gewalttätige Handlungen (1 oder 2 mäßig gewalttätige Handlungen)
- Fortgesetzte oder schwerwiegende frühere Gewaltanwendung (3 oder mehr Handlungen, die als mäßig gewalttätig zu bezeichnen sind) oder jede Art von schwerer oder erheblicher früherer Gewalttätigkeit

In der Schweiz haben Rossegger et al. (2010) das HCR-20-Werte von allen Gewalt- und Sexualstraftätern, die zwischen 1994 und 1999 in einer großen Strafvollzugseinrichtung für erwachsene Männer entlassen wurden, mit ihrem Status nach 7 Jahren in Beziehung gesetzt. Anhand von Strafregisterauszügen konnte festgestellt werden, ob ein Entlassener für das gleiche oder ein anderes Delikt wieder verurteilt worden war. Die allgemeine Rückfallquote lag bei 60 %, die für das gleiche Delikt bei 9 %. Für den Gesamtwert konnte eine hohe Prognosegüte (AUC-Werte; Area Under Curve) von .76 bzw. .79 (spezifisches Delikt) ermittelt werden. Aus Großbritannien liegt eine ähnliche Studie vor (Gray et al. 2008), die sich allerdings mit (ebenfalls männlichen) Straftätern befasst, die aus psychiatrisch-forensischen Einrichtungen entlassen worden waren. Die größte Gruppe stellten mit 56 % Straftäter, bei denen eine Schizophrenie oder eine andere Psychose diagnostiziert worden war. Nach 5 Jahren waren 10,6 % erneut wegen einer Gewalttat verurteilt worden. Für alle Straftaten lag die Basisrate nach 5 Jahren bei 34,2 %. Die AUC-Werte lagen bei .70 bzw. .69 (alle Straftaten).

Studien zur Anwendung des HCR-20

Weltweit wurden viele ähnliche Studien zur Rückfallprognose durchgeführt. Die zur Vorhersage eingesetzten Verfahren variieren, ebenso die Tätergruppen, die bei einem Rückfall begangenen Straftaten und der Beobachtungszeitraum nach der Entlassung. Das Kriterium ist jedoch immer das Gleiche, nämlich ob ein aus der Haft entlassener Mensch innerhalb eines bestimmten Zeitraums wieder straffällig wird oder nicht. Diese Information kann (je nach Staat) in Datenbanken der Justiz oder der Polizei abgerufen werden. Es handelt sich also um ein objektives Kriterium, das jedoch 2 Probleme aufweist: Jemand kann zu Unrecht verurteilt worden sein (ein geringes Risiko), und jemand kann eine Straftat begangen haben, ohne überführt worden zu sein. So lag in Deutschland die Aufklärungsquote über alle Straftaten 2019 bei 57,5 % (Statista GmbH 2020). Begeht jemand eine Straftat, ohne überführt zu werden, zählt sie oder er in diesen Studien automatisch zur Gruppe der Nichtrückfälligen. Dadurch wird die prognostische Validität der Verfahren sehr wahrscheinlich erheblich unterschätzt.

Vorgehen zur Überprüfung der Validität von Prognoseinstrumenten

In einer Metaanalyse werteten Hanson und Morton-Bourgon (2009) Studien zur Rückfallquote explizit bei Sexualdelikten aus 16 Staaten aus. Die Gesamtzahl der entlassenen Strafgefangenen, deren Kriminalgeschichte im Durchschnitt knapp 6 Jahre lang über Datenbanken verfolgt wurde, betrug über 45.000. Die Rückfallquote lag bei 11,5 % (Sexualstraftat), 19,5 %

Metaanalyse zur Rückfallprognose bei Sexualstraftätern

Gute Validität deliktspezifischer Instrumente

(Sexual- oder Gewaltstraftat) bzw. 33,2 % (irgendeine Straftat). Die Ergebnisse sind in □ Tab. 9.5 aufgeführt. Folgende Befunde sind besonders erwähnenswert:

- Der empirisch aktuarische Ansatz (Erläuterung in □ Tab. 9.5) ist den anderen Ansätzen überlegen, führt also zu den besten Vorhersagen.
- Die Verfahren sind besonders gut für den Bereich geeignet, für den sie konstruiert wurden. Verfahren zur Vorhersage erneuter Sexualdelikte sagen Sexualdelikte besser voraus als Gewaltdelikte oder andere Straftaten.
- Die Effektstärken und damit die Validitäten sind relativ groß. Die Effektstärke von Verfahren zur Vorhersage von Sexualdelikten sagen Sexualdelikte mit $d=0,67$ vorher (entspricht $r=.32$); für Gewaltdelikte beträgt die entsprechende Effektstärke .78 ($r=.36$) und für Delikte aller Art .97 ($r=.44$). Diese Werte sind nicht durch irgendwelche Korrekturen nach oben gerechnet worden. Es handelt sich zudem um sehr konservative Schätzungen, weil die Gruppe der „Nichtrückfälligen“ auch Fälle einschließt, in denen ein Rückfall unerkannt geblieben ist (s. o.).

Informationen aus Akten sowie eigener Untersuchung

Ideografische Prognose Fehlen geeignete Instrumente, die auf einer statistischen Vorhersage basieren, oder entscheidet sich die Diagnostikerin oder der Diagnostiker aus grundsätzlichen Überlegungen gegen „starre“ statistische Erklärungsansätze, kann mithilfe des ideografischen Ansatzes (Volbert und Dahle 2010, S. 72 f.) eine Prognose erstellt werden. Ziel dieses Ansatzes ist es, ein individuelles Erklärungsmodell für die Probandin oder den Probanden auszuarbeiten, um damit zu einer Prognose zu gelangen. Gutachterinnen und Gutachter, die diesem Ansatz folgen, werden etwa folgende Fragen stellen (s. auch Dahle 2007):

□ **Tab. 9.5** Vorhersage von erneuten Straftaten ehemaliger Sexualstraftäter nach ihrer Haftentlassung (Effektstärke d)

Art der Risikoschätzung	Kriterium		
	Sexualdelikt	Gewaltdelikt	Irgendeine Straftat
<i>Verfahren zur Vorhersage von Sexualdelikten</i>			
Empirisch aktuarisch ^a	0,67	0,51	0,52
Mechanisch ^b	0,66	0,40	0,37
Strukturiertes Expertenurteil ^c	0,46	0,31	0,26
<i>Verfahren zur Vorhersage von Gewaltdelikten</i>			
Empirisch aktuarisch	0,39	0,78	0,74
Mechanisch	0,33	0,31	–
<i>Verfahren zur Vorhersage von Delikten generell</i>			
Empirisch aktuarisch	0,62	0,79	0,97
Unstrukturiertes Expertenurteil ^d	0,42	0,22	0,11

Quelle: Nach Hanson und Morton-Bourgon (2009, Tab. 1)

^aEmpirisch aktuarische Vorhersage: Die Items werden nach einer vorher festgelegten Regel verrechnet, die empirisch ermittelt wurde (auch die Items wurden nach empirischen Befunden ausgewählt).

^bMechanische Vorhersage: Die Items werden nach einer vorher festgelegten Regel verrechnet, die theoretisch oder anhand der Literatur begründet wird.

^cStrukturiertes Expertenurteil: Dabei stehen strukturierte Listen mit Risiken des Delinquents zur Verfügung, die Kombination dieser Daten wird frei gewählt.

^dUnstrukturiertes Expertenurteil: Weder die Risikofaktoren (Items) noch die Verrechnung werden vorher festgelegt.

Diagnostik in weiteren Anwendungsfeldern

- Unter welchen Bedingungen wurde die Straftat begangen?
- Wie kann die Entstehung der damaligen Straftat erklärt werden?
- Wie haben sich Verhaltensmuster der Delinquentin oder des Delinquenten in der Haft verändert?
- Welche therapeutischen Maßnahmen wurden mit welchem Erfolg durchgeführt?
- Wie gestaltet sich der „soziale Empfangsraum“ nach der möglichen Entlassung (Arbeitsplatz, Unterkunft, soziale Beziehungen)?
- Welche Lebensperspektiven (berufliche Möglichkeiten, Partnerschaft, Familie etc.) hat die Delinquentin oder der Delinquent?
- Wie hoch ist die Wahrscheinlichkeit, dass kritische Umstände auftreten, unter denen bei dieser Person die Gefahr einer erneuten Straftat groß ist?

Die dazu benötigten Informationen finden sich in den Akten über die früheren Straftaten, in früher erstellten Gutachten, in der Dokumentation des Haftverlaufs, in Interviews mit der Täterin oder dem Täter und eventuell auch mit wichtigen Bezugspersonen (z. B. Partnerin bzw. Partner). Zur Beurteilung der Persönlichkeit können auch Persönlichkeitsfragebögen oder projektive Verfahren herangezogen werden. Je nach Fragestellung kann auch ein Intelligenztest oder ein anderer Leistungstest eingesetzt werden.

Ist der ideografische Ansatz dem nomothetischen überlegen? Die Forschung zum Vergleich von statistischer und klinischer Urteilsbildung hat ergeben, dass die statistische Urteilsbildung und damit nomothetische Prognosen generell besser sind (► Abschn. 5.1.3). Speziell für die Vorhersage von erneuten Straftaten von entlassenen Sexualstraftätern erlaubt die oben beschriebene Metaanalyse (Hanson und Morton-Bourgon 2009) eine Einschätzung. Für die Gruppe der Verfahren, die zur Vorhersage von nicht näher spezifizierten Delikten entwickelt wurden, kann ein direkter Vergleich zwischen der nomothetischen und der ideografischen Methode angestellt werden. Die empirisch aktuarische (nur auf empirisch gesicherten Zusammenhängen gründende) Vorhersage war dem unstrukturierten Expertenurteil weit überlegen. Die entsprechenden Effektstärken betrugen $d=0,97$ vs. $0,11$ bei der Vorhersage von Straftaten aller Art.

Nomothetischer Ansatz dem ideografischen überlegen

Liegen Prognosetafeln oder andere empirisch begründete Modelle vor, kann die Gutachterin oder der Gutachter überlegen, ob es gute Gründe gibt, die statistische Vorhersage zu korrigieren. Dazu befasst sie oder er sich mit den Besonderheiten der delinquenten Person und bezieht sie in das Urteil ein. Beispielsweise versichert die delinquente Person, sich während der Haft grundlegend geändert zu haben. Es wäre zu erwarten, dass diese Kombination von nomothetischer und ideografischer Prognose zu besseren Vorhersagen führt, als wenn man starr der Formel vertraut. In 3 Studien wurde dies geprüft (Hanson und Morton-Bourgon 2009). Die rein statistische Vorhersage hatte eine Effektstärke von $d=0,87$, die von Expertinnen und Experten korrigierten statistischen Vorhersagen waren in den 5 Einzelvergleichen immer schlechter; die mittlere Effektstärke betrug nur $d=0,64$. Damit muss das intuitiv „vernünftige“ Konzept, nicht blind auf statistische Vorhersagen zu vertrauen und stattdessen die Weisheit der Formel mit der Weisheit der menschlichen Expertise zu kombinieren, zumindest bei der Kriminalprognose von Sexualstraftätern infrage gestellt werden.

Kombination von nomothetischem und ideografischem Ansatz

Allerdings ist es aus ethischen und juristischen Gründen nicht zu vertreten, Menschen im Rahmen einer Begutachtung nur nach einer Formel zu beurteilen. Die Arbeit der Gutachterin oder des Gutachters würde sich dann darauf reduzieren, die Formel mit Daten zu füttern und die entsprechenden Daten zu erheben. Die einzelnen Prädiktoren stellen zumindest zum Großteil keine Ursachen für ein erhöhtes Rückfallrisiko dar, sondern

es sind nur statistische Indikatoren. Wie oben (s. nomothetische Prognose) ausgeführt wurde, können sich hinter einem Prädiktor wie Hautfarbe ganz andere Ursachen verbergen. Außerdem gehen in die statistischen Prognosemodelle nur Merkmale ein, für die ausreichend Forschungsdaten vorliegen. Dementsprechend wird in den „Mindestanforderungen für Prognosegutachten“ (Boetticher et al. 2007) vorgeschlagen, sowohl ideografische („biografisch fundierte Analyse unter Berücksichtigung der individuellen Risikofaktoren“) als auch nomothetische („Abgleich mit dem empirischen Wissen über das Rückfallrisiko möglichst vergleichbarer Tätergruppen“) Ansätze bei der Begutachtung zu verfolgen.

9.2.4 Familiengericht: Sorgerechtsentscheidungen

Begutachtungsanlässe

Im Rahmen von familiengerichtlichen Verfahren können psychologische Gutachten zu sehr unterschiedlichen Fragestellungen angefordert werden (Salzgeber 2015). Fragestellungen, die z. B. selten vorkommen, sind

- die Beurteilung der Ehemündigkeit von heiratswilligen Minderjährigen,
- die Frage, ob eine Minderjährige die Folgen und die Tragweite eines Schwangerschaftsabbruchs einschätzen kann, oder
- die Frage, ob die Aufrechterhaltung einer Ehe eine besondere Härte darstellt.

Hauptanlass Scheidungsverfahren

Viele Begutachtungen ergeben sich durch Scheidungsverfahren, von denen minderjährige Kinder betroffen sind. Laut Statistischem Bundesamt (2018) wurden im Jahr 2017 in Deutschland insgesamt 153.500 Ehen geschieden; betroffen davon waren insgesamt 124.000 minderjährige Kinder. Allerdings kommt es bei Scheidungsverfahren zum Glück nicht immer zu einem Streit wegen des Sorgerechts.

Bei einvernehmlicher Regelung der elterlichen Sorge kein Handlungsbedarf des Gerichts

Elterliche Sorge Die elterliche Sorge gilt von der Geburt bis zur Volljährigkeit des Kindes und umfasst die Personen- und die Vermögensfürsorge (s. Salzgeber 2015 – auch für die weiteren Ausführungen zum Sorgerecht). Zur Personenfürsorge gehören die Fürsorge für das körperliche Wohl des Kindes, die Erziehung, Aufenthaltsbestimmung, Aufsichtspflicht und die Umgangsbestimmung. Die Vermögensfürsorge betrifft die Vertretung des Kindes in finanziellen Angelegenheiten. Die Ausübung der elterlichen Fürsorge ist nicht nur ein Recht, sondern auch eine Pflicht. Kein Elternteil kann darauf verzichten. Es ist allerdings möglich, die Fürsorge dem anderen Elternteil oder einer dritten Person zu überlassen; dies ist jederzeit widerufbar. In einem Scheidungs- oder Trennungsverfahren müssen die Eltern angeben, ob gemeinsame minderjährige Kinder betroffen sind. Legen sie eine einvernehmliche Regelung zur elterlichen Sorge und zum Umgang mit den Kindern vor, besteht seitens des Gerichts normalerweise kein Handlungsbedarf. Der gemeinsame Elternvorschlag steht sogar über dem Kindeswohl, sofern dieses nicht erkennbar beeinträchtigt ist.

Bei Uneinigkeit sind Diagnostik und Intervention eng verzahnt

Streit um das Sorgerecht Kommt es dagegen zu einem Streit der Eltern um das Sorgerecht für die Kinder, führt das Familiengericht eine Entscheidung herbei, wobei das Gericht auf eine einvernehmliche Regelung der Betroffenen hinwirken sollte. Bevor es jedoch zu einer gerichtlichen Entscheidung kommt, müssen Schlichtungs- und Vermittlungsversuche unternommen werden. Wird eine psychologische Sachverständige oder ein psychologischer Sachverständiger vom Gericht hinzugezogen, gilt dieser Grundsatz auch für sie. Diagnostik und Intervention (Hinwirken auf eine Einigung, Vermittlung, Beratungsangebot etc.) sind in diesem Fall eng verzahnt.

Kindeswohl und Kindeswille Bei der Suche nach einer Lösung der Sorgerechtsfrage haben sowohl Gerichte wie auch hinzugezogene Sachverständige aufgrund gesetzlicher Vorgaben das Kindeswohl und bei über 14-jährigen Kindern auch den Willen des Kindes zu beachten. Das Kindeswohl umfasst das leibliche und das geistige/seelische Wohl des Kindes. Der Begriff ist juristisch nicht definiert; zur Beurteilung des Kindeswohls sind vor allem sozialwissenschaftliche Erkenntnisse anzuwenden. Das Kindeswohl hat in einem Sorgerechtsverfahren eine zentrale Bedeutung. Kommt das Gericht, etwa aufgrund eines psychologischen Gutachtens, zu der Erkenntnis, dass das Kindeswohl gefährdet ist, kann es weitreichende Maßnahmen beschließen. So kann es die Wohnung ausschließlich einem der Elternteile zuweisen, einem Elternteil oder einem Dritten den Zutritt zum Haus bzw. der Wohnung verbieten oder vorschreiben, das Stadtgebiet nicht mehr zu betreten. Es kann sogar einem Elternteil oder auch beiden Eltern das Sorgerecht entziehen. Bei der Sorgerechtsentscheidung ist ferner der Kindeswille zu berücksichtigen. Ein über 14-jähriges Kind darf selbst einen Vorschlag zum Sorgerecht machen. Wenn das Kind einem gemeinsamen Elternvorschlag zur Regelung des Sorgerechts explizit nicht zustimmt, trifft das Gericht eine Entscheidung, die sich am Kindeswohl orientiert.

Beurteilung des Kindeswohls wichtig

Sachverständige In familiengerichtlichen Verfahren, die ein Kind betreffen, hat die Familienrichterin oder der Familienrichter eine Ermittlungspflicht. Es steht in ihrem Ermessen, ein Sachverständigengutachten einzuholen. In der Vergangenheit hatten die Richterinnen und Richter große Freiheiten bei der Auswahl der Expertinnen und Experten.

Im Einzelfall hat dies schon einmal dazu geführt, dass eine Altenpflgerin und Heilpraktikerin, die nie studiert hat, vom Gericht um ein Sorgerechtsgutachten gebeten wurde. In den Medien gab es Berichte über skandalöse Gutachten. Der Koalitionsvertrag der Fraktionen von CDU/CSU und SPD für die 18. Legislaturperiode sah vor, dass die Qualität dieser Gutachten in Zusammenarbeit mit den Berufsverbänden verbessert werden soll. 2016 wurde ein Gesetzentwurf vorgelegt. Im „Gesetz über das Verfahren in Familiensachen und in den Angelegenheiten der freiwilligen Gerichtsbarkeit (FamFG)“ wurde § 163 Sachverständigengutachten wie folgt geändert:

Auswahlkriterien für Sachverständige

§ 163 FamFG Sachverständigengutachten

(1) In Verfahren nach § 151 Nr. 1 bis 3 ist das Gutachten durch einen geeigneten Sachverständigen zu erstatten, der mindestens über eine psychologische, psychotherapeutische, kinder- und jugendpsychiatrische, psychiatrische, ärztliche, pädagogische oder sozialpädagogische Berufsqualifikation verfügen soll. Verfügt der Sachverständige über eine pädagogische oder sozialpädagogische Berufsqualifikation, ist der Erwerb ausreichender diagnostischer und analytischer Kenntnisse durch eine anerkannte Zusatzqualifikation nachzuweisen.

Eine „Arbeitsgruppe Familienrechtliche Gutachten“ (2019), an der verschiedene Fachgesellschaften, Verbände und Kammern beteiligt waren, hat Standards für Sachverständigengutachten im Kinderschaftsrecht vorgelegt, an der sich Gutachterinnen und Gutachter nun orientieren können. Sie helfen auch Familiengerichten und Prozessbeteiligten, die Qualität solcher Gutachten zu beurteilen.

Meist gemeinsame elterliche Sorge

Ausgang von Sorgerechtsverfahren Die gemeinsame elterliche Sorge stellt den Regelfall dar; in über 90 % der Scheidungen mit minderjährigen Kindern einigen sich die Eltern darauf. Dass dieser Entscheidung Vermittlungsbemühungen vorausgegangen sein können, wurde oben erwähnt. In strittigen

Fällen kann der Antrag eines Elternteils jedoch auch anders lauten. In der nachfolgenden Übersicht sind einige Entscheidungsmöglichkeiten in Sorgerechtsverfahren aufgeführt, die oft nicht die Zustimmung eines Elternteils oder die eines über 14-jährigen Kindes finden und die zur Hinzuziehung eines Sachverständigen führen können.

Mögliche Entscheidungen nach Trennung oder Scheidung der Eltern

- Alleinige elterliche Sorge (einem Elternteil wird die Sorge übertragen)
- Aufteilung der Sorge (ein Elternteil ist z. B. zuständig für die schulische Erziehung etc.)
- Aufhebung der gemeinsamen Sorge (die gemeinsame Sorge wird in eine alleinige umgewandelt)
- Entzug der elterlichen Sorge oder Teilen der elterlichen Sorge (evtl. auch Übertragung auf eine dritte Person)
- Rückführung eines Kindes nach dem Haager Übereinkommen (das Kind lebt bei einem Elternteil in einem anderen Staat)
- Regelung des Umgangs des Kindes mit seinen Eltern (das Kind lebt bei einem Elternteil, hat aber auch Umgang mit dem anderen Elternteil oder weiteren Personen)

9

Umgangsregelung eventuell kompliziert

Gerade der letzte Punkt, die Umgangsregelung, kann sich im Detail als schwierig und strittig erweisen. Ziel der Regelung des Umgangs des Kindes mit seinen Eltern ist es, eine harmonische Eltern-Kind-Beziehung mit beiden Elternteilen auch nach deren Trennung zu ermöglichen. Wenn sich ein Elternteil als problematisch erweist, kann das Gericht beispielsweise einen beaufsichtigten Umgang anordnen oder den Kontakt zum Kind für eine bestimmte Zeit untersagen. Auch der Umgang mit weiteren Personen (z. B. neue Partnerin bzw. neuer Partner eines Elternteils, Großeltern, Stiefeltern) kann Gegenstand einer Umgangsregelung sein.

Psychologische Fragen aus Vorinformationen und Rechtsprechung herleiten

Fragestellungen Bei den in der folgenden Übersicht aufgeführten Fragestellungen handelt es sich um juristische Fragen, aus denen zunächst psychologische Fragen abgeleitet werden. Familienrichterinnen und -richter werden in der Regel nur dann eine psychologischen Sachverständige oder einen psychologischen Sachverständigen hinzuziehen, wenn sie Fragestellungen sehen, die sie nicht selbst beantworten können. Für die Ableitung der psychologischen Fragen sind nicht nur Vorinformationen über den individuellen Fall erforderlich, sondern auch Kenntnisse der einschlägigen Gesetze und der Rechtsprechung. Welche konkreten psychologischen Fragen gestellt werden, ergibt sich oft erst nach einem Aktenstudium oder einem ersten Gespräch mit den Eltern. Eine Richterin bzw. ein Richter kann aber auch die Fragestellung von Anfang an auf eine oder mehrere Teilfragen einengen.

Beispiele für psychologische Fragen in Sorgerechtsentscheidungen:

- Ist die Erziehungsfähigkeit durch eine Erkrankung eingeschränkt?
- Liegt sexueller Missbrauch vor?
- Ist die Bereitschaft vorhanden, elterliche Verantwortung zu übernehmen?
- Sind die Betreuungs- und Versorgungsmöglichkeiten ausreichend?
- Wie stark ist die Bindung des Kindes an einen Elternteil?
- Was ist der Kindeswille?
- Wie groß ist die Förderkompetenz des Elternteils?
- Liegt ein Mangel an erzieherischer Kompetenz vor?
- Wendet ein Elternteil unzulässige Erziehungsmaßnahmen an?

Diagnostische Verfahren Auf die psychologischen Fragen sucht die Diagnostikerin oder der Diagnostiker mithilfe von Aktenanalysen, diagnostischen Interviews, Verhaltensbeobachtungen, Persönlichkeitsfragebögen, Leistungstests oder auch projektiven Verfahren eine Antwort. Die Auswahl der Verfahren richtet sich stark nach der spezifischen Fragestellung. Beispielsweise kann sich die Frage stellen, ob die Erziehungsfähigkeit durch eine vorliegende hirnorganische Erkrankung eingeschränkt ist. In diesem Fall liegt es nahe, mithilfe von neuropsychologischen Tests (► Abschn. 9.1) zu versuchen, die Schwere der Funktionsbeeinträchtigungen abzuschätzen.

In einem anderen Sorgerechtsverfahren ist vielleicht der Verdacht aufgekommen, dass ein Elternteil das alleinige Sorgerecht anstrebt, um Unterhaltsforderungen stellen zu können oder um die Partnerin oder den Partner dafür zu „bestrafen“, dass sie oder er die Ehe zerstört hat. Hier ist die Bereitschaft zu hinterfragen, elterliche Verantwortung zu übernehmen. Durch ein diagnostisches Interview kann die Gutachterin oder der Gutachter eruieren, welche konkreten Zukunftspläne bezüglich Kindesbetreuung, Freizeitgestaltung und Umgang mit anstehenden Problemen ein Elternteil hat und wie dieser in der Vergangenheit seine Elternrolle ausgefüllt hat. Zur Beurteilung der Bindung des Kindes an einen Elternteil bietet sich bei Kleinkindern die Verhaltensbeobachtung der Eltern-Kind-Interaktion bei einem Hausbesuch an. Ferner können die Eltern befragt werden. Bei älteren Kindern kommt auch ein diagnostisches Interview mit dem Kind infrage. Es ist auch denkbar, dass jemand versucht, mit einem projektiven Tests wie dem Familien-Beziehungs-Test (FBT; Howells und Lickorish 2017) Einblick in die Beziehung der Familienmitglieder zu bekommen. Jedenfalls besteht Nachfrage nach dem FBT; das englische Original stammt aus dem Jahr 1967, die erste deutsche Auflage erschien 1972, und inzwischen liegt die 8. Auflage vor. In einer Analyse von Gerichtsgutachten (s. u.) zeigte sich, dass meistens auch Tests eingesetzt wurden. Dabei stellen projektive Tests mit einem Anteil von 55 % die größte Kategorie von Tests.

Wie es um die Qualität der Gutachten bestellt ist, zeigt eine Auswertung aller 116 familienrechtspsychologischer Gutachten, die 2010 und 2011 für 4 Amtsgerichte des Oberlandesgerichtsbezirks Hamm erstellt worden waren (Salewski und Stürmer 2015). Die Gutachten hatten einen Umfang von 10 bis 137 Seiten (im Durchschnitt 56 Seiten) und befassten sich mehrheitlich (84,5 %) mit der Erst- oder Neuregelung der elterlichen Sorge bzw. damit verbundenen Rechtsfragen. Die Gutachten erweisen sich in vielen Fällen als mangelhaft:

- In 56 % der Gutachten wurden aus der gerichtlichen Fragestellung keine psychologischen Fragen hergeleitet.
- Bei den 51 Gutachten, in denen psychologische Fragen formuliert wurden, sollte die Auswahl der Verfahren zur Beantwortung der Fragen begründet werden. Das war aber bei den meisten Gutachten (65 %) nicht der Fall.
- Allen Gutachten lagen diagnostische Interviews zugrunde. Aber überwiegend (bei 90 %) wurden keine Angaben dazu gemacht, warum bestimmte Themen angesprochen wurden oder welche Fragen damit geklärt werden sollten.
- Bei 94 Gutachten wurden Verhaltensbeobachtungen durchgeführt. Aber in 97 % der Fälle fehlten Angaben zur Systematik der Beobachtung und der Registrierung von Verhaltensmerkmalen.
- Bei 78 % der Gutachten wurden die Ergebnisse nicht methodenkritisch diskutiert.

Ist die Erziehungsfähigkeit eingeschränkt?

Bindung des Kindes an einen Elternteil

Qualitätsprobleme bei familienpsychologischen Gutachten

Das ernüchternde Fazit der Studie lautet:

- » Wir haben die Studie im Juli 2014 bei einer Expertenanhörung im Bundesministerium für Justiz und Verbraucherschutz vorgestellt. Bei der anschließenden Diskussion mit Vertretern überregionaler psychologischer, psychiatrischer und juristischer Berufsverbände herrschte ebenfalls Einigkeit dahingehend, dass im Bereich der familienrechtspychologischen Begutachtung ein Missstand vorliegt. (Salewski und Stürmer 2015, S. 8).

Weiterführende Literatur

Zur Diagnostik in Strafverfahren bietet das Buch von Volbert und Dahle (2010) in kompakter Form einen guten Überblick. Darüber hinaus finden sich in dem von Volbert und Steller (2008) herausgegebenen Buch jeweils mehrere Beiträge zu den Themen Aussagebeurteilung, Schuld-, Reife und Gefährlichkeitsbeurteilung sowie zur familienpsychologischen Begutachtung. Eindrückliche Fallbeispiele werden bei Kubinger und Ortner (2010) dargestellt.

Zur Begutachtung bei familiengerichtlichen Fragen und den dabei relevanten juristischen Randbedingungen informiert Salzgeber (2020) in einem umfangreichen Werk, dass inzwischen in der 7. Auflage vorliegt.

9.3 Verkehrspychologische Diagnostik

9

Menschliches Fehlverhalten
häufigste Ursache für schwere
Verkehrsunfälle

Der „Idiotentest“ ist emotional
besetzt

In Deutschland gab es 2018 rund 2,6 Mio. Unfälle im Straßenverkehr – mit 3275 Toten, 67.967 Schwer- und 328.051 Leichtverletzten (Statistisches Bundesamt 2019). Für Unfälle mit Personenschäden (Tote oder Verletzte) wurde 2018 in den meisten (240.622) Fällen ein Fehlverhalten der Fahrzeugfahrerinnen oder Fahrzeugführer als Ursache ermittelt. Um die Anzahl der Unfälle, insbesondere der mit Toten und Verletzten, zu reduzieren, gibt es im Wesentlichen 3 Ansatzpunkte: Man kann versuchen, die Straßen, die Kraftfahrzeuge und die Menschen, die ein Fahrzeug führen, sicherer zu machen. Dabei können Psychologinnen und Psychologen mitwirken.

In Deutschland gibt es etwa 600 Psychologinnen und Psychologen, die den Titel „Fachpsychologe Verkehrspychologie“ tragen; die Zahl der Psychologinnen und Psychologen, die in der Verkehrspychologie tätig sind, dürfte aber deutlich größer sein (Vollrath und Krems 2011). Sie können durch Forschung, Schulungen und auch durch Diagnostik zur Verkehrssicherheit beitragen. Der Anteil der Diagnostik an der Berufstätigkeit liegt einer Untersuchung zufolge bei 44 % (Tab. 1.11.1). In bestimmten Fällen müssen Menschen, die im Straßenverkehr auffällig geworden sind, müssen zu einer Medizinisch-Psychologischen Untersuchung (kurz: MPU)– im Volksmund gibt es dafür das herablassende Wort „Idiotentest“. Dass die Betroffenen überwiegend ablehnend reagieren, ist zumindest nachvollziehbar. Schwer zu verstehen ist hingegen, dass sich verkehrsunauffällige Kraftfahrerinnen und Kraftfahrer mit denen solidarisieren, die stark alkoholisiert am Steuer gesessen haben oder etwa durch aggressives Fahrverhalten Leben und Gesundheit ihrer Mitmenschen gefährdet haben.

Im Folgenden befassen wir uns mit der psychologischen Diagnostik im Rahmen der Verkehrspychologie, genauer mit der Begutachtung von Menschen, die am Straßenverkehr teilnehmen, zwecks Feststellung oder Überprüfung ihrer Fahreignung. Am Rande sei angemerkt, dass auch bei der Auswahl von Pilotinnen und Piloten oder Lokfahrerinnen und Lokführern psychologische Eignungstest durchgeführt werden.

9.3.1 Begutachtung der Fahreignung für den Straßenverkehr

Fahrerlaubnis-Verordnung Die rechtliche Grundlage für eine Begutachtung der Fahreignung für den Straßenverkehr stellt die „Verordnung über die Zulassung von Personen zum Straßenverkehr“ (kurz: Fahrerlaubnis-Verordnung, FeV) vom 13. Dezember 2010 (BGBl. I S. 1980) dar, die immer wieder aktualisiert wird; wir beziehen uns auf die Fassung vom 24. Mai 2018 (► <https://www.verkehrsportal.de/fev/fev.php>).

Untersuchungsanlässe Für eine medizinisch-psychologische Begutachtung kommen verschiedene Anlässe infrage. Eine Statistik der Bundesanstalt für Straßenwesen gibt Aufschluss über die Art und die Häufigkeit der einzelnen Untersuchungsanlässe sowie über das Ergebnis der Begutachtung (Tab. 9.6).

Fast in der Hälfte (47 %) der Fälle sind Eignungszweifel aufgrund einer Alkoholproblematik der Anlass zur Untersuchung, wobei ein erheblicher Anteil der zu begutachtenden Klientinnen und Klienten (insgesamt 30 %) zum ersten Mal auffällig geworden sind. Die zweitgrößte Anlassgruppe sind Probleme mit Drogen oder Medikamenten, die zusammen 24 % der Begutachtungen ausmachen. Neben einem positiven oder negativen Ergebnis besteht für die Gutachterinnen und Gutachter in den meisten Fällen auch die Möglichkeit, eine Nachschulung vorzuschlagen und ggf. festzustellen, ob die Klientin oder der Klient nachschulungsfähig ist. Davon wird am häufigsten Gebrauch gemacht, wenn jemand erstmalig wegen Alkohol aufgefallen ist. Über alle Anlassgruppen hinweg sind 2016 insgesamt 59 % der Beurteilungen positiv und 35 % negativ ausgefallen (Abb. 9.1).

Psychologische Fragen und diagnostisches Vorgehen Die Fahreignungsdiagnostik befasst sich mit unterschiedlichen Aspekten der Fahreignung. Je nach Begutachtungsanlass und Fragestellung liegt der Schwerpunkt etwa auf Verhaltensgewohnheiten im Umgang mit Alkohol, der Persönlichkeit

Statistik der Bundesanstalt für Straßenwesen

Alkoholproblematik häufigster Anlass für Begutachtung

Alkoholabhängigkeit und Alkoholmissbrauch

Tab. 9.6 Begutachtungen bei den medizinisch-psychologischen Untersuchungsstellen 2016

Untersuchungsanlass	Anzahl	Anteil (%)	Ergebnis der Begutachtung		
			Positiv (%)	Nachschnüfungsfähig (%)	Negativ (%)
Verkehrsauffälligkeiten	13.900	15,2	62	0,2	38
Sonstige strafrechtliche Auffälligkeiten	2813	3,1	61	0,1	39
Alkoholauffälligkeit, erstmalig	26.966	29,6	55	11	34
Alkoholauffälligkeit, wiederholt	10.820	11,9	48	7	45
Betäubungsmittel- und Medikamentenauffällige	18.336	20,1	65	7	28
Alkohol- und verkehrs- oder strafrechtliche Auffälligkeit	4895	5,4	46	6	48
Alkohol und Betäubungsmittel/Medikamente	1848	2,0	57	6	37
Betäubungsmittel/Medikamente und allgemein verkehrsauffällig	2115	2,3	57	3	39
Sonstige Mehrfachfragestellungen	1791	2,0	53	3	43
FeV § 10: Abweichung vom Mindestalter	4559	5,0	96	–	4
Gesamt^a	91.185	100	59	6	35

Quelle: Bundesanstalt für Straßenwesen (2017, mit freundlicher Genehmigung). Es sind nur Untersuchungsanlässe mit mindestens 1000 Fällen pro Jahr aufgeführt

^aEinschließlich 4933 übrige Anlässe



Abb. 9.1 Alkoholprobleme sind ein häufiger Anlass für eine verkehrseignungsdiagnostische Untersuchung. (© ronstik/► stock.adobe.com)

9

oder auf bestimmten Merkmalen der Leistungsfähigkeit. Im Fall der Fahrerlaubnis zur Fahrgärtbeförderung oder bei Zweifeln an der psychischen Leistungsfähigkeit stehen eindeutig kognitive Leistungsmerkmale im Vordergrund. Bei Straftaten, die im Zusammenhang mit der Kraftfahreignung oder der Teilnahme am Straßenverkehr stehen, können das Aggressionspotenzial, die Neigung zu rücksichtsloser Durchsetzung eigener Anliegen oder die Bereitschaft zu ausgeprägt impulsivem Verhalten begutachtungsrelevant sein. Damit kommt wieder das diagnostische Interview („Exploration“) als Methode infrage, eventuell auch Fragebögen zur Erfassung von Persönlichkeitsmerkmalen (z. B. Aggressivität), die jedoch verfälschbar sind.

§ 13 FeV Klärung von Eignungszweifeln bei Alkoholproblematik

Zur Vorbereitung von Entscheidungen über die Erteilung oder Verlängerung der Fahrerlaubnis oder über die Anordnung von Beschränkungen oder Auflagen ordnet die Fahrerlaubnisbehörde an, dass

1. ein ärztliches Gutachten (§ 11 Absatz 2 Satz 3) beizubringen ist, wenn Tatsachen die Annahme von Alkoholabhängigkeit begründen, oder
2. ein medizinisch-psychologisches Gutachten beizubringen ist, wenn
 - a) nach dem ärztlichen Gutachten zwar keine Alkoholabhängigkeit, jedoch Anzeichen für Alkoholmissbrauch vorliegen oder sonst Tatsachen die Annahme von Alkoholmissbrauch begründen,
 - b) wiederholt Zuwiderhandlungen im Straßenverkehr unter Alkoholeinfluss begangen wurden,
 - c) ein Fahrzeug im Straßenverkehr bei einer Blutalkoholkonzentration von 1,6 Promille oder mehr oder einer Atemalkoholkonzentration von 0,8 mg/l oder mehr geführt wurde,
 - d) die Fahrerlaubnis aus einem der unter den Buchstaben a bis c genannten Gründe entzogen war oder
 - e) sonst zu klären ist, ob Alkoholmissbrauch oder Alkoholabhängigkeit nicht mehr besteht.

Viele andere Fragestellungen ergeben sich aus der Anwendung von § 11 der Fahrerlaubnis-Verordnung, der sich auf die Eignung der Führerscheininhaber/-innen, speziell auf die „notwendigen körperlichen und geistigen Anforderungen“, beziehen. Eine Begutachtung ist u. a. vorgesehen „bei einem erheblichen Verstoß oder wiederholten Verstößen gegen verkehrsrechtliche Vorschriften“ (§ 11 Absatz 3 Punkt 4) oder „bei einer erheblichen Straftat, die im Zusammenhang mit dem Straßenverkehr steht, oder bei Straftaten, die im Zusammenhang mit dem Straßenverkehr stehen“ (§ 11 Absatz 3 Punkt 5).

Für die Erteilung oder Verlängerung einer Fahrerlaubnis der Klassen D, D1, DE, D1E sowie einer Fahrerlaubnis zur Fahrgastbeförderung müssen Bewerberinnen und Bewerber „geistige und körperliche Eignung gemäß § 11 Absatz 9 in Verbindung mit Anlage 5“ nachweisen. Das bedeutet, dass alle Personen, die etwa einen Bus, ein Taxi oder einen Krankenwagen fahren wollen, grundsätzlich ihre Eignung nachweisen müssen. In Anlage 5 der Fahrerlaubnis-Verordnung werden folgende Anforderungen an die geistige Leistungsfähigkeit spezifiziert:

- a) Belastbarkeit
- b) Orientierungsleistung
- c) Konzentrationsleistung
- d) Aufmerksamkeitsleistung
- e) Reaktionsfähigkeit

Sind die „geistigen Anforderungen“ erfüllt? – Aspekte der Fahreignung

Was das für die Diagnostik bedeutet, wird weiter unten erläutert.

Umsetzung der Verordnung in die Praxis Die Fahrerlaubnis-Verordnung ist relativ abstrakt und nicht selbsterklärend. Damit bundesweit einheitliche Standards zur Anwendung kommen, wurden Leitlinien entwickelt, die erstmals im Jahr 2000 erschienen sind. Davor erfüllten die Begutachtungsleitlinien „Krankheit und Kraftverkehr“ und das „Psychologische Gutachten Kraftfahreignung“ diese Aufgabe. Das Bundesministerium für Verkehr, Bau und Stadtentwicklung hat die Bundesanstalt für Straßenwesen beauftragt, die Leitlinien kontinuierlich und kapitelweise zu überarbeiten. Dazu soll sie (externe) Expertinnen und Experten einbeziehen. Eine Expertinnen- und Expertengruppe bearbeitet jeweils ein Kapitel. Nach Genehmigung durch Bund und Länder wird die aktuelle Fassung dann im Internet veröffentlicht. Die am 24. Mai 2018 veröffentlichte „Begutachtungsleitlinien zur Kraftfahreignung“ (Bundesanstalt für Straßenwesen 2018) enthalten also auch Teile, die älter, aber noch immer gültig sind. Für die Begutachtungspraxis erfüllen die Leitlinien vor allem die folgenden beiden Funktionen:

- Zusammenstellung aller wichtigen eignungsausschließenden und -einschränkenden Merkmale.
- Argumentationshilfe: Die Gutachterin bzw. der Gutachter kann sich im Einzelfall auf die Begutachtungsleitlinien beziehen und muss nicht jede gutachterliche Schlussfolgerung eingehend erläutern.

Am Beispiel der Anlage 5 der Fahrerlaubnis-Verordnung (s. o.) lässt sich die Umsetzung in Leitlinien aufzeigen. Die 5 Stichworte (von Belastbarkeit bis Reaktionsfähigkeit) in Anlage 5 werden in ▶ Abschn. 2.5 der Leitlinien auf 2½ Seiten bezüglich ihrer Umsetzung in die diagnostische Praxis erläutert. So wird begründet, warum diese Anforderungen wichtig sind. Psychische Leistungsmängel können sich wie folgt auswirken (hier am Beispiel von

Konzentrations- und Aufmerksamkeitsleistung aufgezeigt; Bundesanstalt für Straßenwesen 2018, S. 11):

- Die Konzentration ist zeitweilig oder dauernd gestört in der Weise, dass die jeweils anstehende Fahraufgabe aufgrund von Abgelenktsein oder Fehldeutungen verkannt oder fehlerhaft gelöst wird.
- Die Aufmerksamkeitsverteilung ist unzureichend, weil nur ein Teilbereich der für den Kraftfahrer bedeutsamen Informationen erfasst wird und/oder bei Situationswechsel, z. B. nach einer Phase der Monotonie, neue Informationen der Aufmerksamkeit entgehen.

Hier wird also eine Art Anforderungsanalyse zur Begründung nachgereicht.

In den Leitlinien wird dann auf die Messung eingegangen:

Leitlinien erläutern die Fahrerlaubnis-Verordnung

- » Die psychische Leistungsfähigkeit wird mit geeigneten, objektivierbaren psychologischen Testverfahren untersucht. Ausschlaggebend ist, ob die Mindestanforderungen erfüllt werden. (Bundesanstalt für Straßenwesen 2018, S. 12)

Als Mindestanforderung wird ein Prozentrang von 16 in allen eingesetzten Leistungstests verlangt. Das würde zwangsläufig zum Ausschluss sehr vieler Personen führen. Vielleicht wird das diagnostische Vorgehen deshalb relativiert, u. a. durch folgende Formulierung:

- » Bei Grenzwertunterschreitungen kann durch Ergebnisse weiterer Verfahren (Ergänzungsverfahren, Verhaltensbeobachtung, Wiederholungsuntersuchung) nachgewiesen werden, dass das aus den Leistungsresultaten zu erschließende Risiko durch das Kompensationspotential (vorausschauendes Denken, ausgeprägtes Risikobewusstsein, sicherheitsbetonte Grundhaltung) angemessen gemindert werden kann. (Bundesanstalt für Straßenwesen 2018, S. 12)

Kommentare zu den Leitlinien

Für die Umsetzung der Leitlinien in die diagnostische Praxis, aber auch für deren eventuelle spätere Überarbeitung oder auch für die Fahrerlaubnis-Verordnung ist ein Standardwerk relevant, dass seit 2018 in 3., überarbeiteter und erweiterter Auflage vorliegt, nämlich die „*Begutachtungsleitlinien zur Kraftfahreignung: Kommentar*“ (Schubert et al. 2018). Das Werk ist wesentlich umfangreicher als die Leitlinien, auf die es sich bezieht. Während beispielsweise die in dem Werk abgedruckte Leitlinie „Anforderungen an die psychische Leistungsfähigkeit“ 1½ Buchseiten umfasst, nehmen die beiden Kommentare dazu insgesamt 33½ Seiten in Anspruch. Wie auch bei den Leitlinien bezieht sich ein Großteil des Werkes auf medizinische Fragen. Für die psychologische Diagnostik sind die Kommentare im allgemeinen Teil zu den „Anforderungen an die psychische Leistungsfähigkeit“, zur „Kompensation von Eignungsmängeln“ und zum Thema „Kumulierte Auffälligkeiten“ relevant. Im speziellen Teil sind es insbesondere die Themen „Alkohol“, „Betäubungsmittel und Arzneimittel“, „Intellektuelle Leistungseinschränkungen“, „Straftaten“, „Verstöße gegen verkehrsrechtliche Vorschriften“, „Auffälligkeiten bei der Fahrerlaubnisprüfung“, „Fahrgastbeförderung“ und „Ausnahmen vom Mindestalter“ sowie die ergänzenden Hinweise zu den Themen „Persönlichkeitsstörungen“ und „ADHS“.

Wir gehen im Folgenden auf 2 Themengebiete näher ein: Die Anforderungen an die psychische Leistungsfähigkeit und Alkoholauffälligkeit.

Anforderungen an die psychische Leistungsfähigkeit Wie bereits oben ausgeführt wurde, verlangt der Gesetzgeber bei bestimmten Personengruppen, die (in der Regel berufsmäßig) andere Menschen mit einem Pkw oder Bus transportieren, dass sie über eine gute Belastbarkeit, Orientierungsleistung, Konzentrationsleistung, Aufmerksamkeitsleistung und Reaktionsfähigkeit verfügen und dies auch durch ein Gutachten nachweisen müssen. Die gleichen

Diagnostik in weiteren Anwendungsfeldern

Merkmale können auch relevant sein, wenn Zweifel an der Kraftfahreignung aufgrund bestimmter Einschränkungen bestehen. Die folgenden Ausführungen basieren überwiegend auf dem Kommentar von Schmidt-Atzert et al. (2018).

Warum gerade diese Merkmale entscheidend sein sollen, lässt sich nicht exakt nachverfolgen. Einer informellen Quelle zufolge wurden sie einmal von 2 Experten auf Nachfrage genannt, die sich dabei an den damals im Einsatz befindlichen Tests und deren Namen (z. B. Aufmerksamkeits-Belastungs-Test) orientiert haben. Jedenfalls findet sich kein Hinweis dafür, dass jemals eine fundierte Anforderungsanalyse durchgeführt worden ist. Die Anforderungen im Straßenverkehr ändern sich u. a. durch die Einführung von technischen Assistenzsystemen. Die Orientierungsleistung (in den Leitlinien als Zielorientierung im Verkehrsraum umschrieben) mag zu Zeiten, als viele Menschen noch mit dem Stadtplan auf dem Schoß und nach Straßennamen suchend unterwegs waren, relevant gewesen sein. Heute werden die Fahrerinnen und Fahrer meist durch ein Navigationssystem zum Ziel geführt. Es ist deshalb fraglich, ob diese Anforderung heute relevant ist. Die Belastbarkeit wird heute in der Psychologie mit Stressresistenz weitgehend gleichgesetzt und nicht als Merkmal der geistigen Leistungsfähigkeit verstanden.

Die Forderung, dass die Testpersonen bei allen 5 Merkmalen mindestens einen Prozentrang von 16 erreichen müssen, um als geeignet zu gelten, führt zu einer sehr hohen Ablehnungsquote. Je mehr Tests jemand bearbeiten muss, desto geringer ist die Chance, bei allen Tests zu bestehen. Bei einem Cut-off von Prozentrang 85 beträgt die Erfolgswahrscheinlichkeit 85 %. Muss man 5 unabhängige (unkorrelierte) Tests mit einem Cut-off von Prozentrang 85 bestehen, beträgt die Wahrscheinlichkeit, in allen Tests als „geeignet“ beurteilt zu werden, 44,37 %. Von 100 Personen werden am Ende also theoretisch nur 44 als geeignet und 56 als ungeeignet gelten. Je höher die Tests jedoch korreliert sind, desto weniger dramatisch ist das Ergebnis; bei einer Korrelation von 1 würden 85 von 100 Personen am Ende positiv beurteilt. Die „Wahrheit“ liegt also irgendwo zwischen diesen beiden Extremen (s. auch ▶ Abschn. 5.1.3.2 für Ausführungen zu multiplen Hürden). In der Praxis lässt sich das Problem der unangemessen hohen Ablehnungsquoten jedoch lösen. In den Leitlinien wird empfohlen, bei Grenzwertunterschreitungen ggf. einen Test zu wiederholen (zu Übungseffekten s. ▶ Abschn. 3.2.1) oder die Fahreignung durch ein anderes Verfahren festzustellen. Hier bietet sich eine Fahrverhaltensbeobachtung an. Damit kann eventuell nachgewiesen werden, dass ein mit einem objektiven Leistungstest entdeckter Eignungsmangel in der Fahrpraxis kompensiert werden kann und somit doch eine positive Eignungsaussage möglich ist (Kranich 2018).

Der Grenzwert bei Prozentrang ≥ 16 führt zwangsläufig zu weiteren Problemen. Erstens führt die vermutlich sehr niedrige Kriteriumsvalidität der Tests zu großen „Kollateralschäden“. Die negativ beurteilten Klientinnen und Klienten haben nur ein geringfügig höheres Unfallrisiko als die Population, was sich leicht mithilfe der Taylor-Russel-Tafeln (▶ Abschn. 5.1.3.3) abschätzen lässt. Nehmen wir an, dass der Zusammenhang zwischen Test und Kriterium nur $r = .10$ beträgt und 1000 Personen untersucht werden. Das Unfallrisiko in der unausgelesenen Population betrage 10 % für einen definierten Zeitraum. Nach dem Kriterium „Prozentrang ≥ 16 “ würden 160 als ungeeignet diagnostiziert. Und davon würden nur 20 in dem definierten Zeitraum einen Unfall verursachen, 140 würden falsch positiv beurteilt. Ohne eine Selektion durch den Test würden von 160 zufällig ausgewählten Personen (entsprechend der Grundrate von 10 %) 16 einen Unfall verursachen. Die Anwendung eines wenig validen Tests in Kombination dem 16 %-Kriterium

Belastbarkeit und
Orientierungsleistung fragliche
Anforderungen

5-mal mindestens
Prozentrang 16 erreichen

Viele falsch positive Entscheidungen
zu erwarten

Messfehlerbedingte Fehlentscheidungen zu erwarten

Eignung der Test wird von unabhängiger Stelle bestätigt

Bestätigungsverfahren

Die Tests müssen nicht kriteriumsvalide sein

hat also nur einen kleinen gesellschaftlichen Nutzen, indem von 1000 untersuchten Personen gerade einmal 4 Unfallfahrer/-innen mehr entdeckt werden als ohne den Test. Auf der anderen Seite ist ein großer Schaden für die untersuchten Menschen zu verzeichnen. Insgesamt 140 der 1000 untersuchten Personen werden negativ begutachtet, ohne dass sie ein Unfallrisiko aufweisen.

Zweitens wird mit dem Grenzwert die Messgenauigkeit der Tests ausgebendet. Berücksichtigt man die Messgenauigkeit, so wird deutlich, dass man in vielen Fällen gar nicht sicher sein kann, dass die getestete Person tatsächlich zu den 16 % schlechtesten gehört: Dem Prozentrang von 16 entspricht exakt ein Standardwert von 90. Nehmen wir an, die Messgenauigkeit des Tests betrage $r_{tt} = .90$ und es sei eine Urteilssicherheit von 90 % erwünscht. Das Konfidenzintervall reicht (bei einseitiger Fragestellung unter Berücksichtigung der Regression zur Mitte) bis zu einem Standardwert von 96,9 (oder einem Prozentrang von 37,8). Im Extremfall wird man Klientinnen und Klienten mit einem beobachteten Standardwert von 90 als ungeeignet klassifizieren, obwohl ihre „wahre“ Testleistung durchschnittlich sein kann.

Die Frage, welche Tests zur Messung der genannten Anforderungsmerkmale geeignet sind, bedarf einer Antwort. Eine schlechte Lösung wäre es, jeder Gutachterin und jedem Gutachter freizustellen, geeignete Tests zu suchen. Dies könnte dazu führen, dass sich unterschiedliche Standards in den Begutachtungsstellen etablieren. Das wiederum hätte zur Folge, dass das Ergebnis einer Begutachtung davon abhängt, welche Stelle begutachtet. Auch könnten negative Gutachten mit dem Argument angefochten werden, es sei ein nicht geeignetes psychologisches Testverfahren eingesetzt worden. In Anlage 5 der Fahrerlaubnis-Verordnung wird deshalb vorgeschrieben, dass die „Eignung der zur Untersuchung dieser Merkmale eingesetzten psychologischen Testverfahren von einer unabhängigen Stelle“ bestätigt werden muss. Früher bestand die Praxis darin, dass für einen Test in einem Bundesland eine Zulassung beantragt wurde. Dafür musste ein Expertengutachten über den Test mit eingereicht werden. War ein Test erst einmal in einem Bundesland zugelassen, haben sich die anderen Länder angeschlossen.

Seit 2017 steht in § 71a der Fahrerlaubnis-Verordnung, dass die Eignung der psychologischen Testverfahren und -geräte von einer amtlich anerkannten unabhängigen Stelle festgestellt werden muss. Die unabhängige Stelle muss dabei bestimmte Vorgaben erfüllen, um die Eignung eines Tests (und ggf. auch anderer Verfahren) zu prüfen. Der Gesetzgeber schreibt ein Bestätigungsverfahren vor, das mehrere Schritte umfasst (TransMIT-Zentrum für wissenschaftlich-psychologische Dienstleistungen (DGPs) 2020):

1. Zwei Gutachterinnen oder Gutachter erstellen getrennt voneinander sog. Einzelbestätigungsunterschriften.
2. Diese werden zu einem sog. vorläufigen Bestätigungsunterlagen zusammengezogen.
3. Der Auftraggeber kann hiergegen sachliche Einwände erheben.
4. Die Gutachterinnen oder Gutachter erstellen ein abschließendes gemeinsames sog. Bestätigungsunterlagen.

Das formale Vorgehen zur Beurteilung von Verfahren zur Kraftfahreignung ist einer Testrezension nach den Standards des Diagnostik- und Testkuratoriums (2018) vergleichbar. Allerdings ist die Begutachtung kostenpflichtig und stärker reglementiert. In einer Veröffentlichung der Bundesanstalt für Straßenwesen (2019) „Fachliche Hinweise zur Begutachtung von Testverfahren und Kursen“ vom 22. Februar 2019 finden sich bemerkenswerte Aussagen zur Validität der zu begutachtenden Verfahren. Eine Prämisse ist: „Ein Nachweis der Verwendbarkeit der Testergebnisse zur Verhaltensprognose in Bezug auf die ‚Fahreignung‘ ist nicht notwendig, da diese nicht hinreichend definiert ist.“

Daraus folgt: „Ein Nachweis der Kriteriumsvalidität in Bezug auf Fahreignung ist nicht nötig“ (Bundesanstalt für Straßenwesen 2019, S. 3).

Mit anderen Worten: Ein Test zur Messung eines der 5 Merkmale der geistigen Leistungsfähigkeit (Belastbarkeit, Orientierungsleistung, Konzentrationsleistung, Aufmerksamkeitsleistung oder Reaktionsfähigkeit; s. o.) muss nicht die Fahreignung messen können. Es genügt der Nachweis, dass er das misst, was er messen soll (also Belastbarkeit etc.). Formal wird dies damit begründet, dass die Fahreignung nicht hinreichend definiert sei. Man kann aber auch vermuten, dass hier zum Tragen kommt, dass die bisherigen Versuche, Fahreignung im Straßenverkehr über das individuelle Unfallrisiko mit Verfahren zu den genannten Merkmalen der geistigen Leistungsfähigkeit vorherzusagen, weitgehend gescheitert sind. In einer Metaanalyse ermittelten Arthur et al. (1991) eine niedrige Korrelation von $r = .26$ zwischen der Leistung in Tests zur selektiven Aufmerksamkeit und Unfällen im Straßenverkehr. In einer Analyse zur Validität der in Deutschland eingesetzten Testverfahren zur psychometrischen Leistungsprüfung der Fahreignung von Poschadel, Falkenstein, Pappachan, Poll und von Hinckeldey (2009) wird für kein einziges Verfahren ein „harter“ Validitätsbeleg in Form von Unfallrisiko berichtet (Fahrverhaltensproben stellen keinen Ersatz für Unfallkriterien dar, weil es sich um diagnostische Verfahren handelt).

Kaum Belege zur Kriteriumsvalidität vorhanden

Ist Aufmerksamkeit überhaupt für die Kraftfahreignung relevant?

In den USA wurde eine extrem aufwendige und aussagekräftige Feldstudie durchgeführt, in der über 100 Fahrerinnen und Fahrer im alltäglichen Straßenverkehr mittels Kameras über 1 Jahr lang beobachtet wurden. In der 100-Car Naturalistic Driving Study (s. auch ▶ Abschn. 3.6.2) ging es nicht um zeitlich stabile Merkmale wie Aufmerksamkeit oder Konzentrationsfähigkeit, sondern um konkretes Verhalten. Und hier zeigte sich, dass unaufmerksames Verhalten wie etwa das Schreiben von Textnachrichten, das Eintippen von Telefonnummern ins Smartphone oder das Auflegen von Make-up das Unfallrisiko deutlich erhöht. Vermutlich hängt dieses gefährliche Verhalten aber nicht mit dem Eignungsmerkmal Aufmerksamkeit zusammen, sondern – wenn überhaupt – mit ganz anderen Persönlichkeitsmerkmalen. Es wird auch niemand ernsthaft fordern, geteilte Aufmerksamkeit sei wichtig, damit man im Auto solche komplexen „Zusatzaufgaben“ gefahrlos erledigen kann. Sich aufmerksam verhalten und eine hohe Aufmerksamkeit im Sinne einer Eigenschaft haben, sind zwei verschiedene Dinge. Der in Deutschland etablierte Ansatz, Kraftfahreignung über stabile Eigenschaften zu operationalisieren, bedarf also nach wie vor einer empirischen Begründung.

Die Diagnostik der laut Fahrerlaubnis-Verordnung 5 relevanten Merkmale der geistigen Leistungsfähigkeit bereitet also viele Probleme. Die Merkmale „Orientierung“ und „Belastbarkeit“ sind vermutlich nicht mehr mit den aktuellen Anforderungen im Straßenverkehr vereinbar. Der Grenzwert „Prozentrang ≥ 16 “ führt zwangsläufig zu sehr vielen diagnostischen Fehlentscheidungen, die aus der vermutlich sehr geringen Kriteriumsvalidität sowie dem Konfidenzintervall bei jeder Messung herrühren. Die Gutachterinnen und Gutachter, soweit sie sich der Problematik bewusst sind, können aber viele Fehlentscheidungen abwenden, indem sie die Kompensierbarkeit der (vermeintlichen) Schwächen der geistigen Leistungsfähigkeit durch andere Verfahren prüfen.

Fazit

Differenzierung zwischen Alkoholmissbrauch und Alkoholabhängigkeit

Alkoholabhängigkeit nach ICD-Kriterien

Kriterien für Alkoholmissbrauch

Alkoholproblematik Erstmalige oder wiederholte Auffälligkeit wegen Trunkenheit am Steuer machen zusammen über 40 % aller Begutachtungsanlässe aus. Rechnet man andere Begutachtungsanlässe hinzu, bei denen Alkohol beteiligt ist, kommt man auf insgesamt 48,9 % (vgl. ▶ Tab. 9.6). In den Leitlinien wird zwischen Alkoholmissbrauch und Alkoholabhängigkeit unterschieden.

Die Diagnostik einer *Alkoholabhängigkeit* (s. Haffner et al. 2018) erfolgt nach den üblichen ICD-10-Kriterien (▶ Abschn. 8.3.1). Dabei wird nicht nur Verhalten exploriert, auch organischen Befunden kommt eine Bedeutung zu. Denn ob eine Alkoholabstinenz vorliegt, lässt sich an einer Reihe von Laborwerten im Blut überprüfen. Marker, die auf Alkoholkonsum hinweisen, finden sich auch im Urin und in den Haaren. In manchen Fällen liefert sogar ein Alkoholtest bei der Begutachtung relevante Erkenntnisse zum Alkoholkonsum. Ein ICD-Kriterium ist „*anhaltender Substanzkonsum trotz schädlicher Folgen*“. Eine körperliche Untersuchung kann Informationen über einschlägige Organschäden liefern. Es ist offenkundig, dass bei Verdacht auf eine Alkoholabhängigkeit, aber auch bei einigen anderen Fragestellungen eine Zusammenarbeit von Medizinerinnen bzw. Medizinern und Psychologinnen bzw. Psychologen erforderlich ist. Die Untersuchungen tragen zu Recht den Namen „*medizinisch-psychologische Untersuchung*“. Menschen, die alkoholabhängig sind, dürfen kein Kraftfahrzeug führen. In diesem Punkt sind die Leitlinien eindeutig. Für die Feststellung, dass keine Abhängigkeit mehr vorliegt, wird der Nachweis verlangt, dass eine stabile Abstinenz besteht. In der Regel sind eine erfolgreiche Entwöhnungsbehandlung und eine 1-jährige Abstinenz nach der Entgiftungs- und der Entwöhnungszeit nachzuweisen. Außerdem dürfen keine sonstigen eignungsrelevanten Mängel vorliegen.

Auch *Alkoholmissbrauch* ist unvereinbar mit dem Führen eines Kraftfahrzeugs. Die folgenden Ausführungen fassen zentrale Argumente des Kommentars von Stephan und Brenner-Hartmann (2018) zum Thema Alkoholmissbrauch zusammen. Alkoholmissbrauch im klinischen Sinne kann nach ICD-10 diagnostiziert werden. Alkoholmissbrauch entspricht am ehesten dem „*schädlichen Gebrauch*“ (F10.1 nach ICD-10). Allerdings ist es nicht das Ziel einer Untersuchung zur Kraftfahreignung, eine klinische Diagnose zu stellen. Vielmehr kommt es darauf an, künftiges Verhalten zu prognostizieren, also die Frage zu klären, ob eine Klientin oder ein Klient künftig unter Alkoholeinfluss am Straßenverkehr teilnehmen wird. Von Missbrauch wird ausgegangen, wenn die Person

- wiederholt ein Fahrzeug mit einer Blutalkoholkonzentration von mindestens 0,5 ‰ geführt hat;
- einmalig ein Fahrzeug mit einer Blutalkoholkonzentration von mindestens 1,0 ‰ geführt hat;
- bei der Verkehrsteilnahme einen Verlust der Kontrolle des Alkoholkonsums gezeigt hat.

Dementsprechend gibt es verschiedene Untersuchungsanlässe. Für die beiden erstgenannten Kriterien genügt es, wenn etwa bei Routinekontrollen mindestens 2 × eine Blutalkoholkonzentration von mindestens 0,5 ‰ oder 1 × eine von mindestens 1,0 ‰ festgestellt wurden. Ob das Fahrverhalten auffällig war oder nicht, spielt dabei keine Rolle. Gerichte entziehen übrigens die Fahrerlaubnis bei einer Verkehrsteilnahme von über 1,1 ‰ (1,0 + 0,1 ‰ für mögliche Messungenauigkeit) und verhängen eine Sperrfrist. Der Kontrollverlust ist dagegen schwer zu operationalisieren. Ein Kriterium ist jedoch, dass die Person mit einer Blutalkoholkonzentration von über 1,6 ‰ ein Fahrzeug geführt hat – das kann auch ein Fahrrad gewesen sein. Dabei muss die Fahrerin oder der Fahrer sich nicht auffällig verhalten

haben. Bei lang anhaltendem starkem Alkoholkonsum kann nämlich eine Gewöhnung eintreten.

Wenn nach den oben genannten Kriterien ein Alkoholmissbrauch und damit die Nichteignung festgestellt worden ist, gilt es in der Untersuchung zu klären, ob inzwischen die Eignung wiederherstellt worden ist. Beispielsweise hat die Behörde jemandem nach einer Trunkenheitsfahrt die Fahrerlaubnis entzogen. Nach Ablauf einer Sperrfrist muss die betroffene Person durch ein Gutachten nachweisen, dass kein Alkoholmissbrauch mehr vorliegt. Erst dann wird die Behörde die Fahrerlaubnis wieder erteilen.

Entscheidend ist die Frage, ob die Person nicht mehr mit erhöhter Wahrscheinlichkeit unter Alkoholeinfluss am Straßenverkehr teilnehmen wird. Dafür gibt es verschiedene Hinweise, die bei der Begutachtung gewürdigt werden:

- Das Alkoholtrinkverhalten wurde ausreichend geändert. Das kann Abstinenz bedeuten oder auch eine zuverlässige Trennung von Alkoholkonsum und Verkehrsteilnahme.
- Die Änderung des Trinkverhaltens ist stabil und motivational gefestigt.
- Die Lebensverhältnisse stehen einer Stabilisierung nicht entgegen.
- Wenn eine „Persönlichkeitsproblematik“ vorlag, wurde diese erkannt und korrigiert.
- Es liegen keine verkehrsrelevanten Leistungs- und Funktionsbeeinträchtigungen als Folge eines früheren Alkoholmissbrauchs mehr vor.

Für diese und weitere Kriterien finden sich in den Leitlinien und vor allem im zuvor angesprochenen Kommentar (Stephan und Brenner-Hartmann 2018) Hinweise zur Operationalisierung. So ist eine stabile und motivational gefestigte Änderung des Trinkverhaltens u. a. daran zu erkennen, dass der Änderungsprozess aus einem angemessenen Problembewusstsein heraus erfolgt ist und die mit der Verhaltensänderung erzielten Wirkungen positiv erlebt werden.

Die Beurteilungsgesichtspunkte lassen bereits erkennen, welche diagnostischen Verfahren bei Annahme einer Alkoholproblematik naheliegen. Zur Feststellung von Alkoholmissbrauch oder -abhängigkeit kommt dem diagnostischen Interview eine zentrale Bedeutung zu. Nur dieses Verfahren kann die nötigen Informationen über Trink- und Verhaltensgewohnheiten sowie über Problembewusstsein, eingeleitete Therapiemaßnahmen etc. liefern. Die Gutachterin oder der Gutachter kann schriftliche Belege über Therapiemaßnahmen verlangen. Dem Interview geht eine Auswertung der Akten voraus. Wichtige Informationen sind die Höhe des Blutalkoholspiegels, das Verhalten unter Alkoholeinfluss (unauffälliges Verhalten spricht für eine hohe Alkoholtoleranz), eventuelle Vorgutachten, medizinische Befunde zu alkoholbedingten Schädigungen, Laborwerte etc. Leistungs- oder Funktionsbeeinträchtigungen nach einer überwundenen Alkoholabhängigkeit werden mit Leistungstests überprüft.

Wiederherstellung der Kraftfahreignung nach Alkoholmissbrauch

Diagnostisches Interview nach Auswertung der Akten

9.3.2 Spezielle Probleme der verkehrspychologischen Diagnostik

Verfälschung Die Klientinnen und Klienten werden bestrebt sein, einen „guten“ Eindruck zu hinterlassen, um den Führerschein zurückzuerhalten. Das diagnostische Interview hat bei vielen Fragestellungen einen hohen Stellenwert. Die Fragen müssen daher so ausgewählt werden, dass diagnostisch relevante Fakten von beschönigenden Darstellungen unterschieden werden können.

Übereinstimmung zwischen Interviewaussagen und Akteninformationen wichtig

Kontrolle der Begutachtungsstellen durch die Bundesanstalt für Straßenwesen

Grundsätze zur Begutachtung in Anlage 4a der Fahrerlaubnis-Verordnung

In Foren und auf Webseiten von Medien findet man viele Tipps, was man im Gespräch mit der Psychologin oder dem Psychologen am besten sagt, um den Führerschein wiederzubekommen. Es gibt sogar Anbieter/-innen von Vorbereitungskursen. Es ist kaum möglich, authentische Antworten von einstudierten zu unterscheiden. Diskrepanzen zwischen verschiedenen Aussagen, zwischen verbalen Angaben und Verhalten (manche Klientinnen bzw. Klienten kommen alkoholisiert zur Untersuchung!) oder Akteninformationen sprechen für eine Verfälschung.

Qualität von Gutachten Für die Erstellung von Gutachten zur Kraftfahreignung gelten die gleichen Anforderungen wie für andere Gutachten (► Abschn. 4.6). Gerade in der verkehrspsychologischen Diagnostik sind große Anstrengungen zu erkennen, die Qualität der Gutachten zu sichern. § 72 der Fahrerlaubnis-Verordnung befasst sich mit der Begutachtung. Dort ist geregelt, dass sich die Träger von Begutachtungsstellen für Fahreignung einer Kontrolle durch die Bundesanstalt für Straßenwesen unterziehen müssen. Dazu gehören eine Erstbegutachtung, regelmäßige Begutachtungen sowie Begutachtungen aus besonderem Anlass. Wichtig ist in dem von uns betrachteten Zusammenhang, dass auch die Gutachten zu überprüfen sind.

Anlage 4a der Fahrerlaubnis-Verordnung zu § 11 Absatz 5 nennt Grundsätze für die Durchführung der Untersuchungen und die Erstellung der Gutachten. Wichtige Punkte, die hier wörtlich wiedergegeben sind, lauten:

Anlage 4a zu § 11 Absatz 5 FeV

1. Die Untersuchung ist unter Beachtung folgender Grundsätze durchzuführen:
 - a) Die Untersuchung ist anlassbezogen und unter Verwendung der von der Fahrerlaubnisbehörde zugesandten Unterlagen über den Betroffenen vorzunehmen. Der Gutachter bzw. die Gutachterin hat sich an die durch die Fahrerlaubnisbehörde vorgegebene Fragestellung zu halten.
 - b) Gegenstand der Untersuchung sind nicht die gesamte Persönlichkeit des Betroffenen, sondern nur solche Eigenschaften, Fähigkeiten und Verhaltensweisen, die für die Kraftfahreignung von Bedeutung sind (Relevanz zur Kraftfahreignung).
 - c) Die Untersuchung darf nur nach anerkannten wissenschaftlichen Grundsätzen vorgenommen werden.
 - d) Vor der Untersuchung hat die Gutachterin bzw. der Gutachter den Betroffenen über Gegenstand und Zweck der Untersuchung aufzuklären.
 - e) Über die Untersuchung sind Aufzeichnungen anzufertigen.
2. Das Gutachten ist unter Beachtung folgender Grundsätze zu erstellen:
 - a) Das Gutachten muss in allgemeinverständlicher Sprache abgefasst sowie nachvollziehbar und nachprüfbar sein. Die Nachvollziehbarkeit betrifft die logische Ordnung (Schlüssigkeit) des Gutachtens. Sie erfordert die Wiedergabe aller wesentlichen Befunde und die Darstellung der zur Beurteilung führenden Schlussfolgerungen. Die Nachprüfbarkeit betrifft die Wissenschaftlichkeit der Begutachtung. Sie erfordert, dass die Untersuchungsverfahren, die zu den Befunden geführt haben, angegeben und, soweit die Schlussfolgerungen auf Forschungsergebnisse gestützt sind, die Quellen genannt werden. Das Gutachten braucht aber nicht im Einzelnen die wissenschaftlichen Grundlagen für die Erhebung und Interpretation der Befunde wiederzugeben.
 - b) Das Gutachten muss in allen wesentlichen Punkten insbesondere im Hinblick auf die gestellten Fragen (§ 11 Absatz 6) vollständig sein.[...]

Auf dem 48. Deutschen Verkehrsgerichtstag im Januar 2010 befasste sich einer von 8 Arbeitskreisen mit der medizinisch-psychologischen Untersuchung. Der provokante Titel des Arbeitskreises lautete „Idiotentest“ auf dem Prüfstand“. Alle Referentinnen und Referenten stellten klar, dass die Begutachtungen grundsätzlich geeignet seien, die Fahreignung festzustellen, und sie einen wichtigen Beitrag zur Verkehrssicherheit darstellten. Eine rege Diskussion entwickelte sich zu der Frage, ob und ggf. wie die Inhalte der Exploration dokumentiert werden sollen. In der Anlage 4a zu § 11 Absatz 5 der Fahrerlaubnis-Verordnung (► Abschn. 9.3.1) steht unter Nummer 1e: „Über die Untersuchung sind Aufzeichnungen anzufertigen.“ Strittig war, ob eine Ton- oder sogar Videoaufnahme zu fordern ist und ob und in welcher Form die Aufzeichnungen den Klientinnen/Klienten bzw. deren Anwälten zur Verfügung gestellt werden sollen.

Betroffene, die aufgrund eines negativen Gutachtens ihren Führerschein nicht wiederbekommen, argumentieren manchmal: „Was im Gutachten steht, habe ich in der Untersuchung so nicht gesagt; die Schlussfolgerungen der Gutachterin bzw. des Gutachters sind deshalb falsch.“ Rechtsanwältinnen bzw. Rechtsanwälte und Richterinnen bzw. Richter, die sich mit dem Fall beschäftigen, wissen nicht, wer die richtige Aussage macht.

„Aufzeichnungen“ von Exploration vorgeschrieben

Ton- und Videoaufzeichnungen kontrovers diskutiert

Geiger (2010, S. 212), Präsident des Verwaltungsgerichts München, argumentiert, dass es Sinn der Regelung sei, im Zweifelsfall nachweisen zu können, wie sich der Betroffene genau geäußert hat. „Das beste Beweismittel – jedenfalls was eine mündliche Exploration angeht – ist eine Tonaufnahme.“

Hillmann (2010, S. 22), Fachanwalt für Verkehrsrecht, stellt fest, dass gegenwärtig nur die Gutachten durch die Bundesanstalt für Straßenwesen überprüft werden, nicht aber die Begutachtung: „Trotz festgelegter Standards bei medizinisch-psychologischen Untersuchungen ist der Untersuchungsablauf nur mangelhaft nachprüfbar, weil das Gespräch nicht als Video- oder Tonbandprotokoll aufgezeichnet wird.“ Er führt außerdem aus: „Zur besseren Nachvollziehbarkeit des Explorationsgespräches, aber auch zum Schutz der Prüfer vor unberechtigten Vorwürfen der Probanden sollte daher ohne jede Ausnahme ein Tonband- und/oder Videomitschnitt erfolgen“ (Hillmann 2010, S. 222). Der Mitschnitt solle mindestens 6 Monate aufbewahrt werden, damit er bei Bedarf auch vom Anwalt angefordert werden kann; eine Abschrift sei nicht nötig. Manche Prozesse würden vermieden, wenn der Anwalt bzw. die Anwältin feststellt, dass die Aussagen im Gutachten korrekt wiedergegeben wurden.

Schubert (2010) weist darauf hin, dass schon jetzt, z. B. bei der DEKRA, das Gespräch auf Wunsch des Probanden (gegen Bezahlung) aufgezeichnet werde. In diesem Fall würde das Gespräch vollständig transkribiert und die Abschrift in das Gutachten integriert. Hillmann (2010) meint dazu, dass der Proband bzw. die Probandin zu Beginn der Untersuchung nicht wisse, ob er/sie positiv oder negativ beurteilt werde. Die Notwendigkeit einer Überprüfung ergebe sich aber nur bei einem negativen Gutachten. Schubert spricht von einem von Vertrauen getragenen Arbeitsbündnis zwischen Gutachter/-in und Proband/-in. „Das Untersuchungsgespräch ist von daher äußeren Einflüssen gegenüber hoch sensibel, weshalb hier dem Vier-Augen-Prinzip absoluter Vorrang zu geben ist“ (Schubert 2010, S. 242). Im Übrigen lasse der Verordnungsgeber in der Fahrerlaubnis-Verordnung Anlage 15 Nr. 1e (s. o.) die „Auswahl der anzuwendenden Methoden bewusst und fachlich gerechtfertigt offen“ (Schubert 2010, S. 243).

Schmidt-Atzert (2010), als Fachvertreter für Psychologische Diagnostik eingeladen, erinnert daran, dass die Beurteilerübereinstimmung selbst bei hoch

standardisierten Interviews nicht perfekt sei. Die Möglichkeiten, die Objektivität der halbstandardisierten Exploration durch Schulung der Interviewer zu erhöhen, seien sehr begrenzt. „Durch eine gute Dokumentation des Interviews kann die Angemessenheit der Durchführung, Auswertung und Interpretation jedoch einer nachträglichen Überprüfung zugänglich gemacht werden. Dies ist vermutlich die einzige konstruktive Lösung für den Umgang mit der begrenzten Objektivität“ (Schmidt-Atzert 2010, S. 260).

In der endgültigen Abstimmung, an der viele Psychologen und Psychologinnen aus medizinisch-psychologischen Untersuchungsstellen teilnahmen, wurde eine Empfehlung an den Gesetzgeber, die Aufzeichnung der Gespräche vorzuschreiben, mehrheitlich abgelehnt.

Einer Empfehlung des Deutschen Verkehrsgerichtstags im Jahr 2014 zufolge können Tonaufzeichnungen die Transparenz der Fahreignungsbegutachtung erhöhen. Es seien aber noch wissenschaftliche und rechtliche Fragen zu klären. Damit wurde die Projektgruppe „MPU-Reform“ beauftragt (Okulicz-Kozaryn et al. 2015, S. 379). Diese Projektgruppe nennt in ihrem Abschlussbericht eine ganze Reihe von Vorbehalten. „Vor einer Einführung von obligatorischen Tonaufzeichnungen wäre insbesondere zu klären: Zugriffsmöglichkeiten und Aufbewahrung, Notwendigkeit der Transkription, Finanzierung, Persönlichkeitsrechte des Gutachters (Schutz des nicht öffentlich gesprochenen Wortes), wissenschaftliche Bestätigung, dass die Exploration/Anamnese und das Ergebnis der MPU nicht nachteilig beeinflusst werden, Nachweis (z. B. im Rahmen eines Pilotversuchs), dass die Qualität der Gutachten nicht nachteilig beeinflusst wird, Möglichkeit des Widerspruchs für die Klienten“ (Albrecht et al. 2015, S. 22). Die Anlage 4a zu § 11 Absatz 5 der Fahrerlaubnis-Verordnung wurde (Stand: Juni 2020) nicht geändert.

Weiterführende Literatur

Für die Praxis verkehrspychologischer Diagnostik und insbesondere die Begutachtung haben sich die Kommentare zu den Begutachtungsleitlinien zur Kraftfahreignung inzwischen als Standardwerk etabliert (Schubert et al. 2018). Wie die Begutachtung ablaufen soll (normativer Ansatz!) wird sehr klar von Schubert und Mattern (2009) dargestellt.

9.4 Zusammenfassung

In diesem Kapitel wurde beschrieben, welche Fragestellungen in den Anwendungsbereichen Neuro-, Rechts- und Verkehrspychologie vorliegen und wie sie mithilfe von psychologischer Diagnostik beantwortet werden.

Die Neuropsychologie befasst sich hauptsächlich mit der Beeinträchtigung psychischer Funktionen (z. B. Gedächtnis) als Folge organischer Störungen im Zentralnervensystem (z. B. Demenz). Manchmal ist die Rolle organischer Ursache noch nicht hinlänglich geklärt, wie etwa bei ADHS oder bei Lese- und Rechtschreibstörungen. Wenn das Gehirn oder allgemeiner das Nervensystem beteiligt ist, kommt der Zusammenarbeit mit der Medizin eine besondere Bedeutung zu. Bestimmte Erkrankungen, etwa ein Schlaganfall oder ein Tumor, können oftmals im Gehirn genau lokalisiert werden. Objektive, durch bildgebende Verfahren nachweisbare körperliche Defekte erlauben Hypothesen darüber, welche Funktionseinschränkungen zu erwarten sind. Diese werden im Rahmen der neuropsychologischen Diagnostik quantifiziert und zumeist auch im Kontext weiterer Beeinträchtigungen, aber auch für eine Kompensation nutzbarer Stärken eingeordnet. Der Weg führt nicht immer von medizinischen Befunden zur neuropsychologischen Diagnostik. Der Verdacht auf eine hirnorganische Störung kann auch durch eine Anamnese, eine Verhaltensbeobachtung oder bestimmte Testergebnisse begründet sein.

und dann durch medizinische Diagnostik abgeklärt werden. Manchmal stellt sich in der neuropsychologischen Diagnostik auch die Frage nach der Verursachung, etwa ob eine Funktionseinschränkung bereits schon vor einem Unfall bestanden hat oder nicht. In diesem Fall müssen die entsprechenden prämorbidien Fähigkeiten oder Fertigkeiten abgeschätzt werden. Wir haben u. a. eine Sozialformel zur Schätzung der prämorbidien Intelligenz, die biografische Daten verwendet, vorgestellt. Prognosen – verbunden mit Empfehlungen – für die Rückkehr ins Alltags- oder Berufsleben sind wichtige Aufgaben. Viele der eingesetzten diagnostischen Verfahren sind nicht spezifisch für die Neuropsychologie; es gibt nur wenige neuropsychologische Tests. Eine Herausforderung für die neuropsychologische Diagnostik ist eine mögliche Verfälschung von Untersuchungsergebnissen durch Simulation oder Aggravation von Symptomen durch die Patientin oder den Patienten. Es wurden verschiedene Möglichkeiten (u. a. Verwendung spezieller Symptomvalidierungstests) vorgestellt, die helfen können, eine Verfälschung zu erkennen.

Die diagnostischen Fragestellungen in der Rechtspsychologie sind sehr heterogen. Viele Themen betreffen Straftaten oder bereits überführte Straftäterinnen und Straftäter. In Gerichtsprozessen kann psychologische Diagnostik helfen, die Schuldfähigkeit von Personen, die eine Tat begangen haben oder die Glaubhaftigkeit von Zeugenaussagen abzuklären. Für die Beurteilung der Schuldfähigkeit ist etwa zu untersuchen, ob eine geistige Behinderung vorliegt, aufgrund derer die Person das Unrecht ihrer Tat nicht einsehen konnte. Eine andere Frage ist, ob die Steuerungsfähigkeit während der Tat durch eine tiefgreifende Bewusstseinsstörung eingeschränkt oder sogar aufgehoben war. Zur Glaubhaftigkeit von Zeugenaussagen liegt mit den sog. „Realkennzeichen“ ein genuin rechtspychologischer Ansatz vor, der über eine Inhaltsanalyse von Aussagen prüft, ob diese vermutlich auf eigenen Erlebnissen beruhen. Auch die Person der Zeugin bzw. des Zeugen und die Befragungsumstände können Gegenstand der Diagnostik sein. Finden sich Motive für eine Falschaussage, ist die Person kognitiv in der Lage, solche Beobachtungen zu machen, richtig einzuordnen und korrekt zu erinnern? Können die Aussagen durch eine suggestive Befragung beeinflusst worden sein? Im Strafvollzug dient Diagnostik u. a. der Klärung der Frage, ob eine Vollzugslockerung angebracht oder mit welchen Risiken eine vorzeitige Entlassung aus der Haft verbunden ist. In diesem Fall wird auch eine Kriminalprognose vorgenommen, für die spezielle diagnostische Verfahren wie das HCR-20 zur Verfügung stehen, die empirische belegte Zusammenhänge zwischen bestimmten Merkmalen des Täters oder der Täterin und der Häufigkeit von Rückfällen nutzen. Solche nomothetische Prognosen, auch empirisch aktuarischer Ansatz genannt, funktionieren in der Regel gut, wie umfangreiche Forschungsergebnisse belegen. Sie haben aber den Nachteil, dass zumeist nur empirische Indikatoren und nicht Ursachen erfasst werden. Deshalb ist es angebracht, zusätzlich einen ideografischen Ansatz zu verfolgen. Damit wird versucht, dem Einzelfall gerecht zu werden und ein individuelles Erklärungsmodell für eine Tat und das Verhalten nach der Entlassung zu entwickeln. Dabei werden auch Informationen genutzt, die im aktuarischen Ansatz fehlen, da sie zu selten vorkommen. Die große Herausforderung besteht darin, das empirisch starke nomothetische mit dem empirisch eher schwachen idiografischen Erklärungsmodell in Einklang zu bringen.

Im zivilrechtlichen Bereich spielen Scheidungsverfahren mit den sich anschließenden Fragen, wie das Sorge- und Umgangsrecht zum Wohle der Kinder am besten geregelt wird, eine große Rolle. Die Qualität von Sachverständigengutachten zu familiengerichtlichen Fragen war lange Zeit problematisch. Die Sachkunde war unzureichend definiert, und selbst Psychologinnen und Psychologen haben Gutachten abgegeben, die einer Studie zufolge oft noch viele weitere Mängel aufwiesen – beispielsweise die Verwendung psychometrisch unzulänglicher Verfahren. Der Gesetzgeber hat nun in diesem Bereich die Anforderungen

an die Personen, die ein Gutachten erstellen dürfen, erhöht. Und eine Arbeitsgruppe hat Mindestanforderungen an die Qualität von Sachverständigengutachten im Kindschaftsrecht festgelegt. Beide Maßnahmen werden der Qualität der Diagnostik im familienrechtlichen Bereich zugutekommen.

Die verkehrpsychologische Diagnostik dient dazu, die Anzahl der Unfälle im Straßenverkehr zu reduzieren. Dazu befasst sie sich anlassbezogen mit dem Risikofaktor Mensch, genauer gesagt mit dessen Fahreignung. Die rechtliche Grundlage dafür ist die „Verordnung über die Zulassung von Personen zum Straßenverkehr“ (kurz Fahrerlaubnis-Verordnung). Jährlich werden rund 90.000 Begutachtungen vorgenommen. Hauptanlass sind Auffälligkeiten im Straßenverkehr, die durch Alkohol-, Betäubungsmittel- und/oder Medikamentenkonsum bedingt sind. Alkoholkonsum stellt dabei die größte Untergruppe dar. Hier ist zu unterscheiden, ob ein Verdacht auf eine Alkoholabhängigkeit im klinischen Sinne oder auf Alkoholmissbrauch vorliegt. Menschen mit Alkoholabhängigkeit dürfen gar nicht am Straßenverkehr teilnehmen, solange sie nicht wieder gesund sind. Bei Verdacht auf Alkoholmissbrauch ist zu prüfen, ob die betroffene Person wieder zuverlässig Alkoholkonsum und Teilnahme am Straßenverkehr trennen kann. Bei beiden Fragestellungen kommt dem diagnostischen Interview eine große Bedeutung zu. Weil viele negative Gutachten vor Gericht angefochten werden, ist es oftmals entscheidend, was die Klientin oder der Klient genau gesagt hat. Die Frage, wie das Gespräch dokumentiert werden soll, wird noch immer kontrovers diskutiert. Ein Ton- oder Videoaufnahme des Gesprächs, mit der im Zweifelsfall geklärt werden könnte, ob eine Aussage im Gutachten begründet ist oder nicht, ist noch immer nicht vorgeschrieben. Besonders bei Verdacht auf eine Alkoholabhängigkeit sind auch medizinische Befunde (u. a. alkoholbedingte körperliche Erkrankungen, Alkoholmarker im Urin) wichtig. Manchmal werden auch Leistungstests eingesetzt, um eventuell vorhandene alkoholbedingte kognitive Defizite nachzuweisen.

Ein anderer Anlass für eine medizinisch-psychologische Untersuchung ist die Überprüfung der Fahreignung von Personen, die eine Fahrerlaubnis zur Fahrgastbeförderung beantragen. Laut Fahrerlaubnis-Verordnung muss überprüft werden, ob die Belastbarkeit, die Orientierungsleistung, die Konzentrationsleistung, die Aufmerksamkeitsleistung und die Reaktionsfähigkeit ausreichend hoch sind. Konkret wird in einschlägigen Tests mindestens ein Prozentrang von 16 gefordert – und zwar in allen Bereichen. Diese Forderungen sind insbesondere aus 2 Gründen problematisch. Dass Belastbarkeit und Orientierungsleistung relevante Anforderungen sind, ist schwer zu begründen. Für die eingesetzten Tests wird nur der Nachweis verlangt, dass sie die Konstrukte erfassen; ein Nachweis, dass die Testwerte mit der Fahrsicherheit (z. B. Unfallhäufigkeit) zusammenhängen, ist nicht erforderlich. Die Eignung der Tests, insbesondere ihre psychometrische Qualität (außer der Kriteriumsvalidität) muss von einer amtlich anerkannten unabhängigen Stelle festgestellt werden, die dazu von Expertinnen und Experten nach festgelegtem Standard erstellte „Bestätigungsgutachten“ einholt.

Die verkehrpsychologische Diagnostik wird stark durch Vorgaben des Gesetzgebers geprägt, die sich im Laufe der Zeit auch ändern können. Sehr hilfreich sind dabei Begutachtungsleitlinien zur Kraftfahreignung und Kommentare zu diesen Leitlinien.

?

Übungsfragen

— Abschn. 9.1:

- Welche Funktionsbereiche können bei einer Hirnschädigung betroffen sein?
- Wozu dient die Quantifizierung von neuropsychologischen Funktionsbeeinträchtigungen?

- Wie lässt sich Verfälschung bei einer neuropsychologischen Untersuchung erkennen?
- Was versteht man unter „Neglect“, und wie äußert sich diese Störung?
- **Abschn. 9.2:**
 - Wozu dienen Realkennzeichen? Nennen Sie 3 Beispiele für Realkennzeichen!
 - Wie lauten nach § 20 StGB die Kriterien für Schuldunfähigkeit?
 - Wann liegt verminderte Schuldfähigkeit vor?
 - Warum sind Prognosen über künftiges delinquentes Verhalten schwer zu stellen?
 - Welche 2 Ansätze werden bei der Kriminalprognose verwendet, und was zeichnet diese Ansätze aus?
 - Nennen Sie 3 mögliche psychologische Fragestellungen in Sorgerechtsentscheidungen!
- **Abschn. 9.3:**
 - Welche geistigen Anforderungen werden für eine Fahrerlaubnis zur Fahrgastbeförderung verlangt?
 - Anhand welcher Kriterien kann nach den Begutachtungsleitlinien zur Kraftfahreignung beurteilt werden, dass ein Alkoholmissbrauch abgestellt wurde? Nennen Sie 3 Kriterien!
 - Warum ist in der verkehrspychologischen Diagnostik oft mit Verfälschung zu rechnen, und wie kann man damit umgehen?
 - Warum soll eine Aufzeichnung der Exploration im Rahmen einer medizinisch-psychologischen Untersuchung angefertigt werden, und welche Argumente sprechen für und gegen eine Tonaufnahme?

Literatur

- Albrecht, M., Evers, C., Klipp, S., & Schulze, H. (2015). Projektgruppe MPU-Reform: Schlussbericht. *Berichte der Bundesanstalt für Straßenwesen: Mensch und Sicherheit, Heft M 257*. Bremen: Carl Schünemann.
- Arbeitsgruppe Familienrechtliche Gutachten. (2019). *Mindestanforderungen an die Qualität von Sachverständigengutachten im Kindshaftrecht* (2. Aufl.). Berlin: Deutscher Psychologen Verlag.
- Arthur, W., Barret, G. V., & Alexander, R. A. (1991). Prediction of vehicular accident involvement: A meta-analysis. *Human Performance* 4, 89–105.
- Benton-Sivan, A., & Spreen, O. (2009). *Der Benton-Test* (8. Aufl.). Bern: Huber.
- Blaskewitz, N., Merten, T., & Kathmann, N. (2008). Performance of children on symptom validity tests: TOMM, MSVT, and FIT. *Archives of Clinical Neuropsychology* 23, 379–391.
- Boetticher, A., Kröber, H.-L., Müller-Isberner, R., Böhm, K. M., Müller-Metz, R., & Wolf, T. (2007). Mindestanforderungen für Prognosegutachten. *Forensische Psychiatrie, Psychologie, Kriminologie* 1, 90–100.
- Bundesanstalt für Straßenwesen. (2017). Begutachtung der Fahreignung 2016. Stand: Juli 2017.
► https://www.bast.de/BAST_2017/DE/Presse/Downloads/2017-10-langfassung-pressemittelung.pdf;jsessionid=14E767B749FA32D1D7AFB71CE4611115.live21304?__blob=publicationFile&v=4. Zugegriffen: 27. Mai 2020.
- Bundesanstalt für Straßenwesen. (2018). *Begutachtungsleitlinien zur Kraftfahreignung, Stand 24. Mai 2018*. Bergisch Gladbach: Bundesanstalt für Straßenwesen.
- Bundesanstalt für Straßenwesen. (2019). Fachliche Hinweise zur Begutachtung der Eignung von Testverfahren und -geräten sowie von Kursen zur Wiederherstellung der Kraftfahreignung.
► https://www.bast.de/BAST_2017/DE/Verkehrssicherheit/Qualitaetsbewertung/Anerkennung/u3-anerkennung/Hinweise.html. Zugegriffen: 27. Mai 2020.
- Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds C. R., et al. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, 20, 419–426.
- Carone, D. A. (2009). Test review of the Medical Symptom Validity Test. *Applied Neuropsychology* 16, 309–311.
- Dahle, K.-P. (2000). Psychologische Begutachtung zur Kriminalprognose. In H.-L. Kröber, & M. Steller (Hrsg.), *Psychologische Begutachtung im Strafverfahren: Indikationen, Methoden und Qualitätsstandards* (S. 77–111). Darmstadt: Steinkopff.

- Dahle, K.-P. (2007). Methodische Grundlagen der Kriminalprognose. *Forensische Psychiatrie, Psychologie, Kriminologie* 1, 101–110.
- Dahle, K.-P. (2009). Kriminal(rückfall)prognose. In R. Volbert, & M. Steller (Hrsg.), *Handbuch der Rechtspsychologie* (S. 444–452). Göttingen: Hogrefe.
- Dahle, K.-P., Schneider, V., & Ziethen, F. (2007). Standardisierte Instrumente zur Kriminalprognose. *Forensische Psychiatrie, Psychologie, Kriminologie* 1, 15–26.
- Diagnostik- und Testkuratorium. (2018). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 03. Jan. 2018. *Psychologische Rundschau* 69, 109–116.
- Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-clinical-risk management-20, version 3 (HCR-20V3): Development and overview. *International Journal of Forensic Mental Health* 13, 93–108.
- Erzberger, C. S., & Engel, R. R. (2010). Zur Äquivalenz der Normen des Wechsler-Intelligenztests für Erwachsene (WIE) mit denen des Hamburg-Wechsler-Intelligenztests für Erwachsene – Revision (HAWIE-R). *Zeitschrift für Neuropsychologie* 21, 25–37.
- Fels, M., & Geissner, E. (1997). *NET: Neglect-Test. Ein Verfahren zur Erfassung visueller Neglektphänomene* (2. Aufl.). Göttingen: Hogrefe.
- Geiger, H. (2010). Die medizinisch-psychologische Untersuchung: Untersuchungsanlässe, inhaltliche Anforderungen, Reformansätze. In Deutsche Akademie für Verkehrswissenschaft (Hrsg.), *Tagungsband zum 48. Deutschen Verkehrsgerichtstag 2010* (S. 203–214). Köln: Luchterhand.
- Gesellschaft für Neuropsychologie, Gauggel, S., & Sturm, W. (2005). Leitlinien der Gesellschaft für Neuropsychologie (GNP) für neuropsychologische Diagnostik und Therapie: Stand: November 2005. *Zeitschrift für Neuropsychologie* 16, 175–199.
- Gesellschaft für Neuropsychologie, Neumann-Zielke, L., Riepe, J., Roschmann, R., Schötzau-Fürwentsches, P., & Wilhelm, H. (2009). Leitlinie „Neuropsychologische Begutachtung“. *Zeitschrift für Neuropsychologie* 20, 69–83.
- Gray, N. S., Taylor, J., & Snowden, R. J. (2008). Predicting violent reconvictions using the HCR-20. *The British Journal of Psychiatry* 192, 384–387.
- Gretzkord, L. (2002). Prognose im Maßregelvollzug (§ 63 StGB) – wie lassen sich die Ergebnisse von Rückfallstudien nutzen? In T. Fabian, G. Jacobs, S. Nowara, & I. Rode (Hrsg.), *Qualitätssicherung in der Rechtspsychologie* (S. 347–360). Münster: LIT-Verlag.
- Haffner, H.-T., Brenner-Hartmann, J., & Musshoff, F. (2018). Abhängigkeit: Kommentar. In W. Schubert, M. Huetten, C. Reimann, M. Graw, W. Schneider, & E. Stephan (Hrsg.), *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl., S. 280–300). Bonn: Kirschbaum.
- Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment* 21, 1–21.
- Hartje, W. (2004). *Neuropsychologische Begutachtung*. Göttingen: Hogrefe.
- Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference, P. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist* 23, 1093–1129.
- Hessler, J., Jahn, T., Kurz, A., & Bickel, H. (2013). The MWT-B as an estimator of premorbid intelligence in MCI and dementia. *Zeitschrift für Neuropsychologie* 24, 129–136.
- Hillmann, F.-R. (2010). „Idiotentest“ auf dem Prüfstand. In Deutsche Akademie für Verkehrswissenschaft (Hrsg.), *Tagungsband zum 48. Deutschen Verkehrsgerichtstag 2010* (S. 215–224). Köln: Luchterhand.
- Howells, J. G., & Lickorish, J. R. (2017). *FBT: Familien-Beziehungs-Test* (8. Aufl.). München: Ernst Reinhardt.
- Huber, W., Poeck, K., Weniger, D., & K. Willmes, K. (1983). *AAT: Aachener Aphasia Test*. Göttingen: Hogrefe.
- Iverson, G. L. (2010). Detecting exaggeration, poor effort, and malingering in neuropsychology. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (2nd ed., pp. 91–135). New York, NY: Springer.
- Jahn, T., Beitlich, D., Hepp, S., Knecht, R., Koehler, K., Ortner, C., Sperger, E., et al. (2013). Drei Sozialformeln zur Schätzung der (prämorbid) Intelligenzquotienten nach Wechsler. *Zeitschrift für Neuropsychologie* 24, 7–24.
- Karnath, H.-O. (2009). Vernachlässigung – Neglect. In W. Sturm, M. Herrmann, & T. F. Münte (Hrsg.), *Lehrbuch der Klinischen Neuropsychologie: Grundlagen, Methoden, Diagnostik, Therapie* (2. Aufl., S. 444–452). Heidelberg: Spektrum.
- Kötter, S., von Franqué, F., Bolzmacher, M., Eucker, S., Holzinger, B., & Müller-Isberner, R. (2014). The HCR-20V3 in Germany. *International Journal of Forensic Mental Health* 13, 122–129.

Diagnostik in weiteren Anwendungsfeldern

- Kranich, U. (2018). Anforderungen an die psychische Leistungsfähigkeit: Kommentar. In W. Schubert, M. Huetten, C. Reimann, M. Graw, W. Schneider, & E. Stephan (Hrsg.), *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl., S. 75–82). Bonn: Kirschbaum.
- Kubinger, K. D., & Ortner, T. (Hrsg.). (2010). *Psychologische Diagnostik in Fallbeispielen*. Göttingen: Hogrefe.
- Lehrl, S. (2005). *MWT-B: Mehrfachwahl-Wortschatz-Tests Form B* (5. Aufl.). Balingen: Spitta.
- Littmann, E. (2000). Forensische Neuropsychologie – Aufgaben, Anwendungsfelder und Methoden. In H.-L. Kröber & M. Steller (Hrsg.), *Psychologische Gutachten im Strafverfahren: Indikationen, Methoden und Qualitätsstandards* (S. 57–75). Darmstadt: Steinkopff.
- Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory* 12, 361–366.
- Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law* 11, 99–122.
- McCaffrey, R. J., & Vanderslice-Barr, J. L. (2010). Estimation of premorbid IQ. In A. M. Horton, Jr., & L. C. Hartlage (Eds.), *Handbook of forensic neuropsychology* (2nd. ed., pp. 365–379). New York, NY: Springer.
- Monahan, J. (2003). Violence risk assessment. In A. M. Goldstein, & I. B. Weiner (Eds.), *Handbook of psychology: Forensic psychology* (Vol. 11, pp. 527–540). New York: Wiley.
- Müller-Isberner, R., Jöckel, D., & Gonzalez-Cabeza, S. (1998). *HCR-20: Die Vorhersage von Gewalttaten mit dem HCR 20 (Version 2 – D 1)*. Haina: Institut für Forensische Psychiatrie.
- Müller-Isberner, R., Schmidbauer, W. M., & Born, P. (Hrsg.). (2014). *Die Vorhersage von Gewalttaten mit dem HCR-20 V3: Benutzerhandbuch Deutsche Version*. Haina: Institut für Forensische Psychiatrie.
- Neumann-Zielke, L., Bahlo, S., Diebel, A., Riepe, J., Roschmann, R., Schötzau-Fürwentsches, P., Wetzig, L. (2015). Leitlinie „Neuropsychologische Begutachtung“. *Zeitschrift für Neuropsychologie*, 26, 289–306.
- Neumann-Zielke, L., Roschmann, R., & Wilhelm, H. (2009). Neuropsychologische Begutachtung. In W. Sturm, M. Herrmann, & T. F. Münte (Hrsg.), *Lehrbuch der Klinischen Neuropsychologie: Grundlagen, Methoden, Diagnostik, Therapie* (S. 329–340). Heidelberg: Spektrum.
- Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior* 40, 440–457.
- Okulicz-Kozaryn, M., Banse, R., Kluck, M.-L., & Schubert, W. (2015). Dokumentation im Explorationsgespräch, Tonband- und Videomitschnitt. In A. Patermann, W. Schubert, & W. Graw (Hrsg.), *Handbuch des Fahreignungsrechts* (S. 378–400). Bonn: Kirschbaum.
- Poschadel, S., Falkenstein, M., Pappachan, P., Poll, E., & von Hinckeldey, K. W. (2009). Testverfahren zur psychometrischen Leistungsprüfung der Fahreignung. *Berichte der Bundesanstalt für Straßenwesen: Mensch und Sicherheit, Heft 203*. Bremen: Carl Schünemann.
- Reichert, J. (1997). Begutachtung des Erinnerungsvermögens einer Zeugin mit mehrjährigem Drogenmissbrauch – Antje F., 20 Jahre. In K. D. Kubinger, & H. Teichmann (Hrsg.), *Psychologische Diagnostik und Intervention in Fallbeispielen* (S. 121 ff.). Weinheim: Psychologie Verlags Union.
- Rosseger, A., Urbanik, F., Elbert, T., Friesa, D., & Endrass, J. (2010). Rückfälligkeit nach Entlassung aus dem Strafvollzug in der Schweiz: Die Validität des HCR-20. *Schweizer Archiv für Neurologie und Psychiatrie* 161, 254–259.
- Salewski, C., & Stürmer, S. (2015). Qualität familienrechtspychologischer Gutachten. *Zeitschrift für Kindschaftsrecht und Jugendhilfe* 10, 4–9.
- Salzgeber, J. (2015). *Familienpsychologische Gutachten: Rechtliche Vorgaben und sachverständiges Vorgehen* (6. Aufl.). München: Beck.
- Salzgeber, J. (2020). *Familienpsychologische Gutachten: Rechtliche Vorgaben und sachverständiges Vorgehen* (7. Aufl.). München: Beck.
- Schellig, D., Drechsler, R., Heinemann, D., & Sturm, W. (Hrsg.). (2009). *Handbuch neuropsychologischer Testverfahren: Aufmerksamkeit, Gedächtnis, exekutive Funktionen* (Bd. 1). Göttingen: Hogrefe.
- Schellig, D., Heinemann, D., Schächtele, B., & Sturm, W. (Hrsg.). (2018). *Handbuch neuropsychologischer Testverfahren* (Bd. 2). Göttingen: Hogrefe.
- Schellig, D., Heinemann, D., Schächtele, B., & Sturm, W. (Hrsg.). (2019). *Handbuch neuropsychologischer Testverfahren* (Bd. 3). Göttingen: Hogrefe.
- Schmidt-Atzert, L. (2010). Die medizinisch-psychologische Untersuchung aus Sicht der wissenschaftlich fundierten Psychologischen Diagnostik. In Deutsche Akademie für Verkehrswissenschaft (Hrsg.), *Tagungsband zum 48. Deutschen Verkehrsgerichtstag 2010* (S. 258–270). Köln: Luchterhand.

- Schmidt-Atzert, L., Stohbeck-Kühner, P., & Schubert, W. (2018). Anforderungen an die psychische Leistungsfähigkeit: Kommentar. In W. Schubert, M. Huetten, C. Reimann, M. Graw, W. Schneider, & E. Stephan (Hrsg.), *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl., S. 48–74). Bonn: Kirschbaum.
- Scholz, O. B., & Schmidt, A. F. (2008). Schuldfähigkeit. In R. Volbert, & M. Steller (Hrsg.), *Handbuch der Rechtspsychologie* (S. 401–411). Göttingen: Hogrefe.
- Schubert, W. (2010). „Die Medizinisch-Psychologische Untersuchung“ auf dem Prüfstand. In Deutsche Akademie für Verkehrswissenschaft (Hrsg.), *Tagungsband zum 48. Deutschen Verkehrsgerichtstag 2010* (S. 225–257). Köln: Luchterhand.
- Schubert, W., & Mattern, R. (Hrsg.). (2009). *Urteilsbildung in der Medizinisch-Psychologischen Fahreignungsdiagnostik: Beurteilungskriterien* (2. Aufl.). Bonn: Kirschbaum.
- Schubert, W., Huetten, M., Reimann, C., Graw, M., Schneider, W., & Stephan, E. (Hrsg.). (2018). *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl.). Bonn: Kirschbaum.
- Slick, D. C., Tan, J. T., Sherman, E. M. S., & Strauss, E. (2011). Malingering and related conditions in pediatric populations. In A. S. Davis (Ed.), *The handbook of pediatric neuropsychology* (S. 457–470). New York, NY: Springer.
- Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: A meta-analytic update and extension. *Archives of Clinical Neuropsychology* 26, 774–789.
- Statista GmbH. (2020). Polizeiliche Aufklärungsquote von Straftaten (insgesamt) in Deutschland von 1993 bis 2019. Veröffentlichungsdatum: März 2020 ► <https://de.statista.com/statistik/daten/studie/2303/umfrage/entwicklung-der-aufklaerungsquote-von-schafftaten-seit-1989/>. Zugegriffen: 26. Mai 2020.
- Statistisches Bundesamt. (2018). Deutlich weniger Ehescheidungen im Jahr 2017. Pressemitteilung Nr. 251 vom 10. Juli 2018. ► https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/07/PD18_251_12631.html. Zugegriffen: 26. Mai 2020.
- Statistisches Bundesamt. (2019). Verkehrsunfälle: Unfälle von Frauen und Männern im Straßenverkehr 2018. Erschienen am 9. Dezember 2019. ► https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/unfaelle-frauen-maenner-5462407187004.pdf?__blob=publicationFile. Zugegriffen: 27. Mai 2020.
- Steller, M., & Volbert, R. (1997). Glaubwürdigkeitsbegutachtung. In M. Steller, & R. Volbert (Hrsg.), *Psychologie im Strafverfahren: Ein Handbuch* (S. 12–39). Bern: Huber.
- Stephan, E., & Brenner-Hartmann, J. (2018). Alkoholmissbrauch: Kommentar. In W. Schubert, M. Huetten, C. Reimann, M. Graw, W. Schneider, & E. Stephan (Hrsg.), *Begutachtungsleitlinien zur Kraftfahreignung: Kommentar* (3. Aufl., S. 245–278). Bonn: Kirschbaum.
- Sturm, W. (2000). Aufgaben und Strategien neuropsychologischer Diagnostik. In W. Sturm, M. Herrmann, & C. W. Wallesch (Hrsg.), *Lehrbuch der Klinischen Neuropsychologie: Grundlagen, Methoden, Diagnostik, Therapie* (S. 265–276). Lisse, NL: Swets & Zeitlinger.
- Sturm, W., Herrmann, M., & Münte, T. F. (2009). *Lehrbuch der klinischen Neuropsychologie: Grundlagen, Methoden, Diagnostik, Therapie*. Heidelberg: Spektrum Akademischer Verlag.
- Suesse, M., Wong, V. W. C., Stamper, L. L., Carpenter, K. N., & Scott, R. B. (2015). Evaluating the clinical utility of the Medical Symptom Validity Test (MSVT): A clinical series. *The Clinical Neuropsychologist* 29, 214–231.
- Sweet, J. J., & Breting, L. M. G. (2013). Symptom validity test research: Status and clinical implications. *Journal of Experimental Psychopathology* 4, 6–19.
- TransMIT-Zentrum für wissenschaftlich-psychologische Dienstleistungen (DGPs). (2020). Fahrerignung: Evaluation psychologischer Testverfahren und -geräte. ► <https://zwpd.transmit.de/zwpd-dienstleistungen/fahreignung>. Zugegriffen: 27. Mai 2020.
- Volbert, R. (2000). Standards der psychologischen Glaubhaftigkeitsdiagnostik. In H.-L. Kröber, & M. Steller (Hrsg.), *Psychologische Begutachtung im Strafverfahren – Indikationen und Qualitätsstandards* (S. 113–145). Darmstadt: Steinkopff.
- Volbert, R., & Dahle, K.-P. (2010). *Forensisch-psychologische Diagnostik im Strafverfahren*. Göttingen: Hogrefe.
- Volbert, R., & Steller, M. (2014). Is this testimony truthful, fabricated, or based on false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist* 19, 207–220.
- Volbert, R., & Steller, M. (Hrsg.). (2008). *Handbuch der Rechtspsychologie*. Göttingen: Hogrefe.
- Vollrath, M., & Krems, J. F. (2011). *Verkehrsprychologie: Ein Lehrbuch für Psychologen, Ingenieure und Informatiker*. Stuttgart: Kohlhammer.
- Wallesch, C.-W. & Herrmann, M. (2000). Klinische Neurologie. In W. Sturm, M. Herrmann & C.-W. Wallesch (Hrsg.), *Lehrbuch der Klinischen Neuropsychologie* (S. 96–125). Lisse, NL: Swets & Zeitlinger.
- Walters, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice and Behavior* 33, 279–304.

Diagnostik in weiteren Anwendungsfeldern

- Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). Assessing risk of violence to others. In C. D. Webster, & M. A. Jackson (Eds.), *Impulsivity: Theory, assessment, and treatment* (pp. 251–277). New York, NY: Guilford Press.
- Wechsler, D. (2012). *WMS-IV: Wechsler Memory Scale – Fourth Edition. Deutsche Bearbeitung A. C. Lepach und F. Petermann*. Frankfurt: Pearson Clinical & Talent Assessment.
- Wells, G. L., & Olsen, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology* 54, 277–295.
- Zimmermann, P., & Fimm, B. (2017). *Testbatterie zur Aufmerksamkeitsprüfung (TAP), Version 2.3.1*. Herzogenrath: Psytest.
- Zimmermann, P., Gondan, M., & Fimm, B. (2002). *Testbatterie zur Aufmerksamkeitsprüfung für Kinder (KITAP)*. Herzogenrath: Psytest.

Serviceteil

Stichwortverzeichnis – 785

Stichwortverzeichnis

360-Grad-Feedback 616

A

Aachener Aphasie Test 737

Abbruchregel 264

Abwehrmechanismus 382

Abweichungsnorm 182, 183

Adaptives Intelligenz Diagnostikum (AID) 267

– AID 3 267

Advanced Progressive Matrices (APM) 283

Affiliationsmotiv 392, 397

Aggravation 740

Akquieszenz 106

Akzept!-Fragebogen 605

Akzeptanz 191, 258, 490, 584, 603

Alcohol, Smoking and Substance Involvement Screening (WHO ASSIST) 710

Alertness 229, 287

Alkoholabhängigkeit 710, 764, 770

Alkoholmissbrauch 445, 710, 764, 770

Alkoholproblematik 763, 764, 770

Allgemeine Depressionsskala (ADS) 710, 711

Allgemeiner Interessen-Struktur-Test (AIST) mit Umwelt-Struktur-Test (UST-3) – Version 3 (AIST-3) 357, 579

Alltagsintelligenz, praktische 286

AMDP-System 708

American College Test (ACT) 655

Analogskala, visuelle 106

Analyse

– empirische 88

– operationale 614

– statistische 532

Anamnese 426, 735

– biografische 696

Änderungssensitivität 365

Anforderungsanalyse 86, 444, 445, 574, 580, 619, 621, 659

Anforderungsdimension 574, 588

Anforderungsmerkmal 431, 488, 578, 580

Anforderungsmodul 348

Anforderungsprofil 348, 488, 580

Angst 363, 710

Angstfragebogen für Schüler (AFS) 711

Angstkonzept 363

Ängstlichkeit 362, 363

Angstmessung 362

Angstreduktion 223

Annahmequote 20

Anorexia-Nervosa-Inventar zur Selbstbeurteilung (ANIS) 711

Ansatz

– dimensionaler 705

– interaktionistischer 14

– kategorialer 704

– konstruktorientierter 586

– psycholexikalischer 306

– systemischer 716

Anschlussmotiv 396

Ant Colony Optimization 131

Antwortalternative 102, 597, 600

Antwortenverrechnung 431

Antwortformat 97

– dichotomes 55, 57

– eingeschränkt freies 101

– freies 97, 101

– gebundenes 99

– geschlossenes 599

– nominales 56

– offenes 98

– ordinales 56, 77

Antwortinkonsistenz 331, 332

Antwortkategorie 78, 447

Antwortprozess 161

Antwortstil 176

Antworttendenz 422, 423

Antwortverzögerung 379

Antwortzeitbegrenzung 319

Anwendungsbereich 46, 96

Anwendungsbereich 4

Aphasie 733

Äquidistanzmodell 80

Äquivalentnorm 182

Äquivalenz 49, 141

Äquivalenzmethode 148

Arbeitsanalyse 573, 619

Arbeitsauftrag 487

Arbeitsbedingungen 502

Arbeitsbezogenes Verhaltens- und Erlebensmuster (AVEM) 80, 305

Arbeitsgestaltung 626, 627

Arbeitshaltung 298

Arbeitshaltungen – Kurze Testbatterie 369, 374

– Auswertung 370

– Bewertung 371

– Durchführung 370

– Gliederung 369

– Normierung 370

– Reliabilität 370

– Validität 370

Arbeitsklima 573

Arbeitskontext 626

Arbeitsplatzmerkmal 578

Arbeitsprobe 582, 622

Arbeitsumgebung 355

Arbeits- und Organisationspsychologie 8, 568

Army Alpha and Beta Examination 23

Assessment, ambulantes 24, 405

Assessment-Center 8, 582, 586, 587, 608, 622

– Dimension 588, 592

– Konstruktvalidität 595

– Kriteriumsreliabilität 591

– Methode 587

– Standardisierung 595

– Standards 587, 590

– Übung 42, 400

– Varianzaufklärung 591

– Varianzeinschränkung 591

Asymmetrie 174

– hybride 175

– vollständige 174

Aufforderungscharakter 383

Aufklärung 497, 501

Aufmerksamkeit 226, 228, 239, 769

– geteilte 230

– konzentrierte 230, 239

– nicht willentliche 228

– selektive 230

– willentliche 228

Aufmerksamkeitsform 229

Aufmerksamkeitsfunktion 228

Aufmerksamkeitstest 170, 225, 228

Auftrag

– formeller 484

– informeller 483

Auftragsannahme 483, 485

Auftragsklärung 17, 485, 590

Augenscheininvalidität 587, 605

Ausbildungserfolg 622

Ausbildungswahl 609, 624

Auskunftsrecht 500

Auslassungsfehler 240

Auspartialisieren 320

Aussagetüchtigkeit 747

Auswahlgespräch 427, 441, 654

Auswahlkriterium 490

Auswahlrichtlinien 33

Auswahlsituation 604

Auswahltest 223

– unqualifizierter 483

Auswertung 447

Auswertungsfehler 504

Auswertungsobjektivität 135

Auswertungsstandardisierung 430

Auswertungsbereinstimmung 436

B

BACO Belastbarkeits-Assessment 374

Barona demographic regression equation 738

Basisrate 534, 550, 551, 612

Baumtest 383

Bayes-Theorem 548

Bayley Scales of Infant and Toddler Development (Bayley-III) 291

– Screening-Test 291

Bearbeitungsgeschwindigkeit 233, 278

Bearbeitungszeit 256, 379

Beck Angst-Inventar (BAI) 710

Beck Depressions-Inventar (BDI) 708

– Revision (BDI-II) 129, 710, 711

Beck Inventar für Kognitive Schemata (B-IKS) 724

Bedeutsamkeit, klinische 722

Bedingungsmodell, funktionales 697

Befangenheit 486

Befindensmaß, mehrdimensionales 364

Befindlichkeit 361

Befund 512
 Befundbogen 513
 Begabung 668
 Begutachtung
 – medizinisch-psychologische 763
 – neuropsychologische 736
 – von Zeugenaussagen 747
 Begutachtungsleitlinien zur Kraftfahreignung 765, 766
 Begutachtungsprozess 479, 512, 515
 – Ablaufschema 479
 Behandlung 716
 – psychotherapeutische 690
 – transdiagnostische 719
 Behandlungsleitlinien 718
 Behandlungsplanung 719
 Behinderung, geistige 751
 Belastbarkeit der Testperson 497
 Belastungstest 239
 Benton-Test 225, 736
 Beobachtendenkonferenz 589
 Beobachtendentraining 424
 Beobachtendenübereinstimmung 421
 Beobachterdrift 423
 Beobachtung 575
 – strukturierte 712
 – teilnehmende 404
 Beobachtungsbogen für 3- bis 6-jährige Kinder (BAV) 711
 Beobachtungs-Checkliste 408
 Beobachtungsergebnis 421
 Beobachtungsfehler 747
 Beobachtungsmethode 3, 692, 712
 Beobachtungsprotokoll 401
 Beobachtungsprotokollbogen 410
 Beobachtungsverfahren 295, 419
 Beratung, begabungsdiagnostische 671
 Berger-Skala zur Erfassung der Selbstzeptanz 715
 Berliner Intelligenzstrukturmodell (BIS-Modell) 239, 277, 286
 Berliner Intelligenzstruktur-Test (BIS-Test) 250
 – Form 4 (BIS-4) 277
 – für Jugendliche – Begabungs- und Hochbegabungsdiagnostik (BIS-HB) 278
 Berner Inventar für Therapieziele (BIT) 724
 Berner Ressourceninventar (BRI) 724
 Berner Therapeuten- und Patientenstundebogen 2000 (BTSTB/BPSTB) 724
 Bertillon'sche Klassifikation der Todesursachen (BCCD) 23
 Berufsanforderung 357
 Berufsauswahl 609, 624
 Berufsberatung 611
 Berufsbild 581
 Berufseignungsdiagnostik 443
 Berufserfolg 392, 622
 Berufserfolgskriterium 585
 Berufsqualifikation 583
 – gesetzliche Anforderungen 482
 Berufsregister 356
 Berufsverband Deutscher Psychologinnen und Psychologen e. V. (BDP) 33
 Berufswahl 355, 581, 609
 Berufswahltheorie von John Holland 355

Berufswissenstest 622
 Bestätigungsgutachten 768
 Bestätigungsverfahren 768
 Beurteilendentraining 424
 Beurteilungsaufgabe 101, 108
 Beurteilungsfehler 311, 432, 440
 Beurteilungsgrundsätze 33
 Beurteilungsskala 77, 104
 Beurteilungsbereinstimmung 579
 Beurteilungsverfahren 419
 Bewältigungsstrategie 305
 Bewegungssensor 406
 Bewerber-Screening 494
 Bewerbungsunterlagen 584
 Bewusstseinsstörung, tiefgreifende 751
 Bezugsrahmen 623
 Big Five 630
 – Inventory-10 (BFI-10) 353
 – Modell 306, 349, 353, 360
 – Persönlichkeitsfragebogen 180
 – Persönlichkeitsmerkmal 378
 Bildungsbereich 673
 – schulischer 644
 – tertärer 654
 Bildungssystemevaluation 677
 Bildungstest 681
 Binet, Alfred 23
 Biografie 434
 – berufliche 436
 Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) 322, 341
 – 6 Faktoren (BIP-6F) 346
 – Anforderungsmodul (BIP-6F-AM) 580
 – Anforderungsmodul (BIP-AM) 580
 – Auswertung 345
 – Bewertung 345
 – Durchführung 345
 – Fremdbeurteilungsversion 347
 – Gliederung 343
 – Gütekriterien 345
 – Hintergrund 342
 – Interpretation 345
 – Konstruktionsprinzip 342
 – Normen 345
 Bochumer Matrizentest (BOMAT)
 – advanced 284
 – advanced – short version 284
 – Standard 283
 Bochumer Veränderungsbogen-2 720
 Borderline-Störung 710
 Borderline-Symptom-Liste (BSL) 710
 Brief Psychiatric Rating Scale 710
 Brief-Symptom-Checklist (BSCL) 710
 Brodgen-Cronbach-Gleser-Nutzenmodell 559
 Broken leg cue 534
 Brunswik'sche Linsengleichung 418
 Brunswik'sches Linsenmodell 416

C

Cambridge Contextual Reading Test (CCRT) 739

Carroll-Horn-Cattell-Modell (CHC-Modell) 270, 285, 296
 Carroll-Modell 276, 285
 Checklistenverfahren 707
 Child Behavior Checklist (CBCL) 711
 Children's Yale-Brown Obsessive Compulsive Scale (CY-BOCS) 711
 Clinical Global Impression (CGI) 710
 Coaching 218
 Cognitron 248
 Cohens Kappa 420
 Coloured Progressive Matrices (CPM) 282
 Composite International Diagnostic Interview (CIDI) 706, 707
 Comprehensive System 387
 Computerauswertungsprogramm 24
 Computerized Assessment of Response Bias (CARB) 742
 Computertest 495
 Conditional-Maximum-Likelihood-Methode 72
 Continuous-Norming-Ansatz 299
 Critical-Incident-Interview 620
 Critical-Incident-Technik (CIT) 577
 Cronbachs Alpha 130, 144
 Cut-off-Wert 194, 539, 541
 C-Wert 185

D

Dark Triad of Personality at Work (TOP) 307
 Daten, personenbezogene 498, 501
 Datenintegration 590
 Datenquellen, objektive 8
 Datenquellendiskrepanz 741
 Datenschutz 498
 Datenschutzgrundverordnung (DS-GVO) 30, 498
 Daueraufmerksamkeit 230
 Day Reconstruction Method (DRM) 405
 Defizitkompensation 738
 DEMAT s. Deutscher Mathematiktest
 Denken, schlussfolgerndes 272, 273
 Depressionsinventar für Kinder und Jugendliche (DIKJ) 711
 Depressionstest für Kinder – II (DTK-II) 711
 Depressivität 710
 Deutsche Gesellschaft für Psychologie e. V. (DGPs) 33
 Deutsche Personality Research Form (PRF) 358
 Deutscher Mathematiktest (DEMAT) 677
 – DEMAT 3 676
 Deutscher Rechtschreibtest (DERET) 675
 Development-Center 614
 Diagnose, falsch positive 743
 Diagnosekriterium 433
 Diagnosesystem 159
 Diagnostik
 – berufsbezogene 277
 – förderrelevante 666
 – klassifikatorische 694
 – klinische 334
 – klinisch-psychologische 690

Stichwortverzeichnis

- kognitiv-verhaltenstherapeutische 696
- medizinische 4, 734
- neuropsychologische 732, 735, 738
- operationale/deskriptive 701
- psychischer Störungen 432
- psychologische 2, 11, 22
- rechtspychologische 746
- verkehrspychologische 33, 762, 771
- Diagnostik-System für psychische Störungen nach ICD-10 und DSM-5 für Kinder und Jugendliche – III (DI-SYPS-III) 711
- Diagnostik- und Testkuratorium 133
- Diagnostik- und Therapieleitlinie 735
- Diagnostischer Rechtschreibtest (DRT) 676
 - DRT 5 676
- Diagnostisches Interview bei psychischen Störungen (DIPS Open Access) 707
 - im Kindes- und Jugendalter (Kinder-DIPS Open Access) 707
- Diagnostisches Kurzinterview bei psychischen Störungen (Mini-DIPS Open Access) 707
- Diagnostisches und Statistisches Manual Psychischer Störungen (DSM) 7, 24, 694, 700, 703
 - DSM-5 7, 13, 701
- Dialog, standardisierter 264
- Differentielle Psychologie 10
- Differentieller Konzentrationstest für Kinder (DKT-K) 235, 248
- Differentielles Leistungsangst Inventar (DAI) 711
- Differenz, kritische 153
- Differenzwert 627
- Diktat 675
- DIN
 - 33430 632
 - Screen 133
- Diskriminationsparameter 73
- Dissimulation 740
- Distraktor 102, 230
- Dokumentation 447
- Dortmunder Entwicklungsscreening für den Kindergarten – Revision (DESK 3-R) 295
- Drei-Spalten-Technik 699
- Drittvariable 176
- DRT s. Diagnostischer Rechtschreibtest
- DSM s. Diagnostisches und Statistisches Manual Psychischer Störungen
- Dunkle Triade 307
- Durchführung, standardisierte 42
- Durchführungsbedingung 518
- Durchführungsobjektivität 134, 421
- Durchführungszeit 257
- Durchstreichtest 235
- Dyskalkulie 663

E

- Eating Attitude Test (EAT) 711
- Eating Disorder Examination (EDE) 434
- Eating Disorder Inventory-2 (EDI-2) 710, 711

- Ecological Momentary Assessment (EMA) 405
- Effektstärke 721
- Ehrengericht 34
- Eichstichprobe 188
- Eigenschaftswörterliste (EWL) 91, 362, 364
 - Auswertung 365
 - Bewertung 366
 - Durchführung 365
 - Gliederung 364
 - Normierung 366
 - Reliabilität 365
 - Validität 365
- Eigenwert 121, 122
- Eigenwerteverlauf 123
- Eignung 584
- Eignungsbeurteilung 482
 - berufsbezogene 632
 - berufliche 342
 - objektive sprachbasierte 381
- Eignungsinterview 432, 435, 441, 626
- Eignungsprüfung 218, 655
- Eignungsprüfungsverfahren, hochschuleigenes 655
- Eignungstest für das Medizin-Studium (EMS) 655
- Eignungsurteil 603, 604
- Eignungszweifel 764
- Eindimensionalität 68
- Einsatzbedingung 97
- Einstellungsgespräch 32, 427, 442, 444
- Einstellungsinterview, multimodales 444
- Einstellungsverfahren 608
- Einwilligung 500
 - informierte 497, 501
 - Widerruf der 500
- Einzelfalldiagnostik 147
- Einzelmessung 553
- Einzeltest 495
- Einzeluntersuchung 255
- Electronically Activated Recorder (EAR) 406
- Elterninterview über Problemsituationen in der Familie (EL-PF) 711
- Eltern-Kind-Interaktion 761
- Emotionen 42
- Empathie 694
- Empfehlung 483
- Entscheidung
 - investigatiorische 21
 - terminale 21
- Entscheidungsbaum 555
- Entscheidungsstrategie
 - disjunktive 537
 - kompensatorische 536, 553
 - konjunktive 538, 553
- Entwicklung
 - altersgerechte 9
 - fröhkindliche 291
 - kognitive 262, 289, 297
 - motorische 301
 - sozial-emotionale 292
- Entwicklungsalter 182, 288, 290, 291
- Entwicklungs-Assessment-Center 614, 626
- Entwicklungsbedarf 613
- Entwicklungsprofil 291–293
- Entwicklungsstörung 702
 - schulische 666
 - umschriebene, schulischer Fertigkeit 661
- Entwicklungstest 288
- 6 Monate bis 6 Jahre – Revision (ET 6-6-R) 295
 - allgemeiner 289
 - spezieller 300
- Entwicklungsverzögerung 288
- Environmental predictability 418
- Erfassungsbogen für aggressives Verhalten in konkreten Situationen (EAS) 713
- Erfolgskontrolle 529, 614, 719
- Erfolgswahrscheinlichkeit 57
- Ergebnisbewertung 512
- Ergebnisdarstellung 508
 - tabellarische 509
- Ergebnisdokumentation 450
- Ergebnisevaluation 558
- Ergebnisinterpretation 505, 512, 518
- Erinnerung, trügerische 315
- Erinnerungseffekt 51
- Erklärungsmodell, individuelles 756
- Erkrankung
 - degenerative 732
 - depressive 332
 - neurologische 733
 - psychische 691, 700
- Erleben 2, 310, 691, 697
- Erscheinungsbild, äußeres 440
- Erwünschtheit, soziale 176, 320
 - Auspartialisieren 320
- Erziehungsfähigkeit 761
- Esstörung 710
- Europäische Menschenrechtskonvention 28
- Evaluation 4, 155, 556, 590, 618
- Evaluationsforschung, nomothetische 721
- Event Reconstruction Method (ERM) 405
- Event Sampling Method 405, 411
- Experience Sampling Method (ESM) 405
- Experiment, wahrnehmungsdiagnostisches 386
- EXPLOJOB 357, 579
- Exploration 426, 695, 735
- EXPLORIX 355, 624
 - Auswertung 356
 - Bewertung 357
 - Durchführung 356
 - Gliederung 355
 - Interpretation 356
 - Normen 356
 - Reliabilität 356
 - Validität 356
- Extraktionskriterium 122, 579
- Extraversion 44
- Extraversionsfaktor 337
- Eysenck-Persönlichkeits-Inventar (EPI) 338

F

- Face-to-Face-Befragung 434
- Fachkompetenz 481
- Fachpsychologe Verkehrspychologie 762
- Fähigkeit 216

- Fähigkeitsparameter 61
 Fähigkeitstest 284
 Fahreignung für den Straßenverkehr 763
 Fahreignungsbeurteilung 9
 Fahreignungsdiagnostik 763
 Fahrerlaubnis-Verordnung 763, 766
 Fahrerlaubnis zur Fahrgastbeförderung 765
 Fahrverhaltensbeobachtung 767
 Fairness 193, 550
 Faking
 – bad 192, 223
 – good 192, 317, 624
 Faktor 119
 – kognitiver 490
 – körperlicher 490
 – orthogonal 122
 Faktorenanalyse 118, 576
 – auf Basis der Skalenwerte 327
 – Definition 118
 – explorative 119, 164, 299
 – konfirmatorische 125, 165, 260, 300
 Faktorladung 120
 Faktorwert 253
 Fall-Crossover-Analyse 412
 Fallstudie 589
 Falschaussage 748
 Familie in Tieren 383, 398
 Familien-Beziehungs-Test (FBT) 761
 Feedback 590
 Fehldiagnose 743
 Fehlentscheidung
 – diagnostische 769
 – messfehlerbedingte 768
 Fehler, logischer 422
 Fehleranalyse 676
 Fehlerprozentwert 242
 Fehlerquelle 422, 504, 621
 Feinnormen 185
 Fertigkeit 216, 581
 – soziale 439, 602
 Fifteen Item Test nach Rey (FIT) 742
 Fit 626
 Forced Choice
 – Antwortformat 109, 319, 624
 – Aufgabe 100, 101
 Förderbedarf, sonderpädagogischer 648
 Förderschule 651
 Förderschwerpunkt 649
 Förderung, sonderpädagogische 648, 651
 Forensische Psychologie 8
 Formdeuteverfahren 383, 386
 Foto-Interessen-Test F-I-T Serie 2020 354
 Frage 445
 – biografische 434, 436
 – geschlossene 452
 – offene 451, 452
 – peinliche 458
 – psychologische 17, 445, 486, 487, 489, 512, 760
 – situative 434, 436
 – suggestive 96, 447, 747
 Fragebogen 42
 – biografischer 622
 – zu Dissoziativen Symptomen (FDS) 710
 – zu Gedanken und Gefühlen (FGG) 710
 – zu körperbezogenen Ängsten, Kognitionen und Vermeidung (AKV) 710
 – zum funktionalen Trinken (FFT) 710
 – zum Gesundheitszustand (SF-36/SF-12) 710
 – zur Arbeitsanalyse (FAA) 575
 – zur frühkindlichen Sprachentwicklung (FRANKIS) 300
 – zur Leistungsmotivation
 – für Schüler der 3. bis 6. Klasse – Revision (FLM 3–6 R) 358
 – für Schüler der 7. bis 13. Klasse (FML 7–13) 358
 – zur Partnerschaftsdiagnostik (FPD) 341
 – zur Psychotherapiemotivation (FMP) 724
 – zur sozialen Unterstützung (F-SozU) 717, 724
 Fragebogenmethode 313
 Fragebogenverfahren 692, 708
 Fragenformulierung 314, 445
 Fragestellung 17, 484, 517
 – Antwort auf die 514
 – diagnostische 7
 – explorative 487
 – globale 485
 – Präzisierung der 486
 Frame of reference 623
 – Training 424
 Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II) 247
 Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2) 246
 Freiburger Persönlichkeitsinventar – revidierte Fassung (FPI-R) 115, 334, 714
 – Auswertung 337
 – Bewertung 340
 – Durchführung 337
 – Gliederung 336
 – Interpretation 337
 – Konstruktionsprinzip 336
 – Normierung 340
 – Reliabilität 337
 – Validität 337
 Fremdanamnese 500
 Fremdbeschreibung 580
 Fremdbeschreibungsinventar 347
 Fremdbeurteilung 48, 309, 310, 321, 349
 Fremdeinschätzung 572
 Fremdeinschätzungsboegen 343
 Fremde-Situations-Test 403
 Frühkindliches Entwicklungsdiagnostikum für Kinder von 0–3 Jahren (FREDI 0–3) 292
 Führungsmotivation 431
 Führungsverhalten 575
 Fünf-Faktoren-Modell der Persönlichkeit 11, 87, 353
- G**
- Galton, Francis 23
 Gedächtnis, autobiografisches 315
 Gedächtnistest 225, 736
 Geheimnisverrat 29
 Generalitätsniveau 175
 Genogramm 716
 Gerechtigkeit
 – distributive 604
 – prozedurale 604
 Gerontopsychologie 9
 Gesamt-IQ 265, 269, 270
 Gesamturteil 19
 Gesellschaft für Neuropsychologie 735
 Gesellungsmotiv 391
 Gesprächsführung 451, 458, 694
 – motivierende 453
 Gesprächsklima 451
 Gesprächspsychotherapie 714
 Gesprächstechnik 451
 Gestaltungsverfahren 383, 398
 Gesundheitspsychologie 7
 Gewissenhaftigkeitsfragebogen 603, 622
 Gießener Beschwerdebogen für Kinder und Jugendliche (GBB-KJ) 711
 Gießen-Test – II (GT-II) 341, 716
 Glaubhaftigkeit von Zeugenaussagen 515, 746, 749
 Glaubhaftigkeitsgutachten 452
 Goal Attainment Scaling (GAS) 720
 Golden Profiler of Personality (GPOP) 83
 Grafologie 48, 582, 622
 Grieshaber, Edmund (Interview) 743, 744
 Griffiths-EntwicklungsSkalen (GES) 289
 – Auswertung 290
 – Bewertung 291
 – Durchführung 290
 – Gliederung 289
 – Normierung 291
 – Reliabilität 290
 – Third Edition (Griffith III) 289
 – Validität 291
 Griffiths Scales of Child Development s.
 – Griffiths-EntwicklungsSkalen (GES)
 Grobnormen 185
 Grundgesetz 28
 Grundintelligenztest
 – Skala 1 – Revision (CFT 1-R) 282
 – Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest (CFT 20-R mit WS-ZF-R) 280
 – Auswertung 281
 – Bewertung 282
 – Durchführung 281
 – Gliederung 280
 – Reliabilität 281
 – Validität 281
 Grundrate 548
 Gruppenuntersuchung 255, 495
 Gültigkeit, individuelle 514
 Gutachten 28, 136
 – Anforderungen 521
 – familienrechtspychologisches 761
 – förderdiagnostisches 653, 655
 – formale Gestaltung 516
 – Gliederung 516
 – psychologisches 512
 – Qualität 520
 Gutachtenerstellung 478
 Gutachtenqualität 772
 Gütekriterium 321, 436, 605

H

Halo-Effekt 422, 594, 713
 Hamburger Zwangsinventar – Kurzversion (HZI-K) 711
 Hamilton Anxiety Scale (HAMA) 710
 Hamilton Depressionsskala (HAMD) 710
 Handlungs-IQ 259
 Hathaway, Starke R. 24
 Hauptgütekriterium 132
 Hauptkomponentenanalyse 123, 329
 Helping Alliance Questionnaire (HAQ) 710
 HEXACO-Modell 307
 HEXACO Personality Inventory-Revised (HEXACO-PI-R) 307
 Hierarchisches Rahmen- bzw. Protomodell der Intelligenzstrukturforschung (HPI) 273
 Hinweisreiz 416
 Hirnschädigung 734
 Historical-Clinical-Risk Management-20 (HCR-20) 753
 Hochbegabung 13, 667, 668, 670, 671
 Hochbegabungsdiagnostik 278, 667
 – durch Lehrkräfte 669
 Hochschulauswahlverfahren 654
 Hochschulrahmengesetz 654
 Hochschulzulassungsberechtigung 654
 Hogrefe, Jürgen (Interview) 214
 Holland-Code 356, 357
 Homogenität 114
 Horoskop 48
 Hospital Anxiety and Depression Scale – Deutsche Version (HADS-D) 710
 Hypothese 487

I

ICD s. Internationale Klassifikation psychischer Störungen
 Impact of Event Scale – revidierte Form (IES-R) 710
 Impliziter Assoziations-Test (IAT) 375
 – Grundannahme 375
 – methodisches Vorgehen 375
 – Reliabilität 375
 – Validität 376
 Impression Management 11, 317, 319, 440
 Indexsystem 407
 Indexwert 262, 265
 Indikation 717
 – adaptive 717
 – behandlungsbezogene 718
 – differentielle 718
 – selektive 717
 Indikationsstellung 697
 Informationsverarbeitung
 – auditive 674
 – eingeschränkte 621
 – visuelle 674
 Informationsverarbeitungsgeschwindigkeit 233, 240
 Inhaltsanalyse 159
 – kriteriumsbasierte 748
 – quantitative 160
 Inhaltsbereich 488

Inhaltsvalidität 158, 422
 Inhibition 298

Inklusion 650
 Instrument
 – diagnostisches 19
 – zur Kodierung von Diskussion (IKD) 414
 Integrität 46
 Integritätstest 603, 622
 Intelligence and Development Scales (IDS) 270
 – 2 (IDS-2) 295
 – Auswertung 299
 – Durchführung 299
 – Gliederung 297
 – Interpretation 299
 – Testgütekriterien 299

Intelligenz 42, 234, 237, 253, 298, 439
 – allgemeine 255, 261, 278, 285
 – fluide 255, 272, 273, 283, 285
 – kristallisierte 255, 272, 273, 281, 286
 – künstliche 377, 379, 429
 – multiple 287
 – prämorbide 739
 – Primärfaktoren 273
 Intelligenzalter 289
 Intelligenzdiagnostik 10
 Intelligenzfaktor 300
 Intelligenzkomponente 255
 Intelligenzminderung 666
 Intelligenzprofil 298
 Intelligenzquotient (IQ) 13, 23
 Intelligenzstruktur 300
 Intelligenzstrukturmodell 285
 Intelligenz-Struktur-Test (I-S-T)
 – 2000 – Revision (I-S-T 2000 R) 115, 272
 – Auswertung 274
 – Bewertung 276
 – Durchführung 274
 – Hintergrund 273
 – Normierung 275
 – Reliabilität 274
 – Validität 275
 – I-S-T 70 243, 272

Intelligenztest 10, 23, 170, 233, 253, 258, 583
 – Systematik 254
 – zur Feststellung von Hochbegabung 670
 Intelligenztestleistung 181
 Interaktionsprozessanalyse 414
 Interessen 42, 354, 610
 Interessentest 354, 603, 624
 International List of Causes of Death (ILCD) 23
 International Personality Disorder Examination (IPDE) 707
 International Personality Item Pool 353
 International Test Commission (ITC) 25, 196, 480, 557
 Internationale Diagnose Checklisten
 – für ICD-10 (IDCL) 707
 – für Persönlichkeitsstörungen (ID-CL-P) 707
 Internationale Grundschul-Lese-Untersuchung (IGLU) 681
 Internationale Klassifikation psychischer Störungen (ICD) 23, 694, 700

– ICD-10 7, 701
 – Symptom-Rating (ISR) 710
 – ICD-11 7, 13
 Internationale Skalen für Hypochondrie (WI-IAS) 710
 Internetinterview 427
 Interpretation 400
 Interpretationsobjektivität 136, 505
 Interrater-Reliabilität 420, 421, 436
 Interview 48, 321, 622, 706
 – diagnostisches 3, 8, 13, 312, 313, 426, 427, 583, 694, 735, 771
 – eignungsdiagnostisches 430, 438
 – halbstandardisiertes 430
 – klinisch-psychologisches 692
 – konventionelles 438
 – standardisiertes 430, 447
 – strukturiertes 430, 432, 582
 – halbstandardisiertes 446
 – klinisches 430, 432, 437, 707
 – unstandardisiertes 430
 – verhaltensbezogenes 438
 Interviewauswertung
 – qualitative 449
 – standardisierte 448
 Interviewertraining 441
 Interviewfrage 445
 Interviewkonstruktion 443
 Interviewleitfaden 443
 – Feinaufbau 445
 – Grobaufbau 443
 Intraklassenkorrelation 420
 Inventar

– Klinischer Persönlichkeitsakzentuierungen (IKP) 334
 – komplexer Aufmerksamkeit (INKA) 235
 – zur Erfassung der Lebensqualität bei Kindern und Jugendlichen (ILK) 711
 – zur Erfassung interpersonaler Probleme – Deutsche Version (IIP-D) 710, 717
 In-vivo-Beobachtung 712
 IQ-Normwerte 148
 Itemanalyse 109, 126
 Itembearbeitung 250
 Itemdiskriminationsparameter 73
 Itemfitmaß 127
 Itemformat 101
 Itemformulierung 96
 Itemgenerierung 85, 87
 Iteminformationsfunktion 76
 Itemladung 118
 Itemleichtigkeitsindex 110
 Itemlösung 250
 Itemparameter 72, 127
 Itempolung 105
 Itempool 90
 Item-Response-Theorie 55, 57, 76
 Itemschwierigkeit 65, 70, 109, 116
 Itemselektion 125, 131
 Itemstreuung 112, 116
 Itemvalidität 126
 Itemvarianz 112
 Itemvorverarbeitung 252

J

- Jingle-Jangle-Irrtum 175
 Job-Demands-Resources-Modell 617
 Joint-Maximum-Likelihood-Methode 72
 Justenhofen, Richard (Interview) 379

K

- Karriereplanungssystem, computerbasiertes 611
 Kategoriensystem 407, 414
 Kategorienwahrscheinlichkeit 128
 Kaufman Assessment Battery for Children (K-ABC) 267
 – Second Edition (KABC-II) 269
 Kersting, Martin (Interview) 634
 Kieler Änderungssensitive Symptomliste (KASSL) 715
 Kieler Schmerz-Inventar (KSI) 710
 Kinder-Angst-Test-III (KAT-III) 711
 Kindeswill 759
 Kindeswohl 759
 Kind-Umfeld-Analyse 653
 Klassifikation 2, 12
 – psychischer Störungen 701, 702
 Klassifizieren von Wörtern 251
 Klassische Testtheorie 49, 109
 Klinische Psychologie 7, 9
 Koeffizient Alpha 144
 Kommunalität 119
 Komorbidität 738
 Komorbiditätsprinzip 705
 Kompetenz 581
 – naturwissenschaftliche 678
 – schulische 298
 – sozial-emotionale 298
 Kompetenzmodellierung 574
 Konditionieren, klassisches 697
 Konfidenzintervall 148, 149, 506, 507, 510
 Konfundierung 176, 234
 Konsistenz, interne 143
 Konsistenzmethode 146
 Konstante 628
 Konstruktionsprinzip 84
 Konstruktvalidität 158, 241, 321, 422
 Kontextualisierung 623
 Kontrollskala 319, 741
 Konzentration 239
 – Definition 232
 Konzentrationsfähigkeit 226, 234, 252
 Konzentrationsfaktor 236
 Konzentrations-Leistungs-Test (KLT) 249
 – revidierte Fassung (KLT-R) 235
 Konzentrationstest 225, 232
 – für 3. und 4. Klassen – Revision (KT 3–4 R) 247
 – für Kinder (KoKi) 247
 Konzentrationstestleistung 234
 – Zerlegung 252
 Konzentrations-Verlaufs-Test (KVT) 235
 Koordinationsleistung 249
 Körperkoordination 674
 Körperkoordinationstest für Kinder (KTK) 288
 Kraftfahreignung 769

– Wiederherstellung nach Alkoholmissbrauch 771

- Kreuzvalidierung 94
 Kriminalprognose 9, 751
 Kriterium 558
 Kriteriumsdefizienz 175
 Kriteriumskontamination 175
 Kriteriumsvalidität 158, 322, 422
 Kriteriumsvariable 576
 Kultur 583
 Kurve, itemcharakteristische 65
 Kurzversion 130

L

- Langversion 130
 Large-Scale-Assessment 25
 Latent-Class-Analyse 55, 81
 Laufbahnplanung 355
 Leadership
 – Judgment Indicator 602
 – Style Assessment 602
 Lebensereignis 304
 Lebensgeschichte 696
 Lebenslauf 633
 Legasthenie 662
 Lehrerurteil 669
 Leistung 216, 668
 Leistungsbeurteilung 615, 626
 Leistungsfähigkeit
 – allgemeine 226
 – intellektuelle 282
 – prämorbide 738
 – psychische 766
 Leistungsmotiv 391, 396, 397
 Leistungsmotivation 44, 217, 358, 391
 – berufsbezogene 359
 – sportbezogene 358
 Leistungsmotivationsfragebogen 358
 Leistungsmotivationsinventar (LMI) 358, 359
 – Auswertung 359
 – Bewertung 361
 – Durchführung 359
 – Gliederung 359
 – Normen 361
 – Reliabilität 359
 – Validität 359
 Leistungsmotivations-TAT 393
 Leistungsprüfsystem (LPS)
 – LPS-2 276
 – LPS 50+ 276
 Leistungstest 45, 216, 605, 735
 – allgemeiner 226
 – kognitiver 285, 582, 622
 Leitlinie
 – fachspezifische 481
 – zur Diagnostik und Behandlung der Rechenstörung 664
 Lernbehinderung 651
 Lernhilfe 652
 Lernschwierigkeit 660
 – als psychische Störung 661
 – individuelle 660
 Lernstörung
 – in der Grundschule 662

– spezifische 662

- Lese-Rechtschreibstörung 662
 Leyton Obsessive Compulsive Inventory (LOI-K) 711
 Liebowitz-Soziale-Angst-Skala (LSAS) 710
 Lienert, Gustav A. 24
 Likelihood
 – Definition 70
 – Funktion 70
 Lincoln-Oseretzkzy-Skala Kurzform (LOS KF 18) 301
 Lindemann, Sara (Interview) 428
 Linsenmodell von Brunswik 416
 Literatursuche 493
 Lizenzierwerb 482
 Lösungswahrscheinlichkeit 62
 Low-Fidelity-Simulation 597
 Lübecker Alkoholabhängigs- und -missbrauchs-Screening (LAST) 710
 Lückentext 675
 Luria-Modell 270

M

- Machiavellismus 307
 Machine Learning 377, 378
 Machtmotiv 391, 396
 Mannheimer Elterninventar (MEI) 711
 Marburger Modell 506, 508
 Marginal-Maximum-Likelihood-Methode 72
 Marlowe-Crowne-Skala 319
 Mathematiktest 676
 Matrizenaufgabe 282
 Maximum-Likelihood-Schätzmethode 72
 Mayer-Salovey-Caruso Test zur Emotionalen Intelligenz (MSCEIT) 287
 McDonalds Omega 145
 McKinley, J. Charnley 24
 MedAT – Aufnahmeverfahren Medizin 655
 Medical Symptom Validity Test (MSVT) 742
 Medizinisch-Psychologische Untersuchung (MPU) 762
 Mehrfachwahl-Wortschatz-Intelligenztest (MWT) 255
 – in der Form B (MWT-B) 739
 Memory Bias 315
 Menschenwürde 28
 Mental Speed 233
 Merkfähigkeit 237, 273, 277
 Merkmal
 – latentes 46, 48
 – mehrdimensionales 90
 Merkmalsprägung 648
 Merkmalskontinuum 76
 Messanspruch 43, 490, 625
 Messäquivalenz 49, 141
 Messfehler 51, 54, 311
 Messgegenstand 42, 84
 Messintention 254
 Messung
 – essenziell parallele 142
 – essenziell tau-äquivalente 142
 – formative 47
 – nichtsequenzielle 553

Stichwortverzeichnis

- parallele 49, 141
 - reflexive 47
 - sequenzielle 553
 - tau-äquivalente 141
 - tau-kongenerische 141
 - Messwertbereich 506
 - Erläuterung 507, 508
 - Festlegung 506, 507
 - Metaanalyse 172
 - Methode
 - arbeitsplatzanalytisch-empirische 574
 - deduktive 85, 87
 - des zirkulären Fragens 716
 - erfahrungsgeleitet-intuitive 574
 - externe 93
 - induktive 85, 89
 - kriteriumsorientierte 86, 92
 - personenbezogen-empirische 575
 - Mildefehler 423
 - Minderungskorrektur 156
 - doppelte 155
 - Mindestanforderung 448
 - Mindestwert 431
 - Mini-Symptom-Checklist (Mini-SCL) 710
 - Minnesota Multiphasic Personality Inventory (MMPI) 24
 - MMPI-2 94, 323, 714
 - Auswertung 326
 - Bewertung 327
 - Durchführung 326
 - Interpretation 326
 - Items 325
 - Konstruktansatz 324
 - Normierung 327
 - Reliabilität 327
 - Struktur 325
 - Validität 327
 - Restructured Form (MMPI-2-RF) 328
 - Auswertung 331
 - Bewertung 333
 - Durchführung 331
 - Interpretation 331
 - Konstruktansatz 328
 - Normierung 333
 - Reliabilität 332
 - Skalen 330
 - Struktur 328
 - Validität 332
 - Misinformation-Effekt 747
 - Modalitäten des Erlebens und des Verhaltens 697
 - Modell
 - dreiparametrisches logistisches (3PL-Modell) 75
 - einparametrisches logistisches (1PL-Modell) 57, 60, 72, 223
 - nach Holland 610
 - respondenten 698
 - zweiparametrisches logistisches (2PL-Modell) 74
 - Modelltest 75
 - grafischer 67, 127
 - Modified PTSD Symptom Scale (MPSS) 710
 - Modifikation 531, 568, 613, 617
 - Motiv 354, 358, 616
 - Motivation 42
 - bei der Testbearbeitung 217
 - Motivmessung 161
 - Motorik 287
 - Motoriktest 287
 - Multiaxiale Schmerzklassifikation – Psychosoziale Dimension (MASK-P) 710
 - Multimethodische Objektive Interessensbatterie (MOI) 374
 - Multi-Motiv-Gitter (MMG) 384, 396, 397
 - Multiple Choice
 - Antwortformat 74, 101, 600
 - Aufgabe 100, 101
 - Multiple-Cut-off-Modell 538
 - Multiple-Hurdle-Modell 538
 - Multiple-Hurdle-Problem 540
 - Multitrait-Multimethod-Ansatz 167, 377, 394
 - Münsterberg, Hugo 23
 - Myers-Briggs-Typenindikator (MBTI) 341
 - myPersonality 378
- N**
- Nachfragen 434, 454
 - Narzissmus 88, 307
 - National Adult Reading Test (NART) 739
 - Naturalistic Driving Study 412, 769
 - Nebengütekriterium 132
 - Neglect 737
 - NEO-Fünf-Faktoren-Inventar (NEO-FFI) 338, 341, 345, 353, 714
 - NEO-Persönlichkeitssinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R) 115, 333, 341, 349
 - Auswertung 350
 - Bewertung 352
 - Durchführung 350
 - Gliederung 350
 - Interpretation 350
 - Normierung 352
 - Reliabilität 351
 - Validität 351
 - Netzwerk
 - nomologisches 165
 - semantisches 375
 - Neurologie 744
 - Neuropsychologie 9, 732
 - Neurotizismusfaktor 337
 - Norm, angemessene 490
 - Normalverteilung 150
 - Normenpyramide 27
 - Normierung 182, 188
 - Normwert 148
 - Nürnberg-Alters-Inventar (NAI) 745
 - Nutzenschätzung 559
 - Nützlichkeit 197
- O**
- Objektiver Leistungsmotivations-Test (OLMT) 217, 371, 374
 - Bewertung 374
 - Gliederung 372
 - Normierung 373
 - Objektivität 372
 - Reliabilität 372
- Validität 373
 - Objektive Testbatterie OA-TB 75 368
 - Objektivität 133, 137, 420, 424
 - der Vergleiche, spezifische 67
 - Obsessive-Compulsive Inventory-Revised (OCI-R) 710
 - Occupational Information Network (O*-NET) 5, 581
 - Odds-Ratio 413
 - Offenbarungspflicht 31
 - Offender Group Reoffending Scale – Version 3 753
 - Offenheit 321
 - Oklahoma premorbid intelligence estimate 3 738
 - Ökonomie 197
 - Online-Assessment-Center 608
 - Online-Self-Assessment 612, 657
 - Ablauf 658
 - Portal 657
 - psychometrische Qualität 659
 - Onlinetest 45
 - Online-Testverzeichnis 213
 - Operanter Multi-Motive-Test (OMT) 397
 - Operationalisierte Fertigkeitsdiagnostik (OFD) 719
 - Operationalisierte Psychodynamische Diagnostik (OPD-2) 715
 - Optionsgewichtung, empirische 108
 - Organigramm 717
 - Organisationsdiagnostik 570
 - Organisationsentwicklung 573, 617
 - Outcome Questionnaire-45.2 (OQ-45.2) 710
 - Overachiever 668
- P**
- Pädagogische Psychologie 8, 9, 644
 - Panel Interview 437
 - Panik- und Agoraphobie-Skala (PAS) 710
 - Papier-und-Bleistift-Test 45, 495
 - Paradaten 379
 - Parallelanalyse 123
 - Paralleltest-Methode 141
 - Parallelversion 141
 - Paraphrasieren 454
 - Partnerschaftsdiagnostik 341
 - Part-whole-Korrektur 114
 - Passung 610, 627, 719
 - Pauli-Test 242, 250
 - Penn State Worry Questionnaire (PSWQ) 710
 - persolog Persönlichkeits-Profil 341
 - Personal Data Sheet 23
 - Personalalauswahl 502, 582, 584
 - Personalalauswahlverfahren 603, 621
 - Personalentwicklung 581, 613
 - Personalentwicklungsbedarf 625
 - Personalfragebogen 32
 - Personality-Psychopathology-Five-Skala 331, 333
 - Personality Research Form (PRF) 309, 338, 397
 - Personalmarketing 607
 - Personenenalyse 614

- Personenauswahl 551
 Personenhomogenität 67
 Personenmerkmal 582, 660
 Personenparameter 71
 Personenselektion 531, 584
 Persönlichkeit 303
 Persönlichkeitsdiagnostik 367
 Persönlichkeitsdimension 308
 Persönlichkeitseigenschaft 369
 Persönlichkeitsfaktor 368
 Persönlichkeitsfragebogen 96, 303, 312, 313, 322, 582, 583, 622
 – für Kinder zwischen 9 und 14 Jahren (PFK 9–14) 353
 – mehrdimensionaler 353
 Persönlichkeitsmerkmal 303
 Persönlichkeits-Stil- und Störungs-Inventar (PSSI) 710
 Persönlichkeitsstörung 707
 Persönlichkeitsstruktur 306
 Persönlichkeitstest 45, 713
 – objektiver 312, 313, 367
 Persönlichkeitstestsystem 322
 Persönlichkeitstheorie von Murray 389
 Persönlichkeitstyp 83
 Phobie 710
 Picture Story Exercise (PSE) 384, 394, 397
 PISA s. Programme for International Student Assessment
 Plananalyse 699
 Politik 308
 Politikberatung 677
 Populismus 308
 Position Analysis Questionnaire (PAQ) 575
 Positive and Negative Syndrome Scale (PANSS) 710
 Powertest 111, 256
 Prädiktionswert 551
 – negativer 547
 – positiver 547
 Prädiktiorwert 561
 Pre-accept-Strategie 554
 PRECIRE JobFit 378, 381
 Pre-reject-Strategie 553
 Primacy-Effekt 422
 Principal Component Analysis (PCA) 123
 Privatsphäre 28
 Proaktivität 45
 Probabilistische Testtheorie 126
 Problem, ethisches 380
 Problemanalyse 697
 Problemaufrechterhaltung 698
 Problemlösen 616
 Problemstrukturierung 696
 Profilauswertung 327
 Profilblatt 351
 Prognose 491
 – ideografische 756
 – nomothetische 752, 757
 Prognosegutachten, Mindestanforderungen für 758
 Prognosekraft 585
 Prognosetafel 752
 Programme for International Student Assessment (PISA) 8, 25, 85, 677
 – PISA-Studie 677
- Kompetenzstufen 679
 – Länderergebnisse 680
 – Testungen 8
 Progress in International Reading Literacy Study (PIRLS) 681
 Projektion 382
 Projektionsbegriff
 – klassischer 382
 – verallgemeinerter 383
 Protokollgültigkeit 332
 Prototypikalität 91
 Prozentrang 185, 188
 Prozess, diagnostischer 2, 15, 478, 512, 529
 Prozessevaluation 556
 Prozessmodell der Itembearbeitung 250
 Prozesswert 262
 Prüfungsangst 502
 Pseudoparalleltest 142
 Psychiatrischer und Sozial-Kommunikativer Befund (PSKB) 716
 Psychoanalyse 715
 Psychologie
 – differentielle 10
 – forensische 8
 – klinische 7, 9
 – pädagogische 8, 9, 644
 Psychologische Diagnostik 2, 11, 22
 Psychologische Rundschau (Zeitschrift) 214
 Psychometrie 47
 Psychomotorik 287, 298
 Psychopathie 307
 Psychopathologie 170
 Psycho-Physiognomik 48
 Psychotechnik 23, 26
 Psychotherapeutengesetz 700
 Psychotherapie 32
 Psychotherapieevaluation 713
 PSYNDEX Tests 492
 PTSD Symptom Scale (PSS) 710
- Q**
- Q-Index 127
 Q-Sort-Technik 715
 Qualität, psychometrische 633
 Qualitätssicherung 719
 Qualitätsstandards 632
 – für berufsbezogene Eignungsbeurteilungen 632
 – für psychologische Gutachten 478–480
 Quantifizierung 407
- R**
- Rahmenbedingung, gesetzliche 27
 Ranking 600
 Rasch-Homogenität 68
 Rasch-Modell 269
 – dichotomes 55
 – ordinale 55
 Rat für Sozial- und Wirtschaftsdaten (RatSDW) 35
 Raten 103
 Rateparameter 75
- Rating 447
 Ratingskala 55, 104, 400, 416, 431
 Ratingskalenmodell 79, 128
 Ratingverfahren 407
 Raven's Progressive Matrices 282
 – Clinical Edition (Raven's 2) 283
 Reaktionsschnelligkeit 229
 – psychomotorische 287
 Reaktivität 423, 425, 713
 Reality Monitoring 748
 Realkennzeichen 748
 Receiver Operating Characteristic 546
 Recency-Effekt 422
 Rechenfertigkeit 237, 249
 Rechenkonzentrationstest 242, 250
 Rechenstörung 663, 664
 – Empfehlungen zur Diagnostik 665
 – Korrelate 666
 Rechtschreibfehlerfinden 243
 Rechtschreibtest 675
 Rechtspychologie 746
 Referenz 582, 622
 Reflexion 454
 Regression
 – multiple 591
 – zur Mitte 154, 646
 Regressionsanalyse
 – logistische 555
 – polynomiale 628
 Regressionsgewicht 628
 Regressionsmethode 148
 Reizselektion 228
 Relevanz, klinische 722
 Reliabilität 52, 138, 147, 155, 705, 706
 – Definition 52
 Reliabilitätschätzung 54, 138
 Reliable Change Index 153
 Response-Surface-Analyse 628
 Retest
 – Effekt 221
 – Methode 138
 – Reliabilität 55, 139, 437
 Revisions-Test 242, 250, 251
 RIASEC
 – Modell 579
 – Typen 355
 Richtlinien
 – berufsethische 481, 519
 – ethische 33
 – zur computer- und internetbasierten Testung 494
 RISIKO – Risikowahlverhalten 374
 ROC-Kurve 546
 Rollenspiel 699
 Rorschach, Hermann 24
 Rorschach-Test 24, 384, 385
 – Auswertung 386
 – Durchführung 386
 – Gütekriterien 387
 Rosenberg Skala zum Selbstwertgefühl (RSES) 724
 Rost, Detlef H. (Interview) 671
 Rückfallprognose 751, 755
 Rückfallquote bei Sexualdelikten 755
 Rückfallrisiko 9
 – Fehleinschätzung 752

Stichwortverzeichnis

– Regressionsgleichung 753

S

Sachkunde 486

Sachverständigungsgutachten 759

Scheidung 760

Scheidungsverfahren 758

Schema, kognitives 699

Schläfrigkeit 412

Schmerzempfindungs-Skala (SES) 710

Schmerztagebuch 712

Scholastic Aptitude Test, später auch Scholastic Assessment Test (SAT) 655

Schulabschlussnote 585

Schuldfähigkeit, verminderte 750

Schuldunfähigkeit 750

Schuleingangsdagnostik 296, 645

Schuleingangstest 301, 645, 673

Schuler, Heinz (Interview) 606

Schulerfolg 648

Schullaufbahnberatung 283, 644, 675

Schulleistung 648

– weltweite Messung 678

Schulleistungstest 85, 282, 301, 675

Schulnote 275, 281, 656

Schulreife 645, 647

– fehlende 646

Schulreifetest 646

Schulschwierigkeit 660

Schultest 225, 301, 673

Schulwechsel 644

Schweigepflicht 29

Schwelle 79

Schwellenabstand 79

Schwellenparameter 129

Scoring 107

Screening 257, 585

– des Entwicklungsstandes bei Einschulungsuntersuchungen (S-ENS) 674

– für Somatoforme Störungen (SOMS) 710

Scree-Plot 123

Segmentierung 407

Selbstauskunft 310

Selbstbeobachtung 404, 425, 712

Selbstbericht 48

Selbstbeschreibung 580

Selbstbeschreibungsinventar 347

Selbstbeschreibungsprofil 348

Selbstbeurteilung 309, 349

Selbstbild 354

Selbstdarstellung 192

– sozial erwünschte 318

– strategische 311

Selbstdarstellungsstrategie 440

Selbsteinsicht 315

Selbstkonzept 311

Selbstselektion 612, 658

Selbsttäuschung 311, 316, 319

Selektion 407, 530, 568

– von Bedingungen 609

Selektionskennwert 117

Selektionsparadigma 20

Selektionsquote 196

Selektionsrate 550, 552

Self-directed Search (SDS) 355

Sensitivität 545, 743, 749

Sexualstraftat 756

Sicherheitswahrscheinlichkeit 150

Signierung 394

Signifikanz, statistische 722

Simon, Theodore 23

Simulation 587, 589, 740, 742

Simultanfaktorisierung 338

Single-Choice-Antwortformat 600

Situation 304

Situational-Judgment-Test 45, 98, 305, 582, 597, 599, 620

– Eindimensionalität 600

– für Teamarbeit (SJT-TA) 602

– Retest-Reliabilität 600

– Vorhersagegüte 600

Situationsklasse 305

Situationsmerkmal 660

Situationsstärke 14

Situationswahrnehmung 14

Skala

– bipolare 104

– klinische 329

– spezielle aus dem Diagnostik-System zur Erfassung psychischer Störungen bei Kindern und Jugendlichen (DI-SYPS-III) 711

– unipolare 104

– verhaltensverankerte 416, 431, 435

– zur Erfassung der Lern- und Leistungsmotivation (SELLMO) 358

– zur Erfassung der Schwere der Alkoholabhängigkeit (SESA) 710

Skalenabstufung 105

Skalenniveau 55

Skalierung 190

Soll-Ist-Diskrepanz 613, 617

Sorge, elterliche 758

Sorgerecht 758

Sorgerechtsentscheidung 758, 760

Sorgerechtsverfahren 758, 759

SORKC-Schema 531

Soziale-Interaktions-Angst-Skala (SIAS) 710

Soziale-Phobie-Skala (SPS) 710

Soziale-Phobie-und-Angst-Inventar (SPAII) 710

– für Kinder (SPAIIK) 711

Sozialformel 739

Spearman-Brown-Formel 143

Speedkomponente 111, 147

Speedtest 111, 256

Spezialbegabung 670

Spezialproblemkala 331

Spezifität 545, 743, 749

Split-Half-Methode 142, 146

Sportbezogener Leistungsmotivationstest (SMT) 358

Sportpsychologie 358

Sprachanalyse 212, 381

Sprachentwicklung 300

Sprachkompetenz 674

Sprachmerkmal 381

Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren (SET 5–10) 300

Standard Progressive Matrices (SPM) 255, 283

Standardabweichung 149

Standardisierung 134, 436, 503

Standardmessfehler 152

Standardmodell 507, 508

Standardschätzfehler 149

Stanford-Binet-Test 23

Stanine-Wert 183, 185, 506

State 361

State-Trait-Angstinventar (STAI) 361, 363

– Auswertung 363

– Bewertung 364

– Durchführung 363

– Gliederung 363

– Normierung 364

– Reliabilität 363

– Validität 364

State-Trait-Ärgerausdrucks-Inventar – 2 (STAXI-2) 711

– für Kinder und Jugendliche (STAXI-2 KJ) 711

Statistik, erschöpfende 72

Status, beruflicher 275

Statusdiagnostik 528

Stellungnahme 514, 518

Sten-Wert 187

Stern, William 23

Stichprobe 177

Stichprobenfehler 177

Störung

– krankhafte seelische 751

– psychische 438, 690, 700

– psychotische 710

– somatoforme 710

Störungsvermeidung 502

Störungswissen 690

Strategie, diagnostische 528

Strengefehler 423

Strengths and Difficulties Questionnaire (SDQ) 711

Stresssituation 305

Stressverarbeitungsfragebogen (SVF) 305

Streuungeinschränkung 178, 593

Strukturanalyse 164

Strukturelle Analyse Sozialer Beziehungen (SASB) 717

Strukturiertes Inventar für Anorektische und Bulimische Essstörungen (SIAB) 711

Strukturiertes Klinisches Interview

– für Dissoziative Störungen (SKID-D) 710

– für DSM-Diagnosen (SKID) 706

– für DSM-5-Persönlichkeitsstörungen

– Klinische Version (SKID-5-CV) 433, 707

– für DSM-5-Persönlichkeitsstörungen (SCID-5-PD) 707

Strukturtest 255

Studiendauer 656

Studienerfolg 656

Studiengang 659

Studiengangeignung 439

Studiennote 656

Studienzufriedenheit 659

Studierfähigkeitstest 655

– spezifischer 656
 Suchaufgabe 246
 Suggestivfrage 96, 447, 747
 Supervision 482
 Symptom
 – neurologisches 733
 – Validity Test 742
 Symptom-Checklist-90-Standard (SCL-90-S) 710
 Symptomstörung 707

T

Tätigkeitsbewertungssystem (TBS) 576
 Teamarbeit 602
 Teamentwicklung 573
 Teamklima 568
 Teamklima-Inventar (TKI) 571
 – Aufbau 571
 – Bewertung 573
 – Hintergrund 571
 – Konstruktionsprinzipien 572
 – psychometrische Kennwerte 572
 Team-Role-Test 598
 Teilleistungsstörung 661
 Telefoninterview 381, 427, 582
 Tendenz
 – zentrale 423
 – zu Extremurteilen 423
 Test
 – Begriff 41
 – computerassistierter 257
 – computergestützter 46, 256
 – computerisierter adaptiver 24
 – d2 – Revision (d2-R) 232, 235, 237, 251
 – Auswertung 240
 – Bewertung 244
 – Computerversion 244
 – Durchführung 240
 – Kennwerte 240
 – Normierung 243
 – Reliabilität 241
 – Validität 241
 – für Medizinische Studiengänge (TMS) 655
 – kriteriumsorientierter 92
 – objektiver 367
 – projektiver 384
 – psychologischer 41
 – semiprojektiver 393
 – zur Praktischen Alltagsintelligenz (PAI 30) 193
 Testangst 222, 503
 Testarchiv, elektronisches 493
 Testauswahl, kompetente 489
 Testauswerteprogramm 504
 Testauswertung 504
 – manuelle 504
 Testbatterie zur Aufmerksamkeitsprüfung (TAP) 229, 736
 Testdauer 496
 Testeinsatz 224
 Testen, adaptives 75, 269
 Testentwicklung 130
 Testentwurf 132
 Testerfahrung 218, 491

Testergebnismitteilung 507
 Testergebnisverfälschung 224
 Testgütekriterium 132
 Testhalbierungsmethode 142
 Testinformationen 493
 Testkompendium 213
 Testkonstruktion 54
 – externe 86
 Testkonstruktionsstrategie 95
 Testleistung 216, 496
 Testleitereffekt 265
 Testmanual 85, 136, 147, 181, 190, 213, 503
 Testmodell, probalistisches 56
 Testmotivation 217
 Testnorm, kontinuierliche 189
 Testrezension 213
 Testschlüsse 103
 Testschutz 211
 Testtheorie 47
 – klassische 49, 109
 – probabilistische 126
 Testtraining 221
 Testverfahren 692
 Testwertschätzung 646
 Testwiederholung 218, 647
 Testzentrale 214, 493
 Testziel 84
 Thematischer Apperzeptionstest (TAT) 383, 384, 389
 – Auswertung 390
 – Durchführung 390
 – Objektivität 390
 – Reliabilität 391
 – Validität 391
 Therapie
 – kognitive 712
 – systemische 716
 Therapieverlaufsdiagnostik 720
 Think-aloud-Technik 163
 Threshold 129
 Thurstone-Intelligenzmodell 272, 277
 Time-Sampling-Methode 409
 T-Normwert 184
 Tonaufzeichnung 773
 Trainingsbedarfsanalyse 613
 Trainingseffekt 221
 Trait 14, 361, 713
 Transparenz 589
 Trefferquote 551
 Trends in International Mathematics and Science Study (TIMSS) 681
 Trennschärfe 108, 113, 116
 Trennung 760
 Trierer Integriertes Persönlichkeitsinventar (TIPI) 340, 714
 Truncated normal distribution 561

U

Übungseffekt 51, 54, 219–221, 244, 252
 Übungsgewinn 154
 Umgangsregelung 760
 Umwelt-Struktur-Test (UST-3) 357
 Unabhängigkeit, lokale stochastische 69
 UN-Behindertenrechtskonvention 650

Unconditional-Maximum-Likelihood-Methode 72
 Underachiever 668
 Unfallrisiko 412, 767
 Uniform Guidelines for Employee Selection Procedures 195
 Unrechtsbewusstsein, eingeschränktes 750
 Unterstützung, soziale 717
 Untersuchung
 – diagnostische 478
 – medizinisch-psychologische (MPU) 762
 – neuropsychologische 735
 Untersuchungsanlass 517, 763
 Untersuchungsbericht 517
 Untersuchungsplanung 489, 494
 Unverfälschbarkeit 192
 Urheberrecht 211
 Urteilsbildung
 – diagnostische 19
 – klinische 532, 738
 – mechanische 532, 603
 – statistische 532, 738
 Urteilsfehler 311, 422, 534
 Urteilsmodell 536
 Urteilsprozess 316, 417

V

Validierung 159, 168, 696
 Validierungstest 224
 Validität 157, 704, 705
 – Definition 157
 – soziale 605
 Validitätsart 158
 Validitätsbeleg
 – diskriminanter 165
 – inkrementeller 171
 – konkurrenzer 170
 – konvergenter 165
 – prädiktiver 170
 – retrograder 170
 Validitätsdilemma 227
 Validitätsskala 331
 Variabilitätsnorm 182, 183
 Varianzeinschränkung 178
 Veränderungsdiagnostik 529, 720
 Veränderungsfragebogen
 – des Erlebens und Verhaltens (VEV) 715
 – für Lebensbereiche (VLB) 715
 Veränderungsmessung
 – direkte 720
 – indirekte 719
 Veränderungswissen 690
 Verarbeitung personenbezogener Daten 501
 Verarbeitungsgeschwindigkeit 233, 253, 285
 Verarbeitungskapazität 278
 Verbaler Lern- und Merkfähigkeitstest (VLMT) 745
 Verbal-IQ 259
 Verfahren
 – biografieorientiertes 584
 – diagnostisches 41, 211, 490
 – eignungsdiagnostisches 607
 – projektives 312, 313, 382

- psychodiagnostisches 744
- psychometrisches 708
- psychophysiologisches/biologisches 692
- rechtlich zulässiges 491
- simulationsorientiertes 586
- suboptimales 492
- Suchstrategie für ein 492
- typologisches 83
- verbal-thematisches 383
- zeichnerisches 383, 398
- Verfahrensabfolge 496
- Verfahrensauswahl 489
- Verfälschbarkeit 584, 624
- Verfälschung 192, 317, 384, 491, 741, 771
- Verhalten 2, 304, 310, 691
 - beobachtbares 697
 - delinquentes 751
 - maximales 593
 - nonverbales 440
 - problematisches 696
 - typisches 593
 - verbales 440
- Verhaltensanalyse 531, 697
 - horizontale 699
 - vertikale 699
- Verhaltensbeobachtung 8, 12, 211, 289, 292, 312, 313, 378, 400, 633, 699
 - direkte 402
 - freie 401, 411, 424
 - im Feld 403
 - nichtteilnehmende 404
 - systematische 402, 407, 424
 - verdeckte 404
- Verhaltensbeurteilung 400, 415, 416, 425
- Verhaltensbeurteilungsbogen für Vorschulkinder (VBV 3-6) 711
- Verhaltendiagnostik, experimentalpsychologische 367, 375
- Verhaltengleichung 488, 698
- Verhaltensstil 354
- Verhaltensverankerung 416
- Verkehrspsychologie 9
- Verkehrsunfall 762
- Verlaufsdiagnostik 529, 735
- Verlaufsprognoze 734
- Vertraulichkeit 427, 498
- Verwechslungsfehler 240
- Videoaufnahme 411, 773
- Videointerview 428, 582
 - zeitversetztes 428

- Visueller und Verbaler Merkfähigkeitstest (VVM) 745
- Visuomotorik 674
- Vorgehen
 - ethisch-verantwortliches 481
 - multimethodales 602
- Vorgeschiede 517
- Vorgesetztenbeurteilung 591

W

- Wahrnehmungsgeschwindigkeit 233
- Wechsler Adult Intelligence Scale (WAIS) 24, 259
 - Fourth Edition (WAIS-IV) 267
- Wechsler-Bellevue Intelligence Scales 24
- Wechsler, David 24, 258
- Wechsler Intelligence Scale for Children (WISC) 258
 - Fifth Edition (WISC-V) 261, 262
 - Auswertung 264
 - Durchführung 264
 - Gliederung 262
 - Interpretation 265
 - Normen 266
 - Objektivität 265
 - Reliabilität 266
 - Testaufbau 262
 - Validität 266
- Wechsler-Intelligenztest 258
- Wechsler Memory Scale 736
- Wechsler Preschool Primary Scale of Intelligence (WPPSI) 259
 - Fourth Edition (WPPSI-IV) 267
- Wende, kognitive 698
- Wert, wahrer 51
- Wertpunkt 265
- Wettquotient 61
 - logarithmierter 62
- Widerstand 456
- Wiener Entwicklungstest (WET) 292
 - Auswertung 293
 - Bewertung 294
 - Durchführung 292
 - Normierung 294
 - Reliabilität 294
 - Validität 294
- Wiener Risikobereitschaftstest Verkehr (WRBTV) 374
- Wilde-Intelligenz-Test 2 (WIT-2) 277

Wirtschaftspsychologie 568

Wissen 216

- berufsrelevantes 217
- berufsspezifisches 602
- erworbenes 285
- prozedurales 285
- psychologisches 2

Wissenstest 272

Wissler, Clark 23

Word Memory Test (WMT) 742

Work Design Questionnaire (WDQ) 576

Working Alliance Questionnaire – short revised (WAI-SR) 710

Wortwahl 452

Wundt, Wilhelm 22

Würzburger Vorschultest (WVT) 674

Y

Yale-Brown Obsessive Compulsive Scale (Y-BOCS) 710

Yale-Tourette-Syndrom-Symptomliste 711

Yerkes, Robert M. 23

Youden-Index 547

Z

Zahlen-Symbol-Test (Z-S-T) 233, 235, 242

Zahlen-Verbindungs-Test (ZVT) 235, 242

Zeichensystem 407, 410

Zeitstichprobe 409

Zentrales Beziehungskonfliktthema (ZBKT) 715

Zeugenaussage 515

Zeugnis 583

Zieldefinition 720

Zielgruppe 46, 95

Zuhören, aktives 453

Zumutbarkeit 191

Zuordnungsaufgabe 99, 101

Zuordnungswahrscheinlichkeit 83

Zusammenfassen 455

Zustand

– aktueller 361

– emotionaler 361

Zustandsmessung 362

Zwangsstörung 710