



Helfried Moosbrugger · Augustin Kelava *Hrsg.*

Testtheorie und Fragebogen- konstruktion

3. Auflage

EXTRAS ONLINE

 Springer

Testtheorie und Fragebogenkonstruktion

Helfried Moosbrugger · Augustin Kelava
Hrsg.

Testtheorie und Fragebogen- konstruktion

3., vollständig neu bearbeitete, erweiterte und aktualisierte
Auflage

Hrsg.

Hefried Moosbrugger
Institut für Psychologie
Goethe-Universität Frankfurt am Main
Frankfurt am Main, Deutschland

Augustin Kelava
Methodenzentrum
Eberhard Karls Universität Tübingen
Tübingen, Deutschland

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

ISBN 978-3-662-61531-7
<https://doi.org/10.1007/978-3-662-61532-4>

ISBN 978-3-662-61532-4 (eBook)

Die Deutsche Nationalbibliothek verzeichnetet 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 2008, 2012, 2020
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: © Worawut / stock.adobe.com

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

Zum Gedenken an OStR. Dr. Gislinde Moosbrugger, die – trotz schwerer Erkrankung – den aufwendigen Entstehungsprozess dieses Lehrbuchs stets unterstützt und bis zu ihrem Tod begleitet hat.

Vorwort zur 3. Auflage

Das von uns herausgegebene Lehrbuch *Testtheorie und Fragebogenkonstruktion* hat seit seinem Erscheinen in der Scientific Community eine so freundliche Aufnahme gefunden, dass wir bereits im Jahr 2015 mit dem Springer-Verlag erste Gespräche führten, um eine 3. Auflage in Angriff zu nehmen, in der wir dem Themengebiet zugleich breiteren Raum geben.

Während es sich bei der 2. Auflage von 2012 um eine Aktualisierung und Überarbeitung der 1. Auflage von 2008 unter Beibehaltung der Grundstruktur gehandelt hatte, war es in der hier vorgelegten vollständig neu bearbeiteten, erweiterten und aktualisierten 3. Auflage nun möglich, eine noch klarere sowie stärker modularisierte Struktur zu verfolgen, um nicht nur innerhalb der bisherigen Kapitel und Themengebiete größere Aktualisierungen und Veränderungen vorzunehmen, sondern auch vollständig neu konzipierte Kapitel und Inhalte zu ergänzen, wie im einleitenden Kapitel (► Kap. 1) dargelegt wird.

Im Ergebnis haben wir das zuvor zweiteilige Lehrbuch in nunmehr drei große Abschnitte unterteilt, die mit Teil I „Konstruktionsgesichtspunkte“, Teil II „Testtheorien“ sowie Teil III „Validität und Möglichkeiten ihrer Überprüfung“ überschrieben sind. Das Lehrbuch umfasst jetzt 27 Kapitel, wobei – auch aus didaktischen Gründen – darauf geachtet wurde, dass die einzelnen Kapitel nicht zu lang geraten, sondern sich auf spezifische Themen konzentrieren.

Teil I „Konstruktionsgesichtspunkte“ (► Kap. 2–11) ist so konzipiert, dass er von einem möglichst breiten Adressatenkreis gewinnbringend gelesen werden kann. Er befasst sich mit jenen „Gütekriterien“ und deren Umsetzung, die sowohl für die traditionelle als auch für die computerbasierte Konstruktion von Fragebogen und Tests als Grundlage dienen. Auch werden in diesem Teil handlungspraktische Hinweise zur Test- und Fragebogenkonstruktion gegeben, für die keine tiefergehenden testtheoretischen Kenntnisse erforderlich sind; vielmehr reichen basale Grundkenntnisse in Deskriptivstatistik aus, um ein sicheres Verständnis zu erreichen.

Teil II „Testtheorien“ (► Kap. 12–20) ist den bedeutenden testtheoretischen Ansätzen der „Klassischen Testtheorie (KTT)“ und der „Item-Response-Theorie (IRT)“ gewidmet, mit denen es möglich wird, nicht direkt beobachtbare, sog. „latente“ Konstrukte zu erfassen. Hierbei wird einerseits gezeigt, wie stark sich die KTT unter Einbeziehung faktorenanalytischer Messmodelle weiterentwickeln konnte und welche Konkretisierungen der Reliabilitätschätzungen damit einhergehen. Auch in die IRT wird nicht nur elementar eingeführt, sondern es wird darüber hinaus ein profunder Überblick über die in der Measurementforschung relevanten Modelle der IRT sowie über die zugehörigen Methoden der Parameterschätzung gegeben, und durch eine Einführung in das computerisierte adaptive Testen abgerundet.

Teil III „Validität und Möglichkeiten ihrer Überprüfung“ (► Kap. 21–27) legt den Fokus auf die Validität von Testwertinterpretationen und thematisiert damit insbesondere die Angemessenheit und Belastbarkeit von testergebnisgestützten Schlussfolgerungen. Die aufgeführten Verfahrensweisen und Methoden bieten fragestellungsorientierte empirische Überprüfungsmöglichkeiten, wobei die exploratorische (EFA) und vor allem die konfirmatorische Faktorenanalyse (CFA) elaborierter als bisher behandelt werden. Die Kapitel über Multitrait-Multimethod-Analysen (MTMM-Analysen), Latent-State-Trait-Theorie (LST-Theorie) sowie deren Integration zur Überprüfung der konvergenten und diskriminanten Validität über die Zeit bilden den Abschluss.

Vorwort zur 3. Auflage

Auch in dieser neuen, 3. Auflage haben wir versucht, dem bereits unserer 1. Auflage vorangestellten Motto

➤ So fundiert wie notwendig, aber so einfach wie möglich!

treu zu bleiben, obschon mit den Aktualisierungen, Erweiterungen und Vertiefungen, die den Weiterentwicklungen der Testtheorie und ihrer Bedeutung für stark expandierende Anwendungsfelder geschuldet sind, höhere Anforderungen an unsere geneigten Leser einhergehen. Auch weiterhin sehen wir uns als Herausgeber gemeinsam mit unseren Autorinnen und Autoren einer größtmöglichen Verständlichkeit unter Beibehaltung der erforderlichen Präzision verpflichtet. Gegenüber der 2. Auflage konnte das zuvor 19-köpfigen Autorenteam auf 28 Expertinnen und Experten erweitert werden, deren Wirkungsstätten weit gestreut sind (Aachen, Berlin, Frankfurt, Genf, Heidelberg, Jena, Kiel, Konstanz, Ludwigsburg, Mainz, München, Potsdam, Pullman, Nashville, Tempe, Tübingen, Zürich). Mit ihrem hervorragenden Kenntnisstand haben sie in nachhaltiger Weise zur Qualitätssicherung und zur Weiterentwicklung unseres Lehrbuches beigetragen.

Mit dem Erscheinen der 3. Auflage sind wir unseren Autorinnen und Autoren nicht nur für ihre zügige Hereingabe der Kapitel und für ihre Flexibilität bei der Anpassung ihrer Texte an den Gesamtduktus des Lehrbuches, sondern auch für zahlreiche Peer-Reviews untereinander sehr zu Dank verpflichtet. Unsere Herausgebertätigkeit wurde nachhaltig von Frau Prof. Karin Schermelleh-Engel und Frau M. Sc. Susanna Suchan (bis 2017) unterstützt. Besonders hervorzuheben ist das abschließende Lektorat, das von Frau Dipl.-Biol. Stefanie Teichert mit der gebotenen Umsicht und Akribie vorgenommen wurde. Die grundlegende Planung unserer 3. Auflage erfolgte beim Springer-Verlag in sehr vertrauensvoller, bewährter Weise mit Herrn Dipl.-Psych. Joachim Coch (Senior Editor), die Projektbetreuung lag in den kundigen Händen von Frau M. A. Judith Danziger (Project Manager). Ihnen allen sei für die stets exzellente und effektive Zusammenarbeit vielmals gedankt!

Als Herausgeber hoffen wir gemeinsam mit allen unseren Autorinnen und Autoren, dass unsere 3. Auflage eine weiterhin anhaltende wohlwollende Resonanz findet und dass die hier zusammengestellten Inhalte für die wissenschaftlichen Fundierung und die daraus resultierende Qualität von Test- und Fragebogenentwicklungen von großem Nutzen sein mögen.

Helfried Moosbrugger

Augustin Kelava

Frankfurt am Main und Tübingen

im Frühjahr 2020

Vorwort zur 2. Auflage

Es ist eine große Freude für die Autoren und Herausgeber, wenn es gelingt, mit einem neuen Lehrbuch trotz starker Konkurrenz einen großen Leserkreis anzusprechen. Diese sehr positive Entwicklung manifestiert sich darin, dass nach weniger als vier Jahren die 1. Auflage unserer *Testtheorie und Fragebogenkonstruktion* nun aufgebraucht ist und eine Neuauflage notwendig wurde.

Da sich die Gesamtstruktur unseres Buches sehr bewährt hatte und die Inhalte den Themenbereich der Testtheorie und Fragebogenkonstruktion auch weiterhin gut abdecken, konnte die Kapitelgliederung der 1. Auflage vollständig beibehalten werden. Von allen Autoren wurden aber entsprechende Aktualisierungen und Ergänzungen vorgenommen, um den jüngsten Entwicklungen gerecht zu werden. Dabei blieben wir auch weiterhin dem Motto unseres Werkes „So fundiert wie notwendig, aber so einfach wie möglich“ (s. Vorwort zur 1. Auflage) kompromisslos treu. Eine breitere Fundierung erwies sich nur in wenigen Bereichen als notwendig; sehr wohl wurde aber von Seiten der Autoren und insbesondere auch von den Herausgebern selbst viel Feinarbeit in eine noch stringenter Detailgliederung und in eine Fülle größerer und kleiner Textpräzisierungen investiert, um unseren Leserinnen und Lesern ein noch leichteres und sichereres Verständnis der Inhalte unseres Lehrbuches zu ermöglichen.

Neben der Korrektur der äußerst geringen Anzahl von Fehlern in der 1. Auflage und den notwendigen Aktualisierungen und Überarbeitungen sind es also vor allem didaktische Verbesserungen und Ergänzungen, die als konkrete Neuerungen der 2. Auflage explizit aufzuführen sind, nämlich

- die Beseitigung von textlichen Unstimmigkeiten, Zweideutigkeiten und anderen Unklarheiten, sofern solche aufgetreten waren,
- die noch konsequentere Verwendung von Marginalien zur Kurzkennzeichnung und leichteren Fassbarkeit der Absatzinhalte,
- die Formulierung von „Kontrollfragen“ am Ende jedes Kapitels als Lern- und Verständnishilfe sowie von dazugehörigen Antworten im Lerncenter unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion),
- die Ergänzung von „EDV-Hinweisen“ zu den einzelnen Kapiteln mit themenspezifischen Informationen zur Nutzung von einschlägigen Softwarepaketen wie SPSS, LISREL und weiteren Auswertungsprogrammen gegeben werden;
- für zahlreiche Kapitel die Bereitstellung von konkreten rechentechnisch aufbereiteten Datenbeispielen und ausführlich dokumentierten Ergebnisdarstellungen im Lerncenter unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion) sowie
- die Überarbeitung des Glossars der wichtigsten Schlüsselbegriffe und die Erweiterung des Sachregisters.

Dem Verlagshaus Springer in Heidelberg gilt es erneut für die hervorragende Zusammenarbeit vielmals zu danken. Insbesondere hat das starke Engagement von Senior Editor Dipl.-Psych. Joachim Coch, auf eine noch bessere Marktposition unseres Lehrbuches hinzuarbeiten, dazu geführt, dass bei der Überarbeitung zur 2. Auflage nicht auf Kostenminimierung, sondern bis zum endgültigen Redaktionsschluss vor allem auf qualitätssichernde Maßnahmen geachtet werden konnte. Die zahlreichen dadurch möglich gewordenen Korrekturzyklen haben die Geduld von Herrn Michael Barton, dem verlagsseitigen Projektmanager, reichlich strapaziert; für sein dennoch stets spürbares Wohlwollen sei ihm an dieser Stelle herzlich gedankt. Auch Frau Dr. Christiane Grosser hat mit ihrem routinierten Copy-Editing beträchtlich zum guten Gelingen des Werkes beigetragen.

Vorwort zur 2. Auflage

Besonders in der Detailarbeit der neuen Auflage waren uns einige herausragende Studierenden und Mitarbeiter sehr behilflich. Namentlich alphabetisch genannt seien aus dem Frankfurter und dem Darmstädter Institut Julia Engel, Sonja Etzler, Daniel Köth, Hannah Nagler, Susanne Penger, Dipl.-Psych. Marlene Schmidt und Dipl.-Psych. Michael Weigand. Ihnen allen gilt unser herzlicher Dank.

Abschließend sei noch eine für ein Vorwort eher ungewöhnliche, aber bemerkenswerte Betrachtung angestellt: Zu unserer sorgfältig ausgewählten Autorengruppe zählten zum Zeitpunkt der Fertigstellung der 1. Auflage 5 Professoren, 1 Habilitierte, 4 Promovierte, 8 Diplomierte und („unter Mitarbeit von“) 1 Studierender (vgl. Autorenverzeichnis 2007). Die Autoren blieben dieselben, bezüglich ihres akademischen Avancements ist aber bei Fertigstellung der 2. Auflage eine eindrucksvolle Steigerung zu verzeichnen: 11 Professoren, 6 Promovierte und 2 Diplomierte (vgl. Autorenverzeichnis 2011). Hieraus ließe sich die erfreuliche Schlussfolgerung ziehen, dass die Autoren nicht nur entscheidend zum Erfolg des Buches beigetragen haben, sondern dass auch das erfolgreiche Buch einen begünstigenden Effekt auf ihre akademischen Karrieren ausgeübt haben könnte.

Möge unserem Lehrbuch *Testtheorie und Fragebogenkonstruktion* auch in der aktualisierten und überarbeiteten 2. Auflage die Gunst unserer Leserinnen und Leser erhalten bleiben, damit möglichst viele Adressaten aus dem Inhalt des Buches den entsprechenden Nutzen ziehen können.

Helfried Moosbrugger

Augustin Kelava

Gargnano sul Garda und Darmstadt

im August 2011

Vorwort zur 1. Auflage

➤ So fundiert wie notwendig, aber so einfach wie möglich!

Unter dieser Maxime schien es Herausgebern und Verlag gleichermaßen lohnend, ein anspruchsvolles Lehrbuch zum Bereich Testtheorie und Fragebogenkonstruktion auf den Markt zu bringen, obwohl zu diesem Thema bereits eine Reihe von Texten mit unterschiedlichen Stärken und Schwächen existieren.

Der mit dem neuen Lehrbuch verfolgte Anspruch möchte jede Überbetonung einer zu spezifischen Sichtweise vermeiden, wie sie bei existierenden Texten z. B. in theoretisch-formalisierte oder in anwendungsorientierte, in statistisch-probablistischer oder auch in computerprogrammlastiger Hinsicht vorgefunden werden kann. Vielmehr sollte ein Werk entstehen, das den theoretischen und anwendungspraktischen Kenntnisstand des Themenbereiches unter Beachtung von leser- und lernfreundlichen Aspekten in inhaltlich ausgewogener Form dokumentiert, und zwar in zwei Teilen, die sich an den neuen Bachelor- und Masterstudiengängen orientieren.

Die „Grundlagen“ (Teil A) richten sich an Leser ohne besondere Vorkenntnisse und befassen sich in acht Kapiteln mit den Inhalten, die typischerweise im Bachelorstudiengang Psychologie behandelt werden sollten: Qualitätsanforderungen, Planung und Entwurf von Tests, Itemanalyse, Klassische Testtheorie, Reliabilität, Validität, Normierung und Testwertinterpretation sowie Teststandards.

Die „Erweiterungen“ (Teil B) richten sich an fortgeschrittene Leser und befassen sich in weiteren sieben Kapiteln mit Inhalten, die typischerweise in psychologischen Masterstudiengängen von Bedeutung sind: Item-Response-Theorie, Adaptives Testen, Latent-Class-Analyse, Exploratorische und Konfirmatorische Faktorenanalyse, Multitrait-Multimethod-Analysen, Latent-State-Trait-Theorie sowie Kombination von MTMM- und LST-Analysen.

Mit gutem Grund wurden die zahlreichen Beispiele, mit denen die spezifischen Themenstellungen durchgängig illustriert sind, weitgehend dem Bereich der Psychologie entnommen. Neben einem soliden Grundverständnis sollen sie dem Leser auch das nötige Wissen für die Einordnung und Interpretation aktueller Forschungsergebnisse vermitteln, so z. B. von internationalen Large-Scale-Assessments (PISA-Studien etc.). Insgesamt wurde die jeweilige Gestaltung so gewählt, dass ein Transfer des „State-of-the-Art“-Wissens auch auf andere inhaltliche Disziplinen jederzeit gelingen sollte. Zu nennen wären hier z. B. die Wirtschaftswissenschaften, in deren Reihen ein zunehmender Bedarf nach einschlägigen Kenntnissen qualitätvoller Fragebogenkonstruktion festzustellen ist. Erwähnt sei hier, dass insbesondere die Grundlagenkapitel 1 bis 3 und 9 auch ohne jede testtheoretische und mathematische Vertiefung gewinnbringend gelesen werden können.

Um eine anspruchsvolle didaktische Sorgfalt zu erzielen, musste kapitelübergreifend auf eine möglichst einheitliche Darstellungsform geachtet werden. Dieser bei einem Herausgeberwerk keinesfalls selbstverständliche Qualitätsaspekt konnte vor allem dadurch sichergestellt werden, dass in der Universität Frankfurt am Main und ihrem engeren Umfeld viele kompetente Autorinnen und Autoren mit einschlägiger Lehrerfahrung und Expertise für die einzelnen Kapitel gewonnen werden konnten. Die räumliche Nähe verkürzte den zeitlichen Aufwand der Peer-Reviewings beträchtlich und ermöglichte es den Herausgebern, die Autoren auch mehrmals um Einarbeitungen von Monita zu bitten, was sich letztlich sehr vorteilhaft auf die Darstellungsqualität und die Homogenisierung der verschiedenen Texte auswirkte. Darauf hinaus haben zur Darstellung der erst kürzlich gelungenen Verschränkung von LST- und MTMM-Modellen (► Kap. 16) auch internationale Autoren beigetragen.

Vorwort zur 1. Auflage

Besonders hervorzuheben ist, dass im Zuge der Qualitätskontrolle nicht nur Peers um ihre Meinung gebeten wurden, sondern auch die wichtigste Zielgruppe des Lehrbuches selbst, nämlich die Studierenden: In einem einwöchigen Intensivseminar in den österreichischen Alpen wurden unter Leitung der Herausgeber die einzelnen Kapitel grundlegend diskutiert und von den Studierenden mit Hilfe eines ausführlichen Qualitätssicherungsschemas kritisch-konstruktiv evaluiert¹. Problematische Stellen und festgestellte Mängel wurden den Autoren rechtzeitig rückgemeldet, so dass die monierten Punkte noch vor Drucklegung bereinigt werden konnten.

All diese Maßnahmen haben dazu beigetragen, dass ein sehr sorgfältig gegliedertes Lehrbuch entstanden ist, das zahlreiche didaktisch-visualisierende Gestaltungshilfen aufweist. Dazu zählen Tabellen, Abbildungen, hervorgehobene Definitionen, Beispiele in Kästen, Marginalien zum besseren Verständnis der Absatzgliederungen, gekennzeichnete Exkurse, Rubriken „Kritisch hinterfragt“, präzise Siehe-Verweise auf andere Kapitel u. v. m. Außerdem enthält das Lehrbuch Zusammenfassungen am jeweiligen Kapitelende, ein ausführliches Sachverzeichnis und ein eigens zusammengestelltes Glossar, das in besonderer Weise auf Leserfreundlichkeit bedacht ist. Zusätzlich zum gedruckten Buch ist auch eine Lernwebsite entstanden: Kapitelzusammenfassungen, virtuelle Lernkarten, Glossarbegriffe und weiterführende Links finden sich unter www.lehrbuch-psychologie.de. Dies alles wäre ohne eine geeignete und geduldige Unterstützung von Seiten des Verlagshauses Springer in Heidelberg nicht möglich gewesen, weshalb der für den Bereich zuständigen Senior-Lektorin, Frau Dr. Svenja Wahl sowie Herrn Michael Barton, der mit der Koordination der Drucklegung befasst war, an dieser Stelle vielmals zu danken ist.

Für die termingerechte Fertigstellung gilt es aber noch weitere Danksagungen auszusprechen:

In erster Linie sind hier die Autorinnen und Autoren zu nennen, die sich trotz ihrer vielfältigen beruflichen Verpflichtungen die Zeit genommen haben, anspruchsvolle Kapitel zu erstellen und gemäß den Wünschen der Herausgeber an das Gesamtwerk anzupassen. Eine nicht zu unterschätzende Leistung wurde auch von unseren studentischen Mitarbeiterinnen und Mitarbeitern erbracht, die sich bei der Literaturarbeit, beim Korrekturlesen, beim Einarbeiten von Textpassagen, beim Vorbereiten der Register und des Glossars, beim Bearbeiten von Abbildungen sowie z. T. auch durch die Formulierung eigener Verbesserungsvorschläge große Verdienste erworben haben. Namentlich seien hier Nadine Malstädt, Benjamin Peters und Yvonne Pfaff genannt, sowie hervorgehoben für ihr besonderes Engagement in der Schlussphase Constanze Rickmeyer, Merle Steinwascher und Michael Weigand.

Allen zusammen gebührt unser herzlicher Dank für ihre Beiträge zum Gelingen dieses schönen Werkes.

Helfried Moosbrugger

Augustin Kelava

Kronberg im Taunus und Frankfurt am Main

Sommer 2007

¹ Für dieses besondere Engagement gilt es neben den mitveranstaltenden Kollegen Prof. Dirk Frank, Dr. Siegbert Reiß und Dipl.-Psych. Wolfgang Rauch folgenden Studierenden des Diplom-Studienganges Psychologie der J. W. Goethe-Universität Frankfurt am Main sehr herzlich zu danken: Miriam Borgmann, Stephan Braun, Felix Brinkert, Bronwen Davey, Ulrike Fehsenfeld, Tim Kaufhold, Anna-Franziska Lauer, Marie Lauer-Schmaltz, Johanna Luu, Dorothea Mildner, Florina Popeanga, Zoe Schmitt, Nadja Weber und Michael Weigand.

Inhaltsverzeichnis

1	Einführung und zusammenfassender Überblick	1
	<i>Helfried Moosbrugger und Augustin Kelava</i>	
1.1	Zielgruppen und Gliederungsüberlegungen.....	2
1.2	Teil I: Konstruktionsgesichtspunkte.....	3
1.3	Teil II: Testtheorien	5
1.4	Teil III: Validität und Möglichkeiten ihrer Überprüfung	8
1.5	Ergänzende Materialien	10
1.6	Zusammenfassung	10
I	Konstruktionsgesichtspunkte	
2	Qualitätsanforderungen an Tests und Fragebogen ("Gütekriterien")	13
	<i>Helfried Moosbrugger und Augustin Kelava</i>	
2.1	Vom Laienfragebogen zum wissenschaftlichen Messinstrument	15
2.2	Unterschiedliche Qualitätsanforderungen	16
2.3	Allgemeine Gütekriterien für Tests und Fragebogen	17
2.4	Spezielle testtheoriebasierte Gütekriterien für wissenschaftliche Tests und Fragebogen.....	27
2.5	Dokumentation der erfüllten Qualitätskriterien	36
2.6	Zusammenfassung	36
2.7	Kontrollfragen	36
	Literatur	37
3	Planungsaspekte und Konstruktionsphasen von Tests und Fragebogen	39
	<i>Holger Brandt und Helfried Moosbrugger</i>	
3.1	Spezifikation des interessierenden Merkmals	41
3.2	Testarten	44
3.3	Geltungsbereich und Zielgruppe	50
3.4	Testlänge und Testzeit	51
3.5	Testadministration	53
3.6	Struktureller Testaufbau	55
3.7	Konstruktionsphasen im Überblick	57
3.8	Zusammenfassung	63
3.9	Kontrollfragen	63
	Literatur	64
4	Itemkonstruktion und Antwortverhalten	67
	<i>Helfried Moosbrugger und Holger Brandt</i>	
4.1	Ziele und Aspekte der Itemkonstruktion	69
4.2	Itemstamm und Zielgruppe	69
4.3	Vorgehensweisen bei der Itemgenerierung	71
4.4	Kategorisierung von Frageformen	73
4.5	Gesichtspunkte der Itemformulierung	75
4.6	Kognitive und motivationale Prozesse bei der Itembearbeitung	78
4.7	Response-Bias als Fehlerquelle beim Antwortverhalten	81
4.8	Computerunterstützte Itemkonstruktion	86
4.9	Zusammenfassung	86
4.10	Kontrollfragen	87
	Literatur	87

5	Antwortformate und Itemtypen	91
	<i>Helfried Moosbrugger und Holger Brandt</i>	
5.1	Antwortformate im Überblick	93
5.2	Aufgaben mit freiem Antwortformat	94
5.3	Aufgaben mit gebundenem Antwortformat	96
5.4	Aufgaben mit atypischem Antwortformat	112
5.5	Entscheidungshilfen für die Wahl des Aufgabentyps	114
5.6	Computerunterstützte Antwortformate	114
5.7	Zusammenfassung	115
5.8	Kontrollfragen	115
	Literatur	115
6	Computerbasiertes Assessment	119
	<i>Frank Goldhammer und Ulf Kröhne</i>	
6.1	Computerbasiertes Assessment: Definition und Übersicht	121
6.2	Itementwicklung: Antwortformat, Stimulus und Antwortbewertung	124
6.3	Testentwicklung: Testzusammenstellung und -sequenzierung	130
6.4	Testadministration	132
6.5	Datenanalyse und Rückmeldung	135
6.6	Zusammenfassung	137
6.7	EDV-Hinweise	137
6.8	Kontrollfragen	138
	Literatur	138
7	Deskriptivstatistische Itemanalyse und Testwertbestimmung	143
	<i>Augustin Kelava und Helfried Moosbrugger</i>	
7.1	Einleitung	145
7.2	Erstellung der Datenmatrix	145
7.3	Schwierigkeitsanalyse	146
7.4	Itemvarianz	151
7.5	Vorläufige Testwertermittlung	153
7.6	Trennschärfe	153
7.7	Itemselektion auf Basis von Itemschwierigkeit, Itemvarianz und Itemtrennschärfe	155
7.8	Testwertbestimmung und Itemhomogenität	156
7.9	Zusammenfassung	157
7.10	EDV-Hinweise	157
7.11	Kontrollfragen	158
	Literatur	158
8	Testwertverteilung	159
	<i>Augustin Kelava und Helfried Moosbrugger</i>	
8.1	Einleitung	160
8.2	Zentrale Tendenz der Testverteilung	160
8.3	Streuung der Testwertverteilung	161
8.4	Beurteilung der Verteilungsform	161
8.5	Ursachen für die Abweichung der Testwertverteilung von der Normalverteilung	163
8.6	Normalisierung der Testwertverteilung	164
8.7	Zusammenfassung und weiteres Vorgehen	168
8.8	EDV-Hinweise	168
8.9	Kontrollfragen	168
	Literatur	168

9	Testwertinterpretation, Testnormen und Testeichung	171
	<i>Frank Goldhammer und Johannes Hartig</i>	
9.1	Testwertbildung und Testwertinterpretation	172
9.2	Normorientierte Testwertinterpretation	173
9.3	Kriteriumsorientierte Testwertinterpretation	179
9.4	Integration von norm- und kriteriumsorientierter Testwertinterpretation ...	187
9.5	Normdifferenzierung	188
9.6	Testeichung	189
9.7	Zusammenfassung mit Anwendungsempfehlungen	193
9.8	EDV-Hinweise	194
9.9	Kontrollfragen	194
	Literatur	195
10	Standards für psychologisches Testen	197
	<i>Volkmar Höfling und Helfried Moosbrugger</i>	
10.1	Ziele von Teststandards	198
10.2	Standards für die Entwicklung und Evaluation psychologischer Tests	198
10.3	Standards für die Übersetzung und Anpassung psychologischer Tests	203
10.4	Standards für die Anwendung psychologischer Tests	204
10.5	Standards für die Qualitätsbeurteilung psychologischer Tests	210
10.6	Zusammenfassung	213
10.7	Kontrollfragen	213
	Literatur	213
11	Standards für pädagogisches Testen	217
	<i>Sebastian Brückner, Olga Zlatkin-Troitschanskaia und Hans Anand Pant</i>	
11.1	Die „Standards for Educational and Psychological Testing“ im Überblick	219
11.2	Domänen, Ziele und Designs pädagogischen Testens	220
11.3	Validitätsstandards und pädagogisches Testen (Standards 1.0–1.25)	229
11.4	Standards zur Reliabilität (Standards 2.10–2.20)	234
11.5	Schwellenwerte und ihre Bedeutung für die Testwertinterpretation	234
11.6	Weitere Implikationen der Standards für pädagogisches Testen	238
11.7	Standards zum Management und zur Archivierung von Daten pädagogischen Testens	242
11.8	Standards für Forschungsethik	244
11.9	Zusammenfassung	245
11.10	Kontrollfragen	245
	Literatur	245

II Testtheorien

12	Testtheorien im Überblick	251
	<i>Helfried Moosbrugger, Karin Schermelleh-Engel, Jana C. Gäde und Augustin Kelava</i>	
12.1	Einleitung	252
12.2	Klassische Testtheorie (KTT)	254
12.3	Item-Response-Theorie (IRT)	260
12.4	Klassische Testtheorie (KTT) vs. Item-Response-Theorie (IRT)	268
12.5	Zusammenfassung	271
12.6	Kontrollfragen	271
	Literatur	271

13	Klassische Testtheorie (KTT)	275
	<i>Helfried Moosbrugger, Jana C. Gäde, Karin Schermelleh-Engel und Wolfgang Rauch</i>	
13.1	Einleitung	277
13.2	Grundannahmen der KTT	277
13.3	Zerlegung einer Itemvariablen in True-Score- und Messfehlervariable	279
13.4	Testwertvariable Y und Testwerte Y_v	280
13.5	Das Gütekriterium der Reliabilität	281
13.6	Messmodelle zur Schätzung der Reliabilität	282
13.7	Empirisches Beispiel	291
13.8	Schätzung individueller Merkmalsausprägungen	294
13.9	Erweiterung der KTT	298
13.10	Zusammenfassung	302
13.11	EDV-Hinweise	302
13.12	Kontrollfragen	302
	Literatur	303
14	Klassische Methoden der Reliabilitätsschätzung	305
	<i>Jana C. Gäde, Karin Schermelleh-Engel und Christina S. Werner</i>	
14.1	Was ist Reliabilität?	307
14.2	Grundlagen	309
14.3	Cronbachs Alpha	314
14.4	Test-Test-Korrelation	322
14.5	Vergleichbarkeit der Reliabilitätsmaße	329
14.6	Einflüsse auf die Reliabilität	330
14.7	Anzustrebende Höhe der Reliabilität	330
14.8	Auswahl eines geeigneten Reliabilitätsmaßes	331
14.9	Zusammenfassung	332
14.10	EDV-Hinweise	333
14.11	Kontrollfragen	333
	Literatur	333
15	Modellbasierte Methoden der Reliabilitätsschätzung	335
	<i>Karin Schermelleh-Engel und Jana C. Gäde</i>	
15.1	Klassische vs. modellbasierte Reliabilitätsschätzung	337
15.2	Eindimensionale Modelle	339
15.3	Mehrdimensionale Modelle	350
15.4	Omega-Koeffizienten im Rahmen weiterer Faktormodelle	360
15.5	Bewertung der modellbasierten Reliabilitätsschätzung	361
15.6	Reliabilitätsschätzung ordinalskalierter Variablen	363
15.7	Erste Empfehlungen zur Beurteilung der Omega-Koeffizienten	364
15.8	Zusammenfassung	365
15.9	EDV-Hinweise	366
15.10	Kontrollfragen	366
	Literatur	366
16	Einführung in die Item-Response-Theorie (IRT)	369
	<i>Augustin Kelava und Helfried Moosbrugger</i>	
16.1	Grundüberlegungen zur IRT	371
16.2	Latent-Trait-Modelle	372
16.3	Dichotomes Rasch-Modell (1PL-Modell)	373
16.4	2PL-Modell nach Birnbaum	399
16.5	3PL-Modell nach Birnbaum	401
16.6	Weitere IRT-Modelle	402
16.7	Zusammenfassung	406

Inhaltsverzeichnis

16.8	EDV-Hinweise	407
16.9	Kontrollfragen	407
	Literatur	407
17	Interpretation von Testwerten in der Item-Response-Theorie (IRT) <i>Dominique Rauch und Johannes Hartig</i>	411
17.1	Vorbemerkungen	412
17.2	Grundlagen kriteriumsorientierter Testwertinterpretation in IRT-Modellen	414
17.3	Definition von Kompetenzniveaus zur kriteriumsorientierten Testwertinterpretation	417
17.4	Verwendung von Post-hoc-Analysen und A-priori-Merkmalen zur Testwertbeschreibung	418
17.5	Zusammenfassung	422
17.6	EDV-Hinweise	423
17.7	Kontrollfragen	423
	Literatur	423
18	Überblick über Modelle der Item-Response-Theorie (IRT) <i>Augustin Kelava, Stefano Noventa und Alexander Robitzsch</i>	425
18.1	Modelle mit eindimensionalen latenten Merkmalen	426
18.2	Modelle mit mehrdimensionalen latenten Merkmalen	438
18.3	Ausblick auf weitere Modelle	443
18.4	Weiterführende Literatur	444
18.5	EDV-Hinweise	444
18.6	Kontrollfragen	445
	Literatur	445
19	Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT) <i>Norman Rose</i>	447
19.1	Verfahren der Parameterschätzung in der IRT: Überblick	449
19.2	Maximum-Likelihood-Schätzung (ML-Schätzung)	450
19.3	Bayes'sche Schätzverfahren	466
19.4	Weitere Schätzverfahren	483
19.5	Personenparameterschätzung in der IRT	484
19.6	Reliabilitätsbeurteilung in der IRT	490
19.7	Zusammenfassung	496
19.8	EDV-Hinweise	497
19.9	Kontrollfragen	498
	Literatur	499
20	Computerisiertes adaptives Testen <i>Andreas Frey</i>	501
20.1	Was ist computerisiertes adaptives Testen?	502
20.2	Grundgedanke	503
20.3	Elementare Bausteine	506
20.4	Auswirkungen des adaptiven Testens	516
20.5	Multidimensionales adaptives Testen	520
20.6	Zusammenfassung und Anwendungsempfehlungen	521
20.7	EDV-Hinweise	522
20.8	Kontrollfragen	522
	Literatur	523

III Validität und Möglichkeiten ihrer Überprüfung

21	Validität von Testwertinterpretationen	529
	<i>Johannes Hartig, Andreas Frey und Nina Jude</i>	
21.1	Einleitung	530
21.2	Validität im fachgeschichtlichen Wandel	530
21.3	Argumentationsbasierter Ansatz der Validierung	535
21.4	Beispiele für Validierungsprozesse	539
21.5	Zusammenfassung	544
21.6	Kontrollfragen	544
	Literatur	544
22	Latent-Class-Analyse (LCA)	547
	<i>Mario Gollwitzer</i>	
22.1	Einleitung und Überblick	549
22.2	Herleitung der Modellgleichung	552
22.3	Parameterschätzung und Überprüfung der Modellgüte	556
22.4	Exploratorische und konfirmatorische Anwendungen der LCA	562
22.5	Erweiterte Anwendungen der LCA	567
22.6	Zusammenfassung	571
22.7	EDV-Hinweise	572
22.8	Kontrollfragen	572
	Literatur	573
23	Exploratorische Faktorenanalyse (EFA)	575
	<i>Holger Brandt</i>	
23.1	Einleitung	577
23.2	Faktormodell (Fundamentaltheorem)	578
23.3	Methoden der Faktorextraktion	585
23.4	Abbruchkriterien der Faktorextraktion	590
23.5	Faktorenrotation	595
23.6	Modellevaluation und Itemauswahl	604
23.7	Neue Verfahren	608
23.8	Abschließende Bemerkungen	610
23.9	Zusammenfassung	611
23.10	EDV-Hinweise	611
23.11	Kontrollfragen	611
	Literatur	612
24	Konfirmatorische Faktorenanalyse (CFA)	615
	<i>Jana C. Gäde, Karin Schermelleh-Engel und Holger Brandt</i>	
24.1	Grundlagen	617
24.2	Spezifikation eines Messmodells	619
24.3	Eindimensionale Modelle: Stufen der Messäquivalenz	629
24.4	Mehrdimensionale Modelle	634
24.5	Parameterschätzung	643
24.6	Modellevaluation	648
24.7	Modifikation der Modellstruktur	651
24.8	Modellvergleiche	652
24.9	Messinvarianztestung	653
24.10	Zusammenfassung	656
24.11	EDV-Hinweise	656
24.12	Kontrollfragen	656
	Literatur	657

25 Multitrait-Multimethod-Analysen (MTMM-Analysen)	661
<i>Karin Schermelleh-Engel, Christian Geiser und G. Leonard Burns</i>	
25.1 Einleitung	663
25.2 Konvergente und diskriminante Validität	663
25.3 Methodeneffekte	664
25.4 Das MTMM-Design	666
25.5 Korrelationsbasierte Analyse der MTMM-Matrix	669
25.6 Faktorenanalytische Ansätze: Klassische CFA-MTMM-Modelle	672
25.7 Faktorenanalytische Ansätze: Neuere CFA-MTMM-Modelle	678
25.8 Zusammenfassung	683
25.9 EDV-Hinweise	684
25.10 Kontrollfragen	684
Literatur	684
26 Latent-State-Trait-Theorie (LST-Theorie)	687
<i>Augustin Kelava, Karin Schermelleh-Engel und Axel Mayer</i>	
26.1 Einleitung	688
26.2 LST-Theorie als Erweiterung der KTT	692
26.3 Modelltypen	697
26.4 Anwendungen der LST-Theorie	703
26.5 Zusammenfassung	708
26.6 EDV-Hinweise	709
26.7 Kontrollfragen	709
Literatur	709
27 Konvergente und diskriminante Validität über die Zeit: Integration von Multitrait-Multimethod-Modellen (MTMM-Modellen) und der Latent-State-Trait-Theorie (LST-Theorie)	713
<i>Fridtjof W. Nussbeck, Michael Eid, Christian Geiser, Delphine S. Courvoisier und David A. Cole</i>	
27.1 Einleitung	715
27.2 Längsschnittliche MTMM-Modelle	721
27.3 Multiconstruct-LST- und Multimethod-LST-Modell in der empirischen Anwendung	730
27.4 Praktische Hinweise zur Analyse longitudinaler multimodaler Modelle	735
27.5 Zusammenfassung	736
27.6 EDV-Hinweise	736
27.7 Kontrollfragen	737
Literatur	737
Serviceteil	739
Übersicht der griechischen Buchstaben	740
Verteilungsfunktion der Standardnormalverteilung (z-Tabelle)	741
Glossar	744
Stichwortverzeichnis	761

Herausgeber- und Autorenverzeichnis

Über die Herausgeber



Univ.-Prof. Dr. Helfried Moosbrugger

geb. 1944, Studium in Graz, Marburg und Innsbruck, Promotion 1969. Professor für Psychologie seit 1977 an der Goethe-Universität Frankfurt am Main. Arbeitsschwerpunkte: Psychologische Forschungsmethoden, Evaluation, Diagnostik und Differentielle Psychologie. Zahlreiche Lehrbücher, Buchkapitel und Zeitschriftenveröffentlichungen zu spezifischen statistischen und diagnostischen Verfahren, zur berufsbezogenen Eignungsbeurteilung und zur Testkonstruktion und Testtheorie. Autor von: Frankfurter Adaptiver Konzentrationsleistungs-Test FAKT-II (auch englisch, koreanisch), Frankfurter Aufmerksamkeits-Inventar FAIR-2 (auch koreanisch), Fragebogen zur Evaluation der Lehre FEL (universitätsweit eingesetzt), Online-Self-Assessment OSA für den Studiengang Psychologie. Vorsitzender (2003–2011) des Testkuratoriums der Föderation deutscher Psychologenvereinigungen (DGPs und BDP), Entwicklung des Testbeurteilungssystems TBS-TK. Emeritus und Seniorprofessor 2012. Hobbys: Skilehrer, Parlamentarier, Oper, Mountainbiking.



Univ.-Prof. Dr. Augustin Kelava

geb. 1979, Studium der Psychologie in Frankfurt, 2004 Diplomprüfung, 2009 Promotion bei Prof. Helfried Moosbrugger am Institut für Psychologie der Goethe-Universität Frankfurt, zwischen 2011 und 2013 Juniorprofessor an der Technischen Universität Darmstadt, seit 2013 Professor an der Universität Tübingen. Nach einem renommierten Auslandsruf 2018 Gründungsdirektor eines neuen Instituts, des „Methodenzentrums“, an der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität Tübingen. Arbeitsschwerpunkte und internationale Publikationen in den Bereichen: Latente Variablenmodelle (inklusive nicht-lineare Effekte, semi- und nichtparametrische Methoden, längsschnittliche dynamische Mehrebenen-Strukturgleichungsmodelle, Frequentistische und Bayes'sche Schätzung, Regularisierung, Knowledge Space Theory, Item Response Theory), Modellierung von Kompetenzen, von Studienabbrüchen in der Mathematik und von inter- und intraindividuellen Differenzen bei der Emotionsregulation. Diverse wissenschaftliche Beirats-, Gutachter- und Kommissionstätigkeiten sowie Verbundprojektleitungen. Hobbys: Familie, Gitarre, Sport, Betonieren.

Autorenverzeichnis

Prof. Dr. Holger Brandt Psychologisches Institut, Universität Zürich, Zürich, Schweiz

Dr. Sebastian Brückner Fachbereich 3, Johannes Gutenberg Universität Mainz, Mainz, Deutschland

Prof. Dr. G. Leonard Burns Department of Psychology, Washington State University, Pullmann, USA

Prof. Dr. David A. Cole Department of Psychology and Human Development, Vanderbilt University, Nashville, USA

PhD Dr. Delphine S. Courvoisier Clinical Epidemiology Service, Geneva University Hospitals, Genève 14, Schweiz

Prof. Dr. Michael Eid FB Erziehungswissenschaften und Psychologie, Freie Universität Berlin, Berlin, Deutschland

Prof. Dr. Andreas Frey Institut für Psychologie – Arbeitsbereich Pädagogische Psychologie, Goethe Universität Frankfurt am Main, Frankfurt am Main, Deutschland

Dipl.-Psych. Jana C. Gäde Methodenzentrum Sozialwissenschaften, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Deutschland

Prof. Dr. Christian Geiser Department of Psychology, Utah State University, Logan, USA

Prof. Dr. Frank Goldhammer Arbeitsbereich: Technologiebasiertes Assessment, Abt. Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main, Deutschland

Prof. Dr. Mario Gollwitzer Department of Psychology, Ludwig-Maximilians-Universität München, München, Deutschland

Prof. Dr. Johannes Hartig Arbeitsbereich: Educational Measurement, Abt. Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main, Deutschland

Dr. Volkmar Höfling Abteilung Klinische Psychologie und Psychotherapie, Universität Potsdam, Potsdam, Deutschland

Dr. Nina Jude Arbeitsbereich: Unterricht und Schule, Abt. Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main, Deutschland

Prof. Dr. Augustin Kelava Methodenzentrum, Eberhard Karls Universität Tübingen, Tübingen, Deutschland

Dr. Ulf Kröhne Arbeitsbereich: Technologiebasiertes Assessment, Abt. Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main, Deutschland

Prof. Dr. Axel Mayer Institut für Psychologie, RWTH Aachen, Aachen, Deutschland

Prof. Dr. Helfried Moosbrugger Institut für Psychologie, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Deutschland

Dr. Stefano Noventa Methodenzentrum, Eberhard Karls Universität Tübingen, Tübingen, Deutschland

Prof. Dr. Fridtjof W. Nussbeck Methods for Intensive Data in Psychology, Universität Konstanz, Konstanz, Deutschland

Prof. Dr. Hans Anand Pant Institut für Erziehungswissenschaften, Humboldt Universität zu Berlin, Berlin, Deutschland

Prof. Dr. Dominique Rauch Institut für Psychologie, Pädagogische Hochschule Ludwigsburg, Ludwigsburg, Deutschland

Prof. Dr. Wolfgang Rauch Fakultät Sonderpädagogik, Pädagogische Hochschule Ludwigsburg, Ludwigsburg, Deutschland

Dipl.-Math. Alexander Robitzsch Pädagogisch-Psychologische Methodenlehre, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Deutschland

Dr. Norman Rose IFB Sepsis und Sepsisfolgen, Universitätsklinik Jena, Jena, Deutschland

Prof. Dr. Karin Schermelleh-Engel Institut für Psychologie, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Deutschland

Dr. Christina S. Werner Psychologisches Institut, Methodenlehre, Evaluation und Statistik, Universität Zürich, Zürich, Schweiz

Prof. Dr. Olga Zlatkin-Troitschanskaia Rechts- und Wirtschaftswissenschaften, Johannes Gutenberg Universität Mainz, Mainz, Deutschland



Einführung und zusammenfassender Überblick

Helfried Moosbrugger und Augustin Kelava

Inhaltsverzeichnis

- 1.1 Zielgruppen und Gliederungsüberlegungen – 2
- 1.2 Teil I: Konstruktionsgesichtspunkte – 3
- 1.3 Teil II: Testtheorien – 5
- 1.4 Teil III: Validität und Möglichkeiten ihrer Überprüfung – 8
- 1.5 Ergänzende Materialien – 10
- 1.6 Zusammenfassung – 10

i Dieses Kapitel informiert zunächst über die Zielgruppen, für die das vorliegende Lehrbuch konzipiert ist und über den Personenkreis, der daraus Nutzen ziehen kann. Sodann folgen kurze Inhaltsangaben über die einzelnen Kapitel dieses Buches, die den drei großen Bereichen Teil I „Konstruktionsgesichtspunkte“ (► Kap. 2–10), Teil II „Testtheorien“ (► Kap. 11–20) und Teil III „Validität und Möglichkeiten ihrer Überprüfung“ (► Kap. 21–27) zugeordnet sind. Abschließend folgen Hinweise auf lehr- und lernergänzende Materialien.

1.1 Zielgruppen und Gliederungsüberlegungen

Es ist eine große Leserschaft, die das vorliegende Lehrbuch *Testtheorie und Fragebogenkonstruktion* von Nutzen sein kann. Zum einen befinden sich in diesem Personenkreis Test- und Fragebogenkonstrukteure, für die der „State of the Art“ der Planung, Entwicklung, Erprobung, Analyse und Dokumentation von Tests und Fragebogen beschrieben wird. Zum anderen sollen Test- und Fragebogenanwender angesprochen werden, die vor der Aufgabe stehen, aus verschiedenen am Markt befindlichen oder innerhalb der Wissenschaft verwendeten Tests und Fragebogen eine qualifizierte, begründete Auswahl zu treffen, die Verfahren sachkundig zum Einsatz zu bringen, die Test-/Merkmalswerte kompetent zu berechnen und zu interpretieren sowie aus den Ergebnissen angemessene, belastbare Schlussfolgerungen zu ziehen.

Für diese zunehmend fachlich breiter werdende Zielgruppe wurde das Buch als praktisches Nachschlagewerk konzipiert. Gleichzeitig soll es im Kontext der Methodenausbildung für Studierende der Psychologie, Erziehungswissenschaft, Wirtschaftswissenschaften, Soziologie und angrenzender Sozial- und Verhaltenswissenschaften als wesentlicher Baustein dienen. Ebenso adressiert es Personen aus der Hochschuladministration, die im Zuge der bundesweit verfassungsrechtlich geforderten Neuregelung von Studierendenauswahl- und Zulassungsprozessen test-theoretisch-diagnostische Expertise benötigen.

Ziel dieses Buches ist es, den gesamten Konstruktionsprozess von Tests und Fragebogen zu beschreiben, wobei allgemeine und spezielle wissenschaftliche Qualitätsanforderungen ebenso Beachtung finden sollen wie zugrunde liegende Testtheorien mitsamt den empirisch-statistischen Überprüfungsmöglichkeiten der theoretischen Annahmen. Auch sollen die wesentlichen Anwendungsfragen für einen kompetenten Einsatz nicht zu kurz kommen (z. B. bei großflächigen Erhebungen).

Das Lehrbuch ist in drei große Teile gegliedert, die sich mit folgenden breiten Themenbereichen befassen:

- Teil I: Konstruktionsgesichtspunkte
- Teil II: Testtheorien
- Teil III: Validität und Möglichkeiten ihrer Überprüfung

Der Teil I „Konstruktionsgesichtspunkte“ (► Kap. 2–11) richtet sich an einen sehr breiten Adressatenkreis und enthält jene praktisch relevanten Informationen, die als Grundlage für die Konstruktion von Fragebogen und Tests anzusehen sind. Für ein sicheres Verständnis sind Grundkenntnisse der Deskriptivstatistik vorteilhaft.

Der Teil II „Testtheorien“ (► Kap. 12–20) ist den Testtheorien und ihren Annahmen gewidmet, und zwar sowohl der Klassischen Testtheorie (KT) mit ihren Erweiterungen und Generalisierungen als auch der Item-Response-Theorie (IRT) und ihren zahlreichen Untermodellen, die zwischenzeitig einen breiten Einsatz finden. Für ein tiefes Verständnis sind solide statistische Kenntnisse notwendig.

Im Teil III „Validität und Möglichkeiten ihrer Überprüfung“ (► Kap. 21–27) liegt der Schwerpunkt auf der Betrachtung der Möglichkeiten zur Validierung von Testergebnissen. Dabei stehen die Angemessenheit von Interpretationen und die

1.2 · Teil I: Konstruktionsgesichtspunkte

Belastbarkeit von Schlussfolgerungen, die sich auf Testergebnisse stützen, im Vordergrund. Die erörterten Verfahren sind methodisch anspruchsvoll und liefern geeignete Überprüfungsmöglichkeiten für die angeschnittenen spezifischen Validitätsfragen.

Die einzelnen Kapitel in den drei Teilen befassen sich mit den folgenden Inhalten¹:

1.2 Teil I: Konstruktionsgesichtspunkte

In ► Kap. 2 *Qualitätsanforderungen an Tests und Fragebogen* („Gütekriterien“) beschäftigen sich Helfried Moosbrugger und Augustin Kelava mit den Unterschieden zwischen einem laienhaft konstruierten Messinstrument (z. B. Laienfragebogen) und einem Test als wissenschaftliches Messinstrument. Dazu werden sog. „Gütekriterien“ eingeführt, die bei der Testkonstruktion zur Sicherstellung der wissenschaftlichen Qualität zu beachten sind. Hierbei werden zunächst jene Gütekriterien erörtert, die insofern „allgemein“ sind, als sie die Rahmenüberlegungen für jede Test- oder Fragebogenkonstruktion umfassen (Objektivität, Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness und Unverfälschbarkeit). Sodann folgen jene Gütekriterien, die insofern „speziell“ sind, als sie auf bestimmten testtheoretischen Annahmen basieren und diese empirisch überprüfbar machen (Reliabilität, Validität). Die Einhaltung dieser speziellen Gütekriterien ist für wissenschaftliche Tests und Fragebogen unerlässlich (► Kap. 2).

Das ► Kap. 3 *Planungsaspekte und Konstruktionsphasen von Tests und Fragebogen* von Holger Brandt und Helfried Moosbrugger startet bei der Spezifikation von interessierenden, d. h. zu erfassenden Merkmalen und stellt verschiedene Testarten (Leistungs- und Persönlichkeitstests, objektive und projektive Verfahren) vor. Es folgt die Festlegung des intendierten Geltungsbereichs und der anvisierten Zielgruppen der Anwendung, der Testlänge und Testzeit sowie der wichtigen Formen der Testadministration (Paper-Pencil- vs. computerbasierte Testung, Einzel- vs. Gruppentestung). Bei der Testkonstruktion sind neben der Itemgenerierung zahlreiche weitere Entwicklungsphasen (Verständlichkeitsprüfung, Erprobung, Revision, Normierung) zu durchlaufen, bis eine endgültige Testform in strukturell typischem Aufbau erreicht ist. Für die vorgestellten Konzepte werden diverse Beispiele gegeben (► Kap. 3).

In ► Kap. 4 *Itemkonstruktion und Antwortverhalten* erläutern Helfried Moosbrugger und Holger Brandt wesentliche Gesichtspunkte der Generierung und Formulierung der einzelnen Aufgabenstellungen/Testitems. Der anspruchsvollste Aspekt bei der Itemgenerierung besteht darin, repräsentative inhalts valide Operationalisierungen des interessierenden Merkmals zu finden, diese in einem entsprechenden Aufgaben-/Itemstamm zu formulieren und mit einem zweckmäßigen Antwortformat abzufragen. In diesem Kapitel wird auf verschiedene Vorgehensweisen bei der Itemgenerierung eingegangen sowie auf wichtige Aspekte, die bei der Formulierung der Items beachtet werden müssen. Basierend auf der Erörterung von typischen kognitiven und motivationalen Prozessen bei der Itembeantwortung werden verschiedene potentielle Störvariablen im Antwortverhalten (Response-Bias) besprochen, die bei der Itemgenerierung mitberücksichtigt werden sollten, da sie das Ergebnis von Tests und Fragebogen verfälschen können (Antworttendenzen, Soziale Erwünschtheit, Akquieszenz, Itemreihenfolge; ► Kap. 4).

In ► Kap. 5 *Antwortformate und Itemtypen* befassen sich Helfried Moosbrugger und Holger Brandt mit verschiedenen Möglichkeiten, wie die Antworten der Testpersonen auf die Testaufgaben/Fragen erfasst und kodiert werden können („Ant-

Kapitel 2

Kapitel 3

Kapitel 4

Kapitel 5

¹ Die Kurzfassungen der Kapitel orientieren sich, teilweise im Wortlaut, an den (E-Book-)Zusammenfassungen der jeweiligen Kapitelautorinnen und -autoren.

wortformate“). Daraus ergeben sich verschiedene Itemtypen. Unter Beachtung von Vor- und Nachteilen wird das freie Antwortformat dem gebundenen Antwortformat gegenübergestellt. Bei Letzterem sind vor allem Ordnungs-, Auswahl- sowie kontinuierliche und diskrete Beurteilungsaufgaben als Itemtypen weit verbreitet. Bezogen auf verschiedene Zielvorgaben werden Entscheidungshilfen für die Wahl des Aufgabentyps und zahlreiche Beispiele gegeben (► Kap. 5).

Kapitel 6

In ► Kap. 6 *Computerbasiertes Assessment* stellen Frank Goldhammer und Ulf Kröhne Vorgehensweisen in den Vordergrund, bei denen die Testentwicklung, die Vorgabe der Items und die Erfassung des Antwortverhaltens der Testpersonen einschließlich der Antwortbewertung computerbasiert erfolgen. Beziiglich der Itementwicklung, -zusammenstellung und -sequenzierung sowie bezüglich der Ablaufsteuerung ergeben sich wichtige zusätzliche Möglichkeiten, vor allem auch für das sog. „Large-Scale-Assessment“ (d. h. sehr große Erhebungen). Das computerbasierte Assessment erlaubt zusätzlich eine zeitkritische Datenerfassung und Modellierung von Antwortzeiten sowie eine sofortige Rückmeldung der Testergebnisse an die Testpersonen (► Kap. 6).

Kapitel 7

In ► Kap. 7 *Deskriptivstatistische Itemanalyse und Testwertbestimmung* beschreiben Augustin Kelava und Helfried Moosbrugger, wie eine erste empirische deskriptivstatistische Evaluation der generierten Testitems vorgenommen werden kann. Die Items werden einer Erprobungsstichprobe von Testpersonen vorgelegt und das Antwortverhalten wird numerisch kodiert. Im Anschluss können (vorläufige) Testwerte ermittelt werden, die zusammen mit den empirisch festgestellten Itemschwierigkeiten, Itemvarianzen und Itemtrennschärfen Auskunft darüber geben, ob die Items ihrer Aufgabe gerecht werden, Differenzierungen zwischen den Testpersonen bezüglich des interessierenden Merkmals zu leisten (► Kap. 7).

Kapitel 8

In ► Kap. 8 *Testwertverteilung* zeigen Augustin Kelava und Helfried Moosbrugger, wie vorläufige (und mit bestimmten Annahmen verbundene) Testwerte einer Erprobungsstichprobe zusammengefasst und mit deskriptivstatistischen Kennwerten (z. B. Maße der zentralen Tendenz, Streuungsmaße) beschrieben werden können. Die empirisch vorgefundene Verteilungsform gibt (erste) Auskünfte darüber, ob die Testwerte einer theoretisch erwarteten Verteilung (z. B. Normalverteilung) entsprechen. Eine Verteilungsabweichung kann in seltenen und stichhaltig zu begründenden Fällen durch eine sog. „nichtlineare Transformation“ in eine erwartete Verteilung überführt werden. Ein solcher Vorgang ist beispielsweise die Normalisierung, die eine schiefe empirische Verteilung in eine Normalverteilung transformiert (► Kap. 8).

Kapitel 9

In ► Kap. 9 *Testwertinterpretationen, Testnormen und Testeichung* beschreiben Frank Goldhammer und Johannes Hartig verschiedene Möglichkeiten, wie ein Testergebnis deskriptivstatistisch interpretiert werden kann. Bei der normorientierten Interpretation kann ein Testergebnis im Vergleich mit den Testwerten anderer Personen einer Bezugsgruppe (den „Testnormen“) interpretiert werden. Die Testnormen werden im Zuge der Testeichung an einer repräsentativen Eichstichprobe gewonnen. Sofern genauere theoretische Vorstellungen bestehen, kann das Testergebnis auch mit einem inhaltlich definierten Kriterium in Bezug gesetzt werden. Dieser Vorgang wird als kriteriumsorientierte Interpretation bezeichnet. Beide Interpretationsarten können außerdem miteinander verbunden werden (► Kap. 9).

Kapitel 10

In ► Kap. 10 *Standards für psychologisches Testen* behandeln Volkmar Höfling und Helfried Moosbrugger allgemein anerkannte nationale und internationale Qualitätsstandards zur Entwicklung, Adaptation, Anwendung und Qualitätsbeurteilung psychologischer Tests. Hierbei kommt den internationalen „Standards for Educational and Psychological Testing“ (SEPT) und den nationalen „Anforderungen zu Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen“ (DIN 33430) eine besondere Bedeutung zu, ebenso dem „Testbeurteilungssystem des Testkuratoriums“ (TBS-TK), einem sehr nützlichen schematischen Vorgehen zum qualitätssichernden Vergleich psychologischer Tests und Fragebogen (► Kap. 10).

In ► Kap. 11 *Standards für pädagogisches Testen* thematisieren Sebastian Brückner, Olga Zlatkin-Troitschanskaia und Hans Anand Pant anhand der internationalen „Standards for Educational and Psychological Testing“ (SEPT) die besonderen Erfordernisse bei der Konstruktion von pädagogisch(-psychologischen) Tests und bei der Durchführung von Untersuchungen in diesem Bereich. Fragen zu der besonderen Zielsetzung solcher Vorhaben und Fragen zu der Validität sind wichtige Inhalte dieses Kapitels. Weitere Implikationen aus den Standards für Anforderungen des pädagogischen Testens lassen sich u. a. zum Standardsetting, zur Fairness, zur Transparenz von Untersuchungsgegenstand und Interpretation, zu Formen der Diagnostik, zum Feedback sowie zum Datenmanagement ableiten. Zusätzlich zu den Standards gibt die American Educational Research Association (AERA) mit dem „Code of Ethics“ Richtlinien zu Fragen der Forschungsethik heraus, denen auch beim pädagogischen Testen eine große Bedeutung zukommt (► Kap. 11).

Kapitel 11

1.3 Teil II: Testtheorien

In ► Kap. 12 *Testtheorien im Überblick* geben Helfried Moosbrugger, Karin Schermelleh-Engel, Jana C. Gäde und Augustin Kelava eine erste Einführung in die Klassische Testtheorie (KTT) sowie die Item-Response-Theorie (IRT), die sich in den vergangenen Jahrzehnten in der psychologischen Forschung und ihren angrenzenden Gebieten etabliert haben. Die Testtheorien werden zur Beurteilung und zur Qualitätssicherung von Testwerten benötigt. Hierbei ist eine Unterscheidung zwischen sog. „manifesten Variablen“ (direkt beobachtbare numerische Ausprägungen der Antworten der Testpersonen) und sog. „latenten Variablen“ (nicht direkt beobachtbare Merkmale der Testpersonen, die dem Verhalten und Erleben zugrunde liegen) sinnvoll. Die KTT liefert die grundlegenden Annahmen, um die wahren Werte (sog. „True-Scores“) der Testpersonen und die Messgenauigkeit (Reliabilität) von Test- und Fragebogenwerten mit kontinuierlichem oder zumindest mehrstufigem Antwortformat zu beurteilen und empirisch überprüfbare Messmodelle über den Zusammenhang zwischen den manifesten Itemvariablen und der latenten Variablen zu formulieren. Demgegenüber wird mit der IRT über Messmodelle (z. B. „Rasch-Modell“) der Zusammenhang von zumeist zweistufigen manifesten Itemvariablen und der dahinterliegenden, zumeist kontinuierlichen latenten Variablen modelliert. Neben diesen Unterschieden im typischerweise verwendeten Skalenniveau der Itemvariablen werden auch wesentliche Übereinstimmungen zwischen den Theorien dargestellt (► Kap. 12).

Kapitel 12

In ► Kap. 13 *Klassische Testtheorie (KTT)* beschreiben Helfried Moosbrugger, Jana C. Gäde, Karin Schermelleh-Engel und Wolfgang Rauch die theoretischen Grundlagen zur Konstruktion von Testverfahren und zur Interpretation von Testwerten, die sich aus der KTT ergeben. Für die meist messfehlerbehafteten manifesten Itemvariablen lassen sich anhand der KTT eindimensionale Messmodelle formulieren, um den Anteil der wahren Werte vom Anteil der Fehlerwerte zu trennen und darauf aufbauend die Reliabilität der Testwertvariablen zu schätzen. Itemvariablen eines Tests können unterschiedliche Messeigenschaften aufweisen und lassen sich daher hinsichtlich ihrer Messäquivalenz differenzieren. Die Messäquivalenz kann anhand verschiedener Messmodelle überprüft werden, die auf unterschiedlich restriktiven, testbaren Annahmen basieren. Abhängig von der gegebenen Messäquivalenz können unterschiedliche Reliabilitätskoeffizienten geschätzt werden, die zusätzlich durch ein Konfidenzintervall ergänzt werden sollten. Neben diesen eindimensionalen Modellen gibt es inzwischen auch verschiedene mehrdimensionale Ansätze, z. B. die Generalisierbarkeitstheorie, die auf der KTT aufbauen und explizit mehrere latente Variablen als systematische Varianzquellen berücksichtigen (► Kap. 13).

Kapitel 13

Kapitel 14

In ► Kap. 14 *Klassische Methoden der Reliabilitätsschätzung* behandeln Jana C. Gäde, Karin Schermelleh-Engel und Christina S. Werner auf Basis der Klassischen Testtheorie (KTT) das Konzept der Messgenauigkeit (Reliabilität) eines Tests oder Fragebogens und der Quantifizierung/Schätzung dieses Gütekriteriums. Im anschaulichen Vergleich werden die traditionellen Verfahren zur Schätzung der Messgenauigkeit von Testwerten (Cronbachs Alpha sowie Retest-, Paralleltest- und Split-Half-Reliabilität) vorgestellt. Diese Reliabilitätsmaße basieren auf strengen Annahmen der Messäquivalenz, die anhand von Messmodellen im Rahmen der konfirmatorischen Faktorenanalyse (CFA) überprüft werden müssen. Wenn die Voraussetzungen gegeben sind, können die beobachteten Kovarianzen zwischen den Itemvariablen oder die Korrelationen zwischen (Halb-)Testwerten zu verschiedenen Messzeitpunkten oder zwischen parallelen Tests zur Schätzung der verschiedenen Reliabilitätskoeffizienten verwendet werden. Die berechneten Reliabilitätskoeffizienten stellen nur dann adäquate Schätzungen der Reliabilität dar, wenn die Modellannahmen erfüllt sind (► Kap. 14).

Kapitel 15

In ► Kap. 15 *Modellbasierte Methoden der Reliabilitätsschätzung* von Karin Schermelleh-Engel und Jana C. Gäde beruhen die vorgestellten Verfahren im Vergleich zu den klassischen Methoden der Reliabilitätsschätzung (► Kap. 14) auf weniger strengen, realitätsnäheren Annahmen. Die Schätzung der Reliabilitätskoeffizienten von Testwerten (oder Subskalenwerten) wird zusammen mit der Überprüfung der Annahmen unter Verwendung von unterschiedlichen Modellen der konfirmatorischen Faktorenanalyse (CFA), z. B. Einfaktormodell, Bikaktormodell, Faktormodell höherer Ordnung, vorgenommen. Für eindimensionale Tests werden Cronbachs Alpha, McDonalds Omega und Bollens Omega und für mehrdimensionale Tests verschiedene Omega-Koeffizienten vorgestellt. Auch Methoden der Reliabilitätsschätzung auf der Basis von kategorialen Variablen mit geordneten Antwortkategorien werden kurz erläutert. Nach einer Abwägung der verschiedenen Vor- und Nachteile der referierten Koeffizienten werden Empfehlungen gegeben, wie die Reliabilitätskoeffizienten und deren Intervallschätzungen beurteilt werden sollten (► Kap. 15).

Kapitel 16

In ► Kap. 16 *Einführung in die Item-Response-Theorie (IRT)* befassen sich Augustin Kelava und Helfried Moosbrugger mit der zwischenzeitig im Kontext der psychologischen/pädagogischen Leistungsdiagnostik und des sog. „Large-Scale-Assessment“ (z. B. großer Schulleistungsstudien) dominanten Item-Response-Modellierung. In der IRT wird das Antwortverhalten einer Person auf ein Item anhand eines probabilistischen Testmodells erklärt. Es wird ein kontinuierliches latentes Merkmal (sog. „Latent Trait“ oder „Ability“) angenommen. Die latente Merkmalsausprägung einer Testperson wird „Personenparameter“ genannt. Mithilfe einer expliziten Annahme über den (meist logistischen) funktionalen Zusammenhang zwischen der latenten Variablen und der Wahrscheinlichkeit einer bestimmten Antwort wird auf die Eigenschaften der Items („Itemparameter“) und auf die Personenparameter geschlossen. Das Kapitel beschreibt sehr basale sog. ein-, zwei- oder dreiparametrische logistische Modelle (1PL-, 2PL- und 3PL-Modell) für dichotome Itemantworten, die sich auf andere Modellarten generalisieren lassen, ferner grundlegende Möglichkeiten der Parameterschätzung sowie den Informationsgehalt von Items und Tests. Anhand von Modelltests lässt sich untersuchen, ob das angenommene Modell und die beobachteten Daten (Responses) in Einklang stehen (► Kap. 16).

Kapitel 17

In ► Kap. 17 *Interpretation von Testwerten in der Item-Response-Theorie (IRT)* zeigen Dominique Rauch und Johannes Hartig am Beispiel von Large-Scale-Assessments auf, welche spezifischen Vorteile die IRT beispielsweise für das Matrix-Sampling von Testaufgaben, für die Erstellung paralleler Testformen oder für die Entwicklung computerbasierter adaptiver Tests aufweist. Als weiteren wesentlichen Vorteil ermöglichen IRT-Modelle eine kriteriumsorientierte Interpretation der Testwerte. Bei Gültigkeit des Rasch-Modells können indi-

viduelle Testwerte durch ihre Abstände zu den Schwierigkeitsparametern der Items interpretiert werden. Zur leichteren Interpretation wird die kontinuierliche Merkmalsskala in Abschnitte (Kompetenzniveaus) unterteilt, die sehr informative kategoriale Vergleiche der Leistungsfähigkeit ermöglichen. Mit einem Beispiel aus der empirischen Bildungsforschung werden die Definition und die Beschreibung von Kompetenzniveaus anhand eines Vorgehens mit Post-hoc-Analysen der Items und der Verwendung von A-priori-Aufgabenmerkmalen veranschaulicht (► Kap. 17).

In ► Kap. 18 *Überblick über Modelle der Item-Response-Theorie (IRT)* von Augustin Kelava, Stefano Noventa und Alexander Robitzsch werden die in ► Kap. 16 vorgestellten Modelle der IRT einerseits für polytome (d. h. mehrstufige) Itemantworten bei einer kontinuierlichen latenten Fähigkeit (Trait oder Ability), und andererseits für den multidimensionalen Fall, bei dem mehrere Fähigkeiten als zugrunde liegende latente Merkmale angenommen werden können, generalisiert. Ziel dieses Kapitels ist es, aufzuzeigen, dass die berichteten Modelle eine gewisse Verwandtschaft zueinander aufweisen und dass aus einer bestimmten sog. „Parametrisierung“, d. h. durch eine spezifische Ausgestaltung der funktionalen Beziehung zwischen Merkmal und Beobachtung, Spezialfälle resultieren, die unterschiedliche Modelltypen definieren. Zu den vorgestellten Modellen gehören u. a. das Rating-Scale-, das Graded-Response-, das multidimensionale Generalized-Partial-Credit- und das Bifaktormodell. Das Kapitel richtet sich als Überblicksartikel an jenen Personenkreis, der im Rahmen seiner Forschung oder der Lektüre angewandter Forschung (z. B. aus der empirischen Bildungsforschung) ein erstes Verständnis dieser Modelltypen entwickeln möchte (► Kap. 18).

Kapitel 18

In ► Kap. 19 *Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT)* macht Norman Rose deutlich, dass verschiedene Verfahren der Item- und Personenparameterschätzung existieren, wobei sich grundsätzlich Maximum-Likelihood-Schätzverfahren (ML-Schätzverfahren) und Bayes'sche Schätzverfahren unterscheiden lassen. Innerhalb beider Verfahrensklassen gibt es wiederum verschiedene Schätzalgorithmen mit unterschiedlichen Eigenschaften. Die wichtigsten werden in diesem Kapitel am Beispiel ein- und zweiparametrischer IRT-Modelle dargestellt. Item- und Personenparameter werden oft (aus gutem Grund) nicht simultan, sondern separat geschätzt. Da die Messgenauigkeit in der IRT in Abhängigkeit der zu schätzenden Personenparameter variiert, gibt es streng genommen nicht nur einen Wert der Reliabilität für einen Test. Aus diesem Grund wurden marginale, d. h. durchschnittliche Reliabilitätskoeffizienten als Gütemaß der Messgenauigkeit eines Tests entwickelt. Die Berechnung und Interpretation der marginalen Reliabilitäten für die verschiedenen Personenparameterschätzer bilden den Abschluss dieses Kapitels (► Kap. 19).

Kapitel 19

In ► Kap. 20 *Computerisiertes adaptives Testen* behandelt Andreas Frey ein spezielles Vorgehen bei der Messung des interessierenden Merkmals, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items adaptiv und maßgeschneidert am vorherigen Antwortverhalten der Testperson orientiert. Der Grundgedanke besteht darin, dass nicht alle Items eines Tests, sondern nur solche Items vorgegeben werden, die möglichst viel diagnostisch relevante Information über die individuelle Ausprägung einer Testperson im interessierenden Merkmal liefern. Dieses Anliegen wird durch die Spezifikation von fünf elementaren Bausteinen umgesetzt. Es handelt sich dabei um den Itempool, die Art, den Test zu beginnen, die Schätzung der individuellen Merkmalsausprägung, die Itemauswahl, die Berücksichtigung nicht-statistischer Einschränkungen (z. B. die Kontrolle relativer Anteile vorgegebener Items je Inhaltsfacette des gemessenen Merkmals) und die Art, den Test zu beenden. Der Hauptvorteil computerisierten adaptiven Testens im Vergleich zum nicht adaptiven Testen besteht in einer Steigerung der Messeffizienz, die in den meisten Fällen beträchtlich ausfällt (z. B. verkürzte Testzeit). Darüber hinaus sind positive Auswirkungen auf die Validität der adaptiv erhobenen Testergebnisse zu verzeichnen (► Kap. 20).

Kapitel 20

1.4 Teil III: Validität und Möglichkeiten ihrer Überprüfung

Kapitel 21

In ► Kap. 21 *Validität von Testwertinterpretationen* betonen Johannes Hartig, Andreas Frey und Nina Jude, dass sich das Verständnis von Validität in den letzten Jahrzehnten deutlich weiterentwickelt hat. Während sich im vergangenen 20. Jahrhundert zunächst eine wenig praktikable Vielzahl „verschiedener Validitäten“ herausgebildet hatte, wird Validität inzwischen als einheitliches Qualitätskriterium betrachtet, das Informationen aus verschiedenen Quellen integriert. Zudem wurde Validität früher als Eigenschaft eines Tests betrachtet, heute bezieht sie sich auf einen Prozess, der die Interpretation von Testwerten inkludiert, und darauf, ob die darauf aufbauenden Schlussfolgerungen gerechtfertigt sind. Im Kontext aktueller internationaler Forschung wird die Validierung von Testwertinterpretationen im Rahmen des sog. „argumentationsbasierten Ansatzes“ beschrieben, bei dem für die zu validierende Testwertinterpretation empirisch überprüfbare Grundannahmen identifiziert werden. Hierzu wird empirische Evidenz gesammelt, anhand derer die Grundannahmen widerlegt oder (vorläufig) gestützt werden können. Eine Testwertinterpretation wird dann als valide betrachtet, wenn die zugrunde liegenden Annahmen nicht widerlegt sind (► Kap. 21).

Kapitel 22

In ► Kap. 22 *Latent-Class-Analyse (LCA)* befasst sich Mario Gollwitzer mit einem Teilbereich probabilistischer Testmodelle, bei dem das diagnostische Interesse darin besteht, die Testpersonen auf Grundlage ihres Antwortverhaltens nicht auf einem kontinuierlichen latenten Trait zu verorten, sondern in qualitative (diskrete) latente Klassen einzuteilen. Die LCA basiert auf der Annahme, dass Personen mit einer gewissen Wahrscheinlichkeit einer von mehreren Klassen (Typen, Gruppen) angehören, wobei die Klassenzugehörigkeit anhand der Antworten auf die Items eines Tests erschlossen wird. Da die Klassenzugehörigkeit nicht direkt beobachtbar ist, spricht man von latenten Klassen. Ziel in diesem Kapitel ist es, in die Grundgedanken der LCA einzuführen und anhand von Beispielen zu verdeutlichen, wann und wie die LCA als Testmodell anwendbar ist. Die Richtigkeit der Zuordnung kann mit entsprechenden Modelltests überprüft werden (► Kap. 22).

Kapitel 23

In ► Kap. 23 *Exploratorische Faktorenanalyse (EFA)* geht Holger Brandt – in Abgrenzung zur Hauptkomponentenanalyse (PCA) – auf die wichtigsten Aspekte bei der Durchführung einer exploratorischen, d. h. struktursuchenden Faktorenanalyse ein. Ausgehend von der allgemeinen Modellvorstellung („Fundamentaltheorem der Faktorenanalyse“) wird die darauf basierende Varianzzerlegung in erklärt, d. h. auf gemeinsame Faktoren rückführbare, und in unerklärte Varianz dargestellt. Anschließend werden die zentralen Begriffe der EFA eingeführt, d. h. die Eigenwerte der Faktoren sowie die Komunalität und Spezifität der Items. Die wichtigsten Extraktionsmethoden, die Principal Axes Factor Analysis (PFA) und Maximum-Likelihood-EFA (ML-EFA) sowie Rotationskriterien (orthogonal vs. oblique) werden diskutiert, bevor auf weitere Aspekte wie die Beurteilung der Modellgüte, alternative Schätzverfahren und die Berechnung von Faktorwerten eingegangen wird (► Kap. 23).

Kapitel 24

In ► Kap. 24 *Konfirmatorische Faktorenanalyse (CFA)* bieten Jana C. Gäde, Karin Schermelleh-Engel und Holger Brandt eine Einführung in die Grundlagen der CFA und beschreiben die Vorgehensweise, wie bestehende und theoretisch begründete Modelle konfirmatorisch auf ihre Passung zu empirischen Daten überprüft werden können. Die CFA stellt ein wichtiges Instrument zur strukturprüfenden Beurteilung der Dimensionalität und der faktoriellen Validität eines Tests dar. Die Hypothesen eines Modells hinsichtlich der Anzahl der Faktoren und der Zuordnung der Items zu den Faktoren werden explizit getestet. Praktische Aspekte der Hypothesenbildung, Modellspezifikation und -identifikation werden ebenso behandelt wie ein kurzer Überblick über Schätzverfahren und Gütekriterien zur Modell-evaluation. Es werden ein- und mehrdimensionale Modelle an einem empirischen Beispiel vorgestellt. Zudem werden der Einsatz der CFA zur Reliabilitätsschätzung

1.4 · Teil III: Validität und Möglichkeiten ihrer Überprüfung

von (Sub-)Skalen eines Tests und zur Überprüfung der Messäquivalenz von Items sowie Möglichkeiten des Modellvergleichs, der Modellmodifikation und der Überprüfung der Messinvarianz eines Tests besprochen (► Kap. 24).

In ► Kap. 25 *Multitrait-Multimethod-Analysen (MTMM-Analysen)* fokussieren Karin Schermelleh-Engel, Christian Geiser und G. Leonard Burns den Sachverhalt, dass sich jede Messung aus einer systematischen Trait-Methoden-Einheit und einem unsystematischen Fehleranteil zusammensetzt, sodass nicht nur der gemessene Trait, sondern darüber hinaus die verwendete (Mess-)Methode/Quelle als Bestandteil des Messwertes berücksichtigt werden muss. Demnach liegt Konstruktvalidität nur dann vor, wenn einerseits Messungen desselben Konstruktts mit verschiedenen Messmethoden zu einer hohen Übereinstimmung führen (konvergente Validität) und andererseits eine Diskrimination zwischen inhaltlich unterschiedlichen Konstrukten sowohl innerhalb einer Messmethode als auch zwischen den Methoden nachgewiesen werden kann (diskriminante Validität). Beim korrelationsbasierten MTMM-Nachweis der Konstruktvalidität werden die Korrelationskoeffizienten in der MTMM-Matrix durch systematische Vergleiche deskriptiv dahingehend beurteilt, ob die Kriterien der konvergenten und der diskriminanten Validität erfüllt sind. Aber erst die konfirmatorische Faktorenanalyse (CFA) ermöglicht es, Trait-, Methoden- und unsystematische Messfehleranteile der gemessenen Variablen unabhängig voneinander zu schätzen und die Gültigkeit der zugrunde liegenden Annahmen inferenzstatistisch zu überprüfen (► Kap. 25).

Kapitel 25

Die in ► Kap. 26 *Latent-State-Trait-Theorie (LST-Theorie)* von Augustin Kelava, Karin Schermelleh-Engel und Axel Mayer vorgestellte LST-Theorie stellt eine Erweiterung der Klassischen Testtheorie (KTT) dar, indem Messungen wiederholt zu mindestens zwei Messgelegenheiten anhand von mindestens zwei Tests (bzw. Testhälften, Items) durchgeführt werden. Die Gesamtvarianz einer Messung lässt sich in einen wahren Anteil und einen Messfehleranteil aufteilen; die wahre Varianz wiederum unterteilt sich in einen personenspezifischen stabilen, zeitlich überdauernden Anteil (Trait, stabile Persönlichkeitsdisposition), einen situationsspezifischen Anteil (inklusive der Interaktion zwischen Situation und Person) und einen zeitlich überdauernden Methodenanteil. Auf der Basis dieser Varianzdekomposition und unter Verwendung verschiedener Äquivalenz-/Stabilitätsannahmen lassen sich dann anhand der konfirmatorischen Faktorenanalyse (CFA) Koeffizienten (als Quantifizierungen dieser Anteile) schätzen, die aufsummiert eine Schätzung der Reliabilität einer Messung ergeben. Aus der LST-Theorie lassen sich verschiedene Modelle ableiten, die anhand eines empirischen Beispiels erläutert werden (► Kap. 26).

Kapitel 26

In ► Kap. 27 *Konvergente und diskriminante Validität über die Zeit: Integration von Multitrait-Multimethod-Modellen (MTMM-Modellen) und der Latent-State-Trait-Theorie (LST-Theorie)* verdeutlichen Fridtjof W. Nussbeck, Michael Eid, Christian Geiser, Delphine S. Courvoisier und David A. Cole, dass Merkmalsausprägungen von Individuen über die Zeit schwanken können und somit auch die konvergente und diskriminante Validität verschiedener Methoden und Konstrukte zeitlichen Veränderungen unterworfen sind. Nur bei gesicherter Qualität der eingesetzten Verfahren können Indikationen für mögliche Interventionen zuverlässig getroffen werden. Besonders bei Kindern, die sich in einem Entwicklungsprozess befinden, aber auch bei Erwachsenen ist es notwendig, die zeitliche Stabilität der gefundenen Testwerte zu untersuchen. Nur bei gegebener Stabilität der Messungen kann von einem stabilen Trait ausgegangen werden. Darüber hinaus ist es wichtig, zu analysieren, wie sich die konvergente Validität verschiedener Messmethoden über die Zeit entwickelt. Drei longitudinale multimethodale Modelle für mehrere Traits werden vorgestellt, die es erlauben, die Konvergenz verschiedener Methoden und die diskriminante Validität von Traits und States zu untersuchen. Die empirischen Anwendungen zeigen deutlich, dass implizite Annahmen über die Übereinstimmung verschiedener Methoden prinzipiell überprüft werden müssen (► Kap. 27).

Kapitel 27

1.5 Ergänzende Materialien

Zur Wiederholung und Festigung der jeweiligen Kapitelinhalte befindet sich am Ende jedes Kapitels eine Zusammenfassung der wesentlichen Inhalte.

Außerdem sind Kontrollfragen angefügt, anhand derer – vor allem für Personen, die das Buch im Selbststudium nutzen – das individuelle Verständnis überprüft werden kann. Die Antworten auf die Fragen finden Sie im Lerncenter zu den jeweiligen Kapiteln unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

Des Weiteren werden EDV-Hinweise (inklusive Software-Skripten) für die rechentechnische Durchführung und Überprüfung der in den Kapiteln angeschnittenen Problemstellungen gegeben.

Schließlich findet sich in jedem Kapitel ein ausführliches Literaturverzeichnis, das spezifische Informationen zur Verbreiterung und Vertiefung der vorgestellten Inhalte enthält.

1.6 Zusammenfassung

Dieses Kapitel zeigte auf, für welche Zielgruppen das vorgelegte Lehrbuch verfasst wurde und welcher Personenkreis daraus Nutzen ziehen kann. Hierzu wurde die dreiteilige Gliederung der Inhalte vorgestellt – Teil I „Konstruktionsgesichtspunkte“, Teil II „Testtheorien“ und Teil III „Validität und Möglichkeiten ihrer Überprüfung“. Teil I behandelt in zehn Kapiteln die Themen Gütekriterien, Planungsaspekte, Itemkonstruktion, Antwortverhalten, Antwortformate, Itemtypen, computerbasiertes Assessment, Itemanalyse, Testwertverteilungen und -interpretation sowie Standards für psychologisches und pädagogisches Testen. In Teil II werden zunächst die Klassische Testtheorie sowie klassische und modellbasierte Methoden der Reliabilitätsschätzung vorgestellt; danach folgt eine Einführung in die Item-Response-Theorie, die Vorstellung ihrer verschiedenen Modelle, die Parameterschätzung und -interpretation, die Messgenauigkeit sowie die Effizienzsteigerung durch adaptives Testen. Teil III widmet sich der Validität von Testwertinterpretationen und fokussiert als Methoden ihrer Überprüfung die Latent-Class-Analyse, die exploratorische und die konfirmatorische Faktorenanalyse sowie die Integration von Multitrait-Multimethod-Analyse und Latent-State-Trait-Theorie. Das Überblickskapitel schließt mit Hinweisen auf lehr- und lernergänzende Materialien ab.

Konstruktionsgesichtspunkte

Inhaltsverzeichnis

- Kapitel 2 Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“) – 13**
Helfried Moosbrugger und Augustin Kelava
- Kapitel 3 Planungsaspekte und Konstruktionsphasen von Tests und Fragebogen – 39**
Holger Brandt und Helfried Moosbrugger
- Kapitel 4 Itemkonstruktion und Antwortverhalten – 67**
Helfried Moosbrugger und Holger Brandt
- Kapitel 5 Antwortformate und Itemtypen – 91**
Helfried Moosbrugger und Holger Brandt
- Kapitel 6 Computerbasiertes Assessment – 119**
Frank Goldhammer und Ulf Kröhne
- Kapitel 7 Deskriptivstatistische Itemanalyse und Testwertbestimmung – 143**
Augustin Kelava und Helfried Moosbrugger
- Kapitel 8 Testwertverteilung – 159**
Augustin Kelava und Helfried Moosbrugger
- Kapitel 9 Testwertinterpretation, Testnormen und Testeichung – 171**
Frank Goldhammer und Johannes Hartig
- Kapitel 10 Standards für psychologisches Testen – 197**
Volkmar Höfling und Helfried Moosbrugger
- Kapitel 11 Standards für pädagogisches Testen – 217**
Sebastian Brückner, Olga Zlatkin-Troitschanskaia und Hans Anand Pant



Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“)

Helfried Moosbrugger und Augustin Kelava

Inhaltsverzeichnis

- 2.1 Vom Laienfragebogen zum wissenschaftlichen Messinstrument – 15**
- 2.2 Unterschiedliche Qualitätsanforderungen – 16**
- 2.3 Allgemeine Gütekriterien für Tests und Fragebogen – 17**
 - 2.3.1 Objektivität – 17
 - 2.3.1.1 Durchführungsobjektivität und Standardisierung – 18
 - 2.3.1.2 Auswertungsobjektivität und Skalierung – 19
 - 2.3.1.3 Interpretationsobjektivität und Normierung (Eichung) – 21
 - 2.3.2 Ökonomie – 23
 - 2.3.3 Nützlichkeit – 24
 - 2.3.4 Zumutbarkeit – 25
 - 2.3.5 Fairness – 25
 - 2.3.6 Unverfälschbarkeit – 26
- 2.4 Spezielle testtheoriebasierte Gütekriterien für wissenschaftliche Tests und Fragebogen – 27**
 - 2.4.1 Reliabilität – 27
 - 2.4.1.1 Klassische Methoden der Reliabilitätsschätzung – 28
 - 2.4.1.2 Modellbasierte Methoden der Reliabilitätsschätzung – 29
 - 2.4.2 Validität – 30
 - 2.4.2.1 Augenschein- und Inhaltsvalidität – 31
 - 2.4.2.2 Kriteriumsvalidität und extrapolierende Testwertinterpretationen – 32
 - 2.4.2.3 Konstruktvalidität – 33
 - 2.4.2.4 Argumentationsbasierter Validierungsansatz von Testwertinterpretationen – 35
- 2.5 Dokumentation der erfüllten Qualitätskriterien – 36**

2.6 Zusammenfassung – 36

2.7 Kontrollfragen – 36

Literatur – 37

2.1 · Vom Laienfragebogen zum wissenschaftlichen Messinstrument

i Laienfragebogen bestehen häufig aus einer Ansammlung von Fragen, die in keinem unmittelbaren Bezug zueinander stehen; Tests und Fragebogen als wissenschaftliche Messinstrumente hingegen erfassen zumeist latente, d. h. nicht direkt beobachtbare Merkmale („latente Konstrukte“), die über mehrere Testitems/Fragen/Aufgabenstellungen erschlossen werden. Bei den Items handelt es sich um Merkmalsindikatoren (Operationalisierungen), die in Zusammenhang mit dem latenten Konstrukt stehen und die das Merkmal messbar machen sollen. Die Bandbreite vom Laienfragebogen bis hin zu einem wissenschaftlichen Test/Fragebogen kann als Kontinuum aufgefasst werden. Ein Fragebogen/Test ist umso wissenschaftlicher, je mehr Qualitätsanforderungen („Gütekriterien“) bei seiner Konstruktion erfüllt werden. Von besonderer Wichtigkeit für Fragebogen/Tests sind die Durchführungs-, Auswertungs- und Interpretationsobjektivität, aber auch weitere Aspekte wie Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness und Unverfälschbarkeit. Die Berücksichtigung dieser Gütekriterien erfordert keine besonderen testtheoretischen Kenntnisse. Für wissenschaftliche Tests/Fragebogen sind die testtheoriebasierten Gütekriterien der Reliabilität und Validität unumgänglich, sie setzen spezielle Kenntnisse und Betrachtungen der Klassischen Testtheorie (KTT) sowie der Item-Response-Theorie (IRT) und faktorenanalytischer Modelle voraus. Die Reliabilität befasst sich mit der Messgenauigkeit eines Tests; sie kann mit verschiedenen Verfahren empirisch überprüft werden. Die Validität beschäftigt sich mit der Frage, ob ein Test das Merkmal, das er messen soll, auch wirklich misst. Hierbei sind einerseits die Aspekte der Augenschein- und Inhaltsvalidität, andererseits der Kriteriums- und Konstruktvalidität von Bedeutung, um feststellen zu können, mit welcher Berechtigung extrapolierende Schlussfolgerungen aus den Testergebnissen gezogen werden können und welche Struktur und Dimensionalität die latenten Konstrukte aufweisen.

2.1 Vom Laienfragebogen zum wissenschaftlichen Messinstrument

Wenn man einzugrenzen versucht, was im Deutschen umgangssprachlich unter dem Begriff „Fragebogen“ zu verstehen ist, so wird man feststellen, dass es sich um einen Sammelausdruck für vielfältige Formen von zumeist nur lose zusammenhängenden Fragen oder Aussagen handelt. Fragebogen sind in verschiedenen inhaltlichen Bereichen weitverbreitet und dienen der Erfassung von z. B. biografischen Daten, wirtschaftlichen Daten, schulischen Daten, medizinisch-anamnestischen Daten, demoskopischen Daten etc.

Wissenschaftlich fundierte (psychologische) Messinstrumente (Tests oder Fragebogen) enthalten hingegen zumeist mehrere thematisch aufeinander abgestimmte Fragen/Aufgabenstellungen/Items, die sich auf verschiedene Erscheinungsformen („Manifestationen“) von nicht direkt beobachtbaren Merkmalen („latente Konstrukte“) beziehen. Bei den Items handelt es sich um Merkmalsindikatoren, mit denen die latenten Konstrukte operationalisiert, d. h. messbar gemacht werden können. Zur Erschließung der latenten Konstrukte werden die Antworten („responses“) auf die Items nicht separat interpretiert, sondern zu einem Testwert verrechnet, der Auskunft über die Ausprägung des interessierenden Merkmals auf einer Skala („scale“) gibt. Eingehende Ausführungen zur Testplanung, zur Itemkonstruktion und zu Aufgabentypen sowie zu Antwortformaten finden sich in ► Kap. 3, 4 und 5.

Um einen Laienfragebogen besser von einem wissenschaftlich fundierten Test/Fragebogen unterscheiden zu können, geben wir zunächst folgende *Definition eines psychologischen Tests*. Hierbei sind *wissenschaftliche Fragebogen* gleichermaßen inkludiert:

Merkmalsindikatoren zur Operationalisierung latenter Konstrukte

Was ist ein „Test“?**Definition**

Ein **Test** ist ein wissenschaftliches Routineverfahren zur Erfassung der Ausprägungen von empirisch abgrenzbaren (psychologischen) Merkmalen mit dem Ziel, möglichst genaue Aussagen über den (relativen) quantitativen Grad oder die qualitative Kategorie der individuellen Merkmalsausprägungen zu gewinnen.

Vier wesentliche Aspekte

In dieser Definition¹ stehen vier Aspekte im Vordergrund:

1. Mit „Routineverfahren“ ist ein Erhebungsverfahren gemeint, das einfach, objektiv (d. h. von Untersuchungsbeteiligten unabhängig) und ökonomisch durchführbar ist und wiederholbare oder nachvollziehbare Ergebnisse liefert.
2. Die „Wissenschaftlichkeit“, erfordert möglichst genaue Vorgaben über die zu messenden (meist latenten, d. h. nicht direkt beobachtbaren) Merkmale. Ebenso werden testtheoretisch-psychometrisch begründbare Qualitätsansprüche an die Testprozeduren und an die Testwerte gestellt; vor allem sollen eine hohe Messgenauigkeit („Reliabilität“) erfüllt sein sowie Evidenzen dafür vorliegen, dass die mit dem Test erzielten Ergebnisse tragfähige Entscheidungen erlauben („Validität“).
3. „Empirisch abgrenzbar“ bedeutet in wissenschaftstheoretischer Hinsicht, dass die untersuchten Merkmale erfahrungswissenschaftlich von anderen Merkmalen unterschiedlich und statistisch gegen den Zufall abgesichert sind.
4. Mit „quantitativer Grad“ bzw. „qualitativer Kategorie“ wird die Zielrichtung der Aussagen über die individuellen Merkmalsausprägungen genauer angegeben. Mit „(relativer) quantitativer Grad“ ist gemeint, dass die Testergebnisse quantifizierbare Einordnungen der Testpersonen bezüglich des untersuchten Merkmals ermöglichen sollen. Diese Einordnungen können im einfachsten Fall aus relationalen Größer-Kleiner-Aussagen bestehen; zudem können sie normorientiert erfolgen, und zwar durch den Vergleich mit den Testergebnissen einer Bezugsgruppe/Eichstichprobe (z. B. „Die Testperson hat einen Intelligenzquotienten [IQ] von 130 und wird damit nur von 2,3 % der Bevölkerung hinsichtlich des IQ übertroffen“); schließlich kann die Einordnung aber auch kriteriumsorientiert erfolgen, d. h. durch den Vergleich mit markanten Merkmalsausprägungen, bei denen es sich etwa um quantitativ gereihte diskrete Kompetenzniveaus handeln kann (z. B. „Die Testperson kann elementare naturwissenschaftliche Modellvorstellungen anwenden“). Mit „qualitativer Kategorie“ ist gemeint, dass die kriteriumsorientierte Einordnung auch durch klassifikatorische Vergleiche mit nominalskalierten qualitativen Kategorien (z. B. Interessenstypen, politischen Parteipräferenzen oder anderen latenten Klassen) erfolgen kann.

2.2 Unterschiedliche Qualitätsanforderungen

Bereits aus dem bisher Gesagten geht deutlich hervor, dass an wissenschaftliche Tests hohe Qualitätsanforderungen gestellt werden müssen. Dennoch lassen sich für verschiedene Verfahren verschiedene Abstufungen der Qualität feststellen, wenn man ein gedankliches Kontinuum bildet, das sich vom Laienfragebogen bis hin zu wissenschaftlichen Tests erstreckt. Je qualitätsvoller ein Verfahren auf diesem Kontinuum angesiedelt sein möchte, desto mehr Qualitätsanforderungen muss das Verfahren erfüllen. Das Bestreben sollte bei der Fragebogen- und Testkonstruktion also dahin gehen, je nach Fragestellung möglichst viele der mittels der Gütekriterien geforderten Qualitätsansprüche zu berücksichtigen und auch zu erfüllen.

¹ Die Definition wurde gegenüber der an Lienert und Raatz (1998) orientierten Fassung der 2. Auflage (Moosbrugger und Kelava 2012) erneut überarbeitet und um den „relativen“ Grad sowie um „qualitative Kategorien“ erweitert.

Unter dem Begriff „Gütekriterien“ versteht man dabei eine Reihe von Gesichtspunkten/Anforderungen, die bei der Test- und Fragebogenkonstruktion zur Qualitätssicherung Berücksichtigung finden sollen. Sie basieren auf international vereinheitlichten Standards für Fragebogen und Tests (s. dazu ▶ Kap. 10 und 11). Als Gütekriterien haben sich zahlreiche Aspekte etabliert (Testkuratorium 1986), die nicht zuletzt auch die Basis der DIN 33430 zur berufsbezogenen Eignungsbeurteilung bilden (DIN 2002, 2016; vgl. auch Westhoff et al. 2010). In der Regel werden folgende zehn Kriterien unterschieden: Objektivität, Reliabilität, Validität, Skalierung, Normierung, Testökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und Fairness (Kubinger 2003).

Von diesen zehn Kriterien werden die ersten drei (Objektivität, Reliabilität, Validität) traditionell als Hauptgütekriterien bezeichnet, weil in erster Linie ihre Berücksichtigung darüber entscheidet, ob es sich auf dem Kontinuum vom Laienfragebogen zum Test um ein fertig entwickeltes wissenschaftliches Messinstrument handelt. Da aber gerade die Erzielung von Objektivität sowie die Erfüllung der weiteren (Neben-)Gütekriterien keine besonderen testtheoretischen Kenntnisse erfordern, sondern auch von weniger spezialisierten Test-/Fragebogenkonstrukteuren in allgemeiner Weise berücksichtigt werden können/sollten, nehmen wir hier eine Umgruppierung vor in „allgemeine Gütekriterien“ für Tests und Fragebogen sowie in „spezielle testtheoriebasierte Gütekriterien“ für wissenschaftliche Messinstrumente. Zugleich stellen wir mit dieser Aufteilung – dem Aufbau und der inneren Kohärenz des vorliegenden Lehrbuches folgend – den Bezug zu jenen Buchkapiteln her, in denen die konstruktiven und evaluativen Maßnahmen beschrieben werden, die erforderlich sind, um den Gütekriterien zu entsprechen:

- In der ersten Gruppe („allgemeine Gütekriterien“), befassen wir uns mit Qualitätsanforderungen, die von allgemein-planerischer Natur sind und (im Unterschied zur zweiten Gruppe) keiner besonderen testtheoretischen Untermauerung bedürfen. Sie betreffen alle Fragen, die bei der Konstruktion von Fragebogen und Tests vor allem in den frühen Stadien der Planung und der Testdurchführung eine wesentliche Rolle spielen (▶ Kap. 3, 4, 5, 6, 7, 8 und 9).
- In der zweiten Gruppe („spezielle testtheoriebasierte Gütekriterien“) beschäftigen wir uns zum einen mit *Fragestellungen zur Reliabilität*, worunter die Messgenauigkeit von Tests für zumeist latente, d.h. nicht direkt beobachtbare Merkmale (▶ Kap. 12) verstanden wird. Die Reliabilitätsbeurteilung erfordert Kenntnisse der Klassischen Testtheorie (KT; ▶ Kap. 13, 14 und 15; s. auch Eid und Schmidt 2014; Steyer und Eid 2001) und der Item-Response-Theorie (IRT; ▶ Kap. 16, 17, 18 und 19; s. auch Eid und Schmidt 2014; Steyer und Eid 2001). Zum anderen müssen *Fragestellungen zur Validität* geklärt sein. Hierunter wird einerseits die Gültigkeit des Tests für extrapolierende Schlussfolgerungen verstanden, die entscheidend ist für die Belastbarkeit/Tragfähigkeit von (diagnostischen) Entscheidungen auf Basis der mit dem Test erzielten Ergebnisse (▶ Kap. 21), und andererseits die Untersuchung von Struktur und Dimensionalität der erfassten latenten Konstrukte. Diese Form der Validitätsbeurteilung erfordert neben testtheoretischen Kenntnissen auch Kenntnisse weiterführender Analysetechniken (▶ Kap. 22, 23, 24, 25, 26 und 27).

Testgütekriterien

Allgemeine Gütekriterien

Spezielle testtheoriebasierte Gütekriterien

2.3 Allgemeine Gütekriterien für Tests und Fragebogen

2.3.1 Objektivität

Um in Test- und Fragebogenuntersuchungen die erforderliche Vergleichbarkeit der Ergebnisse von verschiedenen Testpersonen sicherzustellen, muss notwendigerweise das Gütekriterium der Objektivität erfüllt sein.

Objektivität bedeutet, dass dem Testleiter kein Verhaltensspielraum bei der Durchführung, Auswertung und Interpretation des Tests bzw. Fragebogens eingeräumt wird. Hohe Objektivität wäre also dann gegeben, wenn jeder beliebige Testleiter den Test oder Fragebogen mit einer bestimmten Testperson in identischer Weise durchführt; ebenso müsste jeder beliebige Testauswerter die Testleistung der Testperson genau gleich auswerten und interpretieren.

Objektivität wird wie folgt definiert:

Definition

Ein Test ist dann **objektiv**, wenn das ganze Verfahren, bestehend aus Testmaterialien, Testdarbietung, Testauswertung und Interpretationsregeln, so genau festgelegt ist, dass der Test unabhängig von Ort, Zeit, Testleiter und Auswerter durchgeführt werden könnte und für eine bestimmte Testperson bezüglich des untersuchten Merkmals dennoch dasselbe Ergebnis und dieselbe Ergebnisinterpretation liefert.

Sinnvollerweise werden Tests und Fragebogen hinsichtlich des Gütekriteriums der Objektivität in Bezug auf die folgenden drei wesentlichen Gesichtspunkte separat betrachtet: *Durchführungsobjektivität*, *Auswertungsobjektivität* sowie *Interpretationsobjektivität*.

Um diese drei Gesichtspunkte zu erfüllen, müssen klare und anwenderunabhängige Regeln für die Durchführung, Auswertung und Ergebnisinterpretation vorliegen. Diese möglichst eindeutigen Regelungen werden typischerweise im Testhandbuch („Testmanual“, „Verfahrenshinweise“ etc.) eindeutig dokumentiert.

2.3.1.1 Durchführungsobjektivität und Standardisierung

Von Durchführungsobjektivität kann ausgegangen werden, wenn die Durchführung des Tests/Fragebogens voll standardisiert ist. Die Standardisierung soll sicherstellen, dass Störeinflüsse eliminiert werden, indem die Durchführungsbedingungen nicht von Testung zu Testung variieren, sondern festgelegt sind.

Definition

Durchführungsobjektivität liegt vor, wenn die Durchführungsbedingungen in der Weise standardisiert sind, dass das Testverhalten der Testperson nur von der individuellen Ausprägung des interessierenden Merkmals abhängt. Alle anderen Bedingungen sollen hingegen konstant oder kontrolliert sein, damit sich diese nicht störend und ergebnisverzerrend auswirken können.

Eliminierung von Störeinflüssen

Standardisierung

Festlegung von Testmaterialien, Zeitdauer und Instruktion

Um eine Standardisierung der Durchführungsbedingungen zu erreichen, werden von den Testautoren bzw. Herausgebern eines Tests im Testmanual genaue Anweisungen gegeben. Hierbei müssen mehrere Aspekte berücksichtigt werden, die sich auf die Konstanz des Testmaterials, die Festlegung der Instruktion sowie auf die Angabe von etwaigen Zeitbegrenzungen beziehen:

- **Konstanz der Fragen/Aufgabenstellungen/Testmaterialien:** Um Auswirkungen von Interaktionen mit dem Testleiter zu vermeiden, sollen die Fragen möglichst schriftlich vorgegeben werden. Schon lange sind Variablen bekannt (z. B. Versuchsleitereffekte in Form von „verbal conditioning“ in Einzelversuchen), die als Bestandteil der Testsituation die Testleistung in unkontrollierter Weise beeinflussen (vgl. z. B. Rosenthal und Rosnow 1969); sie können die interne Validität der Testung gefährden und zu Artefakten führen (vgl. Reiß und Sarris 2012). Aus diesem Grund wird – soweit möglich – auf eine Interaktion zwischen Testleiter und Testperson verzichtet oder diese minimiert; nicht zuletzt deshalb ist eine computerbasierte Testdurchführung (► Kap. 6) der Durchführungsobjektivität förderlich.
- **Angabe der zur Beantwortung vorgesehenen Zeitdauer:** Vor allem bei Speedtests (im Unterschied zu Powertests, ► Kap. 3) ist die Angabe der Testzeit von

2.3 · Allgemeine Gütekriterien für Tests und Fragebogen

erheblicher Bedeutung, um die Vergleichbarkeit der Testleistungen zu gewährleisten.

- **Festlegung der Instruktion:** In der Instruktion wird den Testpersonen – möglichst schriftlich – erklärt, was sie im Test zu tun haben; hierbei hat sich die Bearbeitung einiger gleichartiger Probe-Items als sehr hilfreich erwiesen. Eine genau festgelegte Instruktion soll sicherstellen, dass das Testergebnis nicht davon abhängt, welcher Testleiter den Test durchführt. Es muss auch festgelegt werden, ob und wie etwaige Fragen der Testpersonen zum Test behandelt werden sollen. Normalerweise werden Fragen durch Rückverweis auf die Instruktion beantwortet, weshalb dort alles Wesentliche enthalten sein sollte. Die Instruktion wird nach Möglichkeit schriftlich vorgegeben, um Testleitereinflüsse (z.B. unterschiedliche Betonungen) zu vermeiden. ► Beispiel 2.1 veranschaulicht die möglichen Auswirkungen bei einer mündlich unterschiedlich betonten Instruktion.

Beispiel 2.1: Auswirkung der Instruktion bei einem Leistungstest

Im *Frankfurter Aufmerksamkeits-Inventar 2* (FAIR-2; Moosbrugger und Oehlschlägel 2011) lautet die schriftliche Instruktion „Arbeiten Sie möglichst ohne Fehler, aber so schnell Sie können.“ Wenn man sich vorstellt, der Testleiter würde mündlich in einem Fall besonders den ersten Aspekt (also, „möglichst ohne Fehler“) betonen, in einem anderen Fall aber den zweiten Aspekt (also „aber so schnell Sie können“), so wird offensichtlich, dass das Testergebnis durch Versuchsleitereffekte verfälscht werden kann.

Das Ziel der Durchführungsobjektivität besteht also darin, dass die Testleistung nur von der Merkmalsausprägung der Testperson abhängt und nicht von anderen verzerrenden Variablen (s. hierzu ► Kap. 4, ► Abschn. 4.4) beeinflusst ist. Eine absolute Durchführungsobjektivität ist stets anzustreben, aber in der Realität nicht immer erreichbar.

2.3.1.2 Auswertungsobjektivität und Skalierung

Definition

Die **Auswertungsobjektivität** eines Tests/Fragebogens ist dann gegeben, wenn es eine eindeutige Anweisung gibt, wie die Antworten der Testperson auf die einzelnen Testaufgaben hinsichtlich der Unterscheidung von hohen bzw. niedrigen Merkmalsausprägungen zu kodieren sind. Das Ergebnis der Kodierung darf nicht von der Person des Testauswerters abhängig sein.

Die **Auswertungsobjektivität** bezieht sich auf die einzelnen Items und ist in hohem Maße von dem verwendeten Antwortformat (► Kap. 5) abhängig. Bei Tests/Fragebogen mit gebundenem Antwortformat (z.B. bei Multiple-Choice-Tests mit Mehrfachwahltaufgaben, ► Kap. 5) ist die Auswertungsobjektivität bei den einzelnen Testaufgaben im Allgemeinen problemlos zu erreichen. Bei Leistungstests (► Kap. 3, ► Abschn. 3.2.1) kann zwischen richtigen und falschen Antworten einfach unterschieden werden und auch bei Persönlichkeitstests (► Kap. 3, ► Abschn. 3.2.2) kann nach inhaltlichen Gesichtspunkten festgelegt werden, welche Antwortalternative „symptomatisch“ für eine hohe Merkmalsausprägung ist und welche nicht. Somit kann die Vergabe von Punktwerten/Itemwerten für die einzelnen Aufgaben sicher erfolgen. Wenn hingegen ein offenes Antwortformat verwendet wird, bei dem die Testperson nicht zwischen mehreren Antwortalternativen wählen kann, sondern ihre Antwort selbst erzeugen muss, bedarf es zur

Vergabe von Punktwerten/Itemwerten

Gewinnung von Itemwerten detaillierter Kodierungsregeln, deren Anwendung rasch zu Problemen führen kann (► Beispiel 2.2).

2

Beispiel 2.2: Auswertungsobjektivität bei einem Intelligenztest

Es ergeben sich beispielsweise Schwierigkeiten bei der Auswertung einer Intelligenztestaufgabe zum *Finden von Gemeinsamkeiten*, wenn für eine eher „schwache“ Antwort nur ein Punkt, für eine „gute“ Antwort hingegen zwei Punkte gegeben werden sollen. Nennt eine Testperson z. B. für das Begriffspaar „Apfelsine – Banane“ als Gemeinsamkeit „Nahrungsmittel“, eine andere hingegen „Früchte“, so muss der Test klare Anweisungen im Manual dafür enthalten, welche Antwort höher bewertet werden soll als die andere, um die Auswertungsobjektivität zu gewährleisten.

Übereinstimmung verschiedener Testauswerter

Statistische Kontrolle systematischer Abweichungen

Gütekriterium der Skalierung

Adäquate Entsprechung von Merkmalsausprägungen und Testwerten

Bei freien/offenen Antwortformaten (► Kap. 5) und insbesondere bei projektiven Testverfahren (► Abschn. 3.2.4, 4.7.2), bei denen die Testpersonen ihre Antworten völlig ungebunden gestalten können, ist es erforderlich, die Auswertungsobjektivität empirisch nachzuweisen. Dies erfolgt durch den Grad der Übereinstimmung, der von verschiedenen Testauswertern bei der Auswertung erreicht wird. Ein Test ist umso auswertungsobjektiver, je einheitlicher die Auswertungsregeln von verschiedenen Testauswertern angewendet werden. Eine statistische Kennzahl zur Überprüfung der Auswerterübereinstimmung kann z. B. in Form des „Konsordanzkoeffizienten W“ nach Kendall (1962) berechnet werden. (Für weitere Übereinstimmungsmaße sei im Überblick auf Wirtz und Caspar 2002, verwiesen.)

Anmerkung: In bestimmten Situationen können anhand modelltheoretischer Überlegungen mögliche systematische Abweichungen, die bei unterschiedlichen Auswertern/Beurteilern vorkommen (z. B. Merkmalsbeurteilung durch Lehrer, Eltern und Kinder), statistisch kontrolliert werden. Dazu sind starke theoretische Vorüberlegungen über die Variationsquellen erforderlich, die das Zustandekommen der Beurteilungen beeinflussen. In Form einer Multitrait-Multimethod-Matrix können solche Datenlagen erfasst und analysiert werden (s. Eid 2000; ► Kap. 25).

Neben der Gewinnung von adäquaten Itemwerten für die einzelnen Testaufgaben stellt sich die Frage, wie die bei den einzelnen Items/Fragen/Aufgaben erzielten Itemwerte, die sich alle gemeinsam auf das interessierende Merkmal beziehen, zu einem numerischen Testwert zusammengefasst werden können, der die Merkmalsausprägung auf einer „Skala“ widerspiegelt. Für die hierzu erforderliche „Verrechnungsregel“ muss das Gütekriterium der *Skalierung* beachtet werden, welches fordert, dass die jeweiligen Testwerte (Zahlen aus dem sog. „numerischen Relativ“) die tatsächlichen Merkmalsrelationen zwischen den verschiedenen Testpersonen (sog. „empirisches Relativ“) adäquat abbilden.

Definition

Ein Test erfüllt das Gütekriterium der **Skalierung**, wenn die laut Verrechnungsregel resultierenden Testwerte (numerisches Relativ) die tatsächlichen Merkmalsrelationen (empirisches Relativ) adäquat abbilden.

Das Gütekriterium der Skalierung betrifft bei Leistungstests beispielsweise die Forderung, dass eine leistungsfähigere Testperson einen höheren/besseren Testwert als eine weniger leistungsfähige Testperson erhalten muss, d. h., dass sich die (empirische) Relation der Leistungsfähigkeiten (z. B. mehr bzw. weniger gelöste Aufgaben) auch in den resultierenden Testwerten (höhere bzw. niedrigere Zahl im numerischen Relativ) adäquat widerspiegelt. In analoger Form bedeutet die Forderung der Skalierung bei Persönlichkeitstests, dass eine empirisch größere Merkmalsausprägung (z. B. stärkere Extraversion) mit einer größeren Anzahl an

2.3 · Allgemeine Gütekriterien für Tests und Fragebogen

symptomatischen Antworten und einem entsprechend höheren Testwert einhergehen muss.

Die Umsetzbarkeit dieses Gütekriteriums hängt insbesondere vom Skalenniveau des Messinstruments ab. Zunächst ist festzustellen, dass eine Messung des Merkmals auf Nominalskalenniveau nur ausreicht, um Zuordnungen zu ungereihten Klassen vorzunehmen. Um Größer-Kleiner-Relationen zwischen den Testpersonen beschreiben zu können, muss die Messung zumindest Ordinalskalenniveau (Rangskalenniveau) aufweisen, wobei ein höherer Testwert auf eine leistungsfähigere Testperson schließen lässt. Eine Messung auf Intervallskalenniveau erlaubt darüber hinaus eine Beurteilung der Größe von Testwertdifferenzen. Um Proportionen/Verhältnisse zwischen verschiedenen Testwerten bilden zu können, müssen Messungen auf Rationalsskalen- oder Verhältnisskalenniveau vorliegen, was bei der Konstruktion psychologischer Tests nur schwierig erzielt werden kann. (Mehr zum Vorgang des Messens und zu den Skalenniveaus s. z. B. Bortz und Schuster 2010.)

Im Rahmen der KTT (► Kap. 13) wird der Testwert zumeist durch Addition der Itemwerte der gelösten bzw. symptomatisch beantworteten Aufgaben bestimmt. Über das mit diesem Vorgehen einhergehende Skalenniveau äußern sich Lord und Novick (1968, zit. nach Rost 1996, S. 21) wie folgt:

- » Wenn man einen Testwert, z. B. durch Aufsummierung richtiger Antworten bildet und die resultierenden Werte so behandelt, als hätten sie Intervalleigenschaften, so kann dieses Verfahren einen guten Prädiktor für ein bestimmtes Kriterium hervorbringen, muß [es] aber nicht. In dem Ausmaß, in dem diese Skalierungsprozedur einen guten empirischen Prädiktor hervorbringt, ist auch die postulierte Intervall-skala gerechtfertigt.

Im Rahmen der IRT ist man nicht darauf angewiesen, das Skalenniveau eines Tests mittels seiner praktischen Bewährung bei der Prädiktion externer Kriterien („Kriteriumsvalidität“, ► Abschn. 2.4.2.2 sowie ► Kap. 21) zu bestimmen. Vielmehr kann das Gütekriterium der Skalierung statistisch überprüft werden, indem untersucht wird, ob die Verrechnungsvorschrift durch bestimmte theoriebasierte mathematische IRT-Modelle begründbar ist oder nicht (► Kap. 18).

2.3.1.3 Interpretationsobjektivität und Normierung (Eichung)

Neben Durchführungs- und Auswertungsvorschriften erfordert die übergeordnete Definition des Gütekriteriums der Objektivität auch klare, anwenderunabhängige Regeln für die Testwertinterpretation.

Definition

Interpretationsobjektivität liegt vor, wenn verschiedene Testanwender gleiche Testwerte von verschiedenen Testpersonen bezüglich des untersuchten Merkmals in gleicher Weise interpretieren.

Der Fokus der Testwertinterpretation bezieht sich hier ausschließlich auf das untersuchte Merkmal und nicht auf darauf aufbauende Schlussfolgerungen. So wäre es beispielsweise geboten, sich bei der Interpretation von Ergebnissen in einem Intelligenztest auf die erzielten Testwerte des untersuchten Merkmals „Intelligenz“ zu beschränken. Weiterführende Schlussfolgerungen jenseits des untersuchten Merkmals sind nicht Gegenstand der Interpretationsobjektivität, sondern fallen unter das Gütekriterium der Validität (► Abschn. 2.4.2 sowie ► Kap. 21). Übereinstimmen müssen nur die Interpretationen des Testwertes hinsichtlich des untersuchten Merkmals Intelligenz.

Vor dem Hintergrund der Interpretationsobjektivität ist die sog. „normorientierte Interpretation“ von der sog. „kriteriumsorientierten Interpretation“ zu unterscheiden.

Vergleichsmöglichkeiten von Testwerten in Abhängigkeit vom Skalenniveau

Normorientierte Interpretation von Testwerten

Um eine *normorientierte Interpretation* zu ermöglichen, muss der Test normiert werden. Der Zweck der Normierung besteht darin, aussagekräftige „Vergleichswerte“ von entsprechenden Personen der Zielgruppe in Form von Testnormen zu gewinnen.

Definition

Ein Test gilt als **normiert (geeicht)**, wenn für ihn ein Bezugssystem erstellt wurde, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen der Zielgruppe eindeutig eingeordnet und interpretiert werden können.

Erstellung von Normtabellen aus Eichstichproben

Um Testnormen zu gewinnen, muss als Bezugsgruppe eine größere Eichstichprobe untersucht werden, die für die jeweilige Zielgruppe des Tests/Fragebogens repräsentativ ist. Aus den Testwerten der Eichstichprobe können dann Testnormen (Normtabellen) erstellt werden, die eine Einordnung der Testwerte im Vergleich zu jenen der relevanten Zielgruppe ermöglichen.

Bei der normorientierten Testwertinterpretation werden die Testergebnisse der untersuchten Person hinsichtlich ihrer relativen Stellung zu den Ergebnissen der Testpersonen in der Eichstichprobe interpretiert. Hierbei ist darauf zu achten, dass die Vergleichspersonen hinsichtlich relevanter Merkmale (z. B. Alter, Geschlecht, Schulbildung) ähnlich sind; andernfalls müssen für die relevanten Merkmale separate Normtabellen erstellt werden („Normdifferenzierung“, ▶ Kap. 9).

Bei der Relativierung eines Testergebnisses an den Testergebnissen der Eichstichprobe ist es am anschaulichsten, wenn als Normwert der *Prozentrang* der Testwerte herangezogen wird. Der Prozentrang beschreibt den Prozentsatz derjenigen Personen in der Eichstichprobe, die im Test besser bzw. schlechter abgeschnitten haben als die jeweilige Testperson. Beispielsweise bedeutet ein Prozentrang von 73 in einem Intelligenztest, dass 73 % der Personen in der Eichstichprobe gleiche oder schwächere Testleistungen aufweisen; 27 % weisen hingegen bessere Testleistungen auf. Der Prozentrang kumuliert die in der Eichstichprobe erzielten prozentualen Häufigkeiten der Testwerte bis einschließlich zu jenem Testwert, den die gerade interessierende Testperson erzielte.

Weitere Normierungstechniken, die zur Relativierung eines Testergebnisses herangezogen werden, beziehen sich in der Regel auf den Abstand des individuellen Testwertes Y_v vom Mittelwert der Testergebnisse in der entsprechenden Eichstichprobe. Die resultierende Differenz wird in Einheiten der Standardabweichung der Testwertverteilung gemessen. Hierbei ist zu berücksichtigen, ob das interessierende Merkmal in der Population normalverteilt ist. Ist dies der Fall, kann die Interpretation über die Flächenanteile unter der Standardnormalverteilung („ z -Verteilung“) erfolgen.

Die aus der z -Verteilung gewonnenen Normwerte werden als *Standardwerte* bezeichnet; die Normtabellen mit Standardwerten heißen *Standardnormen*. Häufig verwendet werden auch Normwerte, die auf den z -Werten aufbauen, z. B. *IQ-Werte*, *T-Werte*, *Centil-Werte*, *Stanine-Werte*, *Standardschulnoten*, *PISA-Werte*. Auf diese Normwerte gehen Goldhammer und Hartig in ▶ Kap. 9 näher ein.

Standardnormen

Liegt keine Normalverteilung vor, können zur Interpretation lediglich Prozentrangwerte herangezogen werden, da diese nicht verteilungsgebunden sind (unter sehr spezifischen Umständen können nicht normalverteilte Merkmale durch eine „Flächentransformation“ normalisiert werden, ▶ Kap. 8).

Verteilungsform der Eichstichprobe beachten

Anmerkung: Um eine angemessene Vergleichbarkeit der Personen zu gewährleisten, dürfen die Normtabellen nicht veraltet sein. So sieht beispielsweise die DIN 33430 (Westhoff et al. 2010) bei Verfahren bzw. Tests zur berufsbezogenen Eignungsbeurteilung vor, dass spätestens nach acht Jahren die Gültigkeit der Eichung zu überprüfen ist und ggf. eine Aktualisierung oder auch eine Neunormierung vorgenommen werden sollte. Wesentliche Gründe für die Notwendigkeit

Normenaktualisierung und Normenverschiebung

2.3 · Allgemeine Gütekriterien für Tests und Fragebogen

von Neunormierungen können z. B. Lerneffekte in der Population (insbesondere in Form eines Bekanntwerdens des Testmaterials) oder auch im Durchschnitt tatsächlich veränderte Testleistungen in der Population sein. Das nachfolgende ▶ Beispiel 2.3 beschreibt eine empirisch beobachtete Verringerung der Testleistung in der Population.

Beispiel 2.3 Normenverschiebung im Adaptiven Intelligenz Diagnostikum (AID)

(nach Kubinger 2003, S. 201)

In Bezug auf den AID aus dem Jahr 1985 und den AID 2 aus dem Jahr 2000 zeigte sich eine Normenverschiebung im Untertest „Unmittelbares Reproduzieren-numerisch“ (Kubinger 2001): Die Anzahl der in einer Folge richtig reproduzierten Zahlen (z. B.: 8-1-9-6-2-5) lag im Jahr 2000 im Vergleich zu früher, vor ca. 15 Jahren, über das Alter hinweg fast durchweg um eine Zahl niedriger. Waren es 1985 bei den 7- bis 8- bzw. 9- bis 10-Jährigen noch 5 bzw. 6 Zahlen, die durchschnittlich in einer Folge reproduziert werden konnten, so waren es im Jahr 2000 nur mehr 4 bzw. 5 Zahlen. Ein Nichtberücksichtigen dieses Umstands würde bedeuten, dass Kinder in ihrer Leistungsfähigkeit im Vergleich zur altersgemäßen Normleistung wesentlich unterschätzt würden.

Kriteriumsorientierte Testwertinterpretation: Von der zuvor beschriebenen Erzielung von Interpretationsobjektivität durch Normorientierung ist die kriteriumsorientierte Testwertinterpretation zu unterscheiden. Hier geht es nicht um die relative Stellung des Einzelnen im Vergleich zur Zielpopulation, sondern um die Zuordnung von Testleistungen zu inhaltlich begründbaren markanten Merkmalsausprägungen. Die Interpretationsobjektivität wird dadurch erreicht, dass festgelegt wird, welche Testwerte für das Vorhandensein bestimmter Merkmalsausprägungen sprechen und welche dagegen. Im klinisch-diagnostischen Bereich beispielsweise liegt ab einem bestimmten Testwert in einem Depressionsfragebogen die eingehendere Untersuchung einer „Major Depression“ nahe (▶ Kap. 9). Auch bei der Beurteilung von Schulleistungen kann beispielsweise die feste Zuordnung von erzielten Testwerten zu bestimmten „Kompetenzniveaus“ (▶ Kap. 17) einen wichtigen Beitrag zur Interpretationsobjektivität leisten.

Möglichst schon bevor man beginnt, sich auf die Erfüllung aller Aspekte von Objektivität zu konzentrieren, sollten Fragebogen- und Testautoren auch fünf weitere „allgemeine Gütekriterien“ im Auge behalten. Drei dieser Gütekriterien sollen gewährleisten, dass der Test/Fragebogen dem Anspruch eines ökonomischen, nützlichen und zumutbaren Routineverfahrens gerecht wird. Die beiden anderen Kriterien beziehen sich auf die Vermeidung von bewussten Verzerrungen und von „unfairen“ Items/Fragen/Aufgabestellungen, um möglichst genaue Aussagen über den (relativen) Grad der individuellen Merkmalsausprägungen zu erzielen.

Kriteriumsorientierte Testwertinterpretation

Berücksichtigung weiterer allgemeiner Gütekriterien

2.3.2 Ökonomie

Das Gütekriterium der Testökonomie bezieht sich auf die Wirtschaftlichkeit eines Fragebogens/Tests und wird durch die Kosten bestimmt, die bei einer Testung entstehen. In der Regel stimmen die Interessen von Testpersonen, Auftraggebern und Testleitern in dem Wunsch überein, für die Konstruktion und den Einsatz eines Routineverfahrens keinen überhöhten Aufwand zu betreiben. Allerdings lassen sich oftmals die Kosten nicht beliebig niedrig halten, ohne dass andere Gütekriterien darunter leiden.

Wirtschaftlichkeit eines Tests

Definition

Ein Test erfüllt das Gütekriterium der **Ökonomie**, wenn er – gemessen am diagnostischen Erkenntnisgewinn – wenig finanzielle und zeitliche Ressourcen beansprucht.

Finanzieller Aufwand

Im Wesentlichen beeinflussen zwei Faktoren die Ökonomie bzw. die Kosten einer Testung, und zwar der finanzielle Aufwand für das Testmaterial sowie der zeitliche Aufwand für die Testdurchführung.

Der bei einer Testung entstehende *finanzielle Aufwand* kann sich vor allem aus dem Verbrauch des Testmaterials ergeben oder aus der Beschaffung des Tests selbst. Zudem kann bei computerbasierten Tests (► Kap. 3) die Beschaffung aufwendiger Computerhardware und -software einen wesentlichen Kostenfaktor darstellen. Nicht zu vergessen sind anfallende Lizenzgebühren für Testverlage und -autoren, die mit den Beschaffungskosten des Testmaterials einhergehen.

Zeitlicher Aufwand

Das zweite Merkmal der Ökonomie, der *zeitliche Aufwand*, bildet oftmals einen gewichtigeren Faktor als die Testkosten alleine. Die Testzeit umfasst nicht nur die Nettozeit der Durchführung des Tests, durch die sowohl den Testpersonen als auch dem Testleiter Kosten entstehen, sondern auch die Zeit der Vorbereitung, der Auswertung, der Interpretation und der Ergebnisrückmeldung.

Ökonomievorteile durch adaptives Testen

Zusammenfassend kann man also sagen, dass der Erkenntnisgewinn aus dem Einsatz eines Tests größer sein soll als die entstehenden Kosten. Eine Beurteilung der Ökonomie ist oft nur im Vergleich mit ähnlichen Tests bestimmbar. Vor allem Tests, die am Computer vorgegeben werden können (computerbasierte Tests, ► Kap. 3 und 6) erfüllen bestimmte Ökonomieaspekte vergleichsweise leichter. Ein wichtiger Beitrag zur ökonomischeren Erkenntnisgewinnung kann auch durch das adaptive Testen (vgl. ► Kap. 20) erzielt werden, bei dem nur jene Aufgaben von der Testperson zu bearbeiten sind, die jeweils den größten individuellen Informationsgewinn erbringen. Allerdings erfordern computerbasierte Tests mitunter einen wesentlich höheren Entwicklungsaufwand.

Die Fokussierung auf eine hohe Wirtschaftlichkeit darf natürlich nicht zulasten der anderen Gütekriterien gehen. So muss eine geringere Ökonomie eines Tests bei einer konkreten Fragestellung insbesondere dann in Kauf genommen werden, wenn sein Einsatz z. B. aus Validitätsgründen sachlich gerechtfertigt ist. Dies wäre beispielsweise dann der Fall, wenn nur mit dem ausgewählten Test belastbare Ergebnisse zur konkreten Fragestellung erzielt werden können, mit anderen – ökonomischeren – Tests hingegen nicht.

2.3.3 Nützlichkeit

Doch nicht nur die Ökonomie, sondern auch die Nützlichkeit und die Relevanz einer Testung will wohl bedacht sein.

Definition

Das Gütekriterium der **Nützlichkeit** eines Tests ist gegeben, wenn das von ihm gemessene Merkmal praktische Relevanz aufweist und die auf seiner Grundlage getroffenen Entscheidungen (Maßnahmen) mehr Nutzen als Schaden erwarten lassen.

Praktische Relevanz und Nutzen eines Tests

Für einen Fragebogen/Test besteht dann praktische Relevanz, wenn er ein Merkmal misst, das im Sinne der Kriteriumsvalidität (► Abschn. 2.4.2.2 sowie ► Kap. 21) nützliche Anwendungsmöglichkeiten aufweist. Der Nutzen von getroffenen Entscheidungen wird am nachfolgenden ► Beispiel 2.4 veranschaulicht.

Beispiel 2.4: Nützlichkeit des Tests für medizinische Studiengänge (TMS)

Die Konstruktion eines Tests zur Studieneignungsprüfung für ein medizinisches Studium (TMS; Institut für Test- und Begabungsforschung 1988) erfüllte seinerzeit das Kriterium der Nützlichkeit. Da angesichts der Kosten, die mit dem Studium eines medizinischen Faches verbunden sind, ein Bedarf an der korrekten Selektion und Platzierung der potentiellen Medizinstudierenden bestand, wurde damals ein Test konstruiert, der das komplexe Merkmal „Studieneignung für medizinische Studiengänge“ erfassen und eine Vorhersage bezüglich des Erfolgs der ärztlichen Vorprüfung ermöglichen sollte (Trost 1994). Zu diesem Zeitpunkt gab es keinen anderen Test, der dies in ähnlicher Form in deutscher Sprache zu leisten vermochte. Der Nutzen des TMS wurde anhand aufwendiger Begleituntersuchungen laufend überprüft.

2.3.4 Zumutbarkeit

Darüber hinaus müssen Fragebogen/Tests so gestaltet werden, dass die Testung zumutbar ist in dem Sinne, dass die Testpersonen bezüglich des Zeitaufwands sowie des physischen und psychischen Aufwands nicht über Gebühr beansprucht werden. Die Zumutbarkeit eines Tests betrifft dabei ausschließlich die Testpersonen und nicht den Testleiter. Die Frage nach der Beanspruchung des Testleiters ist hingegen ein Aspekt der Testökonomie (► Abschn. 2.3.2).

Definition

Ein Test erfüllt das Kriterium der **Zumutbarkeit**, wenn er hinsichtlich des aus seiner Anwendung resultierenden Nutzens die Testpersonen in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet.

Zeitliche, physische und psychische Beanspruchung der Testpersonen

Im konkreten Fall ist eine verbindliche Unterscheidung zwischen zu- und unzumutbar oft schwierig, da es jeweils um eine kritische Bewertung dessen geht, was unter „Nutzen“ zu verstehen ist. Dabei spielen neben sachlichen Notwendigkeiten auch gesellschaftliche Normen eine wesentliche Rolle. Beispielsweise gilt es als durchaus akzeptabel, einem Anwärter auf den anspruchsvollen Beruf des Piloten einen sehr beanspruchenden Auswahltest zuzumuten (z. B. im Bereich der Aufmerksamkeit). Bei der Auswahl für weniger anspruchsvolle Tätigkeiten würde ein ähnlich beanspruchendes Verfahren hingegen auf wenig Verständnis stoßen.

Zumutbarkeit hängt von der Fragestellung ab

2.3.5 Fairness

Das Gütekriterium der Fairness befasst sich mit dem Ausmaß, in dem Testpersonen verschiedener Gruppen (Geschlecht, Hautfarbe oder Religion etc.) in einem Test oder bei den resultierenden Schlussfolgerungen in fairer Weise, d. h. nicht diskriminierend, behandelt werden.

Definition

Ein Test erfüllt das Gütekriterium der **Fairness**, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen.

Culture-Fair-Tests

Die Frage nach der Fairness eines Tests bzw. der daraus resultierenden Entscheidungen bezieht sich dabei vor allem auf verschiedene Aspekte, die unmittelbar mit den Inhalten der Testitems zu tun haben, und wurde bereits in den 1970er-Jahren insbesondere vor dem Hintergrund der Intelligenzdiagnostik diskutiert (vgl. Stumpf 1996). Eine besondere Rolle spielen in diesem Zusammenhang sog. „Culture-Fair-Tests“ (z. B. Grundintelligenztest Skala 3, CFT 3; Cattell und Weiß 1971), bei denen die Lösung der Aufgaben nicht oder zumindest nicht stark an eine hohe Ausprägung kulturspezifischer sprachlicher Kompetenz gebunden ist. Darunter ist zu verstehen, dass die Aufgaben bei diesen Verfahren derart gestaltet sind, dass die Testpersonen weder zum Verstehen der Instruktion noch zur Lösung der Aufgaben über hohe sprachliche Fähigkeiten verfügen müssen oder – allgemeiner – über andere Fähigkeiten, die mit der Zugehörigkeit zu einer besonderen soziokulturellen Gruppe einhergehen. Dennoch bezeichnet „culture-fair“ bei der Konstruktion von Testitems eher einen Ansatz als eine perfekte Umsetzung. So konnte vielfach gezeigt werden, dass trotz der Intention der Testautoren dennoch ein Rest von „Kulturkonfundierung“ erhalten bleibt (Süß 2003).

Neben der Berücksichtigung der Sprachproblematik bei der Itembearbeitung bezieht sich der Aspekt der *Durchführungsfairness* beispielsweise auch auf die Berücksichtigung von Fähigkeiten beim Einsatz von Computern bei älteren und jüngeren Menschen. Hierbei sind ebenfalls Verzerrungen in Form eines Ergebnisbias zu erwarten, da nach wie vor viele ältere Menschen im Umgang mit Computern weniger vertraut sind als jüngere.

In Hinblick auf die Beurteilung der Fairness eines Tests gilt es ebenfalls die *Testroutine* zu bedenken. Unterschiedliche Testerfahrung oder Vertrautheit mit Testsituationen („test sophistication“), aber auch Übungseffekte bei Testwiederholungen sind ganz allgemein Größen, die das Ergebnis unabhängig vom interessierenden Merkmal beeinflussen können. Da es keine Faustregeln zum Umgang mit diesem Gütekriterium gibt, ist jeder Test individuell auf seine Fairness hin zu beurteilen.

Verschiedentlich ist es möglich, aufgetretene Formen von Unfairness durch „Normdifferenzierung“, zu kompensieren. So können z. B. altersgruppengestaffelte Normtabellen erstellt werden, um den unterschiedlichen Schwierigkeitsgrad von Intelligenzitems für verschiedene Altersgruppen auszugleichen. Hierfür sind aber sorgfältige fachliche Abwägungen erforderlich (► Kap. 9). Auch durch eine differenzierte Normierung nach Erst- bzw. Zweittestung, wie sie z. B. im Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II; Moosbrugger und Goldhammer 2007) zum Ausgleich von Übungseffekten realisiert ist, kann bei der normorientierten Interpretation der Testergebnisse eine höhere Fairness erzielt werden.

2.3.6 Unverfälschbarkeit

Definition

Ein Testverfahren erfüllt das Gütekriterium der **Unverfälschbarkeit**, wenn das Verfahren derart konstruiert ist, dass die Testperson die konkreten Ausprägungen ihrer Testwerte durch gezielte Vortäuschung eines für sie unzutreffenden Testverhaltens nicht verzerren kann.

Während bei Leistungstests allenfalls gezielte Verfälschungen „nach unten“ (nicht hingegen „nach oben“) auftreten können, sind Persönlichkeitsfragebogen prinzipiell anfällig für Verzerrungen (z. B. Minnesota Multiphasic Personality Inventory, MMPI; Heilbrun 1964; Viswesvaran und Ones 1999). Dies gilt insbesondere dann, wenn sie eine hohe Augenscheininvalidität (► Abschn. 2.4.2.1) besitzen.

Mit gezieltem ergebnisverfälschendem Verhalten („faking“) ist vor allem dann zu rechnen, wenn die Testpersonen das Messprinzip des Fragebogens/Tests durch-

Ergebnisverfälschendes Verhalten durch Soziale Erwünschtheit

schauen und somit leicht erkennen können, wie sie antworten müssen, um sich in einem günstigen Licht darzustellen (s. „Soziale Erwünschtheit“, ▶ Kap. 4). Allerdings ist zu beachten, dass nicht alle Persönlichkeitsfragebogen bzw. deren Subskalen gleichermaßen anfällig für Verzerrungen durch Soziale Erwünschtheit sind. So ist beispielsweise von Costa und McCrae (1985) im Rahmen einer Studie zum NEO-Persönlichkeitsinventar (NEO-PI) gezeigt worden, dass lediglich die Skala „Neurotizismus“ bedeutsam von der Sozialen Erwünschtheit beeinflusst wird.

Um Verfälschungstendenzen vonseiten der Testpersonen insbesondere in Richtung der Sozialen Erwünschtheit vorzubeugen, können sog. „Objektive Persönlichkeitstests“ im Sinne von R. B. Cattell (vgl. Kubinger 1997) eingesetzt werden. Hierbei werden die Persönlichkeitseigenschaften nicht durch verfälschungsanfällige Selbstbeurteilungen, sondern über das konkrete Verhalten in standardisierten Situationen erschlossen (▶ Kap. 3, ▶ Abschn. 3.1.2). Da die Testpersonen über das Messprinzip im Unklaren gelassen werden, ist die Verfälschbarkeit von Objektiven Persönlichkeitstests geringer (s. Ortner et al. 2006).

Verfälschbarkeit bei Objektiven Persönlichkeitstests gering

2.4 Spezielle testtheoriebasierte Gütekriterien für wissenschaftliche Tests und Fragebogen

Während die in ▶ Abschn. 2.3 besprochenen Gütekriterien von allgemein-planerischer Natur sind und keine spezielle testtheoretische Untermauerung erfordern, beschäftigt sich die zweite Gruppe von Gütekriterien mit den Fragestellungen der Reliabilität und der Validität, die für wissenschaftliche Tests und Fragebogen unumgänglich sind. Für ihr Verständnis und für die Beurteilung, ob bzw. inwieweit diese Kriterien erfüllt sind, ist eine vertiefte Kenntnis testtheoretischer Modelle und Verfahrensweisen zur Überprüfung ihrer Gültigkeit nötig.

2.4.1 Reliabilität

Das Gütekriterium der Reliabilität (Zuverlässigkeit) befasst sich mit der Messgenauigkeit des Tests zur Erfassung von (zumeist latenten, d. h. nicht direkt beobachtbaren) Merkmalen und ist wie folgt definiert:

Definition

Ein Test erfüllt das Gütekriterium der **Reliabilität/Zuverlässigkeit**, wenn er das Merkmal, das er misst, exakt, d. h. ohne Messfehler, misst.

Messgenauigkeit des Tests

Legt man die KTT (▶ Kap. 13; vgl. Eid und Schmidt 2014; Steyer und Eid 2001) und ihre Annahmen zugrunde, so wird die Ausprägung der Reliabilität formal als Quotient aus wahrer Varianz und Gesamtvarianz der Testwerte definiert. Die wahre Varianz bemisst dabei die Merkmalsstreuung der sog. „wahren Testwerte“ (True-Scores). Aus der Differenz zwischen der wahren Varianz und der Gesamtvarianz der beobachteten Testwerte resultiert die Messfehlervarianz, die die „Unreliabilität“ oder Messfehlerbehaftetheit eines Messinstruments repräsentiert. ▶ Beispiel 2.5 hebt die Bedeutung eines reliablen Messinstruments hervor.

Quotient aus wahrer Varianz und Gesamtvarianz der Testwerte

Beispiel 2.5: Die Auswirkung von Messfehlern

Als Beispiel für ein reliables Messinstrument soll in Analogie der Meterstab betrachtet werden. Mit diesem Messinstrument lassen sich Längen sehr genau bestimmen, z. B. die Körpergröße einer Person.

Nun stelle man sich vor, ein „Maßband“ sei nicht aus einem längenbeständigen Material, sondern aus einem Gummiband beschaffen. Es ist offensichtlich, dass ein solches Maßband etwa bei einem Schneider zu äußerst unzufriedenen Kunden führen würde, die etwa über zu kurze/zu lange Hosen oder zu enge/zu weite Hemden klagen müssten, wenn das Maßband bei der Messung unsystematisch gedehnt worden wäre.

In Übertragung z.B. auf die Intelligenzdiagnostik zur Identifizierung von Hochbegabungen ($IQ > 130$) resultieren bei mangelnder Reliabilität viele Fehleinschätzungen, weil die Intelligenz je nach Größe und Vorzeichen des aufgetretenen Messfehlers häufig über- oder unterschätzt würde.

Reliabilitätskoeffizient

Das Ausmaß der Reliabilität eines Tests wird über den sog. „Reliabilitätskoeffizienten“ (Rel) quantifiziert, der einen Wert zwischen null und eins annehmen kann ($0 \leq Rel \leq 1$; vgl. ► Kap. 13 und 14). Ein Reliabilitätskoeffizient von eins bezeichnet das Freisein von Messfehlern. Eine optimale Reliabilität würde sich bei einer Wiederholung der Messung/Testung an derselben Testperson unter gleichen Bedingungen und ohne Merkmalsveränderung darin äußern, dass der Test zweimal zu dem exakt gleichen Ergebnis führt. Ein Reliabilitätskoeffizient von null hingegen zeigt an, dass das Testergebnis ausschließlich durch Messfehler zustande gekommen ist. Der Reliabilitätskoeffizient eines guten Tests sollte .7 nicht unterschreiten; höhere Werte sollten angezielt werden.

Verschiedene Verfahren zur Reliabilitätsbestimmung

Um das Ausmaß der Reliabilität zu schätzen, sind einerseits mehrere „klassische“ Verfahren (► Kap. 14) gebräuchlich, die auf der KTT (► Kap. 13) aufbauen und die Erfüllung sehr strenger Annahmen erfordern. Sind die Annahmen erfüllt, können die beobachteten Kovarianzen zwischen den Itemvariablen oder die Korrelationen zwischen (Halb-)Testwerten zu verschiedenen Messzeitpunkten oder zwischen parallelen Tests zur Schätzung der Reliabilität verwendet werden. Andererseits sind mehr und mehr auch „modellbasierte“ Verfahren (► Kap. 15) gebräuchlich, die ebenfalls auf der KTT beruhen und weniger strenge Annahmen erfordern als die klassischen Verfahren. Zur Schätzung der Reliabilität werden eindimensionale oder mehrdimensionale Modelle der konfirmatorischen Faktorenanalyse (CFA) verwendet (► Kap. 24).

2.4.1.1 Klassische Methoden der Reliabilitätsschätzung

Zunächst sollen die folgenden „klassischen“ Verfahren kurz besprochen werden:

1. Retest-Reliabilität
2. Paralleltest-Reliabilität
3. Split-Half-Reliabilität
4. Cronbachs Alpha

■ ■ Retest-Reliabilität

Um die Reliabilität nach der Retest-Methode zu bestimmen, wird ein und derselbe Test (unter der idealen Annahme, dass sich das zu messende Merkmal selbst nicht verändert hat) zu zwei verschiedenen Messzeitpunkten vorgelegt. Die Reliabilität wird dann als Korrelation zwischen den Testwerten aus der ersten und zweiten Messung ermittelt (► Kap. 14).

Bei der Retest-Reliabilität ist zu beachten, dass die ermittelte Korrelation in Abhängigkeit vom Zeitintervall zwischen beiden Testungen variieren kann, da – je nach Zeitabstand – eine Vielzahl von Einflüssen auf die Messungen denkbar ist, die sich reliabilitätsverändernd auswirken können. Hierbei handelt es sich insbesondere um Übungs- und Erinnerungseffekte oder auch um ein sich tatsächlich veränderndes Persönlichkeitsmerkmal. Veränderungen der wahren Testwerte über die zwei Situationen hinweg können als „Spezifität“ mittels der Latent-State-Trait-

Zwei Messzeitpunkte, derselbe Test

Theorie (LST-Theorie; Steyer 1987; Steyer et al. 2015) explizit identifiziert und berücksichtigt werden (s. auch ► Kap. 26).

■■ Paralleltest-Reliabilität

Etliche Reliabilitätsverändernde Einflüsse (z. B. Übungs- und Erinnerungseffekte, aber auch Merkmalsveränderungen) können eliminiert bzw. kontrolliert werden, wenn die Reliabilität nach dem Paralleltestverfahren bestimmt wird. Hierfür wird die Korrelation zwischen den beobachteten Testwerten aus zwei „parallelen Testformen“ berechnet. Dabei handelt es sich um Testformen, bei denen man nach empirischer Prüfung davon ausgehen kann, dass sie (trotz unterschiedlicher Items) zu gleichen wahren Werten und gleichen Varianzen der Testwerte führen. Parallelle Testformen können durch Aufteilung von inhaltlich und formal möglichst ähnlichen Items (sog. „Itemzwillingen“) auf die zwei Testformen erstellt werden. Ob Parallelität gegeben ist, kann mit faktorenanalytischen Verfahren (► Kap. 24) geprüft werden.

Oftmals ist es nicht möglich, einen Test zu wiederholen oder parallele Testformen herzustellen (sei es, dass die Verzerrungen durch eine Messwiederholung zu hoch wären, dass die Testpersonen zu einem zweiten Termin nicht zur Verfügung stehen oder dass der Itempool nicht groß genug war, um zwei parallele Testformen herzustellen). In solchen Fällen ist es angebracht, den Test in zwei parallele Testhälften zu teilen und die Korrelation der beiden Testhälften zu bestimmen. Ob Parallelität gegeben ist, kann mit faktorenanalytischen Verfahren (► Kap. 24) geprüft werden. Da diese Halbtestkorrelation gewöhnlich niedriger ausfällt als die Gesamtreliabilität des ungeteilten Tests wird eine Korrekturformel (Spearman-Brown-Formel, ► Kap. 14) benötigt, um die Halbtestkorrelation wieder auf eine Gesamtrelia-bilität der ursprünglichen Testlänge („Split-Half-Reliabilität“) hochzurechnen.

Zwei Messzeitpunkte, parallele Testformen

Ein Messzeitpunkt, zwei Testhälften

■■ Cronbachs Alpha (α)

Die Beurteilung der Messgenauigkeit erfolgt auch häufig anhand des Reliabilitätskoeffizienten *Cronbachs Alpha* (Cronbach 1951; Moosbrugger und Hartig 2003, S. 412; vgl. ► Kap. 14). Diese Reliabilitätsschätzung stellt eine Verallgemeinerung der Testhalbierungsmethode in der Weise dar, dass jedes Item eines Tests als eigenständiger Testteil betrachtet wird. Je stärker die Testteile untereinander positiv korrelieren, desto höher ist die Reliabilität der Testwertvariable. Voraussetzung ist die strenge – aber häufig nicht erfüllte – Annahme, dass die Kovarianzen zwischen allen Items identisch sind, was anhand der CFA (► Kap. 24) geprüft werden kann.

Verallgemeinerung der Testhalbierungsmethode

Auf die genaue Herleitung dieser und weiterer Reliabilitätsmaße und detaillier-te Möglichkeiten ihrer Berechnung wird von Gäde, Schermelleh-Engel und Werner in ► Kap. 14 näher eingegangen.

2.4.1.2 Modellbasierte Methoden der Reliabilitätsschätzung

Auf der Basis der KTT wurden neuere Möglichkeiten der Reliabilitätsbestimmung entwickelt und in der Fachwelt diskutiert. Im Vergleich zu den klassischen Methoden (► Kap. 14) beruhen modellbasierte Methoden der Reliabilitätsschätzung (s. Bollen 1980; McDonald 1999; Revelle und Zinbarg 2009; Zinbarg et al. 2005) auf weniger strengen Annahmen, die leichter erfüllt werden können. Sie bilden die Voraussetzung für die Schätzung der Reliabilitätskoeffizienten anhand von eindimensionalen und mehrdimensionalen Modellen der CFA (► Kap. 15).

Weniger strenge Annahmen erforderlich

So wurde erst in jüngerer Zeit genauer erkannt, dass Cronbachs Alpha die Erfüllung (zu) strenger Annahmen voraussetzt und zudem eine Reihe von problemati-schen Eigenschaften bei der Schätzung der Reliabilität aufweist, sodass inzwischen von einer unkritischen Verwendung des Koeffizienten Alpha zur Schätzung der Re-liabilität eher abzuraten ist. Sind die Annahmen jedoch erfüllt (s. z. B. ► Kap. 14; Raykov und Marcoulides 2011; Revelle und Condon 2018), so kann Cronbachs Alpha sowohl klassisch als auch modellbasiert geschätzt werden (► Kap. 15).

Kritik an Cronbachs Alpha

Omega-Koeffizienten

Modellbasiert wurden verschiedene Reliabilitätskoeffizienten entwickelt, die als Omega-Koeffizienten bezeichnet werden (► Kap. 15). Diese Koeffizienten sind nicht nur auf eindimensionale Konstrukte beschränkt, wie dies bei den klassischen Maßen der Fall ist, sondern umfassen auch mehrdimensionale Konstrukte. Modelltests anhand der CFA ermöglichen eine Beurteilung, ob die Voraussetzungen zur Schätzung der Reliabilitätskoeffizienten erfüllt sind. Die Omega-Koeffizienten können – wie auch die Alpha-Koeffizienten – als Punktschätzungen vorteilhaft durch Intervallschätzungen ergänzt werden.

Pauschale vs. testwertabhängige Genauigkeitsbeurteilung**Hinweis**

Während bei Tests, die nach der KTT konstruiert wurden, der Reliabilitätskoeffizient eine pauschale Genauigkeitsbeurteilung der Testwerte ermöglicht (s. Konfidenzintervalle, ► Kap. 13), ist bei Tests, die nach der IRT (► Kap. 16) konstruiert worden sind, darüber hinaus eine speziellere Genauigkeitsbeurteilung der Testwerte mithilfe der „Informationsfunktion“ der verwendeten Testitems möglich. Diese berücksichtigt die konkrete Merkmalsausprägung der Testperson, während in der KTT eine Reliabilität für alle Testpersonen gilt.

2.4.2 Validität

Beim Gütekriterium der Validität (vgl. ► Kap. 21) handelt es sich hinsichtlich der praktischen Anwendung von Tests und der theoretischen Diskussion von Merkmalszusammenhängen um das wichtigste Gütekriterium überhaupt, wobei eine hohe Objektivität und eine hohe Reliabilität zwar notwendige, aber keine hinreichenden Bedingungen für eine hohe Validität darstellen.

Das Gütekriterium der Validität befasst sich generell mit der inhaltlichen Übereinstimmung zwischen dem, was der Test misst, und dem Merkmal, das man mit dem Test messen möchte, und insbesondere auch mit der Belastbarkeit von Testwertinterpretationen sowie den Schlussfolgerungen, die auf der Basis von Testergebnissen hinsichtlich eines Außenkriteriums gezogen werden.

In nicht näher differenzierter Weise wird die Validität (Gültigkeit) häufig wie folgt definiert:

Definition

Validität/Gültigkeit eines Tests liegt vor, wenn der Test das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes.

Verschiedene Validitätsaspekte

Für eine differenziertere Beurteilung dessen, was ein Test misst, können und sollten verschiedene Aspekte herangezogen werden (► Beispiel 2.6).

Beispiel 2.6: Validitätsaspekte der Schulreife

- a. Wenn man beispielsweise die Validität eines Tests für das zu messende Kriterium „Schulreife“ beurteilen will, wäre als erster Aspekt zu prüfen, ob die Testpersonen (und ihre Eltern) per Augenschein akzeptieren können, dass hier etwas überprüft wird, das für Schulreife ausschlaggebend erscheint. Die konstruierten Items würden vor allem dann eine hohe Akzeptanz erfahren, wenn sie Verhaltens- und Erlebensweisen überprüfen, die auch dem Laien als für das Merkmal relevant erscheinen. Dies ist dann der Fall, wenn diese Items eine hohe sog. *Augenscheininvalidität* aufweisen. Jedem Laien ist beispielsweise intuitiv

einsichtig, dass Schulreife u. a. dadurch gekennzeichnet ist, dass Kinder mit niedrigen Zahlen umgehen und verbalen Ausführungen (z. B. einer lehrenden Person) aufmerksam folgen können. Insofern kann man vom bloßen Augenschein her jenen Items, die solche Fähigkeiten erfassen, diese Form von Validität zusprechen.

- b. Augenscheininvalidität darf aber nicht mit inhaltlicher Validität verwechselt werden. Man darf sich nicht einfach darauf verlassen, welchen *Eindruck* die Items vermitteln; vielmehr gilt es zu überprüfen, ob das interessierende Merkmal „Schulreife“ tatsächlich durch geeignete Testaufgaben (Items) operationalisiert wurde. Die Forderung, die dabei an die Items zu richten ist, besteht in erster Linie darin, dass die Items das interessierende Merkmal *inhaltlich* in seiner vollen Breite *repräsentativ* abbilden. Man spricht hierbei von *Inhaltsvalidität*. Im Falle der Schulreife wären insbesondere Testaufgaben für niedrige Zahlen, für das Sprachverständnis und für die sprachliche Ausdrucksfähigkeit zu konstruieren; aber auch soziale Kompetenzen sowie motivationale und emotionale Variablen sollten dabei Berücksichtigung finden.
- c. Ein weiterer Problembereich beschäftigt sich mit der *Berechtigung* und der *Belastbarkeit* von *diagnostischen Entscheidungen*, die durch *extrapolierende Verallgemeinerungen* von Testergebnissen auf das Verhalten der Testpersonen außerhalb der Testsituation („Kriterium“) vorgenommen werden. Ob solche extrapolierenden Schlussfolgerungen gerechtfertigt sind und zu tragfähigen Entscheidungen führen, kann als *Kriteriumsvalidität* durch Untersuchung der Zusammenhänge zwischen Testwerten und Kriterien außerhalb der Testsituation empirisch überprüft werden. Als Maß der Kriteriumsvalidität werden z. B. Korrelationen zwischen dem Testwert (Schulreifetest) und dem Kriterium (tatsächliche Schulreife, z. B. in Form des Lehrerurteils) berechnet.
- d. Eine berechtigte Frage besteht aber auch darin, ob „Schulreife“ als ein eindimensionales Merkmal mit heterogenen Iteminhalten oder als mehrdimensionales Merkmal mit jeweils homogeneren Iteminhalten aufgefasst werden kann und sollte. Um hierbei nicht auf Spekulationen oder ideologische Standpunkte angewiesen zu sein, wird die Frage der Ein- bzw. Mehrdimensionalität der zur Merkmalserfassung konstruierten Items sowie die Abgrenzung zu anderen Merkmalen als *Konstruktvalidität* empirisch untersucht. Hierbei kommen sowohl struktursuchende als auch strukturprüfende faktorenanalytische Verfahren zum Einsatz.²

Wie ► Beispiel 2.6 zeigt, sollte man für ein differenziertes Bild über die Validität eines Tests also sinnvollerweise zumindest die aufgeführten Validitätsaspekte

- a. Augenscheininvalidität,
- b. Inhaltsvalidität,
- c. Kriteriumsvalidität und
- d. Konstruktvalidität

berücksichtigen.

2.4.2.1 Augenschein- und Inhaltsvalidität

Vor dem Hintergrund der Akzeptanz vonseiten der Testpersonen kommt der *Augenscheininvalidität* eines Tests eine erhebliche Bedeutung zu.

Akzeptanz des Tests vonseiten der Testperson

² Im Fall der Schuleingangsuntersuchung, die sehr unterschiedliche Merkmale erfasst (u. a. Seh- und Hörvermögen, Aufmerksamkeit), ist zu erwarten, dass sich eine mehrdimensionale Lösung ergeben würde.

Definition

Augenscheininvalidität gibt an, inwieweit der Gültigkeitsanspruch eines Tests vom bloßen Augenschein her einem Laien gerechtfertigt erscheint.

Nicht zuletzt auch wegen der Bekanntheit der Intelligenzforschung haben z. B. Intelligenztests eine hohe Augenscheininvalidität, da es Laien aufgrund des Testinhalts und der Testgestaltung für glaubwürdig halten, dass damit Intelligenz gemessen werden kann. Dies kommt auch der Vermittelbarkeit der Testergebnisse zugute.

Aus wissenschaftlicher Perspektive ist allerdings festzustellen, dass Augenscheininvalidität nicht mit Inhaltsvalidität verwechselt werden darf, obwohl dies leicht passieren kann (vgl. Tent und Stelzl 1993), da augenscheininvaliden Tests oftmals zugleich auch Inhaltsvalidität zugesprochen wird.

Die *Inhaltsvalidität* wird in der Regel nicht numerisch anhand eines Maßes bzw. Kennwertes bestimmt, sondern aufgrund „logischer und fachlicher Überlegungen“ (vgl. Cronbach und Meehl 1955; Michel und Conrad 1982), die bei der Planung (► Kap. 3) und bei der Itemgenerierung (► Kap. 5) ihren Niederschlag finden müssen.

Definition

Inhaltsvalidität liegt vor, wenn die Testitems im Zuge der Operationalisierung so konstruiert und ausgewählt werden, dass sie das interessierende Merkmal repräsentativ abbilden.

Repräsentationsschluss

Zur Erfüllung der Inhaltsvalidität sollen die Items eines Tests/Fragebogens eine repräsentative Stichprobe an Verhaltens- und Erlebensweisen aus jenem Itemuniversum (d. h. allen merkmalsrelevanten Verhaltens- und Erlebensweisen) darstellen, mit dem das interessierende Merkmal vollständig erfasst werden könnte. Bei der Beurteilung, inwieweit die Inhalte der Items das interessierende Merkmal repräsentativ erfassen, spielt die Bewertung der Items durch Experten eine maßgebliche Rolle.

Am einfachsten ist die Frage nach der Inhaltsvalidität eines Tests dann zu klären, wenn eine „simulationsorientierte Zugangsweise“³ gewählt wird (s. Moosbrugger und Rauch 2010), bei der die einzelnen Items unmittelbar Auskunft über den Verhaltensbereich geben, über den eine Aussage getroffen werden soll. Dies ist z. B. dann der Fall, wenn Rechtschreibkenntnisse anhand eines Diktats überprüft werden oder die Eignung eines Autofahrers anhand einer Fahrprobe ermittelt wird. Dabei ist die Eignung des Autofahrers besser (inhaltsvalider) zu ermitteln, wenn er in einer Prüfung länger fährt (z. B. 45 Minuten), als wenn er nur kurz „um die Ecke“ fährt und wieder aussteigt. So wird der Autofahrer während einer längeren Fahrt zahlreichen Entscheidungssituationen ausgesetzt sein (z. B. „Rechts-vor-links“-Situationen, Kreisverkehr, Einparken, Autobahn), während er bei einer sehr kurzen Fahrt vielleicht nur vier Mal rechts abgebogen wäre.

Differenziertere Überlegungen zu Aspekten der Inhaltsvalidität werden von Brandt und Moosbrugger in ► Kap. 3 (insbesondere in ► Abschn. 3.1 zur Spezifikation des interessierenden Merkmals) besprochen.

2.4.2.2 Kriteriumsvalidität und extrapolierte Testwertinterpretationen

Extrapolierte Interpretationen

Die Kriteriumsvalidität bezieht sich auf die Frage, welche *extrapolierenden Interpretationen* von Testergebnissen auf das Verhalten der Testpersonen außerhalb der Testsituation („Kriterium“) zulässig sind. Kriteriumsvalidität liegt z. B. bei einem

³ Der Begriff des „simulationsorientierten Zugangs“ unterscheidet sich vom gleichnamigen Begriff, wie er beispielsweise in der Pädagogik verwendet wird, wenn eingeweihte Akteure (etwa Schauspieler) in einer Untersuchungssituation eine Situation selbst „simulieren“.

2.4 · Spezielle testtheoriebasierte Gütekriterien für wissenschaftliche Tests und Fragebogen

„Schulreifetest“ vor allem dann vor, wenn jene Kinder, die im Test leistungsfähig sind, sich auch im Kriterium „Schule“ als leistungsfähig erweisen und umgekehrt, wenn jene Kinder, die im Test leistungsschwach sind, sich auch in der Schule als leistungsschwach erweisen.

Definition

Ein Test weist **Kriteriumsvalidität** auf, wenn von einem Testwert (gewonnen aus dem Verhalten innerhalb der Testsituation) erfolgreich auf ein „Kriterium“, d. h. auf ein Verhalten außerhalb der Testsituation, extrapoliert werden kann. Die Enge dieser Beziehung und ihre Belastbarkeit bestimmen das Ausmaß der Kriteriumsvalidität.

Liegt eine hohe Kriteriumsvalidität vor, so erlauben die jeweiligen Testergebnisse die Extrapolation des in der Testsituation beobachteten Verhaltens auf das interessierende Verhalten außerhalb der Testsituation. Man bezeichnet die Testergebnisse dann auch als valide hinsichtlich des jeweiligen Kriteriums. Empirisch kann man die Kriteriumsvalidität eines Testwertes im einfachsten Fall durch die Berechnung der Korrelation der Testwerte in der Testsituation mit einem interessierenden Verhalten außerhalb der Testsituation (Kriterium) überprüfen.

Obwohl es von Vorteil ist, wenn für extrapolierende Schlussfolgerungen der inhaltlich-theoretische Hintergrund der mit dem Test erfassten Konstrukte und vor allem deren Dimensionalität genau untersucht sind („Konstruktvalidität“, ► Abschn. 2.4.2.3), ist die Überprüfung der Kriteriumsvalidität im Prinzip an keine besonderen testtheoretischen Annahmen gebunden. Somit können bei praktischen Anwendungen auch Testwerte mit (noch) nicht geklärter inhaltlich-theoretischer Fundierung eine empirisch festgestellte Kriteriumsvalidität aufweisen. Ein Beispiel hierfür findet sich bei Goldhammer und Hartig in ► Kap. 9. Auch die „kriteriumsorientierte Strategie der Itemgenerierung“ (► Kap. 4) basiert auf diesem Sachverhalt.

Anwendungspraktisch wird man die Kriteriumsvalidität eines Tests nicht nur mit einer einzigen Korrelation ausdrücken können, sondern vor allem über das Ausmaß, in dem die Angemessenheit und die Güte von Interpretationen auf Basis von Testwerten oder anderen diagnostischen Verfahren durch empirische Belege und theoretische Argumente gestützt sind. Insbesondere dieser Aspekt wird in ► Kap. 21 ausführlich vertieft.

Abhängig von der zeitlichen Verfügbarkeit des Kriteriums, d. h., ob es bereits in der Gegenwart oder erst in der Zukunft vorliegt, spricht man gelegentlich auch von *Übereinstimmungsvalidität* (sog. „konkurrenter Validität“) bzw. von *Vorhersagevalidität* (sog. „prognostischer Validität“). Im ersten Fall ist also der Zusammenhang eines Testwertes mit einem Kriterium von Interesse, das zeitgleich existiert, im zweiten Fall steht die Prognose einer zukünftigen Ausprägung eines Merkmals im Vordergrund.

Feststellung der Kriteriumsvalidität als Test-Kriterium-Korrelation

Stützung der Interpretation durch theoretische Annahmen und empirische Belege Zeitliche Verfügbarkeit des Kriteriums

2.4.2.3 Konstruktvalidität

Unter dem Aspekt der Konstruktvalidität beschäftigt man sich mit der theoretischen Fundierung (vor allem mit der Dimensionalität und der Struktur) des mit dem Test gemessenen Merkmals.

Definition

Ein Test weist **Konstruktvalidität** auf, wenn die Zusammenhangsstruktur zwischen den Testitems und den interessierenden (Persönlichkeits-)Merkmälern („Konstrukte“, „latente Variablen“, „Traits“, „latente Klassen“, z. B. Fähigkeiten, Dispositionen, Charakterzüge oder Einstellungen) wissenschaftlich fundiert ist.

Gemeint ist, ob z. B. von den Testaufgaben eines „Intelligenztests“ wirklich auf die Ausprägung eines latenten Persönlichkeitsmerkmals „Intelligenz“ geschlossen

Struktursuchendes vs. strukturprüfendes Vorgehen

werden kann oder ob die Aufgaben eigentlich ein anderes Konstrukt (etwa „Konzentration“ anstelle des interessierenden Konstrukts „Intelligenz“) messen.

Die Beurteilung der Konstruktvalidität erfolgt häufig unter Zuhilfenahme *struktursuchender* und *strukturprüfender* methodischer Ansätze. Während die struktursuchenden Verfahren dabei helfen, geeignete Hypothesen über die Dimensionalität des interessierenden Merkmals zu gewinnen, dienen die strukturprüfenden Verfahren der statistischen Absicherung der vermuteten Dimensionalität.

- ■ **Struktursuchende faktorenanalytische Verfahren zur Konstruktvalidierung**
- Zur Gewinnung von Hypothesen über die Ein- bzw. Mehrdimensionalität der Merkmalsstruktur der Testitems werden vor allem *exploratorische Faktorenanalysen* (EFA) zum Einsatz gebracht (► Kap. 23).
- Innerhalb der einzelnen Faktoren geben die *Faktorladungen* in Analogie zu den Trennschärfekoeffizienten der deskriptivstatistischen Itemanalyse (► Kap. 7) Auskunft über die Homogenität der Testitems.

Exploratorische Faktorenanalyse

Die solchermaßen gewonnenen Merkmalsdimensionen erlauben eine erste deskriptive Einordnung in ein bestehendes Gefüge hypothetischer Konstrukte. Dabei kann z. B. die Bildung eines „nomologischen Netzwerks“ nützlich sein (► Kap. 21), wobei die Betrachtung der Zusammenhänge zu anderen Tests/Merkmalen im Vordergrund steht. Dazu formuliert man inhaltliche Überlegungen über den Zusammenhang des vorliegenden Tests bzw. des/der von ihm erfassten Merkmals/Merkmale mit konstruktverwandten bzw. konstruktfremden bereits bestehenden Tests/Merkmalen. Danach werden die Testergebnisse empirisch mit denen anderer Tests hinsichtlich Ähnlichkeit bzw. Unähnlichkeit verglichen, wobei zwischen *konvergenter* und *diskriminanter* Validität unterschieden wird.

Zur Feststellung der *konvergente[n] Validität*, die Hinweise dafür liefert, dass ein Test tatsächlich das interessierende Merkmal und nicht irgendein anderes misst, kann das Ausmaß der Übereinstimmung mit Ergebnissen aus Tests für gleiche oder konstruktverwandte Merkmale ermittelt werden. So sollte z. B. die Korrelation eines neuartigen Intelligenztests mit einem etablierten Test, z. B. dem Intelligenz-Struktur-Test 2000R (I-S-T 2000R; Liepmann et al. 2007), zu einer hohen Korrelation führen, um zu zeigen, dass auch der neue Test das Konstrukt „Intelligenz“ misst und nicht irgendein anderes Konstrukt.

Neben der konvergenten Validität ist aber auch die *diskriminante Validität* wichtig, die Hinweise dafür liefert, dass das Testergebnis des interessierenden Tests/Merkmales von Testergebnissen in anderen, konstruktfremden Tests/Merkmalen abgrenzbar ist. So soll beispielsweise ein Konzentrationsleistungstest ein diskriminierbares eigenständiges Konstrukt, nämlich „Konzentration“, erfassen und nicht das Gleiche wie andere Tests für andere Konstrukte. Wünschenswert sind deshalb niedrige korrelative Zusammenhänge zwischen Konzentrationstests und Tests für andere Variablen. Zum Nachweis der diskriminanten Validität ist es aber nicht hinreichend, dass der zu validierende Test nur mit den Ergebnissen aus irgendwelchen offensichtlich konstruktfremden Tests verglichen wird, sondern dass er auch zu relativ konstruktnahen, aber nicht konstruktverwandten Tests in Beziehung gesetzt wird. So wäre z. B. eine niedrige Korrelation zwischen einem Konzentrationstest und einem Intelligenztest zur Feststellung der Existenzberechtigung von eigenständigen Konstrukten wünschenswert (so z. B. FAKT-II, Moosbrugger und Goldhammer 2007). Hierdurch soll gewährleistet werden, dass ein Konstrukt/Merkmal wirklich abgrenzbar ist und nicht schon früher unter anderem Namen vorgeschlagen worden war (man denke z. B. an die Debatte Depressivität vs. Burn-out).

Um zu vermeiden, dass „Methodenfaktoren“ irrtümlich für abgrenzbare inhaltliche Merkmale gehalten werden, können korrelationsbasierte Multitrait-Multimethod-Analysen (MTMM-Analysen) erste deskriptivstatistische Anhaltspunkte liefern (► Kap. 25).

■■ Strukturprüfende statistische Verfahren zur Konstruktvalidierung

Die **strukturprüfende Vorgehensweise** erlaubt es, inferenzstatistische Schlüsse bezüglich der Konstruktvalidität zu ziehen. Dies ist nur auf der Basis von testtheoretischen Annahmen möglich, die eine explizite und inferenzstatistisch überprüfbare Beziehung zwischen zuvor genau definierten, latenten Merkmalen (beispielsweise Intelligenz) und ihren Indikatorvariablen (den Testitems) herstellen.

- Mit *Latent-Trait-Modellen* (Roskam 1996) können die Beziehungen zwischen den Testitems und quantitativen latenten Konstrukten statistisch überprüft werden (► Kap. 16).
- Mit *Latent-Class-Analysen* (LCA; Lazarsfeld und Henry 1968) können die Beziehungen zwischen den Testitems und qualitativen latenten Klassen statistisch überprüft werden (► Kap. 22).
- Mit *konfirmatorischen Faktorenanalysen* (CFA; Jöreskog und Sörbom 1996) können die in exploratorischen Faktorenanalysen gefundenen dimensionalen Strukturen der Testitems (► Kap. 23) inferenzstatistisch abgesichert werden (► Kap. 24). Dies ist allerdings nur dann sinnvoll, wenn die Absicherung nicht an den Datensätzen der exploratorischen Analysen, sondern an neuen Datensätzen erfolgt. Man spricht dann von einer „Kreuzvalidierung“.
- Eine weitere konfirmatorische Vorgehensweise der Konstruktvalidierung ermöglichen faktorenanalytische *Multitrait-Multimethod-Analysen* (MTMM-Analysen) im Rahmen latenter Strukturgleichungsmodelle (Eid 2000). Dabei wird der Zusammenhang zwischen verschiedenen Merkmalen (Traits) unter Herauspartialisierung der Methodeneinflüsse strukturprüfend untersucht (► Kap. 25).
- Als weitere Frage der Konstruktvalidierung stellt sich, ob ein Merkmal als ein zeitlich überdauerndes Merkmal oder als ein hinsichtlich situativer (d. h. nicht messfehlerbezogener) Einflüsse temporär variierendes Merkmal zu betrachten ist. Mithilfe der *Latent-State-Trait-Theorie* (LST-Theorie; Steyer 1987; Steyer et al. 2015) kann eine Zerlegung in zeitlich variierende State- und zeitlich stabile Trait-Anteile vorgenommen werden, die eine Überprüfung dieses Aspekts der Konstruktvalidität erlaubt (► Kap. 26).
- Schließlich kann die *LST-Theorie* mit der faktorenanalytischen *MTMM-Analyse* in ein Modell zusammengefasst werden. Dieses erlaubt eine Überprüfung der konvergenten und diskriminanten Validität über die Zeit (► Kap. 27).

Strukturprüfendes Vorgehen

2.4.2.4 Argumentationsbasierter Validierungsansatz von Testwertinterpretationen

In den letzten Jahren hat sich das Verständnis von Validität deutlich weiterentwickelt (► Kap. 21). Während früher die Validität als Eigenschaft eines Tests betrachtet wurde, bezieht sich die Validität heute auf die Interpretation von Testwerten und die aus ihnen abgeleiteten Handlungen (vgl. Messick 1989). Validität wird somit inzwischen verstärkt als einheitliches Qualitätskriterium betrachtet, das Informationen aus verschiedenen Quellen integriert und einen fortwährenden argumentativen Prozess darstellt. Da Tests für unterschiedliche Zwecke eingesetzt werden, erfordert jede intendierte Testwertinterpretation eine separate Validierung.

Im Rahmen des sog. „argumentationsbasierten Ansatzes“ ist es zunächst notwendig, festzulegen, auf welche Interpretationen der Testwerte sich die intendierte Validität beziehen soll. Dann werden die zu validierende Testwertinterpretation präzise formuliert und empirisch überprüfbare Grundannahmen identifiziert. Hierauf wird empirische Evidenz gesammelt, anhand derer die Grundannahmen widerlegt oder vorläufig gestützt werden können. Wichtige Evidenzquellen sind die Testinhalte, die bei der Testbeantwortung ablaufenden kognitiven Prozesse, die interne Struktur der Testdaten und die Beziehungen der Testwertvariablen zu anderen Konstrukten. Bei der abschließenden zusammenfassenden Bewertung werden Testwertinterpretationen dann als valide betrachtet, wenn keine der zugrunde liegenden

Annahmen widerlegt werden konnte. Dieser Ansatz wird ausführlich von Hartig, Frey und Jude in ► Kap. 21 behandelt.

2

2.5 Dokumentation der erfüllten Qualitätskriterien

Im „Testmanual“, der Handanweisung eines fertig entwickelten Tests, sollte in geeigneter Weise dokumentiert sein, welche Testgütekriterien in welcher Weise erfüllt sind.

Die beschriebenen Qualitätsanforderungen werden darüber hinaus nach Möglichkeit durch weitere Teststandards ergänzt, die sich u. a. auf Evaluationsfragen, Übersetzungen und Adaptionen der Tests beziehen. Ausführungen hierzu finden sich zum psychologischen Testen bei Höfling und Moosbrugger in ► Kap. 10, zum pädagogischen Testen bei Brückner, Zlatkin-Troitschanskaia und Pant in ► Kap. 11.

2.6 Zusammenfassung

Laienfragebogen bestehen häufig aus einer Ansammlung von Fragen, die in keinem unmittelbaren Bezug zueinander stehen; wissenschaftliche Messinstrumente (Tests und Fragebogen) hingegen erfassen zumeist einzelne latente, d. h. nicht direkt beobachtbare Merkmale, die mit mehreren Operationalisierungen dieses Merkmals in Form der Testitems erschlossen werden.

Die Bandbreite von einem Laienfragebogen bis hin zu einem wissenschaftlichen Fragebogen/Test kann als Kontinuum aufgefasst werden. Ein Fragebogen/Test ist umso wissenschaftlicher, je mehr Qualitätsanforderungen („Gütekriterien“) bei seiner Konstruktion Beachtung finden. Von grundlegender Wichtigkeit für Fragebogen und Tests sind die Durchführungs-, Auswertungs- und Interpretationsobjektivität, aber auch weitere Aspekte wie Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness und Unverfälschbarkeit. Die Berücksichtigung dieser Gütekriterien erfordert keine besonderen testtheoretischen Kenntnisse.

Für wissenschaftliche Tests ist die Erfüllung der Gütekriterien der Reliabilität und Validität unumgänglich, für deren genaue Beurteilung spezielle testtheoretische Kenntnisse (KTT bzw. IRT und faktorenanalytische Modelle) vorausgesetzt werden. Die Reliabilität befasst sich mit der Messgenauigkeit eines Tests; sie kann mit verschiedenen Verfahren empirisch überprüft werden. Die Validität beschäftigt sich mit der Frage, ob ein Test das Merkmal, das er messen soll, auch wirklich misst. Hierbei sind die Aspekte der Augenschein-, Inhalts-, Kriteriums- und Konstruktvalidität von Bedeutung. In jüngerer Zeit verschiebt sich der Betrachtungsfokus mehr und mehr auf den „argumentationsbasierten Ansatz“, um festzustellen, mit welcher Berechtigung extrapolierende Schlussfolgerungen aus den Testergebnissen gezogen werden können.

2.7 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Welche Formen von Objektivität kennen Sie?
2. Was versteht man unter „Normierung“ (Testeichung)?
3. Erklären Sie bitte eine Möglichkeit, einen Test zu normieren.
4. Wie kann man die Testökonomie erhöhen?
5. Was versteht man unter Testfairness?

6. Worin unterscheiden sich die verschiedenen Verfahren zur Reliabilitätsbestimmung?
7. Welche wesentlichen Validitätsaspekte sollten Berücksichtigung finden und warum?
8. Warum ist nicht nur die konvergente, sondern auch die diskriminante Validität wichtig?

Literatur

- Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, 45, 370–390.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.
- Cattell, R. B. & Weiß, R. H. (1971). *Grundintelligenztest Skala 3 (CFT 3)*. Göttingen: Hogrefe.
- Costa, P. T. & McCrae, R. R. (1985). *The NEO Personality Inventory Manual*. Odessa, FL: Psychological Assessment Resources.
- 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.
- Deutsches Institut für Normung e.V. (DIN). (2002). *DIN 33430:2002-06: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Deutsches Institut für Normung e.V. (DIN). (2016). *DIN 33430:2016-07: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Heilbrun, A. B. (1964). Social learning theory, social desirability, and the MMPI. *Psychological Bulletin*, 61, 377–387.
- Institut für Test- und Begabungsforschung (Hrsg.). (1988). *Test für medizinische Studiengänge (aktualisierte Originalversion 2)*. Herausgegeben im Auftrag der Kultusminister der Länder der BRD (2. Aufl.). Göttingen: Hogrefe.
- Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Chicago: Scientific Software International.
- Kendall, M. G. (1962). *Rank Correlation Methods*. London, UK: Griffin.
- Kubinger, K. D. (1997). Zur Renaissance der objektiven Persönlichkeitstests sensu R. B. Cattell. In H. Mandl (Hrsg.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* (S. 755–761). Göttingen: Hogrefe.
- Kubinger, K. D. (2001). Zur Qualitätssicherung psychologischer Tests – Am Beispiel des AID 2. *Psychologie in Österreich*, 21, 82–85.
- Kubinger, K. D. (2003). Gütekriterien. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik*. (S. 195–204). Weinheim: Beltz PVU.
- Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. (6. Aufl.). Weinheim: Psychologie Verlags Union.
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *I-S-T 2000R Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.
- Lord, F. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: American Council on Education and Macmillan.
- Michel, L. & Conrad, W. (1982). Testtheoretische Grundlagen psychometrischer Tests. In K.-J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie* (Bd. 6, S. 19–70). Göttingen: Hogrefe.
- McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. New York, NY: Psychology Press.
- Moosbrugger, H. & Goldhammer, F. (2007). *Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT II): Grundlegend neu bearbeitete und neu normierte 2. Auflage des FAKT von Moosbrugger und Heyden (1997)*. Göttingen: Hogrefe.
- Moosbrugger, H. & Hartig, J. (2003). Klassische Testtheorie. In K. Kubinger und R. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 408–415). Weinheim: Psychologie Verlags Union.
- Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Berlin, Heidelberg: Springer.
- Moosbrugger, H. & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2)*. Bern, Göttingen: Huber.

- Moosbrugger, H. & Rauch, W. (2010). Grundkenntnisse über Verfahren der Eignungsbeurteilung. In K. Westhoff, C. Hagemeister, M. Kersting, F. Lang, H. Moosbrugger, G. Reimann, G. Stemmler (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Aufl., S. 146–148). Lengerich: Pabst.
- Ortner, T. M., Proyer, R. T. & Kubinger, K. D. (Hrsg.). (2006). *Theorie und Praxis Objektiver Persönlichkeitstests*. Bern, Stuttgart, Hans Huber.
- Raykov, T. & Marcoulides, G. A. (2011). *Psychometric Theory*. New York, NY: Routledge.
- Reiß, S. & Sarris, V. (2012). *Experimentelle Psychologie*. München: Pearson.
- Revelle, W. & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth & D. Hughes (Eds.), *The Wiley Blackwell Handbook of Psychometric Testing* (pp. 709–749). Chichester, West Sussex, UK: Blackwell Publishing Ltd.
- Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 145–154.
- Rosenthal, R. & Rosnow, R. L. (1969). The volunteer subject. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 59–118). New York, NY: Academic Press.
- Roskam, E. E. (1996). Latent-Trait-Modelle. In E. Erdfelder, R. Mausfeld, Th. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 431–458). Weinheim: Psychologie Verlags Union.
- Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.
- Steyer, R. (1987). Konsistenz und Spezifität: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 8, 245–258.
- Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin, Heidelberg: Springer.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- Stumpf, H. (1996). Klassische Testtheorie. In: E. Erdfelder, R. Mansfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 411–430). Weinheim: Beltz PVU.
- Süß, H.-M. (2003). Culture fair. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 82–86). Weinheim: Beltz.
- Tent, L. & Stelzl, I. (1993). *Pädagogisch-psychologische Diagnostik. Band 1: Theoretische und methodische Grundlagen*. Göttingen: Hogrefe.
- Testkuratorium (der Föderation deutscher Psychologenverbände). (1986). Mitteilung. *Diagnostica*, 32, 358–360.
- Trost, G. (1994). *Test für medizinische Studiengänge (TMS): Studien zur Evaluation (18. Arbeitsbericht)*. Bonn: ITB.
- Viswesvaran, C. & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59, 197–210.
- Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G. & Stemmler, G. (Hrsg.). (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Aufl.). Lengerich: Pabst.
- Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω_H : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 1–11.



Planungsaspekte und Konstruktionsphasen von Tests und Fragebogen

Holger Brandt und Helfried Moosbrugger

Inhaltsverzeichnis

3.1	Spezifikation des interessierenden Merkmals – 41
3.1.1	Merkmaldefinition – 41
3.1.2	Merkmalsindikatoren – 41
3.1.3	Arten von Merkmalen – 42
3.2	Testarten – 44
3.2.1	Leistungstests – 44
3.2.2	Persönlichkeitsfragebogen und -tests – 47
3.2.3	Objektive Persönlichkeitstests – 47
3.2.4	Projektive Verfahren – 49
3.3	Geltungsbereich und Zielgruppe – 50
3.4	Testlänge und Testzeit – 51
3.5	Testadministration – 53
3.5.1	Paper-Pencil- und computeradministrierte Tests – 53
3.5.2	Computerbasierte Tests – 53
3.5.3	Einzel- und Gruppentestung – 54
3.5.4	Selbst- und Fremdeinschätzung – 54
3.6	Struktureller Testaufbau – 55
3.6.1	Testteil 1: Instruktion – 55
3.6.2	Testteil 2: Konkrete Testaufgaben – 56
3.6.3	Testteil 3: Demografische Angaben – 56

3.7	Konstruktionsphasen im Überblick – 57
3.7.1	Itemgenerierung – 57
3.7.2	Qualitative Verständlichkeitsprüfung der Items – 59
3.7.3	Empirische Erprobung der vorläufigen Testversion – 60
3.7.4	Revision und Abschluss der Testentwicklung – 62
3.7.5	Normierung der endgültigen Testform – 62
3.8	Zusammenfassung – 63
3.9	Kontrollfragen – 63
	Literatur – 64

3.1 · Spezifikation des interessierenden Merkmals

i Psychologische Tests und Fragebogen haben das Ziel, Merkmalsträger (Testpersonen) hinsichtlich ihrer Merkmalsausprägungen einer metrisch vergleichenden Beurteilung zugänglich zu machen. Vor und während der Konstruktion eines Tests sind zahlreiche Aspekte zu berücksichtigen, die es erlauben, die Merkmalsausprägung zu quantifizieren und jeder Person einen (numerischen) Testwert zuzuordnen. Zur Beurteilung, ob eine solche Zuordnung von Testwerten zu Personen angemessen ist, werden verschiedene testtheoretisch basierte psychometrisch-statistische Maße herangezogen. Dieses Kapitel bietet einen Überblick über den Entwicklungsprozess eines Tests oder Fragebogens, angefangen von der ersten Testplanung, über die Testkonstruktion bis hin zur Erstellung und Erprobung einer vorläufigen Version (Pilotstudie) mit dem Ziel der Revision hin zur fertig entwickelten endgültigen Testfassung, die auch eine Normierung der Testwerte mit einschließt. Das Ziel der Darstellungen in diesem Kapitel besteht darin, einen Überblick über die wichtigsten Aspekte zu geben, die bei der Planung von Tests und Fragebogen berücksichtigt werden müssen. Die Kenntnis und das Verstehen der Schritte, die dabei durchlaufen werden müssen, sind nicht nur für Testkonstrukteure von Bedeutung; vielmehr sind sie auch für Testanwender von Nutzen, wenn sie vor dem Problem stehen, verschiedene Testverfahren hinsichtlich ihrer Qualität sowie ihrer Anwendungs- und Aussagemöglichkeiten zu vergleichen und adäquat zu beurteilen.

3.1 Spezifikation des interessierenden Merkmals

3.1.1 Merkmalsdefinition

Am Beginn der Planung und der Entwicklung eines Tests oder Fragebogens steht die Spezifikation eines „interessierenden“ Merkmals, das erfasst werden soll. In einem ersten Schritt ist es hierfür notwendig, das interessierende Merkmal zu definieren und möglichst genau einzuzgrenzen. Hierfür sollte auch eine Literaturrecherche durchgeführt werden, um vorhandene theoretische und empirische Befunde zum interessierenden Merkmal, ggf. aber auch bereits existierende Tests in die Überlegungen mit einzubeziehen.

Definition

Bei einem **Merkmal** handelt es sich um eine (numerisch erfassbare) Variable, hinsichtlich derer sich verschiedene Personen (allgemeiner: Merkmalsträger) unterscheiden. Mit **Merkmalsausprägung** bezeichnet man eine quantitative oder qualitative Angabe darüber, welche Größe das Merkmal bei einer untersuchten Person aufweist.

Während mit Fragebogen im umgangssprachlichen Sinn zumeist direkt erfassbare Merkmale untersucht werden (z. B. Kaufverhalten, Essgewohnheiten), steht ein psychologischer Test vor der Herausforderung, dass es sich bei den interessierenden Merkmalen zumeist um (hypothetische) Konstrukte handelt, und zwar um nicht direkt erfassbare latente Variablen (z. B. „Intelligenz“, „Perfektionismus“), die zur Erklärung interindividueller Verhaltensunterschiede dienen sollen.

Direkt erfassbare Merkmale vs. latente hypothetische Konstrukte

3.1.2 Merkmalsindikatoren

Da sich latente Merkmale auf das Verhalten auswirken, wird davon ausgegangen, dass die Merkmalsausprägungen aus dem beobachtbaren Verhalten der Testperson (manifeste Variablen, „Indikatorvariablen“, z. B. Lösen oder Nichtlösen einer

Studienbox 3.1

Psychometrische Modellvorstellung: Die latente Variable „Intelligenz“ wird durch mehrere Testitems operationalisiert

Die richtige bzw. falsche Antwort auf das jeweilige Item (beobachtbares Verhalten) lässt durch Anwendung eines psychometrischen Modells einen Rückschluss auf die latente Variable zu (Lösungen: Schraube, b, 25).

Ziel psychometrischer Tests ist eine indirekte Erfassung latenter Merkmale

Das Ziel psychometrischer Tests besteht in der indirekten Erfassung latenter Merkmale anhand von Merkmalsindikatoren (sog. „Items“), von denen angenommen werden kann, dass sie einen Rückschluss auf die latenten Merkmale erlauben.

Ein Indikator für einen ausgeprägten Perfektionismus wäre z. B. eine hohe Zufriedenheit bei perfekter Erledigung einer Aufgabe, was mit dem Item „Ich bin mit meiner Arbeit erst zufrieden, wenn sie perfekt ist“ erfragbar wäre.

Um ein latentes Merkmal anhand von Items inhalts valide erfassen zu können (zum Begriff Inhaltsvalidität ► Kap. 2, ► Abschn. 2.4.2.1, 21), muss überlegt werden, welche Eigenschaften für das Merkmal charakteristisch sind und wie diese Eigenschaften aus Handlungen der Testperson in bestimmten Situationen, d. h. aus Reaktionen auf Testaufgaben bzw. -fragen erschlossen werden können (► Studienbox 3.1). Eine *operationale Merkmalsdefinition* gibt somit an, welche beobachtbaren Verhaltensweisen für das interessierende Merkmal charakteristisch sind. Die theoriebasierte Verbindung zwischen den beobachtbaren Verhalten und den latenten Konstrukten wird von psychometrischen Modellen geleistet, die in Teil II dieses Bandes („Testtheorien“) vorgestellt werden.

Operationale Definition von Merkmalen

3.1.3 Arten von Merkmalen

Qualitative vs. quantitative Merkmale

Je nach der zugrunde liegenden *Messintention* kann das interessierende Merkmal in qualitativer oder in quantitativer Form erfasst werden. Man spricht dann von einem *qualitativen Merkmal*, wenn sich Personen bezüglich ihrer Merkmalsausprägung in verschiedene ungeordnete Kategorien (auch „Klassen“) einteilen lassen; ein angemessenes Skalenniveau für die Testwerte wäre das Nominalskalenniveau. Lassen sich die Merkmalsausprägungen darüber hinaus zumindest in geordnete Kategorien einteilen, so spricht man von einem *quantitativen Merkmal*. (Beispielsweise bei sog. „Wissensstrukturen“ Doignon und Falmagne 1999; Heller und Repitsch

3.1 · Spezifikation des interessierenden Merkmals

2008), wäre für die Testwerte das Ordinalskalenniveau angemessen. Kann man des Weiteren davon ausgehen, dass sich das Merkmal als eine kontinuierliche Dimension darstellen lässt (z. B. Ängstlichkeit) und sich Personen darin graduell unterscheiden, wäre für die Testwerte das Intervall- oder auch Verhältnisskalenniveau angemessen. Hierbei unterscheidet sich das Verhältnis- vom Intervallskalenniveau dadurch, dass die Verhältnisskala einen (natürlich/nicht willkürlich) festgelegten Nullpunkt aufweist.

Eine weitere Differenzierung von Merkmalen betrifft das Ausmaß ihrer Veränderung pro Zeit. Es lassen sich zeitlich stabile Merkmale, sog. „Traits“, von zeitlich veränderbaren Merkmalen unterscheiden, deren Ausprägungen als sog. „States“ bezeichnet werden (vgl. ► Kap. 26). Unter Traits im engeren Sinn werden zumeist zeitüberdauernde Persönlichkeitsmerkmale verstanden. States beziehen sich auf Zustände und sind von den jeweiligen Situationen abhängig. Ein Beispiel für einen Test, in dem beide Arten von Merkmalen erhoben werden, ist das State-Trait-Ärgerausdrucks-Inventar (STAXI; Schwenkmezger et al. 1992) bzw. STAXI-2 (Rohrmann et al. 2013; Spielberger 1999), ein Verfahren zur Messung von vier dispositionellen Ärgerdimensionen (Traits) sowie der Intensität von situationsbezogenen Ärgerzuständen (States).

Des Weiteren lassen sich Merkmale in *unidimensionale* und *multidimensionale* Merkmale einteilen. Unidimensional bedeutet, dass sich die Merkmalsausprägungen auf einer Dimension (oft auch als *Skala* bezeichnet) abbilden lassen (vgl. Döring und Bortz 2016). Ein multidimensionales Merkmal liegt hingegen vor, wenn das interessierende Merkmal mehrere (zumindest zwei) Facetten aufweist und zur Abbildung der Merkmalsausprägungen mehrere Dimensionen/(Sub-)Skalen erforderlich sind. In der Praxis sind Merkmale häufig multidimensional (z. B. Intelligenz, Extraversion, Psychopathie, numerische Basiskompetenzen). Beispielsweise ist es bei der Erfassung des multidimensionalen Merkmals „Impulsivität“ (Eysenck 1993) notwendig, die Facetten einzeln zu spezifizieren; die resultierenden Subskalen (nach Eysenck: „Impulsivität“ im engeren Sinne, „Waghalsigkeit“ und „Empathie“) sind dann jeweils wieder unidimensional.

Diesen Überlegungen folgend lassen sich uni- oder multidimensionale Tests entwickeln. Bei einem multidimensionalen Test wird für jede der Facetten des Merkmals ein unidimensionaler Subtest erstellt, wobei die Items jeweils einer einzigen Dimension/Facette zugeordnet sind. Bei den sog. „Persönlichkeitsstrukturtests“ handelt es sich beispielsweise um multidimensionale Tests, die den Anspruch haben, die Persönlichkeit strukturiert nach den zentralen Dimensionen (z. B. „Big-Five“; Stemmler et al. 2011) möglichst umfassend zu beschreiben (► Abschn. 3.2.2).

In neueren psychometrischen Modellen (z. B. ► Kap. 18) wird die Idee aufgegeben, dass ein einzelnes Item nur einem Merkmal zugeordnet ist (unidimensionale Items). Vielmehr ist es auch zulässig, dass einzelne Items simultan als Indikatoren für mehrere Merkmale fungieren (multidimensionale Items). So hängt z. B. die Lösung einer Subtraktionsaufgabe mit Brüchen sowohl von der Fähigkeit ab, subtrahieren zu können, als auch vom Umgang mit Brüchen selbst (Kürzen, Erweitern etc.). Eine Klasse moderner psychometrischer Modelle zur Analyse solcher multidimensionalen Items zur Leistungsmessung stellen die „kognitiven Diagnosemodelle“ dar (de la Torre 2009; von Davier 2014). Im Persönlichkeitsbereich finden sich hierzu Beispiele im Bereich der Multitrait-Multimethod-Analysen (MTMM-Analysen; ► Kap. 25), bei denen untersucht wird, inwieweit die Beantwortung von Items einerseits auf die interessierenden Merkmale selbst („Traits“), andererseits aber auch auf die jeweilige Art der Erfassungsmethode („Methoden“) zurückzuführen ist.

Bevor mit der konkreten Itemkonstruktion zur Erfassung von interessierenden Merkmalen (► Abschn. 3.7.1) begonnen werden kann, sind in der Testplanungsphase eine Reihe weiterer Aspekte zu beachten, die im Folgenden erörtert werden.

Zeitlich stabile vs. zeitlich veränderbare Merkmale

Uni- vs. multidimensionale Merkmale

Uni- vs. multidimensionale Tests

Uni- vs. multidimensionale Items

3.2 Testarten

Je nach Art des zu erfassenden Merkmals werden verschiedene Testarten unterschieden, auf die im Folgenden kurz eingegangen werden soll. Detaillierte Beschreibungen findet der interessierte Leser in Lehrbüchern zur psychologischen Diagnostik (z. B. Beauducel und Leue 2014; Fissen 2004; Kubinger 2009; Petermann und Eid 2006; Pospeschill und Spinath 2009; Schmidt-Atzert und Amelang 2012; Stemmler und Margraf-Stiksrued 2015) oder auch im *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (Brähler et al. 2002).

Obwohl es sich bei allen Merkmalen/Konstrukten, hinsichtlich derer sich Personen unterscheiden, um Persönlichkeitsvariablen handelt, ist es im Zusammenhang mit ihrer Erfassung durch psychologische Tests sinnvoll, zwischen *Leistungstests* und *Persönlichkeitstests* zu unterscheiden.

3.2.1 Leistungstests

Leistungstests

Psychologische Leistungstests beziehen sich in der Regel auf die Erfassung von Dimensionen der kognitiven Leistungsfähigkeit. Den Testpersonen wird „die Lösung von Aufgaben oder Problemen [...], die Reproduktion von Wissen, das Unterbeweisstellen von Können, Ausdauer oder Konzentrationsfähigkeit“ abverlangt (Rost 2004, S. 43).

Definition

Leistungstests sind dadurch gekennzeichnet, dass sie sich aus Testaufgaben zusammensetzen, deren Bearbeitung/Beantwortung in inhaltlich-logischem Sinn als richtig oder falsch bewertet werden kann.

Maximale Verhaltensleistung und Leistungsverfälschung

Das Gemeinsame aller Leistungstests ist, dass in der Regel die maximale Verhaltensleistung gefordert wird. Eine größere als die maximale Leistung kann also nicht vorgetäuscht werden, d. h., es ist keine Verfälschung des Verhaltens nach oben (*Leistungssimulation*, „faking good“), sondern allenfalls eine Verfälschung des Verhaltens nach unten (*Leistungsdisimulation*, „faking bad“) möglich. Üblicherweise kann davon ausgegangen werden, dass die Testpersonen motiviert sind und um die an sie gestellten Anforderungen wissen.

Leistungstests lassen sich hinsichtlich ihrer zeitlichen und schwierigkeitsbezogenen Anforderungen in zwei Gruppen aufteilen (auch andere, differenziertere Aufteilungen sind denkbar, s. z. B. Brähler et al. 2002; Schmidt-Atzert und Amelang 2012):

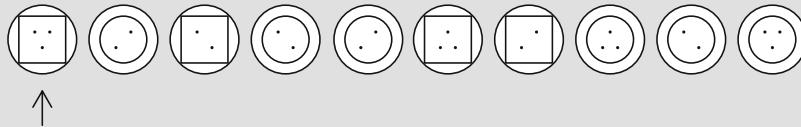
Speed- oder Geschwindigkeitstests verwenden in der Regel einfache Aufgaben, die zumeist von allen Testpersonen gelöst werden können. Die Differenzierung der Leistungen erfolgt üblicherweise durch eine Begrenzung der Bearbeitungszeit bei einer hohen Zahl an zu bearbeitenden Aufgaben. Es wird erfasst, wie viele Aufgaben die Testperson in der begrenzten Bearbeitungszeit richtig bearbeiten kann. Dieses Prinzip wird vor allem zur Feststellung von basalen kognitiven Fähigkeiten genutzt (beispielsweise im „Frankfurter Aufmerksamkeits-Inventar“, FAIR-2; Moosbrugger und Oehlschlägel 2011, das verschiedene Facetten der Aufmerksamkeit misst). Bei computerisierten und insbesondere computerbasierten Tests (► Abschn. 3.5.2) kann aber auch das Bearbeitungstempo selbst („Testing Speed“) zur Leistungserfassung herangezogen werden, wie beispielsweise im Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II; Moosbrugger und Goldhammer 2007; s. auch ► Beispiel 3.1).

Speed- oder Geschwindigkeitstests

3.2 · Testarten

Beispiel 3.1: FAKT-II

Die an die Testperson gestellte Aufgabe besteht darin, am Bildschirm in einer Folge von runden Zeichen die „Zielitems“ zu erkennen, und zwar jene, die innen einen Kreis mit 3 Punkten oder ein Quadrat mit 2 Punkten besitzen:



Beispielitems im FAKT-II (Moosbrugger und Goldhammer 2007, mit freundlicher Genehmigung von Hogrefe)

Der Pfeil zeigt zunächst auf das erste Item, sodann auf das zweite Item usw. Immer, wenn das Zeichen, auf das der Pfeil zeigt, ein Zielitem ist, soll von der Testperson auf der Tastatur die „1“ gedrückt werden, bei den anderen Zeichen die „0“. Die Bearbeitungsaufforderung lautet: „Arbeiten Sie möglichst ohne Fehler und so schnell Sie können.“ Das individuell unterschiedliche Bearbeitungstempo (durchschnittlicher Zeitverbrauch pro Antwort) dient als Testwert für die Konzentrationsleistung.

Power- oder Niveautests hingegen verwenden Aufgaben mit breit gestreuten Schwierigkeitsniveaus bis hin zu Aufgaben, die auch bei theoretisch unbegrenzter Zeitvorgabe nur sehr schwer und auch nur von manchen Testpersonen richtig gelöst werden können.

Zur Differenzierung der Leistung wird geprüft, welches Schwierigkeitsniveau die jeweilige Testperson ohne Zeitbegrenzung bewältigen kann. Verfahren dieses Typs werden primär zur Feststellung komplexerer kognitiver Fähigkeiten genutzt, beispielsweise im Snijders-Oomen Non-verbalen Intelligenztest (SON-R 2-8; Tellegen et al. 2018). Bei diesem Test handelt es sich um ein nonverbales Verfahren zur Intelligenzdiagnostik bei Kindern im Vorschulalter und im Einschulungsalter (► Beispiel 3.2).

Häufig wird auch eine Mischform der beiden genannten Testarten angewendet, z. B. im Wechsler-Intelligenztest für Erwachsene (WAIS-IV; Petermann 2012), bei der einerseits schwierige Aufgaben vorgegeben werden; andererseits wird auch die Zeit zum Lösen der Aufgaben gemessen und bei der Auswertung berücksichtigt. Durch die Mischung beider Testarten können verschiedene Facetten der Intelligenz erfasst werden, beispielsweise die verbale Intelligenz oder die Verarbeitungsgeschwindigkeit. Weitere Hinweise auf Aspekte, die bei einer Entscheidung für bzw. gegen eine Verwendung von Speed- bzw. Powertests zu beachten sind, lassen sich z. B. in Lienert und Raatz (1998, S. 34 ff.) finden.

Apparative Tests stellen eine Gruppe von Verfahren dar, die insbesondere zur Erhebung sensorischer und motorischer (Leistungs-)Merkmale geeignet ist; mit ihnen können z. B. kognitive oder körperliche Fähigkeiten erfasst werden. Typische Vertreter dieser Klasse sind sensumotorische Koordinationstests, Tests zur Messung der Muskelkraft als Indikator für Willensanstrengung, Montage- und Handtiertests sowie Testverfahren vom Typ der Hand-Augen-Koordinationstests, z. B. Zweihand-Koordination (2HAND; Schuhfried 2007) oder der Doppellabyrinthtest (B19; Bonnardel 1946, 2001), die sensumotorische Koordinationsfähigkeiten erfordern (► Beispiel 3.3).

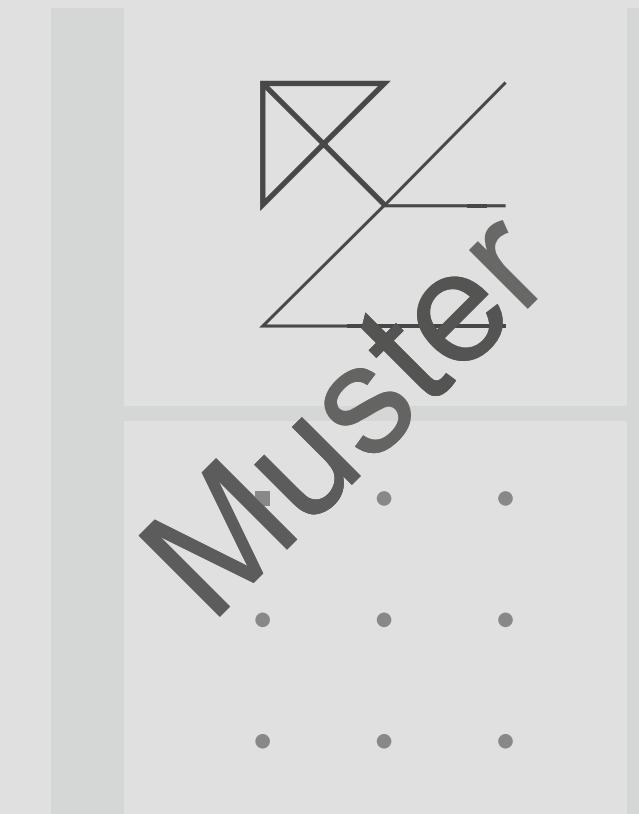
Power- oder Niveautests

Mischformen

Apparative Tests

Beispiel 3.2: SON-R 2-8

Die Aufgabe der Testperson besteht darin, das vorgegebene Muster (oben) auf dem freien Feld (unten) nachzuzeichnen.



Beispielitem in Anlehnung an den SON-R 2-8. (Aus Tellegen et al. 2018, © by Hogrefe Verlag GmbH & Co. KG, Göttingen ● Nachdruck und jegliche Art der Vervielfältigung verboten. Bezugsquelle des Non-verbaler Intelligenztest (SON-R 2-8): Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, www.testzentrale.de)

Beispiel 3.3: Apparative Tests

Durchgang 1



Bewegen Sie den roten Punkt innerhalb der blauen Bahn zu Feld B.

Beispielitem aus dem 2HAND. (Aus Schuhfried 2007, mit freundlicher Genehmigung von Schuhfried)

3.2.2 Persönlichkeitsfragebogen und -tests

Unter dem Begriff „Persönlichkeitsfragebogen und -tests“ werden verschiedene Erhebungsverfahren subsumiert. Neben Instrumentarien und Fragebogen zur Erfassung von stabilen Eigenschaften bzw. temporären Zuständen (Trait vs. State), Symptomen oder Verhaltensweisen existieren auch zahlreiche Verfahren zur Messung von Motivation, Interessen, Meinungen und Einstellungen. Persönlichkeitsstrukturtests und Persönlichkeitstestsysteme dienen der psychometrischen Erfassung von mehreren Persönlichkeitsdimensionen, beispielsweise das Minnesota Multiphasic Personality Inventory 2 (MMPI-2; Hathaway et al. 2000) bzw. dessen revidierte Fassung MMPI-2-RF (Ben-Porath und Tellegen 2011) oder das NEO-Fünf-Faktoren-Inventar (NEO-FFI; Borkenau und Ostendorf 2008), NEO-FFI-3 (McCrae et al. 2005) bzw. das NEO-Persönlichkeitsinventar (NEO-PI-R; Ostendorf und Angleitner 2004) und NEO-PI-3 (McCrae und Costa 2010) zur Erfassung der Big-Five-Persönlichkeitsdimensionen.

Im Gegensatz zu Leistungstests, bei denen es um die Erfassung der maximalen Verhaltensleistung geht, enthalten Persönlichkeitstests und -fragebogen keine mit „richtig“ oder „falsch“ bewertbaren Aufgaben, sondern erfordern in der Regel eine Selbstauskunft über das für die Testperson typische Verhalten und Erleben in Abhängigkeit von der Ausprägung ihrer Persönlichkeitsmerkmale (Verhaltensdispositionen). Die Testaufgaben erfassen das charakteristische Verhalten, wobei es keine optimale Ausprägung der interessierenden Persönlichkeitsmerkmale gibt; vielmehr werden die Antworten danach bewertet, ob sie für eine hohe oder für eine niedrige Ausprägung des interessierenden Merkmals symptomatisch sind.

Definition

Persönlichkeitstests sind dadurch gekennzeichnet, dass sie sich nicht aus Testaufgaben zusammensetzen, die „richtig“ oder „falsch“ bearbeitet werden können; vielmehr werden die Items so gewählt, dass sie für das interessierende Merkmal **charakteristisch** sind und die Antworten als **symptomatisch** für eine hohe bzw. für eine niedrige Merkmalsausprägung bewertet werden können.

Persönlichkeitstests und Persönlichkeitstestsysteme

Typisches Verhalten und Erleben, symptomatische Antworten

Da es sich bei den gegebenen Antworten um subjektive Angaben handelt, sind (von der Testperson beabsichtigte) Verfälschungen in beide Richtungen („faking good“/„faking bad“) möglich, d. h., es kann sowohl eine Simulation einer scheinbar höheren als auch eine Dissimulation einer scheinbar niedrigeren Merkmalsausprägung auftreten (s. auch „Antworttendenzen“, ▶ Abschn. 4.7).

3.2.3 Objektive Persönlichkeitstests

Um dem Problem subjektiver Verfälschungsmöglichkeiten bei Selbstauskunftsfragebogen zu begegnen, werden zunehmend „Objektive Persönlichkeitstests“ konstruiert. Ihre spezielle Form von Objektivität wird zum einen dadurch erreicht, dass keine Selbstbeurteilungen der Testpersonen erfasst werden, sondern (Stich-)Proben von konkretem Verhalten in (experimentell erzeugten) Anforderungssituationen, sodass eine bewusste subjektive Verfälschung der gegebenen Antworten, z. B. im Sinne der „Sozialen Erwünschtheit“ (▶ Kap. 4, ▶ Abschn. 4.7.2), zumindest im Vergleich zu einer Selbstauskunft in traditionellen Persönlichkeitsfragebogen erheblich reduziert ist. Zum anderen werden die Aufgabenstellungen häufig so gewählt, dass sie keine Augenscheininvalidität (▶ Kap. 2) besitzen. Dies bedeutet, dass die Testpersonen den Zusammenhang zwischen Messintention und Messprinzip aus den Testaufgaben nicht oder nur schwer erkennen können. Während die Messintention selbst in den meisten Fällen für die Testpersonen erkenntlich ist, zeichnen

Objektive Persönlichkeitstests begrenzen den Spielraum für individuelle Verhaltensverfälschungen

sich moderne Objektive Persönlichkeitstests dadurch aus, dass das Messprinzip verschleiert ist (z. B. die Verrechnungsvorschriften; s. Kubinger 2006). Deshalb können Testpersonen das Testergebnis – insbesondere im Sinne eines „faking-good“ – nicht bewusst verfälschen (Kubinger 2006). Objektive Persönlichkeitstests sind somit dem klassischen Gütekriterium der Objektivität (► Kap. 2) förderlich, indem sie die Standardisierung der Durchführung, Auswertung und Interpretation von Tests in der Weise ergänzen, dass sie auch den Testpersonen keinen Spielraum für individuelle Verhaltensverfälschungen einräumen.

Reduktion von Verfälschungstendenzen

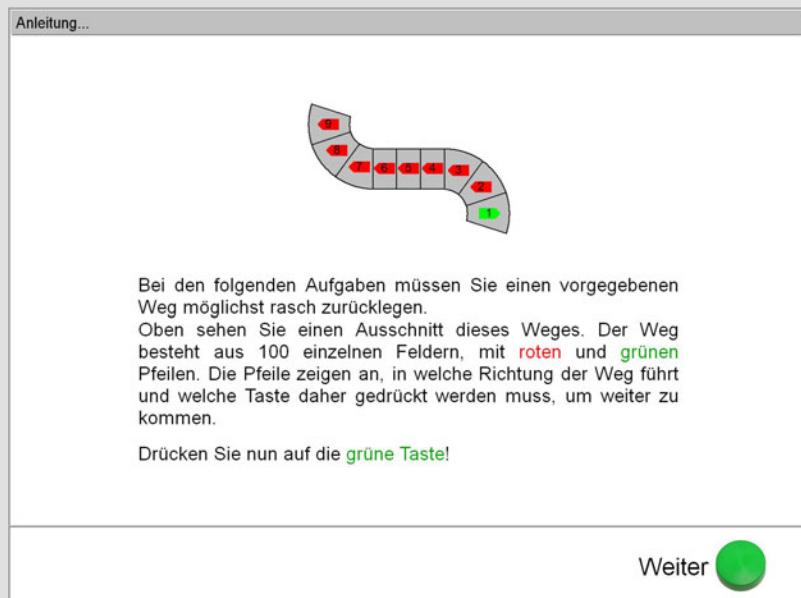
Definition

Als **Objektive Persönlichkeitstests** werden Verfahren bezeichnet, die das Messprinzip so verschleiern, dass den Testpersonen kein Spielraum für bewusste individuelle Verfälschungen des Testergebnisses (insbesondere „faking-good“) eingeräumt wird.

In Objektiven Persönlichkeitstests wird ein Merkmal also zumeist nicht durch individuelle Beurteilungen der eigenen Person, sondern über das Verhalten in standardisierten Situationen erschlossen. Ein Vorteil solcher Verfahren kann auch darin gesehen werden, dass der Überrepräsentation verbaler Konzepte (Fragen, Antworten) entgegengewirkt wird. Ein Merkmal, das auf diese Weise gut erfasst werden kann, ist z. B. Leistungsmotivation, die beispielsweise mit dem Objektiven Leistungsmotivations-Test (OLMT; Schmidt-Atzert 2007) ermittelt werden kann (► Beispiel 3.4).

Beispiel 3.4: OLMT

Die Aufgabe der Testperson besteht darin, unter verschiedenen motivationalen Bedingungen durch Drücken der Cursor- bzw. Pfeiltasten mit den nach links bzw. rechts weisenden Pfeilen den auf dem Bildschirm dargebotenen Weg möglichst rasch zu verfolgen.



Anleitung aus dem OLMT. (Aus Schmidt-Atzert 2007, mit freundlicher Genehmigung von Schuhfried)

Als verschleiertes Messprinzip wird nicht das Arbeitstempo an sich erfasst, sondern die Relation zwischen eigenen und – durch einen fingierten Gegner – fremd vorgegebenen Zielsetzungen. Bereits zurückgelegte Felder werden dunkelgrau markiert.

Zum Bereich der Objektiven Testverfahren zählen auch die *impliziten Assoziationstests* (IAT; Greenwald et al. 1998), bei denen die Merkmalsausprägung durch die Assoziationsstärke gemessen wird, die als Reaktionszeit auf implizite Merkmalsindikatoren operationalisiert ist. Anwendungen objektiver Tests finden sich in verschiedensten Bereichen. Sie sind vor allem dann angezeigt, wenn bei den Testpersonen von einem starken Motiv zu Verfälschungstendenzen ausgegangen werden muss, u. a. in der Delinquenzforschung (z. B. sexuelle Devianz bei Sexualstraftätern; Briken et al. 2013; Schmidt 2013) oder bei der Erfassung von rassistischen Einstellungen (Greenwald et al. 1998; Hofmann et al. 2008). Ausführliche Informationen zum Thema „Objektive Tests“ finden sich in Ortner et al. (2006).

Der Einsatz von Objektiven Persönlichkeitstests ist immer häufiger zu beobachten, obwohl es mitunter schwierig ist, für einige Merkmale implizite Merkmalsindikatoren zu identifizieren, die es eindeutig – also ohne Konfundierung mit anderen Merkmalen – und sensitiv (zum Begriff ► Kap. 9) erlauben, Rückschlüsse auf das interessierende Merkmal zu ziehen (vgl. Borkenau et al. 2005). Ein Grund für den häufigeren Einsatz ist die im Zunehmen begriffene Verwendung von computerbasierten Tests, auf die in ► Abschn. 3.5.2 genauer eingegangen wird.

Implizite Assoziationstests und Reaktionszeitmessung

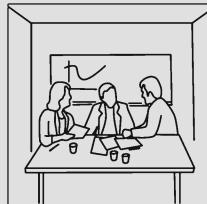
3.2.4 Projektive Verfahren

Eine weitere Gruppe von Verfahren zur Erfassung von Persönlichkeitsmerkmalen stellen projektive Verfahren dar. Sie sind dadurch gekennzeichnet, dass sie auf eine qualitative Erfassung der Gesamtpersönlichkeit unter Berücksichtigung der Einmaligkeit von Erlebnis- und Bedürfnisstrukturen ausgerichtet sind. Bei diesen Verfahren kommt in der Regel mehrdeutiges Bildmaterial zum Einsatz, und es wird angenommen, dass die Testpersonen unbewusste oder verdrängte Erfahrungen oder Erlebnisse (im Sinne des tiefenpsychologischen Abwehrmechanismus der „Projektion“) in dieses Bildmaterial hineinprojizieren. Erfasst wird die individuelle Deutung des Bildmaterials, wodurch ebenfalls Rückschlüsse auf Persönlichkeitsmerkmale möglich sind. Da von projektiven Verfahren die einschlägigen Gütekriterien (► Kap. 2) nur schwer erfüllt werden können, genügen sie den erforderlichen Qualitätsansprüchen zumeist nicht (Lilienfeld et al. 2000; s. zu den Problemen freier Antwortformate auch ► Kap. 5). Projektive Tests wie der Rosenzweig Picture Frustration Test für Kinder (PFT-K) von Duhm und Hansen (1957) oder der Rorschach-Test von Rorschach (1954) können aber als Explorationshilfen im Bereich der Diagnostik z. B. von Kindern dienen. Jedenfalls bedarf es einer besonders sorgfältigen Schulung bzw. ausführlicher Anweisungen in diesen Verfahrensweisen (z. B. Exner 2010), bevor sie zur Anwendung kommen dürfen. Als eine Alternative wurden kürzlich sog. „semiprojektive Verfahren“ entwickelt (Multi-Motiv-Gitter; Schmalt et al. 2000; ► Beispiel 3.5), die zwar ähnliches Stimulusmaterial wie die projektiven Verfahren verwenden, aber ein geschlossenes Antwortformat vorgeben, wodurch die Durchführungs- und die Auswertungsobjektivität deutlich gesteigert werden können.

Individuelle Deutung von Bildmaterial

Beispiel 3.5: (Semi-)projektive Verfahren

Hier sehen Sie eine Beispielaufgabe. Bitte beantworten Sie dazu die folgenden Aussagen mit Ja oder Nein.



	ja	nein
Angst vor schwierigen Aufgaben haben	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Hierbei Stolz empfinden, weil man etwas kann	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Das macht Spaß	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Man will sich auf ein gemeinsames Ziel verständigen	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Man will einen positiven Abschluss herbeiführen	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Beispielaufgabe in Anlehnung an das Multi-Motiv-Gitter von Schmalt et al. (2000, mit freundlicher Genehmigung von Pearson Assessment)

3.3 Geltungsbereich und Zielgruppe

Gewinnung belastbarer Informationen

Unter Geltungsbereich sind die Einsatz- und Anwendungsmöglichkeiten von Tests und Fragebogen zu verstehen. Vor allen anderen Aspekten muss darauf geachtet werden, dass der Test/Fragebogen belastbare Informationen zur Beantwortung von interessierenden Fragestellungen liefert, sei es zur Identifikation bestimmter Merkmale in der Persönlichkeitsdiagnostik, zur Prädiktion von Entwicklungsverläufen in der Schul- bzw. Laufbahnberatung oder zur Klassifikation von Verhaltensdefiziten in der klinischen Psychologie. Die Erschließung des Geltungsbereichs betrifft Aspekte der Validität (► Kap. 21) und erfolgt durch eine Beurteilung des Zusammenhangs des Tests mit externen Kriterien oder vorhandenen Tests.

Neben diesen grundsätzlichen Anforderungen an die Validität gilt es schon in der Planungsphase auch abzuklären, wie (mess-)genau das Verfahren sein soll; ein Verfahren, das für eine *Individualdiagnostik* verwendet werden soll (z. B. Test zur Diagnose von Rechenschwäche – TEDI-Math; Kaufmann et al. 2009), muss wesentlich genauer messen, als ein *Screeningverfahren*, das bei einer breiten Anwendung nur eine erste grobe Beurteilung des interessierenden Merkmals erbringen soll (z. B. das Depressions-Screening PRIME-MD; Spitzer et al. 1999).

Ein weiterer abzuklärender Aspekt des Geltungsbereichs besteht in der *Enge/Weite* des interessierenden Merkmals (► Abschn. 3.1.1). Je enger ein Merkmal definiert ist, desto genauer kann es erfasst werden, aber desto enger ist auch sein inhaltlicher Geltungsbereich. Möchte man z. B. beruflichen Erfolg beim Verkauf von Versicherungen messen, so kann dieses Merkmal sehr eng gefasst sein, indem man es z. B. anhand der Zahl der abgeschlossenen Verträge operationalisiert. In diesem Fall würde es aber nicht auch andere Aspekte des Berufserfolgs abdecken (Arbeitszufriedenheit, Kundenbindung etc.); der Geltungsbereich dieses Tests wäre also entsprechend eng.

Grundsätzlich gilt: Je breiter und je genauer ein Test ein Merkmal erfassen soll, desto mehr Items werden benötigt und desto länger wird der Test. Entsprechend genügen für den PRIME-MD, dessen Einsatzbereich sich im Sinne des Diagnostic and statistical manual of mental disorders (DSM-5; APA 2013) nur auf die „Major Depression“ bezieht, 9 Items; ein breit angelegtes Persönlichkeitsinventar erfasst hingegen mehrere Facetten und benötigt entsprechend viele Items; beispielsweise umfasst das Freiburger Persönlichkeitsinventar (FPI-R; Fahrenberg et al. 2010) 12 verschiedene Facetten und 138 Items.

Merkmalsbreite und Itemanzahl

Homogene vs. heterogene Zielgruppe

Im Rahmen der Überlegungen zum Geltungsbereich eines Tests muss vor Konstruktionsbeginn auch eine Entscheidung über die *Zielgruppe* gefällt werden, d. h. über den Personenkreis, für den mit dem Test Aussagen getroffen werden sollen.

3.4 · Testlänge und Testzeit

Eine solche Zielgruppe kann entweder sehr spezifisch und somit homogen sein, z. B. Grundschulkinder der 1. Klasse oder Anwärter zur Fluglotsenausbildung, oder sie kann sehr breit und als Konsequenz heterogen sein, z. B. Erwachsene im Alter zwischen 20 und 60 Jahren. Heterogenität bedeutet hierbei, dass sich die Personen nicht nur hinsichtlich des interessierenden Merkmals, sondern auch hinsichtlich anderer Eigenschaften (Bildungsstand, biografischer Hintergrund etc.) unterscheiden können, wodurch die Beantwortung der Testaufgaben beeinflusst sein kann.

Je breiter (und somit heterogener) eine Zielgruppe definiert wird, desto höher sind die Anforderungen, die an das Testverfahren gestellt werden müssen. Vor allem müssen die Testaufgaben/Fragen über einen breiteren Bereich von Merkmalsausprägungen gestreut und ggf. auch inhaltlich breiter gefächert sein, um möglichst viele Aspekte des Merkmals abdecken zu können. So muss beispielsweise ein Intelligenztest breiter angelegt sein, wenn er in der Allgemeinbevölkerung eingesetzt werden soll. Der Test kann hingegen wesentlich enger konstruiert werden, wenn die Zielgruppe sehr spezifisch ist (z. B. Identifizierung von Hochbegabungen bei Mittelstufenschülern).

In Abhängigkeit von der Zielgruppe muss zudem berücksichtigt werden, dass zur adäquaten Beantwortung der Items – neben dem eigentlich interessierenden Merkmal – auch andere Eigenschaften hinreichend ausgeprägt sein müssen wie z. B. die Intelligenz, der Bildungsstand oder die Lesefähigkeit. So müsste beispielsweise bei einem Test zur Erfassung numerischer Kompetenzen, der für Erst- bis Vierklässler einsetzbar sein soll, insbesondere die erst in Entwicklung begriffene und deshalb unterschiedliche Lesefähigkeit berücksichtigt werden. Würde der Test ein hohes Leseverständnis voraussetzen, wäre insbesondere für Erstklässler eine unerwünschte Konfundierung der interessierenden numerischen Leistung und der nicht hinreichend entwickelten Lesefähigkeiten zu erwarten. Es bedarf entsprechend eines wesentlich höheren Konstruktionsaufwands, um Items für heterogene Zielgruppen zu generieren als für homogene. Sollen heterogene Zielgruppen getestet werden, so ist darauf zu achten, dass möglichst viele der nicht intendierten Einflüsse konstant gehalten werden. Da das nicht immer möglich ist, erweist es sich mitunter als notwendig, insbesondere nach Alter und Geschlecht entsprechend differenzierte Testnormen zu erstellen (► Kap. 9), um die nicht konstant gehaltenen Einflüsse bei der Interpretation der Testergebnisse auszugleichen.

Schließlich muss bei der Konstruktion der Items neben der grundsätzlichen Lesefähigkeit der Testpersonen auch berücksichtigt werden, dass der Iteminhalt/-stamm bezüglich der sprachlichen Verständlichkeit (► Kap. 4, ► Abschn. 4.5.1) zielgruppenadäquat formuliert wird; beispielsweise sollten spezifische Fachbegriffe nur dann verwendet werden, wenn davon ausgegangen werden kann, dass sie in der Zielgruppe bekannt sind. Auch muss bei der Festlegung des Antwortformats (► Kap. 5) die Differenzierungsfähigkeit der Testpersonen berücksichtigt werden, z. B. bei der Wahl der Anzahl von Skalenstufen bei Ratingskalen.

3.4 Testlänge und Testzeit

Unter dem Begriff „Testlänge“ wird in Übereinstimmung mit Lienert und Raatz (1998, S. 33) die Anzahl der Items in einem Test verstanden. Im Unterschied dazu bezeichnet „Testzeit“ die Zeitdauer, die für die Bearbeitung der Testaufgaben vorgesehen ist.

Welche *Testlänge* angemessen erscheint, hängt vom Geltungsbereich und der Zielgruppe des Tests ab. Beispielsweise benötigen Persönlichkeitstestsysteme, die mehrere, breiter definierte Konstrukte erfassen, ziemlich viele Items (z. B. 12–14 Items pro Merkmalsdimension und insgesamt 138 Items im FPI-R; Fahrenberg et al. 2010). In Verfahren, mit denen hingegen zum Teil sehr spezifische und eng definierte Konstrukte erfragt werden sollen, genügen oft ökonomische

Inhaltliche Breite der Testaufgaben muss für Zielgruppe adäquat sein

Berücksichtigung nicht intendierter Einflüsse bei heterogenen Zielgruppen

Sprachliche Angemessenheit

Testlänge in Abhängigkeit von Anzahl und Definitionsbreite der Konstrukte

Testlänge beeinflusst Messgenauigkeit

Kurzskalen (z. B. mit 3–4 Items pro Merkmalsdimension zur Erfassung von Narzissmus, Machiavellismus und Psychopathie; Küfner et al. 2015).

Allgemein gilt, dass mit zunehmender Anzahl von Items zur Erfassung eines Merkmalsbereichs das Testergebnis präziser wird, denn bei einer hohen Itemanzahl geht der Mittelwert der Messfehler gegen null und der Mittelwert der Messungen entspricht dem wahren Wert der Merkmalsausprägung (► Kap. 13). Mit zunehmender Testlänge steigt auch die Reliabilität, ein Gütekriterium, das über die Messgenauigkeit des Tests informiert (► Kap. 14). Je höher also die Reliabilität sein soll, desto mehr Testaufgaben sind zur Informationsgewinnung notwendig. Entscheidet man sich beispielsweise für ein Screeningverfahren, also für ein Instrument, das bei vielen Personen (ggf. aber auch bei Einzelpersonen für eine grobe Vordiagnostik) eingesetzt werden soll, um eine nur grobe Beurteilung eines Merkmals zu erhalten, so reicht eine kürzere Testlänge aus, da es nicht so vieler Testaufgaben bedarf wie bei einem Test, der für eine differenzierte Individualdiagnose konzipiert wurde. Zu beachten ist aber, dass ab einer gewissen Itemanzahl kein bedeutender Reliabilitätszuwachs mehr zu erwarten ist (► Kap. 14). Zudem kann bei sehr langen Tests die Validität der Messung abnehmen, da die Beantwortung der Items zunehmend von testfremden Variablen (z. B. Absinken der Konzentration oder Rückgang der Motivation, ► Kap. 4, ► Abschn. 4.6) beeinträchtigt wird.

Unabdingbar ist die Qualität der Items

Die für eine bestimmte Fragestellung notwendige Testlänge hängt auch von der Qualität der einzelnen Items ab. Die Konstruktion und die Auswahl/Selektion „guter“ Items ist somit von zentraler Bedeutung, um den Test möglichst reliabel und valide, aber gleichzeitig auch hinreichend kurz und ökonomisch zu gestalten. Wie „gute“ Items zu konstruieren sind bzw. woran man gute Items zum Zweck der Itemauswahl/-selektion erkennen kann, ist Inhalt nachfolgender Kapitel (► Kap. 4 und 7).

Bedeutung der Testzeit

Auch die zur Bearbeitung der Aufgaben vorgesehene *Testzeit* muss sorgfältig überlegt werden. Hierbei sind vor allem die Teststart, die Zielgruppe und der intendierte Geltungsbereich zu berücksichtigen. Bei der Festlegung der Testzeit darf die Praktikabilität des Tests und insbesondere die Motivationslage der Testpersonen nicht aus den Augen verloren werden. Je länger der Test dauert, desto mehr ist damit zu rechnen, dass die Items nicht mehr nur merkmalsadäquat bearbeitet werden. Auf dieses Problem wird von Moosbrugger und Brandt unter „Fehlerquellen“ (► Kap. 4, ► Abschn. 4.6) im Rahmen des Optimizing-Satisficing-Modells (Krosnick 1999) gesondert eingegangen; es macht deutlich, dass bei einem von den Testteilnehmern subjektiv als zu lang empfundenen Test die Bearbeitungsqualität sinkt, sodass nicht mehr von einer adäquaten Testbearbeitung, sondern nur von einem verfälschten Testergebnis ausgegangen werden kann.

Zielgruppenadäquate Testzeit

Darüber hinaus müssen zielgruppenbedingte Einschränkungen der Testzeit beachtet werden. So sollte z. B. ein Schulleistungstest zweckmäßigerweise nicht länger als eine Unterrichtsstunde dauern; älteren Schülern, z. B. 15-Jährigen im Programme for International Student Assessment (Organisation for Economic Co-operation and Development 2014) bzw. Studienbewerbern in einem Hochschulzulassungstest können hingegen durchaus auch längere Testzeiten zugemutet werden.

Testzeit in Leistungstests

Bei Leistungstests hängt die Testzeit zudem davon ab, ob der Test als Speed- oder als Powertest konzipiert ist (► Abschn. 3.2.1), bzw. auch davon, zu welchen Anteilen Geschwindigkeits- und Niveaukomponenten im Test vertreten sind. Eine – aber großzügig bemessene – Zeitbegrenzung in einem primär als Niveautest konzipierten Verfahren dient vor allem dazu, eine überflüssig lange Bearbeitungsdauer zu unterbinden. Ein Beispiel für die Begrenzung der Testzeit bei Niveautests stellt der Intelligenztest WAIS-IV (Petermann 2012) dar.

3.5 Testadministration

Eine wichtige Entscheidung bei der Planung eines Tests oder Fragebogens ist bezüglich der Testadministration zu treffen, und zwar bezüglich der Art und Weise, wie der Test dargeboten und durchgeführt werden soll. Dazu gehört die Entscheidung über die grundsätzliche Durchführungsart des Tests/Fragebogens (als Paper-Pencil- oder computerisierte Tests), über die Form der Testung (als Einzel- bzw. als Gruppentestung) sowie über die Form der Bearbeitung/Beantwortung der Testaufgaben (als Selbst- bzw. als Fremdeinschätzung).

3.5.1 Paper-Pencil- und computeradministrierte Tests

Sogenannte „Paper-Pencil-Tests“ stellen bis heute eines der am weitesten verbreiteten Standardverfahren für empirische Untersuchungen dar. Sie sind dadurch gekennzeichnet, dass für ihre Bearbeitung nur Papier (zur Vorgabe des Tests) und ein Bleistift/Schreibstift (zur Bearbeitung/Beantwortung der Items) notwendig sind. Paper-Pencil-Tests können auch in computerisierter Form als computeradministrierte Verfahren implementiert und von den Testpersonen direkt am Computer bearbeitet werden. Ein Vorteil besteht darin, dass computeradministrierte Verfahren eine hohe Durchführungs- und Auswertungsobjektivität aufweisen, da sie von der Testleitung völlig unabhängig sind. Weiterhin sind sie insofern ökonomischer, als die Testauswertung wegen der mittels Computer direkt erfassten Antworten der Testpersonen wesentlich vereinfacht ist.

Verschiedentlich werden Tests im Handel sowohl in computeradministrierter als auch in der Paper-Pencil-Version angeboten. So existiert beispielsweise zum Aufmerksamkeits-Belastungstest d2 (Brickenkamp 2002) neben der PC- auch eine Paper-Pencil-Version.

Computeradministrierte Tests steigern die Objektivität

3.5.2 Computerbasierte Tests

Computerisierte Tests, die in ihrer Durchführung auf den Computer angewiesen sind, werden als computerbasierte Tests bezeichnet. Solche Verfahren eröffnen neue Möglichkeiten der Itemgenerierung (beispielsweise durch eingespielte Audio- oder Videodateien, ▶ Kap. 6). Vor allem erlauben sie eine Erfassung zeitkritischer Daten, die auch für Validierungsstudien oder spätere Einsatzgebiete des Tests relevant sein können. Hierzu zählen z. B. Reaktionszeiten oder Arbeitsgeschwindigkeit, aber auch physiologische Parameter wie Hautleitfähigkeit, Herzfrequenz oder Daten von bildgebenden Verfahren, die in zunehmendem Maße zur Validierung für kognitive Leistungen oder Persönlichkeitseigenschaften herangezogen werden (Ranger und Kuhn 2012; Schnipke und Scrams 2002; van der Linden 2009, 2011). Beispielsweise kann die Aktivität in bestimmten Hirnarealen, die mit spezifischen Kompetenzen wie Wortflüssigkeit oder Arbeitsgedächtniskapazität in Verbindung stehen, zur Validierung eines Tests herangezogen werden, der solche Kompetenzen erfassen soll (z. B. Allen und Fong 2008; Kane et al. 2007; Miller et al. 2009).

Computerbasierte Möglichkeiten der Itemgenerierungen und Erfassung zeitkritischer Daten

Eine spezielle Sorte von computerbasierten Tests stellen sog. „adaptive Tests“ dar, bei denen das individuelle Antwortverhalten der Testpersonen zur Steuerung/Auswahl der jeweils nachfolgenden Testaufgaben herangezogen wird. Durch die (computergesteuerte) Wahl von individuell passenden, d. h. auf das jeweilige Leistungsniveau der Testpersonen zugeschnittenen („tailored“) Testaufgaben kann eine ökonomische Testdurchführung erzielt werden (▶ Kap. 20). Der FAKT-II (Moosbrugger und Goldhammer 2007) ist ein Beispiel für ein solches Verfahren.

Adaptive Tests

Bedeutung computerisierter Verfahren nimmt zu

3

In den letzten Jahren haben die Bedeutung und der Einsatz von computerisierten Verfahren immer weiter zugenommen (z. B. als Teilerhebung auch in PISA; OECD 2014). Die mit ihnen einhergehende Möglichkeit von Online-Befragungen erlaubt zudem in vielen Fällen eine ökonomische Rekrutierung einer relativ bevölkerungsrepräsentativen Stichprobe (z. B. Bandilla 1999; Lefever et al. 2007; ► Kap. 6).

3.5.3 Einzel- und Gruppentestung

Vor- und Nachteile von Einzel- vs. Gruppentestungen

Prinzipiell können Tests entweder als Einzel- oder als Gruppentestung durchgeführt werden. Beide Verfahren haben Vor- und Nachteile. *Einzeltestungen*, in denen jeweils nur eine Testperson getestet wird, erlauben es, die Testdurchführung sehr sorgfältig zu überwachen und ggf. auch noch weitere Verhaltensdaten zu erfassen (z. B. bei der Durchführung des WAIS-IV; Petermann 2012). Einzeltestungen sind allerdings sehr aufwendig, da für jede Testung eine Testleitung zur Verfügung stehen muss. Sind Kinder die Zielgruppe des Tests oder ist der Test auf eine Individualdiagnostik ausgelegt, so sind Einzeltestungen häufig unumgänglich. *Gruppentestungen*, in denen mehrere Testpersonen zeitgleich getestet werden, sind wesentlich ökonomischer und werden vor allem auch in „Large-Scale-Assessments“ eingesetzt. Sie sind jedoch insoweit für Fehler anfälliger, als die Testpersonen nicht so gut überwacht werden können. Letzteres trifft auch insbesondere für Online-Befragungen zu.

3.5.4 Selbst- und Fremdeinschätzung

Selbst- und Fremdeinschätzungen in Leistungstests

Ein letzter wichtiger Entscheidungspunkt bezüglich der Testadministration besteht in der Frage, ob die Testpersonen den Test/Fragebogen selbst bearbeiten/beantworten und somit über sich selbst Auskunft geben, was als *Selbsteinschätzung* bezeichnet wird. Es kann aber auch eine andere Person sein, die das beobachtete Verhalten der Testperson in das jeweilige Antwortformat überträgt; dies wird als *Fremdeinschätzung* bezeichnet.

In Leistungstests ist es üblich, dass Personen den Test selbst bearbeiten/beantworten, insbesondere in Gruppentestungen. Für Individualdiagnosen oder in Situationen, in denen nicht davon ausgegangen werden kann, dass die Testpersonen in der Lage sind, ihre Antworten selbst schriftlich zu fixieren, ist eine Fremdeinschätzung angezeigt (z. B. im Wechsler Intelligenztest für Kinder, WISC-IV; Petermann und Petermann 2011).

Selbst- und Fremdeinschätzungen in Persönlichkeitstests

Auch im Bereich der Persönlichkeitstests existieren beide Formate. Selbsteinschätzungen erlauben einen besonderen Zugang zu Merkmalen, die nur durch eine Innensicht erfasst werden können, also z. B. Gefühle, Motivationen, Einstellungen. Fremdeinschätzungen sind hingegen immer dann eine mögliche Alternative, wenn davon ausgegangen werden muss, dass die Testpersonen nicht über ein ausreichendes Einsichtsvermögen oder Verständnis verfügen (z. B. psychiatrische Patienten, Testpersonen mit deutlich unterdurchschnittlicher Intelligenz). Weiterhin können Fremdeinschätzungen angebracht sein, wenn bei den Testpersonen im Fall von Selbstauskünften starke Antwortverzerrungen (► Kap. 4) zu erwarten sind (z. B. bei Begutachtungsprozessen).

3.6 Struktureller Testaufbau

Psychologische Tests haben zumeist einen typischen strukturellen Aufbau, der aus drei Teilen besteht, d. h. aus der Testanweisung (Instruktion), aus dem Testteil mit den konkreten Aufgabenstellungen (Items) und aus einem ergänzenden Teil, in dem zumeist demografische Informationen erfasst werden.

3.6.1 Testteil 1: Instruktion

Die Instruktion (Testanweisung) ist die „Eintrittskarte“ zu einem Fragebogen oder Test. Die Instruktion soll die Testpersonen zu einer adäquaten Testteilnahme/-durchführung motivieren und darüber informieren, welche Art von Aufgaben zu bearbeiten sind. Die Instruktion soll des Weiteren eine klare Handlungsanweisung darüber enthalten, in welcher Weise die Antworten auf die Items gegeben werden sollen (Antwortmodus, ▶ Kap. 5). Zweckmäßigerweise soll auch zumindest ein Item mit einer Musterantwort konkret vorgestellt werden. Das Musteritem und die Musterantwort sollen das Bearbeitungsprinzip und den Antwortmodus verdeutlichen (► Beispiel 3.6). Falls das Antwortformat im Verlauf des Tests gewechselt wird, muss speziell darauf hingewiesen werden.

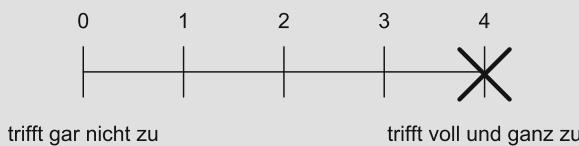
Die Instruktion muss präzise Angaben zur Art der Aufgabenstellung und zur Bearbeitungsweise der Aufgaben enthalten

Beispiel 3.6: Instruktion in einem Persönlichkeitstest

Ein Basistext für eine Instruktion in einem Persönlichkeitstest könnte beispielsweise folgenden Inhalt haben:

Ihre Aufgabe wird darin bestehen, verschiedene Aussagen dahingehend zu beurteilen, in welchem Ausmaß die Aussagen für Sie zutreffend sind.

Einmal angenommen, die zu beurteilende Aussage lautet „Entscheidungen treffe ich generell schnell“ und Sie finden, dass diese Aussage auf Sie völlig zutrifft, dann markieren Sie dies bitte auf der nachfolgenden fünfstufigen Antwortskala mit einem Kreuz an der entsprechenden Position:



Zusätzlich ist bei Persönlichkeitstests die Aufforderung wichtig, spontan und wahrheitsgetreu zu antworten, eindeutige Antworten zu geben und möglichst keine Aufgaben auszulassen. Bei Leistungstests ist eine sehr präzise Instruktion hinsichtlich der Testlänge und der Testzeit wichtig, um eine adäquate Bearbeitung sicherzustellen.

Weitere Instruktionsaspekte

Sofern die mit einem Test erhobenen Daten nicht (nur) individualdiagnostisch verwendet werden, sondern (auch) als Datenbasis für eine gruppenstatistische wissenschaftliche Untersuchung dienen, ist es unerlässlich, explizit einen Hinweis auf die Anonymität der Testpersonen zu formulieren. Dies dient einerseits der Sicherstellung des Datenschutzes und andererseits kann davon ausgegangen werden, dass die Testpersonen wahrheitsgetreuer antworten, wenn die erhobenen Daten nicht mit den einzelnen Testpersonen in Verbindung gebracht werden. Sollen sensible personenbezogene Daten einander zugeordnet werden (z. B. Schulnoten und Testergebnisse) oder soll eine Person mehrfach befragt werden (z. B. in einer Längsschnittstudie), so kann die personenbezogene Zuordnung anhand von personspezifischen Codes (Pseudonymisierung) erfolgen, die keine Rückschlüsse auf

Anonymitätszusicherung

die Testpersonen selbst zulassen, wohl aber eine Zuordnung der Informationen, die von derselben Person stammen. Weiterhin sollte bei wissenschaftlichen Untersuchungen eine verantwortliche Person oder Institution benannt werden, bei der man eine nähere Auskunft über den Sinn und Zweck der Untersuchung einholen kann. Selbstverständlich haben Testpersonen auch das Recht, nach Abschluss der Untersuchung die Löschung ihrer Daten zu verlangen.

3.6.2 Testteil 2: Konkrete Testaufgaben

Begründung für mehrere Testitems

Dieser Testteil besteht in der Regel aus mehreren konkreten Testaufgaben („Testitems“), die entsprechend der jeweiligen Zielsetzung konstruiert wurden. Einer der Gründe, warum ein Test mehrere Aufgaben/Items aufweisen muss, besteht darin, dass eine einzelne Testaufgabe zur zuverlässigen, messgenauen Erfassung eines Merkmals in der Regel nicht ausreichen würde; erst mehrere Items erlauben die Abschätzung der Messgenauigkeit (Reliabilität, ▶ Kap. 14) und eine entsprechend genaue Angabe der Merkmalsausprägung. Sollen mit einem Test mehrere Merkmale/Merkmalsfacetten erfasst werden, so müssen für jede Facette separate Items konstruiert werden (▶ Abschn. 3.1).

Für die Reihenfolge der Items gelten unterschiedliche Überlegungen:

- Bei Leistungstests (beispielsweise bei Intelligenztests) werden die Items in der Regel mit aufsteigender Schwierigkeit gereiht. Kämen die Items in zufälliger Schwierigkeitsreihung, wäre mit fehlerhaften Testergebnissen zu rechnen. Es bestünde nämlich die Gefahr, dass die Testpersonen schon am Beginn zu viel Testzeit für zu schwierige Items verbrauchen, die ihnen bei der Bearbeitung später folgender leichterer Items dann fehlt. Eine elegante Variante für den Umgang mit dieser Problematik stellt auch die „maßgeschneiderte“ Itemauswahl beim adaptiven Testen dar (▶ Kap. 20). Hierbei wird Itemhomogenität im Sinne der Item-Response-Theorie (IRT, ▶ Kap. 16) vorausgesetzt, wodurch reihenfolgeunabhängige „spezifische objektive“ Messungen möglich werden.
- Bei Persönlichkeitstests hingegen werden die Items üblicherweise in randomisierter Anordnung über den Test verteilt, und zwar auch im Fall von mehreren Facetten, um die Augenscheininvalidität und das Auftreten von Antworttendenzen zu verringern (▶ Kap. 4).

3.6.3 Testteil 3: Demografische Angaben

Itemreihenfolge bei Persönlichkeitstests

Am Anfang oder am Ende des Tests oder Fragebogens werden zumeist demografische Angaben über die jeweilige Testperson erhoben. Sie sind auf notwendige Angaben zu beschränken, wobei die Erfassung von Alter, Geschlecht, Schulbildung und Beruf üblich ist. Diese Angaben sollten vor allem dann erfasst werden, wenn sie relevant für die untersuchte Fragestellung sind, z. B. um beurteilen zu können, ob der Test wirklich das interessierende Merkmal misst oder ob Variablen vermengt („konfundiert“) sind mit anderen (demografischen) Variablen. In letzterem Fall wird es notwendig sein, für die jeweiligen demografischen Gruppen separate Testnormen (▶ Kap. 9) zu erstellen, um ggf. den Konfundierungseinfluss (z. B. von verschiedenen Altersstufen bei der normorientierten Interpretation von Intelligenzleistungen) auszugleichen. Bereits bei der Itemkonstruktion kann und sollte darauf geachtet werden, dass konfundierende Variablen möglichst keinen störenden Einfluss auf das Antwortverhalten ausüben. Im Zuge der Überprüfung der differentiellen Itemfunktionsweise („Differential Item Functioning“, DIF, ▶ Kap. 16; s. z. B. in Holland und Wainer 1993) kann getestet werden, ob das Antwortverhalten der Testpersonen z. B. von Störvariablen (z. B. demografischen Variablen), die auch als Biasvariablen bezeichnet werden, abhängt. Items mit DIF sollten nach Möglichkeit

Kontrolle von konfundierenden Variablen

3.7 · Konstruktionsphasen im Überblick

nicht in den Test aufgenommen werden. Es sollte somit deutlich sein, dass es wichtig ist, relevante Kontrollvariablen als zusätzliche Angaben zu erfassen. Allerdings ist auch darauf zu achten, dass von den Testpersonen nicht zu viele demografische Angaben abgefragt werden, um die Anonymität nicht zu gefährden (bei sehr vielen Angaben ließe sich eine Zuordnung der Daten zu den einzelnen Testpersonen rekonstruieren).

Über die formale Strukturierung des Tests hinaus soll die äußere Gestaltung der Testanweisung und des Testmaterials sprachlich und optisch ansprechend sowie an die Zielgruppe angepasst sein. Das gesamte Layout des Tests sollte die Durchführung/Bearbeitung erleichtern und potentielle Testpersonen zur adäquaten Testteilnahme motivieren. Im Vordergrund sollten daher die Einfachheit und die Übersichtlichkeit stehen. Zur Vermeidung von Fehlern sollte beispielsweise auf eine gut lesbare Schrift sowie auf optische Hilfen (z. B. alternierend unterschiedliche Schattierungen bei der Itemabfolge) geachtet werden (► Beispiel 3.7).

Layout des Tests/Fragebogens

Beispiel 3.7: Layoutentwurf für die Mehrdimensionale Perfektionismuskala

(Altstötter-Gleich und Bergemann 2006)

	trifft gar nicht zu	trifft wenig zu	trifft mittelmäßig zu	trifft überwiegend zu	trifft völlig zu
Wenn ich für mich selbst nicht die höchsten Maßstäbe setze, besteht die Gefahr, dass ich zweitklassig werde.	<input type="radio"/>				
Es ist wichtig für mich, in allem, was ich tue, äußerst kompetent zu sein.	<input type="radio"/>				
Wenn ich bei der Arbeit/im Studium versage, bin ich als Mensch ein Versager.	<input type="radio"/>				
Wenn ich nur zum Teil versage, ist das genauso schlecht, als wenn ich im Ganzen versagt hätte.	<input type="radio"/>				

(Mit freundlicher Genehmigung von Hogrefe)

Details zum Fragebogen finden sich in ► Kap. 24.

3.7 Konstruktionsphasen im Überblick

Unter Berücksichtigung der oben genannten Planungsaspekte hinsichtlich der Spezifizierung des interessierenden Merkmals, der Testart, der Eingrenzung des Gelungsbereichs und der Zielgruppe, der Festlegung von Testlänge und Testzeit sowie der Administrationsart soll nun das konkrete Vorgehen bei der Konstruktion eines Testverfahrens skizziert werden. Hierbei ist eine Reihe von aufwendigen Arbeitsschritten zu vollziehen, wie sie in ► Studienbox 3.2 aufgelistet sind. Am Ende dieses Prozesses steht dann ein fertig entwickelter Test, der den erforderlichen psychometrischen Qualitätsansprüchen (► Kap. 2) entspricht und zum Einsatz bereitsteht.

3.7.1 Itemgenerierung

Das Ziel der Itemgenerierung besteht darin, geeignete Aufgabenstellungen („Items“) zu erzeugen, die das interessierende Merkmal erfassbar machen. Die Aufgabenstellungen sollen so geartet sein, dass aus dem aufgabenbezogenen Verhalten (Antwortverhalten) der Testpersonen auf die individuellen Merkmals-

Studienbox 3.2**Konstruktionsphasen**

1. Konstruktion/Generierung einer geeigneten Menge von Testaufgaben/Items („Itempool“), einschließlich der Instruktion und der Wahl eines geeigneten Antwortformats
2. Qualitative Verständlichkeitsanalyse der Instruktion und der Items mit erforderlichen Nachbesserungen (erste Revision)
3. Erste empirische Erprobung der vorläufigen Testfassung („Pilotstudie“) an einer kleineren Stichprobe mit Itemanalyse und -selektion (zweite Revision)
4. Zweite empirische Erprobung („Evaluationsstudie“) an einer größeren, repräsentativen Stichprobe („Analysesstichprobe“) mit psychometrischen Analysen und anschließender dritter Revision (ggf. sind weitere Revisionsschleifen erforderlich)
5. Abschließende Normierung der endgültigen Testform

Aufgabenstellung, Aufgabenstamm und Antwortformat

ausprägungen zurückgeschlossen werden kann. Jede Aufgabenstellung setzt sich prinzipiell aus zwei Teilen zusammen, und zwar aus der Aufgabe selbst, die als *Aufgabenstamm* bezeichnet wird, und aus dem *Antwortformat* der Aufgabe (vgl. Rost 2004); je nach Antwortformat (z. B. frei vs. gebunden) lassen sich verschiedene Aufgabentypen („Itemtypen“) unterscheiden (► Kap. 4 und 5).

Der wichtigste und mitunter auch aufwendigste Teil der Testentwicklung besteht in der inhaltlichen Festlegung des jeweiligen Aufgabenstamms, der als Stimulus für die Reaktionen der Testpersonen dient. Hierzu müssen Überlegungen angestellt werden, in welchen charakteristischen Situationen sich Personen mit einer hohen Merkmalsausprägung von Personen mit einer niedrigen Merkmalsausprägung in ihrem Verhalten unterscheiden. Solche Situationen müssen in Form eines Aufgabenstamms abgebildet werden; das erfasste Antwortverhalten in den ausgewählten Situationen dient dann als Indikator für die Ausprägung des interessierenden Merkmals. Diese Umsetzung (Operationalisierung) von Merkmalseigenschaften in Testaufgaben ist eine sehr anspruchsvolle Tätigkeit, bei der die inhaltliche Validität, d. h. die inhaltliche Übereinstimmung des interessierenden Merkmals mit seinen Operationalisierungen gewährleistet sein muss (► Kap. 21). Hierfür ist eine explizite Merkmalsspezifikation notwendig, die anschließend in ihrer Enge bzw. Breite von den Aufgaben abgebildet werden muss.

Sofern es sich um einen Leistungstest handelt, enthält der Aufgabenstamm jeweils eine Frage oder eine Problemstellung, die von der Testperson zu bearbeiten ist. Soll der Test eine Aussage über die allgemeine Intelligenz liefern, so müssen die enthaltenen Aufgaben – je nach Theorie – eine Reihe von Facetten der Intelligenz abbilden (z. B. Alltagswissen, angewandtes Rechnen, Kodieren und Assoziieren im Adaptiven Intelligenz Diagnostikum 3, AID 3; Kubinger und Holocher-Ertl 2014).

Handelt es sich um einen Persönlichkeitstest, so enthält der Aufgabenstamm jeweils eine Frage oder eine Aussage (Statement), wobei die Testperson meist graduell beurteilen soll, inwieweit die Aussage auf sie zutrifft. Die Formulierung des Aufgabenstamms und die Wahl des Aufgabentyps bzw. des Antwortformats sind für die Testkonstruktion sehr relevant; diese beiden Themen werden daher in den nachfolgenden Kapiteln (► Kap. 4 und 5) ausführlich besprochen.

Bei der Wahl der Aufgabenstellungen können verschiedene Konstruktionsstrategien herangezogen werden, so z. B. die intuitive, die rationale, die kriteriumsorientierte und die faktorenanalytische Strategie, auf die im Folgekapitel ausführlich eingegangen wird (► Kap. 4). Die Entscheidung zugunsten einer der Strategien erfolgt in Abhängigkeit von dem interessierenden Merkmal, dem Geltungsbereich

Operationalisierung von Merkmalseigenschaften in Testaufgaben**Formulierung des Aufgabenstamms und Auswahl des Antwortformats****Konstruktionsstrategien**

und der Zielgruppe. In der Praxis folgt die Itemgenerierung nur selten einer einzelnen Strategie, meist wird eine gemischte, mehrstufige Vorgehensweise gewählt.

Am Ende der Itemgenerierung, die auch die Wahl von geeigneten Antwortformaten und einer geeigneten Instruktion umfasst, steht ein „Itempool“ zur Verfügung, der üblicherweise deutlich mehr Items beinhalten soll, als für die finale Testversion benötigt werden. Um aus dem Itempool eine vorläufige und später eine endgültige Testversion erstellen zu können, werden die Items zunächst einer qualitativen Verständlichkeitsprüfung und sodann einer empirisch-psychometrischen Erprobung unterzogen.

Itempool

3.7.2 Qualitative Verständlichkeitsprüfung der Items

In diesem Schritt wird zunächst die Eignung und Klarheit der Instruktion überprüft. Sodann sollen jene Items ausgesondert/nachgebessert werden, die nicht den Anforderungen entsprechen, z. B. weil der Aufgabenstamm zu Verständnisschwierigkeiten führt oder weil der Aufgabenstamm und das Antwortformat nicht zusammenpassen. Auch technische Probleme können bereits in dieser ersten Erprobungsphase aufgedeckt werden; beispielsweise kann die Art des Schreibgeräts (Filzstift etc.) beim Einscannen von Antwortbögen Probleme bereiten. Von Bedeutung ist, dass die Ersterprobung des Tests unter möglichst realistischen Bedingungen mit Personen aus der Zielgruppe stattfindet. Wurde die Itemkonstruktion sehr sorgfältig durchgeführt, so genügen für diese Ersterprobung auch kleine Stichproben.

Identifikation von inhaltlichen und praktischen Schwierigkeiten

Die einfachste und zeitlich effektivste Erprobungsmethode zur Verständlichkeitsüberprüfung besteht in der *retrospektiven Befragung* der Testpersonen. Nachdem die vorläufige Testversion bearbeitet wurde, werden die Testpersonen befragt, bei welchen Items die Bearbeitung mit Problemen verbunden war. Bei dieser Art der Testerprobung bleiben naturgemäß mehrere Fehlerquellen verborgen: So können sich die Testpersonen nicht immer an alle problematischen Aufgaben erinnern; auch sind sie oft nicht in der Lage, Gedankenabläufe bei problembehafteten Aufgaben bewusst zu erkennen und adäquat zu formulieren.

Retrospektive Befragung

Als weitere Erprobungsform hat sich im Anschluss an den Test die Durchführung von Interviews in Form eines sog. „*Debriefings*“ bewährt. In einer solchen Sitzung werden die Probleme erörtert, die von den Testleitenden beobachtet wurden, beispielsweise bei welchen Aufgaben es deutlichen Klärungsbedarf gab oder welche Items am häufigsten nicht bearbeitet wurden. Problematisch beim Debriefing ist, dass hierbei keine standardisierten Instrumente zur Beurteilung der Testqualität zum Einsatz kommen. Damit wird die Definition dessen, was an einer Situation bzw. einer Aufgabe als „Problem“ anzusehen sei, der Testleitung überlassen, und es hängt maßgeblich von ihr ab, was in der Sitzung besprochen wird und was nicht.

Debriefing

Als alternatives bzw. ergänzendes Verfahren zum Debriefing wurde die Technik der *Verhaltenskodierung* („behavior coding“) entwickelt (Canell et al. 1981). Die Testsituation wird von einer dritten Person beobachtet oder aufgezeichnet und anschließend dahlingshend analysiert, ob und wann sich die Testleitenden oder die Testpersonen nicht instruktionsgemäß verhalten haben, z. B. bei welchen Fragen Testleitende oder Testpersonen Verständnisprobleme geäußert haben. Items, die sich in dieser Hinsicht als auffällig erweisen, werden dann nachgebessert bzw. vom Test ausgeschlossen. Die Verhaltenskodierung ist eine recht zuverlässige Methode, wenn es darum geht, Schwierigkeiten aufseiten der Testpersonen und/oder der Testleitung aufzudecken.

Verhaltenskodierung

Sind die Testentwickler vor allem an Gedanken, die den Testpersonen während der Testbearbeitung durch den Kopf gehen, interessiert, so hat sich das *kognitive Vortesten* („cognitive pretesting“), verbunden mit der Technik des *lauten Denkens* („think aloud“) als eine weitere Methode der Testerprobung bewährt. Diese

Kognitives Vortesten, lautes Denken

Technik wurde ursprünglich zur möglichst lückenlosen Offenlegung gedanklicher Prozesse bei Problemlöseaufgaben verwendet. Die/der Testleitende liest Items vor und bittet die Testpersonen, alle Überlegungen, die zur Beantwortung der Frage führen, zu formulieren. Diese Äußerungen werden zumeist auf Video aufgezeichnet. Die Methode liefert Einsichten in die Art und Weise, wie jedes Item verstanden wird, sowie in die Strategien, welche zur Bearbeitung angewendet werden. Verständnis- und Interpretationsschwierigkeiten sowie Probleme bei der Anwendung bestimmter Itemformate können so leicht aufgedeckt werden. Die Technik des lauten Denkens ist allerdings recht aufwendig in Bezug auf die Durchführung und die Auswertung.

Jede der aufgeführten Techniken zur Verständlichkeitsüberprüfung hat ihre Vorteile und Nachteile; welche Techniken bei der Erprobung der vorläufigen Testversion zur Anwendung kommen, ist von dem als vertretbar angesehenen Aufwand, der Teststart, der Aufgabenkomplexität und dem Aufgabentyp abhängig.

Erste Revision des Itempools

! Die qualitative Überprüfung der Items soll in einer ersten Revision des Itempools münden, bei der alle Items mit Verständnisschwierigkeiten ausgesondert oder nachgebessert werden. Wird diese erste Revision nicht sorgfältig durchgeführt, so resultieren Mängel in der Testkonstruktion, die sich zu einem späteren Zeitpunkt auch nicht mit ausgefeilten statistischen Analysetechniken beheben lassen.

3.7.3 Empirische Erprobung der vorläufigen Testversion

Pilotstudie und zweite Revision des Itempools

Unverzichtbare Grundlage für alle Fragebogen und Tests ist, dass für die Durchführung, Auswertung und Interpretation eindeutige Anweisungen formuliert sind, damit die Objektivität des Verfahrens gewährleistet ist (Näheres hierzu findet sich in ► Kap. 4, 5 und 9). Sobald diese Anweisungen klar verständlich vorliegen und der vorläufige Itempool festgelegt ist, kann eine erste empirische Erprobung der vorläufigen Testversion („Pilotstudie“) an einer kleineren Stichprobe durchgeführt werden. Diese hat zum Ziel, quantifizierte Informationen über die Qualität der Items zu gewinnen, um diejenigen Items zu identifizieren, die den Konstruktionsansprüchen besonders gut genügen. Diese Informationen bestehen insbesondere aus deskriptiven Statistiken zu den Items wie Schwierigkeitsindizes, Itemvarianzen und Itemtrennschärfen („Itemanalyse“, ► Kap. 7). Stellen sich an dieser Stelle Probleme heraus, die z. B. darin bestehen können, dass Items keine geeignete Itemvarianz aufweisen, weil die Itemschwierigkeiten extrem hoch oder extrem niedrig sind, oder dass die Itemtrennschärfen nicht den Anforderungen entsprechen, sollten diese Items aus dem Itempool entfernt werden, weil sie keine Informationen über die Verschiedenheit der Merkmalsausprägungen der Testpersonen liefern können. Aus dieser empirisch begründeten „Itemselektion“ resultiert eine zweite, revidierte vorläufige Testversion.

Evaluationsstudie zur psychometrischen Überprüfung des Tests

In der darauffolgenden *Evaluationsstudie* müssen anhand einer größeren, möglichst repräsentativen Analysestichprobe die ersten Ergebnisse der Itemanalyse sichergestellt und umfangreichere weitere Analysen vorgenommen werden. Für einen psychometrischen Test im engeren Sinn beinhalten diese Konstruktionsansprüche eine Reihe von Voraussetzungen, die erfüllt sein müssen, damit aus den gegebenen Antworten der jeweiligen Testperson ein zusammenfassender Testwert für das interessierende Merkmal bzw. die untersuchte Dimension gebildet werden kann (► Studienbox 3.3).¹ Dies bedeutet nicht nur, dass die Items nach den in ► Kap. 4 und 5 beschriebenen Prinzipien und Strategien konstruiert werden sollen. Vielmehr sind noch weitere Überlegungen erforderlich, und zwar zur Stichproben-

¹ Bei einem Fragebogen zur Erhebung einzelner Fakten, die untereinander in keiner dimensionalen Beziehung stehen, kann es durchaus ausreichend sein, auf einige Entwicklungsschritte zu verzichten.

Studienbox 3.3**Aspekte zur Evaluation eines psychometrischen Tests**

- Die Items sollen unterschiedlich schwierig sein; nur, wenn sie sich auf viele unterschiedliche Abstufungen des Merkmals beziehen und ein interindividuell variierendes Antwortverhalten („Itemvarianz“, s. ▶ Kap. 7, ▶ Abschn. 7.4) erzeugen, sind sie geeignet, verschiedene Ausprägungsgrade des Merkmals zu messen. Auskunft darüber liefern die Schwierigkeitsindizes (▶ Kap. 7, ▶ Abschn. 7.3).
- Die Items sollen „trennscharf“ sein, d.h., sie sollten Personen mit höheren Merkmalsausprägungen von Personen mit niedrigeren Merkmalsausprägungen möglichst eindeutig unterscheidbar machen. Auskunft darüber liefern die Trennschärfeindizes (▶ Kap. 7, ▶ Abschn. 7.6).
- Die Itemanzahl soll so gewählt werden, dass die innerhalb einer Dimension zusammengefassten Items eine zuverlässige, messgenaue Erfassung der Ausprägungen des interessierenden Merkmals ermöglichen. Auskünfte über die Messgenauigkeit liefern die Reliabilitätskoeffizienten (▶ Kap. 14).
- Es soll sichergestellt sein, dass die Items – um sie zu einer Dimension zusammenfassen zu können – einem testtheoretischen (psychometrischen) Modell genügen. Zu diesem Zweck können Modelltests auf Basis der IRT (▶ Kap. 16) oder faktorenanalytische Dimensionalitätsuntersuchungen (▶ Kap. 23 und 24) herangezogen werden.
- Der Test soll valide sein, d.h., die erfassten Merkmale sollen eine inhaltliche Bedeutsamkeit haben und zu anderen Tests und Kriterien in einem sinnvollen Zusammenhang stehen (▶ Kap. 21 und 25).

wahl und zum testtheoretischen Modell, das für die psychometrische Beurteilung angewendet werden soll, sowie zur Testbatterie, die zu Validierungszwecken des neuen Verfahrens eingesetzt werden soll.

Die *Analysestichprobe* für die psychometrische Erprobung des Tests sollte hinsichtlich ihrer Zusammensetzung eine repräsentative Stichprobe aus der Zielgruppe darstellen. Hierbei ist darauf zu achten, dass die Spannweite der Merkmalsausprägungen in der gewählten Stichprobe nicht ergebnisverzerrend eingeschränkt ist („shrinkage of range“). Will man beispielsweise einen Studierfähigkeitstest erproben, so sollte die Analysestichprobe nicht nur aus bereits zugelassenen Studierenden bestehen, da in der Stichprobe dann nicht die volle Merkmalsbreite von Studienbewerbern vorhanden wäre, sondern nur noch jene der ausgewählten Studierenden, die den Studierfähigkeitstest bereits erfolgreich bestanden haben.

Um die Entsprechung zwischen der Messung mit dem in Entwicklung befindlichen Test/Fragebogen und dem interessierenden Merkmal genauer untersuchen zu können, muss auf psychometrische Modelle zurückgegriffen werden, die in Teil II und Teil III dieses Bandes in mehreren Kapiteln ausführlich erörtert werden. Zum einen ist es die Klassische Testtheorie (KTT, ▶ Kap. 13), auf deren Basis Fragen zur Messgenauigkeit (Reliabilität) beantwortet werden können (▶ Kap. 14 und 15); zum anderen sind es die IRT (▶ Kap. 16 und 18) sowie faktorenanalytische Verfahren (▶ Kap. 23 und 24), mit denen Fragen der Itemhomogenität bzw. der Dimensionalität der Items beurteilbar sind. Möglichst schon bei der Testplanung sollten Überlegungen einfließen, wie der Test letztendlich psychometrisch überprüft werden soll, da dies u.a. die Wahl des Antwortformats beeinflussen kann. Empfehlungen zur Wahl testtheoretischer Modelle finden sich in ▶ Kap. 12.

Zum Zweck der Validierung eines neuen Testverfahrens, d.h. zur Beurteilung, welche belastbaren diagnostischen oder prognostischen Aussagen auf Basis der Testwerte getroffen werden können, ist es mitunter notwendig, von denselben Test-

Analysestichprobe soll keine Merkmalseinschränkungen enthalten

Wahl des psychometrischen Modells zur Dimensionalitätsüberprüfung

Validitätsprüfung: konvergente und diskriminante Validität

personen auch eine „Testbatterie“ weiterer Testverfahren bearbeiten zu lassen. Typischerweise werden zum einen solche Tests eingesetzt, die ähnliche, verwandte Merkmale messen, um die konvergente Validität zu beurteilen. Zum anderen benötigt man aber auch Tests für andere Merkmale, und zwar für solche, von denen das eigentlich interessierende Merkmal abgegrenzt werden soll, damit dem Test eine diskriminante Validität attestiert werden kann. *Konvergente Validität* liegt vor, wenn das neue Verfahren hoch mit ähnlichen, verwandten Merkmalen korreliert; *diskriminante Validität* liegt vor, wenn das neue Verfahren nicht oder nur gering mit abzugrenzenden Merkmalen korreliert (Genaueres s. MTMM-Analyse; ► Kap. 25). Die Auswahl dieser zumeist schon etablierten Testverfahren sollte vorab – am besten schon bei der Literaturrecherche im Rahmen der Eingrenzung des Merkmals (► Abschn. 3.1) – erfolgen.

3.7.4 Revision und Abschluss der Testentwicklung

Eventuelle weitere Revisionen

Die psychometrische Beurteilung des Tests soll zeigen, dass die an das Verfahren zu richtenden Qualitätsansprüche erfüllt sind. Solange die Ergebnisse aus der Itemanalyse und den weiteren psychometrischen Überprüfungen nicht zufriedenstellend ausfallen, muss die vorläufige Testfassung erneut revidiert werden. Mitunter kann es notwendig sein, wieder einen Entwicklungsschritt zurückzugehen und z. B. diejenigen Items, die sich als ursächlich für die Mängel herausstellen, nachzubessern oder auszusondern oder auch durch neu generierte Items zu ersetzen.

Abschluss der Testentwicklung

Sofern am Ende dieser empirischen Erprobung keine weiteren Revisionen des Tests und seiner Items erforderlich sind, ist die Entwicklungsarbeit bis auf die Normierung (► Abschn. 3.7.5) abgeschlossen. Zu den vielfältigen Anwendungsbereichen eines neuen Tests ergeben sich aber auch späterhin noch neue Erkenntnisse hinsichtlich der Validität des Verfahrens und zur Belastbarkeit der auf den Testergebnissen basierenden Schlussfolgerungen.

3.7.5 Normierung der endgültigen Testform

Normentabellen

Um eine objektive, metrische Beurteilung der Testergebnisse von einzelnen Personen in Relation zu den Merkmalsausprägungen der Zielpopulation vornehmen zu können, muss anhand einer großen, repräsentativen Stichprobe („Eichstichprobe“) die für die Zielpopulation geltende Verteilung der Testwerte (► Kap. 8) ermittelt werden, auf deren Basis Normentabellen erstellt werden können. Normentabellen (Testnormen) geben darüber Auskunft, welchen Prozentrang (PR) das Testergebnis einer einzelnen Testperson in der Verteilung aller Testwerte in der Eichstichprobe einnimmt. So bedeutet z. B. ein PR = 75, dass 75 % der Merkmalsträger in der Eichstichprobe eine Testleistung aufweisen, die niedriger/schwächer/kleiner oder gleich ausgeprägt ist als die der Testperson; lediglich von 25 % wird die Merkmalsausprägung übertroffen.

Eichstichprobe

Bezüglich der Zusammensetzung der Eichstichprobe müssen sehr genaue Überlegungen angestellt werden, damit die Zielpopulation repräsentativ abgebildet wird. Wenn sich bei der Normierung bedeutsame Unterschiede für einzelne Subgruppen zeigen (z. B. Alters- oder Geschlechtseffekte), die im Sinne der Test-fairness (zum Begriff ► Kap. 2) berücksichtigt werden sollen, ist es angezeigt, separate Normen für die jeweiligen Subgruppen zu erstellen. Zeigen sich hingegen merkmalsrelevante Unterschiede zwischen den zu differenzierenden Gruppen (z. B. zwischen klinisch auffälligen depressiven Patienten und einer unauffälligen Vergleichsgruppe bei einem Depressionstest), sollte die Normierung nicht in separaten Subgruppen, sondern über alle Personen hinweg erfolgen; der beobach-

3.8 · Zusammenfassung

tete Effekt ist dann nämlich ein Nachweis der erwünschten Kriteriumsvalidität des Tests. Nähere Ausführungen zur Normierung finden sich in ▶ Kap. 9. Nach DIN 33430 (▶ Kap. 10) sollte alle acht bis zehn Jahre eine Aktualisierung der Normentabellen vorgenommen werden.

Abschließend sei betont, dass die einzelnen, in diesem Abschnitt skizzierten Erprobungsschritte jeweils an separaten Stichproben durchgeführt werden sollten. Ein wiederholtes Heranziehen derselben Stichprobe für die empirische Beurteilung der Items, des Tests und für die Erstellung der Normentabellen würde systematische Verzerrungen und fehlerhafte Schlussfolgerungen nach sich ziehen, die dringend vermieden werden müssen.

Verwendung separater Stichproben bei den einzelnen Schritten der Testkonstruktion

3.8 Zusammenfassung

In diesem Kapitel wurde der aufwendige Prozess einer Testentwicklung skizziert, bei dem zwischen einer Planungs- und einer Konstruktionsphase unterschieden wird. Die Planungsphase beginnt mit der Eingrenzung und Definition des Merkmals, das erfasst werden soll. Ein Überblick über verschiedene Testarten gibt einen Einblick in die Breite möglicher Tests. Wesentliche Entscheidungen über die geplante Testkonstruktion betreffen sodann den Geltungsbereich und die Zielgruppe, für die der Test entwickelt werden soll, die Testlänge und -zeit, auf die der Test ausgelegt werden soll, sowie den strukturell typischen Testaufbau.

In der Konstruktionsphase werden die wesentlichen Aspekte, die bei der Planung Berücksichtigung gefunden haben, konkret umgesetzt: Sie beginnt mit der Formulierung der Instruktion, der Testaufgaben/Items und der Wahl des Antwortformats. Eine erste qualitative Beurteilung erlaubt einen Einblick in die Verständlichkeit der Items; eine „Pilotstudie“ ermöglicht die Berechnung erster statistischer Kennwerte; die darauffolgende empirische Erprobung („Evaluationsstudie“) liefert Aussagen zur Passung mit dem zugrunde gelegten psychometrischen Modell und mit externen Validitätskriterien, die zusammen die Voraussetzungen für die Anwendbarkeit des Tests und für die Belastbarkeit der Testergebnisse bilden. Die abschließend zu erstellenden Normentabellen erlauben eine standardisierte Prozentrangausgabe über die Merkmalsausprägung einer Testperson im Vergleich zu den Ausprägungen in der Eichstichprobe/Zielpopulation. In den folgenden Kapiteln dieses Bandes wird im Detail auf die hier skizzierten Konzepte eingegangen.

3.9 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ▶ <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was ist der Unterschied zwischen Testlänge und Testzeit? Welche Aspekte sollten Sie bei der Entscheidung darüber, wie lang ein Test sein sollte, berücksichtigen?
2. Welche Testarten unterscheidet man prinzipiell?
3. Nehmen Sie an, Sie wollen einen Studierfähigkeitstest für das Psychologiestudium entwerfen. Welche Aspekte bei der Rekrutierung der Analysestichprobe sollten Sie beachten? Welche externen Kriterien zur Validierung Ihres Tests könnten Sie verwenden?
4. Worauf sollten Sie bei der Testnormierung achten, wenn Sie einen kognitiven Test zur Erkennung von Demenzpatienten entwickeln möchten?
5. Versuchen Sie, eine Definition für das Merkmal „Extraversion“ zu formulieren. Charakterisieren Sie Personen mit hohen bzw. niedrigen Ausprägungen in diesem Merkmal. Würden die beiden Items „Ich würde gern einmal Fallschirmspringen“ und „Ich bin begeisterter Wildwasserbahnfahrer“ geeignet und

ausreichend sein, um eine adäquate Operationalisierung des Konstrukts zu ermöglichen?

Literatur

- Allen, M. D. & Fong, A. K. (2008). Clinical application of standardized cognitive assessment using fMRI. II. Verbal fluency. *Behavioural Neurology*, 20, 141–152.
- Altstötter-Gleich, C. & Bergemann, N. (2006). Testgüte einer deutschsprachigen Version der Mehrdimensionalen Perfektionismus Skala von Frost, Marten, Lahart und Rosenblatt (MPS-F). *Diagnostica*, 52, 105–118.
- American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- Bandilla, W. (1999). WWW-Umfragen – eine alternative Datenerhebungstechnik für die empirische Sozialforschung? In B. Batinic, A. Werner, L. Gräf & W. Bandilla (Hrsg.), *Online Research. Methoden, Anwendungen und Ergebnisse*. Göttingen: Hogrefe.
- Beauducel, A. & Leue, A. (2014). *Psychologische Diagnostik*. Göttingen: Hogrefe.
- Ben-Porath, Y. S. & Tellegen, A. (2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form): Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Bonnardel, R. (1946). Le test du double labyrinth B. 19-D.L. *Le Travail Humain*, 9, 212–218.
- Bonnardel, R. (2001). *B19 Doppelabyrinthtest*. Mödling: Schuhfried.
- Borkenau, P., Egloff, B., Eid, M., Hennig, J., Kersting, M., Neubauer, A. C. & Spinath, F. M. (2005). Persönlichkeitspsychologie: Stand und Perspektiven. *Psychologische Rundschau*, 56, 271–290.
- Borkenau, P. & Ostendorf, F. (2008). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae (NEO-FFI)*. Manual (2. Aufl.). Göttingen: Hogrefe.
- Brähler, E., Holling, H., Leutner, D. & Petermann, F. (Hrsg.) (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (3. Aufl.). Göttingen: Hogrefe.
- Brickenkamp, R. (2002). *Test d2 – Aufmerksamkeits-Belastungs-Test*. Göttingen: Hogrefe.
- Briken, P., Rettenberger, M., Dekker, A. (2013). Was sagen „objektive“ Messverfahren über Sexualstraftäter? *Forensische Psychiatrie, Psychologie, Kriminologie*, 7, 28–33.
- Canell, C. F., Miller, P. V. & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 389–437). San Francisco, CA: Jossey-Bass.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115–130.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Human- und Sozialwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer.
- Doignon, J.-P. & Falmagne, J.-C. (1999). *Knowledge spaces*. New York, NY: Springer.
- Duhm, E. & Hansen, J. (1957). *Rosenzweig Picture Frustration Test für Kinder. PFT-K*. Göttingen: Hogrefe.
- Eysenck, S. B. G. (1993). The I7: development of a measure of impulsivity and its relationship to the superfactors of personality. In W. G. McCown, J. L. Johnson & M. B. Shure (Eds.), *The impulsive client: theory, research and treatment* (pp. 134–152). Washington, DC: American Psychological Association.
- Exner, J. E. (2010). *Rorschach-Arbeitsbuch für das Comprehensive System: Deutschsprachige Fassung von A Rorschach Workbook for the Comprehensive System – Fifth Edition*. Göttingen: Hogrefe.
- Fahrenberg, J., Hampel, R. & Selg, H. (2010). *Freiburger Persönlichkeitseinventar (FPI-R)* (8. Aufl.). Göttingen: Hogrefe.
- Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.
- Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Hathaway, S. R., McKinley, J. C. & Engel, R. (Hrsg.) (2000). *MMPI-2. Minnesota Multiphasic Personality Inventory 2*. Göttingen: Hogrefe.
- Heller, J. & Repitsch, C. (2008). Distributed skill functions and the meshing of knowledge structures. *Journal of Mathematical Psychology*, 52, 147–157.
- Hofmann, W., Gschwendner, T., Castelli, L. & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes and Intergroup Relations*, 11, 69–87.
- Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Kane, M. J., Conway, A. R. A., Miura, T. K. & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 615–622.
- Kaufmann, L., Nuerk, H.-C., Graf, M., Krinzinger, H., Delazer, M. & Willmes, K. (2009). *Test zur Erfassung numerisch-rechnerischer Fertigkeiten vom Kindergarten bis zur 3. Klasse (TEDI-MATH)*. Bern: Huber.

Literatur

- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537–567.
- Kubinger, K. D. (2006). Ein Update der Definition von Objektiven Persönlichkeitstests: Experimental-psychologische Verhaltensdiagnostik. In T. M. Ortner, R. Proyer & K. D. Kubinger (Hrsg.), *Theorie und Praxis Objektiver Persönlichkeitstests* (S. 38–52). Bern: Huber.
- Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.
- Kubinger, K. D. & Holocher-Ertl, S. (2014). *AID 3: Adaptives Intelligenz Diagnostikum 3*. Göttingen: Beltz-Test.
- Küfner, A. C. P., Dufner, M. & Back, M. (2015). Das Dreckige Dutzend und die Niederträchtigen Neun-Zwei Kurzskalen zur Erfassung von Narzissmus, Machiavellismus, und Psychopathie. *Diagnostica*, 61, 76–91.
- Lefever, S., Dal, M. & Matthíassdóttir, Á. (2007). Online data collection in academic research: advantages and limitations. *British Journal of Educational Technology*, 38, 574–582.
- Lienert, G. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz PVU.
- 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.
- McCrae, R. R. & Costa, P. T. (2010). *NEO inventories for the NEO Personality Inventory-3 (NEO-PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3), NEO Personality Inventory-Revised (NEO PI-R): professional manual*. Lutz, FL: PAR.
- McCrae, R. R., Costa, P. T. & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, 84, 261–270.
- Miller, K. M., Price, C. C., Okun, M. S., Montijo, H. & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, 24, 711–717.
- Moosbrugger, H. & Goldhammer, F. (2007). *Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT II): Grundlegend neu bearbeitete und neu normierte 2. Auflage des FAKT von Moosbrugger und Heyden (1997)*. Göttingen: Hogrefe.
- Moosbrugger, H. & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2)*. Bern, Göttingen: Huber.
- Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Ergebnisse: Was Schülerinnen und Schüler wissen und können (Band I, überarbeitete Ausgabe): Schülerleistungen in Lesekompetenz, Mathematik und Naturwissenschaften*. Bielefeld: W. Bertelsmann.
- Ortner, T. M., Proyer, R. & Kubinger, K. D. (Hrsg.) (2006). *Theorie und Praxis Objektiver Persönlichkeitstests*. Bern: Huber.
- Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R)*. Göttingen: Hogrefe.
- Petermann, F. (Hrsg.) (2012). *WAIS-IV: Wechsler Adult Intelligence Scale – Fourth Edition. Deutschsprachige Adaption nach David Wechsler*. Frankfurt am Main: Pearson Assessment.
- Petermann, F. & Eid, M. (Hrsg.) (2006). *Handbuch der Psychologischen Diagnostik*. Göttingen: Hogrefe.
- Petermann, F. & Petermann, U. (Hrsg.) (2011). *WISC-IV. Wechsler Intelligence Scale for Children – Fourth Edition*. Frankfurt am Main: Pearson Assessment.
- Pospeschill, M. & Spinath, F. M. (2009). *Psychologische Diagnostik*. München: Reinhardt.
- Ranger, J. & Kuhn, J.-T. (2012). A flexible latent trait model for response time in tests. *Psychometrika*, 77, 31–47.
- Rohrmann, S., Hodapp, V., Schnell, K., Tibubos, A. N., Schwenkmezger, P. & Spielberger, C. D. (2013). *Das State-Trait-Ärgerausdrucks-Inventar-2 (STAXI-2)*. Bern: Huber.
- Rorschach, H. (1954). *Psychodiagnosistik*. Bern: Huber.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.
- Schmalt, H. D., Sokolowski, K. & Langens, T. A. (2000). *Das Multi-Motiv-Gitter für Anschluss, Leistung und Macht MMG*. Frankfurt: Pearson.
- Schmidt, A. F. (2013). Indirekte Messverfahrenen pädophiler sexueller Interessen – Ein Überblick über empirische Ergebnisse und methodische Implikationen. In P. Briken, J. L. Möller, M. Rösler, M. Rettenberger, V. Klein & D. Yoon (Hrsg.). *EFPPO Jahrbuch 2013 – Empirische Forschung in der forensischen Psychiatrie, Psychologie und Psychotherapie* (S. 65–75). Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.
- 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. & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl., unter Mitarbeit von T. Frydrich & H. Moosbrugger). Berlin, Heidelberg: Springer.
- Schnipke, D. & Scrams, D. (2002). Exploring issues of examinee behavior: insights gaines from response-time analyses. In C. Mills, M. Potenza, J. Fremer & W. Ward (Eds.), *Computer-based testing: building the foundation for future assessments* (pp. 237–266). Mahwah: Lawrence Erlbaum.
- Schuhfried, G. (2007). *Zweihand Koordination – 2HAND*. Mödling: Schuhfried.

- Schwenkmezger, P., Hodapp, V. & Spielberger, C. D. (1992). *State-Trait-Ärgerausdrucks-Inventar (STAXI)*. Bern: Huber.
- Spielberger, C. D. (1999). *The State-Trait Anger Expression Inventory-2 (STAXI-2): Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Spitzer R., Kroenke, K., Williams, J. (1999). Validation and utility of a self-report Version of PRIME-MD: the PHQ Primary Care Study. *Journal of the American Medical Association*, 282, 1737–1744.
- Stemmler, G., Hagemann, D., Amelang, M. & Bartussek, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (7. Aufl.). Stuttgart: Kohlhammer.
- Stemmler, G. & Margraf-Stiksrud, J. (2015). *Lehrbuch Psychologische Diagnostik*. Göttingen: Hogrefe.
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2018). *SON-R 2-8. Non-verbaler Intelligenztest. Testmanual mit deutscher Normierung und Validierung*. Göttingen: Hogrefe.
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46, 247–272.
- van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, 53, 334–358.
- von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, 67, 49–71.



Itemkonstruktion und Antwortverhalten

Helfried Moosbrugger und Holger Brandt

Inhaltsverzeichnis

- 4.1 Ziele und Aspekte der Itemkonstruktion – 69**
- 4.2 Itemstamm und Zielgruppe – 69**
- 4.3 Vorgehensweisen bei der Itemgenerierung – 71**
 - 4.3.1 Intuitive vs. rationale Strategie – 71
 - 4.3.2 Kriteriumsorientierte Strategie – 71
 - 4.3.3 Faktorenanalytische Strategie – 72
- 4.4 Kategorisierung von Frageformen – 73**
- 4.5 Gesichtspunkte der Itemformulierung – 75**
 - 4.5.1 Sprachliche Verständlichkeit – 75
 - 4.5.2 Eindeutigkeit des Iteminhalts – 76
 - 4.5.3 Vermeidung bestimmter Itemarten – 77
- 4.6 Kognitive und motivationale Prozesse bei der Itembearbeitung – 78**
 - 4.6.1 Kognitive Prozesse bei der Itembearbeitung – 79
 - 4.6.2 Motivationale Prozesse bei der Itembeantwortung – 79
- 4.7 Response-Bias als Fehlerquelle beim Antwortverhalten – 81**
 - 4.7.1 Antworttendenz vs. Antwortstil – 81
 - 4.7.2 Soziale Erwünschtheit – 82
 - 4.7.3 Akquieszenz – 83
 - 4.7.4 Tendenz zur Mitte – 84
 - 4.7.5 Effekte der Itemreihenfolge – 85
- 4.8 Computerunterstützte Itemkonstruktion – 86**

4.9 Zusammenfassung – 86

4.10 Kontrollfragen – 87

Literatur – 87

4.1 · Ziele und Aspekte der Itemkonstruktion

i Hat man die Entscheidungen bezüglich des Merkmals, der Testart, des Geltungsbereichs, der Zielgruppe etc. getroffen (► Kap. 3), besteht der nächste Schritt der Testkonstruktion in der Generierung und Formulierung der einzelnen Aufgabenstellungen/Testitems. Der wichtigste und zugleich anspruchsvollste Aspekt bei der Itemgenerierung besteht darin, inhalts valide Operationalisierungen des interessierenden Merkmals zu finden, diese in einem entsprechenden Aufgabenstamm/Itemstamm zu formulieren und mit einem zweckmäßigen Antwortformat (► Kap. 5) abzufragen. In diesem Kapitel soll auf wichtige Aspekte eingegangen werden, die bei der Generierung und Formulierung der Items beachtet werden müssen. Neben sprachlichen Aspekten soll hier auch auf Modelle zur Beschreibung des Prozesses des Antwortverhaltens und Fehlerquellen im Antwortverhalten eingegangen werden.

4.1 Ziele und Aspekte der Itemkonstruktion

Das Ziel eines Tests oder Fragebogens besteht in der Regel darin, die Unterschiedlichkeit von Personen (Merkmisträgern) hinsichtlich ihrer Ausprägungen in einem interessierenden Merkmal zu erfassen. Um interindividuelle Unterschiede in interessierenden Merkmalen erfassen zu können, ist es notwendig, inhalts valide, d.h. für das Merkmal repräsentative Aufgabenstellungen zu generieren, aus deren Beantwortung die individuellen Merkmalsausprägungen erschlossen werden können. Es ist hierbei wesentlich, dass die Items den gesamten Umfang des interessierenden Merkmals repräsentieren, damit das Merkmal in angemessener Weise abgebildet werden kann. Falls die Repräsentativität an dieser Stelle der Testkonstruktion zu wenig umfänglich ausfällt, entsteht ein Mangel, der in keinem der weiteren Schritte der Testkonstruktion ausgeglichen werden kann.

Die Aufgabenstellungen setzen sich prinzipiell aus zwei Teilen zusammen: aus der Aufgabe selbst, die auch als Aufgabenstamm oder Itemstamm, zumeist aber verkürzt als das „Item“ bezeichnet wird, und dem Antwortformat der Aufgabe, d.h. einer Festlegung, auf welche Art und Weise die Testpersonen ihre Antworten geben sollen.

Was unter einem Item zu verstehen ist, ließe sich mit Rost (2004, S. 18) wie folgt näher beschreiben:

- » Als **Item** (das Wort wird üblicherweise englisch gesprochen und dekliniert) bezeichnet man die Bestandteile eines Tests, die eine Reaktion oder Antwort hervorrufen sollen, also die Fragen, Aufgaben, Bilder etc. Wenn auch die Items von Test zu Test sehr unterschiedlich aussehen können, sind sie innerhalb eines Tests sehr ähnlich (homogen), da sie dasselbe Merkmal der Person ansprechen.

Während sich dieses Kapitel vordergründig mit dem Itemstamm und der Itemgenerierung/-konstruktion und der Vermeidung typischer Fehlerquellen im Antwortverhalten befasst, wird das Thema der Antwortformate Gegenstand des nachfolgenden Kapitels sein (► Kap. 5).

Aufgabenstellungen müssen das Merkmal repräsentativ abbilden

Itemstamm

Was ist ein Item?

Antwortformat

4.2 Itemstamm und Zielgruppe

Damit ein Test seiner Aufgabe gerecht werden kann, Differenzierungen von Personen aus einer definierten Zielpopulationen hinsichtlich eines interessierenden Merkmals zu leisten, müssen geeignete Testaufgaben/-items konstruiert werden, die im Itemstamm an die Merkmalsbreite und an die Merkmalsausprägungen in der Zielgruppe angepasst sind. Der Itemstamm, der als Stimulus für das Antwortverhalten dient, muss so gewählt werden, dass Testpersonen mit unterschiedlichen Merkmalsausprägungen bei der Bearbeitung der Items auch Unterschiede in Bezug auf das Lösungs- bzw. Zustimmungsverhalten aufweisen.

Itemstamm und Testlet bei Leistungstests

Sofern es sich um einen Leistungstest handelt, enthält der Itemstamm eine Problemstellung, die von den Testpersonen gelöst werden soll. Wenn an einen gemeinsamen Itemstamm oder einen Teil davon eine Gruppe von mehreren Detailfragen angefügt ist, spricht man von einem Testlet. Beispielsweise kann den Testpersonen ein Textabschnitt vorgelegt werden, zu dem anschließend mehrere Fragen gestellt werden, die von den Testpersonen beantwortet werden sollen.

Itemstamm bei Persönlichkeitstests

Handelt es sich hingegen um einen Persönlichkeitstest, so enthält der Itemstamm in der Regel eine einzelne Aussage (Statement), die von den Testpersonen hinsichtlich des Zutreffens auf die eigene Person beurteilt werden soll. Hierbei muss festgelegt sein, ob eine zustimmende bzw. ablehnende Antwort im Sinne einer hohen bzw. einer niedrigen Ausprägung des interessierenden Merkmals zu bewerten ist.

Bei der Itemkonstruktion muss stets darauf geachtet werden, dass die einzelnen Items bei Personen, die sich hinsichtlich des interessierenden Merkmals unterscheiden, auch tatsächlich verschiedene Antworten hervorrufen. Extrem leicht oder extrem schwer lösbar Aufgaben bzw. Statements, die extrem leicht oder extrem schwer bejaht werden können, sollten folglich vermieden werden, da sie von den allermeisten Testpersonen gelöst/bejaht bzw. nicht gelöst/verneint würden und fast keine Unterschiede bei der Beantwortung erzeugen. Zur Feststellung interindividueller Unterschiede wären sie folglich kaum geeignet (► Beispiel 4.1).

Antwortverhalten soll interindividuelle Unterschiede aufzeigen

Beispiel 4.1: Extrem leicht oder extrem schwer lösbar Aufgaben bzw. Statements

Sehr leichte Items sind Items, die von (fast) niemandem verneint bzw. bejaht werden, weshalb (fast) keine Variation in den Itemantworten zu erwarten wäre:

- „Ich halte Umweltverschmutzung für schädlich.“
- „Fallschirmspringen ist mein Lieblingshobby.“
- „Ich vertraue niemandem.“

Itemschwierigkeit passend zur Zielgruppe wählen

Von der Regel der Vermeidung extremer Schwierigkeit/Leichtigkeit der Items muss abgewichen werden, wenn insbesondere sehr starke oder sehr schwache Merkmalsausprägungen, die nur selten auftreten, mit sehr selten auftretenden Verhaltensweisen erfasst werden sollen. So würden z. B. in einer Zielgruppe von Gesunden dem Item „Die Bewältigung des Alltags macht mir Angst“ nur sehr wenige Testpersonen zustimmen; in einem klinischen Test zur Differenzierung einer Zielgruppe von depressiven Patienten wäre das Item aber nützlich, da schwer depressive Personen dem Item voraussichtlich vermehrt zustimmen würden, wodurch eine Differenzierung zwischen unterschiedlichen Depressivitätsgraden möglich wird.

Auch bei der Konstruktion von Leistungstests für eine breite Zielgruppe muss bei der Itemgenerierung auf die *Korrespondenz von Aufgabenschwierigkeit und Merkmalsausprägung* geachtet werden. Diese Korrespondenz wird bei Niveautests und bei Speedtests (► Kap. 3) in unterschiedlicher Weise realisiert: Bei Niveautests muss sichergestellt sein, dass auf allen Schwierigkeitsstufen genügend Items vorhanden sind, um nicht nur im mittleren, sondern auch im unteren und oberen Merkmalsbereich differenzieren zu können. Bei Speedtests hingegen erfolgt die Leistungsdifferenzierung in der Regel über die für die Aufgabenlösung benötigte Zeit bzw. über die Anzahl der Items, die bei Begrenzung der Bearbeitungszeit richtig bearbeitet werden. Dafür müssen viele einfache Items mit geringer Itemschwierigkeit generiert werden, die zumeist von allen Testpersonen gelöst werden können.

Nach dieser allgemeinen Einführung sollen nun zuerst Vorgehensweisen zur Itemgenerierung vorgestellt werden (► Abschn. 4.3); im Anschluss werden verschiedene Kategorien von Frageformen vorgestellt (► Abschn. 4.4).

Überblicksartig sollen hier allgemeine Vorgehensweisen/Strategien vorgestellt werden, anhand derer die Itemkonstruktion/-generierung erfolgen kann. Die Entscheidung zugunsten einer der Strategien erfolgt in Abhängigkeit vom interessierenden Merkmal, dem Geltungsbereich und der Zielgruppe. In der Praxis folgt die Itemkonstruktion nur selten einer einzelnen Strategie, meist wird eine gemischte, mehrstufige Vorgehensweise gewählt.

- !** Bei allen Konstruktionsstrategien muss darauf geachtet werden, dass das interessierende Merkmal durch die Testitems in seiner vollen Breite abgebildet wird.

Items müssen das Merkmal in voller Breite abbilden

4.3.1 Intuitive vs. rationale Strategie

Die *intuitive Konstruktionsstrategie* wird verwendet, wenn der theoretische Kenntnisstand bezüglich des interessierenden Merkmals gering ist. Anstelle einer theoriegeleiteten Formulierung der Items ist die Konstruktion von der Intuition und Erfahrung des Testkonstrukteurs getragen. Diese Strategie wird vor allem zu Beginn neuer Forschungszweige/Forschungsthemen angewendet, in denen neue Merkmale erschlossen werden sollen.

Dank intensiver Forschung liegen zu sehr vielen psychologischen Merkmalen aber bereits mehr oder weniger ausgereifte Theorien vor, die als Basis für eine rationale Testkonstruktion dienen können. Im Unterschied zur intuitiven bedient sich die *rationale Konstruktionsstrategie* der Methode der Deduktion. Voraussetzung ist das Vorhandensein einer elaborierten Theorie über die Differenziertheit von Personen hinsichtlich des interessierenden Merkmals. Innerhalb des Merkmals orientiert sich die differenzierte Abstufung an der Häufigkeit/Intensität von beobachtbaren Merkmalsindikatoren, in denen sich die unterschiedlichen Merkmalsausprägungen manifestieren.

Zu den klassischen Verfahren, die rational entwickelt wurden, zählt im Leistungsbereich z. B. der Intelligenz-Struktur-Test (I-S-T 2000R; Liepmann et al. 2007), der insbesondere auf dem Intelligenzmodell von Thurstone (1931, 1938) beruht, in der aktuellen Auflage aber auch Bezüge zum Ansatz von Horn und Cattell (1966) und dem Berliner Intelligenzmodell von Jäger (1984) herstellt. Im Persönlichkeitbereich wäre z. B. das Persönlichkeitssinventar NEO-PI-R (McCrae und Costa 2010) zu nennen, das auf den Big-Five-Persönlichkeitsmerkmalen basiert. Hierbei handelt es sich um jene fünf Persönlichkeitsdimensionen, die sich international zur Differenzierung von Individuen als besonders leistungsfähig erwiesen haben.

Intuitive Strategie baut auf Erfahrung auf

Rationale Strategie orientiert sich an inhaltlich-theoretischen Vorgaben

Beispiele rational konstruierter Tests

4.3.2 Kriteriumsorientierte Strategie

Bei der *kriteriumsorientierten* bzw. *externalen Konstruktionsstrategie* werden Items danach ausgewählt, ob sie zwischen Gruppen mit unterschiedlichen Ausprägungen/Abstufungen eines externen Merkmals (eines „Kriteriums“) eindeutig differenzieren können. Theoriegeleitete Aufgabeninhalte im Sinne der rationalen Konstruktionsstrategie sind hier nicht von vordergründigem Interesse. Entscheidend ist vielmehr der Nützlichkeitseffekt, der dann vorhanden ist, wenn die Items das gewählte Kriterium geeignet vorhersagen können. Um Items zu finden, die gute Differenzierungseigenschaften bezüglich des Kriteriums aufweisen, wird zunächst ein großer Itempool zusammengestellt und an Personengruppen erprobt, die sich hinsichtlich des Kriteriums möglichst stark unterscheiden. Aus dem Itempool

Kriteriumsorientierte Strategie optimiert Unterscheidung von relevanten Gruppen

Beispiele für kriteriumsorientiert konstruierte Tests

werden dann diejenigen Items ausgewählt, die diese Differenzierung bestmöglich leisten; einer rationalen Erklärung darüber, worauf diese Differenzierungsleistung basiert, bedarf es nicht.

Ein klassisches Beispiel für die kriteriumsorientierte Konstruktionsstrategie stellt das bis heute weitverbreitete Testsystem Minnesota Multiphasic Personality Inventory (MMPI) dar (MMPI-2, Hathaway et al. 2000; MMPI-2-RF, Ben-Porath und Tellegen 2011). Die Subtests des MMPI-2 erlauben eine breite Differenzierung zwischen verschiedenen klinisch-psychiatrischen Gruppen. Bei der Entwicklung des Verfahrens wurden zunächst 1000 Items von den zu differenzierenden psychiatrisch auffälligen Gruppen (Schizophrene, Hypochondre usw.) und einer Kontrollgruppe von Unauffälligen bearbeitet. Übrig blieben 566 Items, die signifikant zwischen den Kriteriumsgruppen der psychiatrisch unauffälligen und den psychiatrisch auffälligen Gruppen differenzierten konnten. Für ein jüngeres Anwendungsbeispiel aus dem Bereich der Studierendenauswahl mit dem Kriterium Studienerfolg s. Moosbrugger et al. (2006). Zur Absicherung gegen rein situative Effekte, z. B. gegen Auswahlverzerrungen durch spezifische Personengruppen, sollten die Ergebnisse der kriteriumsorientierten Itemauswahl möglichst auch an anderen Stichproben auf ihre Belastbarkeit überprüft werden (► Kap. 21).

4.3.3 Faktorenanalytische Strategie

Faktorenanalytische Strategie fokussiert Dimensionalität

Die *faktorenanalytische* bzw. *interne Konstruktionsstrategie* lässt sich von dimensionsanalytischen Überlegungen leiten. Das Ziel besteht darin, Teilgruppen von Items (Subtests) zu finden, die im Sinne einer faktorenanalytischen „Einfachstruktur“ unidimensional sind und mit den anderen Teilgruppen/Subtests jeweils nicht oder nur geringfügig korrelieren. Bei diesem Konstruktionsprinzip wird zu einem breiten Merkmalsbereich mit möglicherweise mehreren Facetten eine Anzahl von Items konstruiert und einer Stichprobe von Testpersonen vorgelegt. Anhand der exploratorischen Faktorenanalyse, einem statistischen Verfahren zur Dimensionalitätsprüfung (► Kap. 23), werden diejenigen Items identifiziert, die untereinander hohe Zusammenhänge aufweisen und sich zu jeweils eindimensionalen Subtests („Faktoren“) zusammenfassen lassen. Die faktorenanalytisch gefundenen Dimensionen bedürfen einer sorgfältigen Interpretation im Hinblick darauf, durch welche Verhaltensweisen/Items sich die jeweils gefundenen Dimensionen konstituieren.

Ein klassisches Beispiel für die faktorenanalytische Konstruktionsstrategie im Leistungsbereich stellt der von Thurstone und Thurstone (1941) entwickelte mehrdimensionale Intelligenztest dar. Das Ziel bei der Entwicklung dieses Tests bestand darin, isolierbare Facetten der Intelligenz (induktiv) zu identifizieren. Hierzu setzten die Autoren eine explorative Faktorenanalyse ein und extrahierten insgesamt sechs Faktoren („Primary Mental Abilities“), deren inhaltliche Bedeutung posthoc bestimmt wurde (z. B. verbales Verständnis und Wortflüssigkeit). Ein Beispiel für die faktorenanalytische Konstruktionsstrategie im Persönlichkeitsbereich stellt das Freiburger Persönlichkeitsinventar (FPI) dar. Seinen Ursprung hatte das Verfahren in einer 1968 entstandenen Vorform (ALNEV; Fahrenberg und Selg 1968), mit der die Bereiche Emotionalität, Extraversion/Introversion, Aggressivität und psychovegetative Labilität gemessen wurden. In der jetzigen revidierten Form des Freiburger Persönlichkeitsinventars (FPI-R; Fahrenberg et al. 2010) existieren zwölf Subtests, mit denen u. a. auch die Dimensionen Lebenszufriedenheit und Soziale Orientierung erfasst werden können.

In aktuellen Testentwicklungen hat nicht nur die exploratorische, sondern auch die konfirmatorische Faktorenanalyse (CFA, ► Kap. 24) als Konstruktionsstrategie eine starke Bedeutung, weil sie eine Verknüpfung der faktorenanalytischen mit der rationalen Konstruktionsstrategie ermöglicht. Mit dem konfirmatorischen Verfah-

Beispiele für faktorenanalytisch konstruierte Tests

Testevaluation durch konfirmatorische Faktorenanalyse

4.4 · Kategorisierung von Frageformen

ren lässt sich nämlich überprüfen, ob die zuvor (rational) postulierten Dimensionen (Faktoren/Facetten) einer empirischen Überprüfung standhalten. So wurde z. B. der in ► Abschn. 4.3.1 erwähnte, rational konstruierte I-S-T 2000R (Liepmann et al. 2007) auch konfirmatorisch hinsichtlich seiner Struktur überprüft. Sowohl eine Unidimensionalität der Items innerhalb der einzelnen Facetten als auch eine Bestätigung der postulierten Faktorenstruktur bilden einen zentralen Bestandteil einer erfolgreichen Testevaluation.

4.4 Kategorisierung von Frageformen

Eine wesentliche Rolle bei der Itemgenerierung spielt die Frageform, d. h. die Art und Weise, wie die Fragen/Aufgaben gestellt werden, um Informationen zu dem interessierenden Merkmal zu gewinnen. Hierzu seien einige Kategorisierungssaspekte aufgeführt.

Ein erster Aspekt betrifft das *direkte* oder *indirekte* Erfragen des interessierenden Merkmals (► Beispiel 4.2). Während direkte Fragen das interessierende Merkmal direkt ansprechen (z. B. Ängstlichkeit), wählen indirekte Fragen für den Rückschluss auf das Merkmal spezifische Indikatoren, indem sie z. B. nach Verhalten in bestimmten Situationen fragen. Bei der direkten Befragung kann nicht immer von einer interindividuellen Übereinstimmung bezüglich der Bedeutung der Frage ausgegangen werden, weil z. B. ein bewusster Zugang zu dem Merkmal schwierig ist (Beispiel: „Sind Sie ängstlich?“). Im ungünstigen Fall wird das direkt angebrochene Merkmal von Testpersonen sehr unterschiedlich aufgefasst („Ängstlich ist eine Person, die nicht den Mut hat, Fallschirm zu springen“ vs. „Ängstlich ist jemand, der sich nicht traut, vor Publikum zu sprechen“). Deshalb erleichtern gut gewählte indirekte Verhaltensindikatoren einen sichereren Rückbezug bei der Interpretation des interessierenden Merkmals.

Beispiel 4.2: Direkte vs. indirekte Frageform

- Direkte Frage: „*Sind Sie ängstlich?*“
- Indirekte Frage: „*Fühlen Sie sich unsicher, wenn Sie nachts allein auf der Straße sind?*“

Ein anderer Aspekt betrifft die Art, wie der Sachverhalt erfragt wird. Eine Möglichkeit besteht darin, dass im Itemstamm ein *hypothetischer Sachverhalt* geschildert wird (► Beispiel 4.3). Das Gegenteil dazu stellt ein *biografiebezogenes* Item dar, mit dem das individuelle Verhalten in bestimmten Situationen erfragt wird, wobei man davon ausgeht, dass die meisten Testpersonen solche Situationen bereits erlebt haben. Das Erfragen hypothetischer Sachverhalte birgt die Gefahr von Fehleinschätzungen. Das Erfragen biografiebezogenen Verhaltens gilt als zuverlässiger; allerdings enthält es außer dem untersuchten Einfluss des interessierenden Merkmals der Person immer auch eine Situationskomponente. Darüber hinaus sind die entsprechenden Fragen hinsichtlich interindividuell passender Situationen beschränkt und nicht immer sinnvoll (z. B. die Frage „Welche Erfahrungen mit Forschungsprojekten können Sie vorweisen?“ in einem Fragebogen für Studienbewerber, die ihre Forschungslaufbahn erst beginnen).

Direkte vs. indirekte Frageform

Hypothetische vs. biografiebezogene Frageform

Konkrete vs. abstrakte Frageform

Des Weiteren lassen sich Items in solche mit einem *konkreten* und solche mit einem *abstrakten* Inhalt einteilen (► Beispiel 4.4). Konkrete Fragen informieren meist besser über die situationalen Bedingungen (Arbeitstätigkeit, Verhältnis zu den Kollegen); abstrakte Items hingegen lassen Interpretationsfreiräume zu und beinhalten die Gefahr von Fehleinschätzungen.

Beispiel 4.4: Konkrete vs. abstrakte Frageform

- Konkrete Frage: „Wie verhalten Sie sich, wenn Sie einen Streit zwischen Kollegen schlichten müssen?“
- Abstrakte Frage: „Wie belastend schätzen Sie die Arbeit in einem konfliktreichen Arbeitsumfeld ein?“

Personalisierte vs. depersonalisierte Frageform

Auch muss man entscheiden, ob man die Fragen in *personalisierter* oder *depersonalisierter* Form stellt. Eine personalisierte Frage (► Beispiel 4.5) liefert, vorausgesetzt, sie wird ehrlich beantwortet, sehr zuverlässige Informationen. Von einigen mag sie aber als eine Verletzung der Privatsphäre empfunden werden, insbesondere wenn es sich um „heikle“ Fragen handelt (► Abschn. 4.5.3). Wenn hingegen auf depersonalisierte Fragen ausgewichen wird, kann es passieren, dass nur allgemeine, nichtssagende Antworten gegeben werden.

Beispiel 4.5: Personalisierte vs. depersonalisierte Frageform

- Personalisierte Frage: „Ich bemühe mich, rücksichtsvoll im Straßenverkehr zu sein.“
- Depersonalisierte Frage: „Autofahrer sollten rücksichtsvoll im Straßenverkehr sein.“

Emotionsneutrale vs. emotionalisierende Frageform

Darüber hinaus können Items in Bezug auf ihre *Stimulusqualität* unterschiedlich formuliert werden (► Beispiel 4.6). Gemeint ist damit die emotionale Intensität, mit der Reaktionen bei Testpersonen hervorgerufen werden sollen (z. B. emotionsneutral vs. emotionalisierend).

Beispiel 4.6: Emotionsneutrale vs. emotionalisierende Frageform

- Emotionsneutrale Frage: „Halten Sie sich für einen ängstlichen Menschen?“
- Emotionalisierende Frage: „Wenn mir nachts jemand auf der Straße folgt, geht mein Puls rapide hoch.“

Kategorisierung nach Aufgabeninhalten

Eine weitere Kategorisierung von Fragen in Persönlichkeitstests lässt sich in Anlehnung an Angleitner et al. (1986) nach den abgefragten *Aufgabeninhalten* aufstellen. Diese Kategorisierung ist bei der Itemgenerierung besonders wesentlich,

4.5 · Gesichtspunkte der Itemformulierung

da das Vermischen von Items aus unterschiedlichen der im Folgenden aufgeführten Kategorien innerhalb eines Tests zu methodischen Artefakten („Methodeneffekte“, ▶ Kap. 25) führen kann:

- Fragen zur *Selbstbeschreibung*, d.h., wie ich selbst mein Verhalten wahrnehme, z.B. „Ich lache oft“ oder „Vor einem mündlichen Vortrag bekomme ich schwitzige Hände“.
- Fragen zur *Fremdbeschreibung*, d.h., wie ich meine, dass andere mein Verhalten wahrnehmen, z.B. „Meine Freunde halten mich für eine tüchtige Person“.
- Fragen zu *biografischen Fakten*, d.h. zu Sachverhalten, die sich auf den individuellen Lebenslauf beziehen, z.B. „Ich habe mehrmals Abenteuerurlaube gemacht“.
- Fragen zu *Trait-/Eigenschaftszuschreibungen*, d.h. zu Persönlichkeitsmerkmalen, die man sich selbst zuordnet, z.B. „Ich bin ein sehr spontaner Typ“.
- Fragen nach *Motivationen*, d.h. nach Begründungen für das individuelle Handeln, z.B. „Ich habe eine besondere Vorliebe für Aufgaben, die schwer zu knacken sind“.
- Fragen zu *Wünschen und Interessen*, d.h. zu Zielvorstellungen und Neigungen, z.B. „Ich schaue gerne wissenschaftliche Sendungen an“.
- Fragen zu *Einstellungen und Meinungen*, d.h. zu Werthaltungen und Überzeugungen, z.B. „Es gibt im Leben Wichtigeres als beruflichen Erfolg“.

Vermischung von Itemkategorien kann zu methodischen Artefakten führen

4.5 Gesichtspunkte der Itemformulierung

Wenn die Entscheidungen bezüglich der zu wählenden Frageformen der Items getroffen sind, stellt die sorgfältige Ausformulierung der Items einen besonders wichtigen weiteren Schritt in der Test- und Fragebogenentwicklung dar. Damit der Test eine objektive, reliable und valide Messung ermöglicht, sollen einige Grundregeln hinsichtlich der *sprachlichen Verständlichkeit*, der *Eindeutigkeit des Iteminhalts* sowie der *Vermeidung bestimmter Itemarten* beachtet werden.

4.5.1 Sprachliche Verständlichkeit

Die Items sollten für die Testpersonen ohne große Mühe bereits nach einmaligem Durchlesen verständlich sein. Ist dies nicht gewährleistet, besteht die Gefahr von Fehlinterpretationen und Motivationseinbußen seitens der Testpersonen, woraus Verzerrungen der Antworten resultieren können.

! Die Klarheit des sprachlichen Ausdrucks hat bei der Itemformulierung oberste Priorität.

Sprachliche Klarheit hat oberste Priorität

Folgende Verständlichkeitsaspekte sind bei der Itemformulierung zu beachten:

Items sollen *positiv formuliert* sein, d.h. Verneinungen (Negationen) sollen vermieden werden. Die Beantwortung von positiven Formulierungen mit „Ja“ oder „Nein“ führt in der Regel zu keinen Problemen (▶ Beispiel 4.7). Die Bewertung von verneinenden Statements kann hingegen missverständliche Antworten erzeugen, da eine Zustimmung sowohl mit „Ja“ („Ja, ich finde keinen Gefallen ...“) als auch mit „Nein“ („Nein, ich finde keinen Gefallen ...“) ausgedrückt werden könnte. Insbesondere sollen doppelte Verneinungen vermieden werden, da sich hierbei die geschilderte Problematik noch verstärkt (s. auch ▶ Abschn. 4.7 zum Thema „Response-Bias“).

Items positiv formulieren, Negationen vermeiden

Beispiel 4.7: Positive vs. negative Formulierung

- Positive Formulierung: „*Die Abholzung des Regenwaldes beobachte ich mit Sorge.*“
- Negative Formulierung: „*Finden Sie keinen Gefallen an der Abholzung des Regenwaldes?*“
- Doppelte Verneinung: „*Ich finde nicht, dass ich an der Abholzung des Regenwaldes keinen Gefallen finde.*“

Klare Satzkonstruktionen

Komplizierte und umständliche *Satzkonstruktionen* sollten vermieden werden, insbesondere wenn die Zielgruppe aus Kindern besteht oder aus Testpersonen, die kognitiv als nicht sehr leistungsfähig eingeschätzt werden können (► Beispiel 4.8). Einfache Sätze ohne Verschachtelungen sind deutlicher und verständlicher.

Beispiel 4.8: Einfache vs. komplizierte Satzkonstruktion

- Komplizierte Satzkonstruktion: „*Es kommt vor, dass ich mir darüber Sorgen mache, in Konfliktsituationen nicht zu wissen, was ich sagen könnte.*“
- Einfache Satzkonstruktion: „*In Konfliktsituationen fällt es mir schwer, etwas zu sagen.*“

Fachbegriffe und Abkürzungen vermeiden

Begriffe und Formulierungen, insbesondere *Fachbegriffe und Abkürzungen*, die nur einem kleinen Teil der in Aussicht genommenen Zielgruppe geläufig sind, sollten ebenfalls vermieden werden (► Beispiel 4.9).

Beispiel 4.9: Verwendung von Fachbegriffen und Abkürzungen

- Viele Fachbegriffe und Abkürzungen: „*Das DSM sollte gegenüber dem ICD präferiert werden, weil dessen Differenzierungspotential der Repräsentation der Probandendiversität adäquater ist.*“
- Einfachere Formulierung: „*Bei klinischen Beurteilungssystemen sollte jenes bevorzugt werden, das feinere Unterscheidungen liefert.*“

4.5.2 Eindeutigkeit des Iteminhalts**Eindeutigkeit des Iteminhalts als Voraussetzung für interindividuelle Vergleichbarkeit von Itemantworten**

Eng verknüpft mit der sprachlichen Verständlichkeit ist die Eindeutigkeit des Iteminhalts. Eindeutigkeit liegt vor, wenn alle Testpersonen den Iteminhalt in gleicher Weise auffassen, sodass ihre Antworten einen eindeutigen Rückschluss auf die individuelle Ausprägung des interessierenden Merkmals erlauben. Die Eindeutigkeit ist erforderlich, um eine intersubjektiv gemeinsame Vergleichsbasis zu schaffen. Denn nur dann, wenn alle Testpersonen dasselbe Verständnis des Iteminhalts haben, nehmen sie unter denselben Bedingungen an der Testung teil, und nur dann kann von einer Vergleichbarkeit der Messungen ausgegangen werden. Folgende Regeln sollten hierzu beachtet werden:

Universalausdrücke wie „immer“, „nie“ oder „alle“ sollten nicht unkritisch verwendet werden, da schon ein einzelnes Gegenbeispiel genügen würde, um das Item anders beantworten zu müssen. Beispielsweise wäre es günstiger, anstatt „Mein Kind kann sich nie auf eine Aufgabe konzentrieren“ die Formulierung „Mein Kind kann sich nur schwer auf eine Aufgabe konzentrieren“ zu wählen.

Keine Universalausdrücke

4.5 · Gesichtspunkte der Itemformulierung

Falls es notwendig ist, *Definitionen* zu geben, sollten diese gegeben werden, bevor die eigentliche Frage gestellt wird, z. B. „Soziale Intelligenz ist die Fähigkeit, die Gefühle anderer zu erkennen und zu interpretieren. Glauben Sie, dass Sie eine hohe soziale Intelligenz besitzen?“

Mehrdeutigkeit sollte vermieden werden: Es darf keine Möglichkeit geben, den Iteminhalt in unterschiedlicher Weise zu interpretieren. Beispielsweise könnte das Item „Meine Stimmung verändert sich schnell“ nicht zuverlässig zur Erfassung von emotionaler Labilität verwendet werden. Man kann sich durchaus vorstellen, dass auch Personen, die einen geringen Labilitätswert besitzen, ihre Stimmung schnell verändern können, wenn die Situation das verlangt, beispielsweise wenn man von dem Tod eines Anverwandten erfährt. Deshalb wäre eine situative Einengung vorzuziehen, z. B. „Meine Stimmung verändert sich schnell, auch wenn es dafür keinen äußeren Anlass gibt.“

Ein Item sollte *nicht mehrere Aussagen* enthalten, sondern sich nur auf eine einzelne Aussage konzentrieren. Bei zwei Aussagen in einem Item ist nämlich nicht klar, ob die Testperson auf die eine, auf die andere oder auf beide Aussagen geantwortet hat, wodurch der Rückschluss auf das interessierende Merkmal unsicher wird. Beispielsweise würde das Item „Bei vielen Aufgaben in Englischtests bin ich mir schon im Voraus sicher, dass ich sie nicht lösen kann, weil ich sprachlich nicht begabt bin“ besser in zwei Items aufgeteilt: „Ich weiß schon im Voraus sicher, dass ich viele Aufgaben in Englischtests nicht lösen kann“ und „Ich bin sprachlich nicht begabt“, da es verschiedene Gründe dafür geben könnte, dass man sich sicher ist, die Aufgaben nicht lösen zu können (z. B. weil man sich nicht genügend geeignet auf die Tests vorbereitet oder Prüfungsangst hat).

Der Aufgabeninhalt muss dem *Vorwissen* und dem *Sprachniveau der Zielgruppe* angepasst sein; anderenfalls informieren die Antworten nicht über das interessierende Merkmal. Beispielsweise ist ein Item „Ich setze mich für gesetzliche Maßnahmen zur Regulierung der Industrieemissionen ein, selbst wenn dies die Produktionskosten erhöht“ für Kinder bzw. Jugendliche eher inadäquat. Besser wäre die Formulierung „Ich unterstütze umweltfreundliche Produkte, selbst wenn ich mehr dafür bezahlen muss“.

Statements zur *Häufigkeit* oder *Intensität* sind ohne nähere Spezifizierung eher verwirrend, da ihre Beantwortung uneindeutig bleibt und keinen sicheren Rückschluss auf das interessierende Merkmal erlaubt. So kann z. B. die Antwort „Trifft nicht zu“ auf die Frage „Ich rauche häufig“, bedeuten, dass die Testperson entweder nur selten oder dass sie nie raucht, sodass der Rückschluss auf das Merkmal „Rauchgewohnheiten“ nicht eindeutig wäre. Um eine Differenzierung von Häufigkeiten oder Intensitäten zu ermöglichen, wäre es sinnvoller, die Häufigkeit nicht als Statement, sondern in Form einer direkt gestellten Frage zu formulieren, z. B. „Wie häufig rauchen Sie?“ oder „Wie viele Zigaretten rauchen Sie pro Tag?“; die Intensität kann dann in den Antwortalternativen (► Kap. 5) abgebildet werden.

Auch der *Zeitpunkt* oder die *Zeitspanne*, auf die Bezug genommen wird, sollte eindeutig definiert sein; z. B. wäre das Item „In letzter Zeit war ich häufig in der Oper“ nicht so genau wie „Wie oft waren Sie innerhalb des letzten halben Jahres in der Oper?“.

4.5.3 Vermeidung bestimmter Itemarten

Ferner ist bei der Konstruktion eines Tests darauf zu achten, dass man bestimmte Itemarten vermeidet, weil sie zu Beantwortungsproblemen führen können.

Zunächst sind hier *Items mit kurzlebigem Inhalt* zu nennen, die es zu vermeiden gilt, weil sie zu schnell veralten. Beispielsweise wäre die Wissensfrage zur politischen Bildung „Wie heißt der Wirtschaftsminister von Deutschland?“ mit der Richtigantwort „Peter Altmaier“ (Stand: Dezember 2019) ggf. nur temporärer

Schwieriges definieren

Mehrdeutigkeit vermeiden

Keine Verknüpfung mehrerer Aussagen

Vorwissen und Sprachniveau an Zielgruppe anpassen

Häufigkeitsangaben geeignet spezifizieren

Zeitspannen eindeutig machen

Kurzlebige Inhalte vermeiden

Implizite Wertungen vermeiden

Art, weil sich die Richtigantwort in Abhängigkeit von der politischen Entwicklung rasch ändern kann.

Problematisch sind auch Items, die *Wertungen* enthalten, beispielsweise „Warum ist es im Allgemeinen besser, einer Wohltätigkeitsorganisation Geld zu geben als einem Bettler?“ (aus dem Hamburg-Wechsler-Intelligenz-Test für Kinder, HAWIK; Hardesty und Priester 1963). Als Beispiel für eine „gute“ Antwort ist aufgeführt, dass man so sicher sein könne, dass das Geld tatsächlich den Bedürftigen zukäme; eine „schlechte“ Antwort hingegen wäre, dass der Bettler das Geld nur vertrinken würde. Die durchaus nachvollziehbare Antwort, dass man auch bei einer Wohltätigkeitsorganisation nicht immer sicher sein könne, dass das gespendete Geld bei den Bedürftigen angelangt, ist im Auswertungsheft allerdings nicht vorgesehen. Außerdem wird implizit in abwertender Weise angenommen, „Bettler“ seien Alkoholiker.

Suggestion vermeiden

Zudem ist es wichtig, dass die Fragen *keinen suggestiven Formulierungen/Inhalte* aufweisen, damit den Testpersonen die Antwort nicht in den Mund gelegt wird. So würde z. B. für das Statement „Würden Sie nicht auch von sich sagen, dass Sie den Zuzug von Migranten eigentlich immer schon problematisch fanden“ bei den Testpersonen eine höhere Zustimmung zu finden sein, als wenn die Frage ohne *suggerierende Antwortrichtung* gestellt wird, z. B. mit dem ergebnisoffenen Statement „Den Zuzug von Migranten halte ich für...“ mit verschiedenen Antwortalternativen (► Kap. 5).

Des Weiteren sollten die Iteminhalte so gewählt werden, dass das Antwortverhalten nur vom interessierenden Merkmal selbst abhängt und *nicht von merkmalsfremden Einflüssen mitbestimmt* wird. Solche Merkmalsvermengungen (Konfundierungen) sind insbesondere bei „heiklen“ Fragen aus den Inhaltsbereichen Religiosität, Pubertät, Sexualität, Weltanschauung, Politik etc. zu erwarten, weil hier – neben den interessierenden individuellen Merkmalsausprägungen – auch Gebote, Normen, Moralvorschriften sowie die Political Correctness einen wesentlichen Einfluss auf das Antwortverhalten ausüben; eine wahrheitsgemäße Beantwortung von heiklen Fragen (z. B. „Ich befriedige mich regelmäßig selbst“) kann durch solche Einflüsse erheblich verzerrt sein.

Merkmalskonfundierung vermeiden**Heikle Fragen in der Instruktion ankündigen**

Sofern es wegen des zu untersuchenden Merkmals unumgänglich sein sollte, *heikle Fragen* zu stellen, ist es – einer Empfehlung von Porst (2008) folgend – ratsam, die Testpersonen darüber zu informieren, dass die zu stellenden Fragen zwar unangenehm sein könnten, aufgrund der zugesicherten Anonymität aber keine nachteiligen Folgen durch wahrheitsgemäße Antworten zu befürchten seien. Erst nach dieser Vorwarnung sollte mit der Befragung gestartet werden.

4.6 Kognitive und motivationale Prozesse bei der Itembearbeitung

Dieser Abschnitt soll einen Überblick über die grundlegenden kognitiven und motivationalen Prozesse vermitteln, die bei der Itembearbeitung eine Rolle spielen und deshalb bei der Itemgenerierung ebenfalls bedacht werden sollten. Hierzu sollen zunächst die kognitiven Stadien und die potentiellen Fehlerquellen bei der Itembearbeitung dargestellt werden (► Abschn. 4.6.1); danach wird das Optimizing-Satisficing-Modell (Krosnick 1999) als motivationales Modell der Itembeantwortung vorgestellt (► Abschn. 4.6.2). Sowohl die kognitiven als auch die motivationalen Prozesse können als Grundlage für viele der in ► Abschn. 4.7 beschriebenen Fehlerquellen beim Antwortverhalten herangezogen werden.

4.6.1 Kognitive Prozesse bei der Itembearbeitung

Die *kognitiven Prozesse* bei der Bearbeitung von Items beinhalten nach Tourangeau et al. (2000) üblicherweise vier (bzw. fünf) Stadien (► Studienbox 4.1):

1. Verständnis und Interpretation des Iteminhalts
2. Abruf von Erinnerungen
3. Integration dieser Informationen zu einem einzelnen Urteil
4. Übersetzung des Urteils in eine Antwort, wobei Podsakoff et al. (2003) hierbei außerdem zwischen der Antwortwahl und der Antwortabgabe unterscheiden

Kognitive Anforderungen bei der Itembeantwortung

Mit jedem der Stadien gehen spezifische Fehlerquellen einher (► Studienbox 4.1), die in ► Abschn. 4.7 genauer beschrieben werden.

Zumeist kann nicht davon ausgegangen werden, dass die Testpersonen in jedem einzelnen Stadium eine explizite (bewusste) Entscheidung treffen; vielmehr kann die Bearbeitung als ein schneller (automatisierter) Prozess verstanden werden. Jedes der Stadien stellt hohe kognitive Anforderung an die Testpersonen (vgl. Krosnick und Fabrigar 1998). Das Ausmaß der tatsächlichen kognitiven Anstrengung und Gründlichkeit bei der Itembeantwortung hängt jedoch vom motivationalen Status der Testpersonen ab (► Abschn. 4.6.2).

4.6.2 Motivationale Prozesse bei der Itembeantwortung

Krosnick (1999) stellt im sog. *Optimizing-Satisficing-Modell* ein motivationales Modell der Itembeantwortung dar, das auf die kognitiven Prozesse direkt Bezug nimmt. Das Modell nimmt zwei unterschiedliche Beantwortungsmotive an, die den Antwortprozess beeinflussen.

Das *Optimizing* stellt das (positive) Motiv der Testperson dar, bei der Beantwortung der Items optimal mitzuhelfen und Informationen zu den interessierenden Merkmalen zu liefern. Positive Gründe können im Selbstbild der Testperson oder in einem Interesse am Selbstverständnis liegen; auch können zwischenmenschliche Verantwortung oder Altruismus, aber auch der Wille, einem Unternehmen oder dem Staat etc. zu helfen, als positive Motive gelten; weiterhin kommt auch eine (finanzielle) Belohnung als Beweggrund für eine gründliche Bearbeitung in Betracht. Als Konsequenz kann erwartet werden, dass Testpersonen in allen Stadien des Antwortprozesses eine hohe Motivation einbringen und eine gründliche Itembeantwortung erfolgt.

Optimizing: Testpersonen sind positiv motiviert

Im Gegensatz dazu handelt es sich bei *Satisficing* (engl. zusammengezogen aus „satisfying“ und „sufficing“) um ein Verhalten, das auftritt, wenn eine Testperson keine positiven Gründe hat, an der Testung teilzunehmen, oder sie im Laufe der Testung ermüdet bzw. eine abnehmende Motivation aufweist. Die Testung erfolgt dann nur beiläufig oder infolge einer angeordneten Teilnahmeverpflichtung. In solchen Situationen steht die Testperson vor einem Dilemma: Sie soll kognitiv anspruchsvolle Aufgaben bewältigen, will sich – mangels positiver Gründe – aber gleichzeitig nicht anstrengen.

Satisficing: Ausweichstrategien von unmotivierten Testpersonen

Als Konsequenz können zwei Verhaltensweisen bei der Itembeantwortung auftreten:

- Zum einen kann die Testperson die in die Itembearbeitung involvierten kognitiven Stadien – Verstehen, Abrufen, Urteilen, Antwortwahl und Antwortangabe – nur halbherzig ausführen und statt einer gründlich-optimalen eine nur oberflächliche Antwort wählen (*schwaches Satisficing*).
- Zum anderen können die Stadien des Abrufens und Urteilens vollständig von der Testperson ausgelassen werden. Die Testperson gibt dann eine Antwort, die ihr als eine vernünftige Antwort für den Testleiter erscheint. Diese erfüllt aber nicht mehr die Testintention, weil die Antwort unabhängig von tatsächli-

Schwaches vs. starkes Satisficing

Studienbox 4.1**Kognitive Stadien bei der Beschäftigung mit Testaufgaben (vgl. Podsakoff et al. 2003)**

1. Verständnis (*Comprehension*): Dieses Stadium beinhaltet erstens, dass die Testperson ihre Aufmerksamkeit auf die Aufgaben richtet, und zweitens, dass sie den Aufgabeninhalt und die Instruktion versteht.
Fehlerquelle: Eine etwaige Mehrdeutigkeit des Itemstamms und/oder der Instruktion stellt in diesem Stadium die Hauptfehlerquelle dar. Aufgrund der Mehrdeutigkeit (vgl. ► Abschn. 4.5.2) sucht die Testperson nach Zusatzinformation aus anderen Items bzw. im Kontext oder sie antwortet willkürlich.
2. Abruf (*Retrieval*): Nachdem die Aufgabe verstanden wurde, müssen nun Informationen durch Schlüsselreize aus dem Langzeitgedächtnis abgerufen werden und es wird eine Abrufstrategie entwickelt.
Fehlerquelle: Verschiedene Aspekte können diesen Prozess beeinflussen, z. B. die momentane Stimmungslage, die den Zugriff auf Erinnerungen und Informationen verändert. Eine negative Stimmungslage produziert vermehrt negative Erinnerungen (► Abschn. 4.6.2).
3. Urteil (*Judgment*): Nun bewertet die Testperson die abgerufenen Informationen hinsichtlich ihrer Vollständigkeit und Richtigkeit und entscheidet sich für ein Urteil.
Fehlerquelle: Aufgrund der vorher bearbeiteten Items könnte sich die Testperson bereits eine globale Meinung zu dem erfragten Themengebiet gebildet haben und jedes Item in einer mit dem globalen Urteil konsistenten Weise bewerten, obwohl sie eigentlich unabhängige Antworten geben sollte (► Abschn. 4.6.2).
4. Antwortwahl (*Response Selection*): Nachdem die Testperson ihr Urteil getroffen hat, überprüft sie nun die angebotenen Antwortmöglichkeiten und versucht, ihr Urteil durch die Wahl einer geeigneten Antwortmöglichkeit entsprechend abzubilden.
Fehlerquelle: In diesem Stadium werden Fehler durch Antworttendenzen hervorgerufen, z. B. indem die Testperson Zustimmung bevorzugt oder extreme Antworten scheut (► Abschn. 4.7.3 und 4.7.4).
5. Antwortabgabe (*Response Reporting*): Den letzten Schritt stellt die Überprüfung auf inhaltliche Konsistenz zwischen der getroffenen Entscheidung und der tatsächlichen Abgabe der Antwort (also z. B. dem Kreuz auf dem Fragebogen) dar.
Fehlerquelle: Einer der möglichen Fehler kann hier sein, dass die Testperson ihre Antwort noch dahingehend verändert, dass sie eine sozial erwünschte Antwort gibt (► Abschn. 4.7.1), also eine Antwort, die sie in den Augen anderer positiver erscheinen lässt, als es bei wahrheitsgemäßer Antwort der Fall wäre.

Jedes dieser Stadien kann selbst wiederum in komplexe Schritte unterteilt werden (s. z. B. Clark und Clark 1977; Krosnick 1999).

chen Einstellungen, Meinungen und Interessen der Testperson gegeben wird. Das Antwortverhalten der Testperson wird dadurch gänzlich arbiträr und verliert jeden Bezug zum interessierenden Merkmal (*starkes Satisficing*). Typische Formen der Antwortabgabe können dann „sichere“ Antworten beinhalten (wie die Wahl der mittleren Kategorie in Ratingskalen, ► Kap. 5), das Ankreuzen von „Weiß-nicht“-Kategorien oder – in Extremfällen – die Abgabe von Zufallsantworten.

4.7 · Response-Bias als Fehlerquelle beim Antwortverhalten

Das Auftreten von Satisficing wird insbesondere durch drei Faktoren begünstigt: Es tritt verstärkt auf, wenn die Aufgabenschwierigkeit hoch ist, wenn die (kognitiven) Antwortfähigkeiten der Testperson eingeschränkt sind und wenn die Gründe zum Optimizing gering sind. Unter diesen Gesichtspunkten ist eine einfache und eindeutige Formulierung der Items, wie sie in den vorherigen Abschnitten dargestellt wurde, unerlässlich. Des Weiteren treten beim Satisficing verstärkt Antworttendenzen auf, auf die im folgenden Abschnitt eingegangen werden soll.

4.7 Response-Bias als Fehlerquelle beim Antwortverhalten

Wenn Testpersonen bei der Itembearbeitung nicht die Antworten geben, die den Ausprägungen des interessierenden Merkmals entsprechen, sondern sich z. B. im Sinne des (starken) Satisficing von merkmalsfremden Gesichtspunkten leiten lassen, so kann dies an Erscheinungsformen des *Response-Bias* liegen. Paulhus (1991) spricht dann von Response-Bias, wenn es eine *systematische Tendenz* gibt, auf Items in einem Fragebogen Antworten zu geben, die nichts mit dem eigentlich interessierenden Merkmal zu tun haben. Tritt ein Response-Bias auf, so wirkt dieser als antwortverfälschende Störvariable; die Validität der Items und des Tests insgesamt wird durch die Existenz von unkontrollierten Störvariablen erheblich beeinträchtigt.

4.7.1 Antworttendenz vs. Antwortstil

Die Entstehung des Response-Bias lässt sich insbesondere auf die beiden folgenden Sichtweisen zurückführen:

- Die eine Sichtweise geht davon aus, dass ein Response-Bias nur in bestimmten Situationen auftritt und z. B. durch bestimmte Itemformate oder durch Zeitdruck bei der Testbearbeitung hervorgerufen wird (Cronbach 1946, 1950). Der Bias wird also insbesondere als eine Eigenschaft bzw. ein Artefakt der Testsituation und der Items selbst angesehen und wird in diesem Zusammenhang als Antworttendenz (*Response Set*) bezeichnet (Moors 2008). Dies impliziert, dass eine Veränderung der Testsituation oder der Items auch zu einer Verringerung des Response-Bias führen kann.
- Einer anderen Sichtweise zufolge kann der Response-Bias als eine Persönlichkeitseigenschaft verstanden werden, die dann vorliegt, wenn sich das spezifische Antwortverhalten über verschiedene Items, Methoden und Situationen hinweg konsistent zeigt (van Herk et al. 2004). In diesem Fall spricht man von Antwortstil (*Response Style*; Baumgartner und Steenkamp 2001; Messick 1991) und geht davon aus, dass diese Persönlichkeitseigenschaft im Prinzip messbar ist (z. B. Couch und Keniston 1960).

In vielen Untersuchungen konnten Belege für beide Varianten gefunden werden: In kulturvergleichenden Studien zeigte sich, dass ein Response-Bias in Abhängigkeit der Kulturgehörigkeit stärker oder schwächer ausgeprägt sein kann. Dies lässt eher auf eine Persönlichkeitseigenschaft schließen (z. B. van Herk et al. 2004), da diese Unterschiede trotz Konstanthaltung der Messinstrumente auftraten. In anderen experimentellen Studien wurde hingegen gezeigt, dass Änderungen insbesondere im Antwortformat der Ratingskalen zu Veränderungen im Response-Bias führen (z. B. Weijters et al. 2010; vgl. Empfehlungen in ▶ Kap. 5).

Unabhängig davon, welcher Sichtweise man den Vorzug gibt, ist die Berücksichtigung des Response-Bias für die Validität eines Tests unerlässlich (Schuman und Presser 1981). Dies kann durch eine gezielte Berücksichtigung beim Testdesign (also bei der Itemformulierung, dem Antwortformat und der Testart) oder

Ursachen und Konsequenzen von Satisficing

Response-Bias ist die systematische Verzerrung von Antworten

Response Set – Antworttendenz

Response Style – Antwortstil

Belege für Response Set und Response Style

Berücksichtigung von Response-Bias beim Testdesign

durch eine statistische Kontrolle erfolgen (Podsakoff et al. 2003, 2012). Im Allgemeinen kann davon ausgegangen werden, dass es mehrere, konfundierte Störvariablen/Fehlerquellen in einem Test geben kann, die sich in ihrer Wirkungsintensität – je nach Testsituation – unterscheiden können. Sie sind zumeist am stärksten ausgeprägt in Tests mit vagen bzw. uneindeutigen sowie schwierigen Items (Moors 2008) sowie in Situationen, in denen Personen Items im Sinne des Satisficing bearbeiten.

Nachfolgend soll auf die wichtigsten Formen von Response-Bias und seiner Vermeidung eingegangen werden. Es soll hierbei auch Bezug auf die zuvor eingeführten kognitiven und motivationalen Modelle genommen werden (► Abschn. 4.6).

4.7.2 Soziale Erwünschtheit

Testergebnisse haben für die Teilnehmer häufig weitreichende Konsequenzen, z. B. wenn man an Auswahlentscheidungen im Bildungsweg, an die Aufnahme einer Arbeitstätigkeit oder an die Indikation einer Therapieform denkt. Auf diesem Hintergrund neigen Testpersonen dazu, sich in einem möglichst günstigen Licht darzustellen. Als Soziale Erwünschtheit (*Social Desirability*) wird ein Antwortverhalten bezeichnet, das in vielen Fragebögen verzerrende Effekte produziert und viele Testergebnisse verfälscht. Hierbei handelt es sich eher um einen Antwortstil als um eine Antworttendenz.

Soziale Erwünschtheit setzt sich aus zwei Aspekten zusammen, und zwar aus Selbsttäuschung (*Self-deceptive Enhancement*) und Fremdtäuschung (*Impression Management*; Paulhus 1984). In Bezug auf den Aspekt des *Impression Management* geht man davon aus, dass Menschen bemüht sind, den Eindruck, den sie auf andere machen, in eine günstige Richtung zu lenken und zu kontrollieren (Fremdtäuschung); *Impression Management* ist kein Verhalten in Ausnahmesituationen, sondern ein ganz wesentliches Element unseres Verhaltens im alltäglichen sozialen Kontext (s. z. B. Crowne und Marlowe 1964). *Self-deceptive Enhancement* stellt die eher unbewusste Tendenz dar, vorteilhafte Selbsteinschätzungen zu produzieren, die man selbst aber als ehrlich ansieht (Selbsttäuschung).

Personen, die sich „sozial erwünscht“ verhalten, äußern eher Meinungen und Einstellungen, von denen sie annehmen, dass sie mit den sozialen Normen und Werten der Gesellschaft übereinstimmen. Aussagen zu Verhaltensweisen, die zwar weitverbreitet sind, aber auf soziale Ablehnung stoßen, werden von den Testpersonen verneint, beispielsweise „In manchen Fällen komme ich ohne Begründung zu spät zur Arbeit“ oder „Ich reagiere unfreundlich, wenn ich um einen Gefallen gebeten werde“.

Im Sinne des Optimizing-Satisficing-Modells kann davon ausgegangen werden, dass eine verstärkte Verzerrung durch die Soziale Erwünschtheit beim Satisficing auftritt, da die Testpersonen eher oberflächliche Antworten geben und eher sozial erwünschte Antworten liefern, da dies ihren Vorstellungen von einer „korrekten“ oder plausiblen Antwort entspricht (Krosnick 1999). Der Effekt der Sozialen Erwünschtheit ist bei mündlichen Interviews im Allgemeinen stärker ausgeprägt als bei schriftlichen oder telefonischen Befragungen, da bei Letzteren keine Face-to-Face-Begegnung stattfindet; eine subjektive Anonymität ist dadurch eher gewährleistet (Holbrook et al. 2003; Kreuter et al. 2008).

Zur Verringerung des Effekts der Sozialen Erwünschtheit hilft bei wissenschaftlichen Studien eine Aufklärung über den Untersuchungsgegenstand sowie eine Zusicherung der Anonymität der Testpersonen. Auch in anderem Kontext, z. B. bei Arbeitszufriedenheitsuntersuchungen, sind Erklärungen darüber nötig, dass der Arbeitgeber keine personalisierten Daten erhält. Bei sog. „Self-Assessments“ (Selbsttestungen) liegt die Anonymitätsbedingung per se vor. Diese stellt z. B. beim Online Self Assessment (OSA) Psychologie (Reiß und Moosbrugger 2008) einen

Selbst- und Fremdtäuschung

Orientierung an sozialen Normen

Satisficing begünstigt die Verzerrung durch Soziale Erwünschtheit

Anonymität schwächt sozial erwünschte Antworten ab

4.7 · Response-Bias als Fehlerquelle beim Antwortverhalten

großen Vorteil dar, bei dem Studienplatzinteressierte im Numerus-Clausus-Fach Psychologie anhand ihrer Testergebnisse sehr konkrete Hinweise für sich selbst erhalten, ob sie für ein erfolgreiches Psychologiestudium geeignet sind, ohne sich gegenüber der Testleitung oder einer Auswahlkommission „sozial erwünscht“ verhalten zu müssen.

Eine andere Technik, mit dem Effekt des sozial erwünschten Antwortverhaltens umzugehen, besteht darin, dass man das Ausmaß der Verhaltensverfälschung durch Verwendung von sog. „Kontrollskalen“ explizit erfasst. Mit einer Kontrollskala, auch „Lügenskala“ genannt, kann man die Tendenz der Testperson zu sozial erwünschtem Verhalten ermitteln. Hierbei muss die Testperson Stellung zu sozial unerwünschten Verhaltensweisen beziehen, die in der Bevölkerung aber sehr verbreitet sind, z. B. „Als Kind habe ich manchmal gelogen“. Eine Verneinung dieses Items würde bedeuten, dass die Testperson eine starke Neigung besitzt, sich vorteilhaft darzustellen, also sozial erwünscht zu antworten. Eines der ersten Inventare, in denen Kontrollskalen verwendet wurden und heute noch werden, ist das auf Hathaway und McKinley (1943) zurückgehende MMPI (MMPI-2; Hathaway et al. 2000; MMPI-2RF; Ben-Porath und Tellegen 2011), das insgesamt drei solche Kontrollskalen („Lügenskalen“) enthält. Im FPI-R (Fahrenberg et al. 2010) wird die Kontrollskala als „Offenheitsskala“ bezeichnet. Je stärker die diagnostizierte Soziale Erwünschtheit ausfällt, desto mehr Umsicht ist bei der Interpretation der eigentlichen Testergebnisse angezeigt.

Auch die bei Brandt und Moosbrugger (► Kap. 3) vorgestellten Objektiven Persönlichkeitstests und Impliziten Testverfahren verfolgen das Ziel, durch die Verschleierung des Messprinzips bzw. durch das anstelle einer Selbstauskunft erhobene unmittelbare Verhalten eine Verfälschung des Testergebnisses in Richtung von Sozialer Erwünschtheit von vorneherein zu unterbinden. Über weitere Methoden zur Kontrolle der Tendenz zu Sozialer Erwünschtheit berichten Döring und Bortz (2016, S. 437 ff.).

4.7.3 Akquieszenz

Als Akquieszenz (Zustimmungstendenz) bezeichnet man die Tendenz, den vorgegebenen Fragen oder Statements unkritisch, d. h. unabhängig vom Iteminhalt zuzustimmen (Winkler et al. 1982). Die Akquieszenz einer Testperson würde sich z. B. darin äußern, dass diese sowohl einer positiv formulierten Aussage als auch der gegenteiligen Aussage zustimmt. Durch Akquieszenz wird die tatsächliche Merkmalsausprägung somit verzerrt erfasst. Untersuchungen haben gezeigt, dass Testpersonen im Durchschnitt lieber einem positiv formulierten Statement zustimmen, als dasselbe Statement, wenn es negativ formuliert ist, abzulehnen (Krosnick 1999; Krosnick et al. 2014). Die Zustimmungstendenz ist auch bei Leistungstests beobachtbar, und zwar in der Form, dass Testpersonen, wenn sie z. B. bei dichotomen Aufgaben raten, häufiger die Option „richtig“ wählen als die Option „falsch“ (Krosnick und Fabrigar 1998). Vereinzelt ist auch eine zur Akquieszenz gegenläufige sog. „Ablehnungstendenz“ zu beobachten.

Im Sinne des Satisficing ist die Akquieszenz dadurch zu erklären, dass Testpersonen typischerweise eher nach Gründen für Zustimmung als für Ablehnung suchen, was bei einer nur oberflächlichen Bearbeitung der Items (im Sinne eines schwachen Satisficing) entsprechend verstärkt auftritt (Krosnick 1999). Tritt ein starkes Satisficing bei Testpersonen auf, so kann erwartet werden, dass die Zustimmung auch einfach nur Ausdruck von Höflichkeit gegenüber dem Testleiter ist.

Akquieszenz tritt verstärkt bei Testpersonen mit geringen kognitiven Fähigkeiten auf, ebenso bei Testpersonen, die ermüdet sind, sowie bei schwierigen Items und langen Testbatterien (Krosnick et al. 2014). Sie ist weiterhin eher in unpersön-

Kontroll-/Lügen-/Offenheitsskalen

Objektive Testverfahren beugen vor

Zustimmungstendenz

Akquieszenz als Konsequenz von Satisficing

Aufdeckung durch Iteminversion**Situationen verstärkter Zustimmungstendenz****Wechselnde Itempolung kann Faktorenstruktur beeinflussen****Effekte der Itempolung abhängig von der Zielgruppe****Ungerechtfertigte Bevorzugung der mittleren Antwortkategorie**

lichen Befragungen (z. B. Telefoninterviews) zu beobachten als in Face-to-Face-Befragungen.

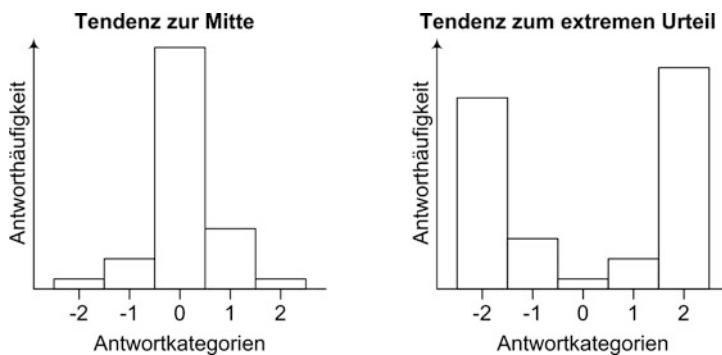
Ein oft empfohlenes Mittel, um Verzerrungseffekte durch Akquieszenz aufzudecken, ist die Verwendung invertierter Items, d. h., es werden Items in positiv formulierter Form dargeboten, wobei Zustimmung einen hohen Wert in dem interessierenden Merkmal anzeigt; die gleichen Items werden aber auch durch variiertes *Item-Wording* in negierter Form („invertiert“, „negativ gepolt“) dargeboten, wobei die Zustimmung einen niedrigen Wert in dem interessierenden Merkmal anzeigt. Testpersonen, die den positiv formulierten Items zugestimmt haben, müssten die invertierten Items eigentlich ablehnen; eine erneute Zustimmung weist hingegen auf Akquieszenz hin. Beispielsweise würde eine Zustimmung zu dem Statement „Atomkraftwerke sollten in Deutschland abgeschaltet werden“ sowie eine Zustimmung zur Aussage „Atomkraftwerke sollten in Deutschland weiterbetrieben werden“ deutlich für einen Akquieszenzeffekt sprechen. Bei der Iteminversion sollte bedacht werden, dass das alleinige Einfügen eines „nicht“ in den Itemstamm nicht unbedingt zu einer eindeutigen Inversion des Items führt und unter Berücksichtigung der in ► Abschn. 4.5.1 aufgeführten Gesichtspunkte als problematisch anzusehen ist; auch sollte beachtet werden, dass negativ formulierte Items häufig eine höhere Schwierigkeit aufweisen (s. auch Eifermann 1961; Wason 1961).

Bei der Verwendung von unterschiedlich gepolten Items kann allerdings als Effekt der *Itempolung* eine artificielle Faktorenstruktur entstehen, weil trotz des Vorhandenseins eines homogenen, eindimensionalen Merkmals positiv und negativ gepolte Items dazu tendieren, zwei verschiedene Faktoren zu bilden (Greenberger et al. 2003). Mithilfe von „Methodenfaktoren“ (► Kap. 25) kann diese Tendenz genauer untersucht werden (s. z. B. Rauch et al. 2007 oder Höfling et al. 2011).

Dass Effekte der Itempolung in unterschiedlichen Subgruppen der Gesellschaft unterschiedlich stark sind, weist eine Studie von Krebs und Matschinger (1993) aus. Die Ergebnisse zeigen, dass die Zustimmungstendenz mit zunehmendem Alter und sinkender Schulbildung steigt und dass das Ausmaß der Asymmetrie zwischen Zustimmung zu negativen und Ablehnung von positiven Items am höchsten bei Subgruppen mit niedriger Schulbildung und hohem Alter ist. Benachteiligte Gruppen der Gesellschaft (niedrige Schulbildung, hohes Alter) mögen zwar insgesamt negativere Einstellungen haben als andere; durch die Verwendung negativ (im Sinne der interessierenden Einstellung) gepolter Items wird dieser Eindruck jedoch verstärkt, weil das Antwortverhalten auf negativ gepolte Items mit soziografischen Merkmalen variiert und darüber hinaus innerhalb der Merkmalskategorien unterschiedlich starke Asymmetrien aufweist.

4.7.4 Tendenz zur Mitte

Als *Tendenz zur Mitte* wird die bewusste oder unbewusste Bevorzugung der mittleren (neutralen) Antwortkategorien unabhängig vom Iteminhalt verstanden (Paulhus 1991; van Herk et al. 2004). Die umgekehrte, seltene Tendenz der Vermeidung der mittleren Antwortkategorie wird als „Tendenz zum extremen Urteil“ bezeichnet. Die Bevorzugung mittlerer Antwortkategorien kann auf ein subjektiv unzureichendes Wissen zurückzuführen sein („Ich bin mir nicht ganz sicher in meiner Einschätzung, ich weiß zu wenig für ein sicheres Urteil – in der Mitte kann ich am wenigsten falsch machen!“) oder aber auf die Ansicht, dass sich die Antwortalternativen zur Beurteilung nicht eignen. Wenn Testpersonen dazu neigen, ihre Entscheidungen auf die mittleren Kategorien zu beschränken, führt dies zu einer verringerten Itemvarianz (vgl. ► Kap. 7) und zu Verzerrungen. □ Abb. 4.1 zeigt typische Antwortverteilungen bei „Tendenz zur Mitte“ und bei „Tendenz zum extremen Urteil“.



■ Abb. 4.1 Mögliche Häufigkeitsverteilungen der Kategorienwahl bei Antworttendenzen

Um der Tendenz zur Mitte entgegenzuwirken, sollte zum einen keine neutrale Mittelkategorie angeboten werden (vgl. ► Kap. 5); zum anderen sollten möglichst keine zu extremen sprachlichen Bezeichnungen für die Pole der Beurteilungsskalen gewählt werden. Auch das Anbieten einer eigenen „Weiß-nicht“-Kategorie kann dieser Tendenz vorbeugen (► Kap. 5).

Möglichkeiten zur Verringerung der Tendenz zur Mitte

4.7.5 Effekte der Itemreihenfolge

Auch die Reihenfolge der Items, aus denen der Fragebogen bzw. Test aufgebaut ist, kann einen ergebnisverfälschenden Einfluss auf die Itembeantwortung ausüben. Hierbei sind vor allem Anker-, Konsistenz-, Subtraktions- und Testlet-Effekte zu nennen.

Als *Ankereffekte* werden Effekte bezeichnet, die durch ein Aufeinanderfolgen von Informationen und Items auftreten und somit im weiteren Sinne einen *Priming-Effekt* darstellen (Salancik und Pfeffer 1977). Bei Tests können diese Effekte auftreten, wenn ein vorangegangenes Item Hinweise auf die Beantwortung des aktuellen Items liefert oder einen Rahmen (Anker) darstellt. Im kognitionspsychologischen Sinne können diese Ankereffekte als urteilsheuristische Effekte verstanden werden, die häufig auch im Alltag zu beobachten sind (Tversky und Kahneman 1974; s. auch Critcher und Gilovich 2008; Mussweiler und Strack 1999).

Ankereffekte

Konsistenzeffekte (oder auch *Assimilationseffekte*; Salancik und Pfeffer 1977) können als eine Facette der Ankereffekte angesehen werden, die dadurch entstehen, dass Testpersonen versuchen, ein kohärentes bzw. konsistentes Bild in ihrem Antwortmuster zu erzeugen. Dies geschieht dadurch, dass sie aufeinanderfolgende Items ähnlicher beantworten, als dies zu erwarten wäre, wenn die Fragen unabhängig voneinander gestellt worden wären. *Kontrasteffekte* stellen analog dazu Effekte dar, die auftreten, wenn die Beantwortung zweier Items unterschiedlicher ausfällt, wenn sie gemeinsam anstatt getrennt erfragt werden. Konsistenzeffekte können auch als Antwortstile aufgefasst werden, die verstärkt auftreten, wenn Testpersonen ein Satisficing-Motiv haben (vgl. Krosnick 1999; Podsakoff et al. 2003).

Konsistenz- und Kontrasteffekte

Eine Möglichkeit, Ankereffekte im Allgemeinen und Konsistenz- oder Kontrasteffekte im Speziellen zu verringern, besteht in mehrdimensionalen Tests z. B. darin, die Position der jeweils zu einer Dimension gehörenden Items über die beteiligten Dimensionen hinweg zu randomisieren, sodass nicht Items derselben Dimension direkt aufeinander folgen. Die durch die Randomisierung erzielte vergrößerte Distanz zwischen den Items erhöht die Wahrscheinlichkeit, dass Testpersonen die Items unabhängig voneinander beantworten (Podsakoff et al. 2003). Weiterhin erlauben moderne psychometrische Verfahren die Berücksichtigung oder zumindest eine Beurteilung von Reihenfolgeeffekten (z. B. Debeer und Janssen 2013; Weirich et al. 2014).

Randomisierung als Maßnahme zur Verringerung von Ankereffekten

Subtraktionseffekte

Subtraktionseffekte treten dadurch auf, dass nach einer (oder mehreren) spezifischen Fragen eine allgemeinere Frage folgt. Beispielsweise könnte nach der Frage „Sind Sie sehr gesprächig in der Gesellschaft Ihrer Familie?“ eine Frage „Sehen Sie sich im Allgemeinen als gesprächig an?“ dazu führen, dass die Testperson die zweite Frage in dem Sinne beantwortet, dass sie sich nur auf Situationen außerhalb der Familie bezieht. Auch hier erscheint es intuitiv plausibel, dass ein separates Beantworten der beiden Fragen ein anderes Antwortmuster ergäbe als bei einer direkten Aufeinanderfolge beider Fragen.

Testlet-Effekte

Testlet-Effekte können dann auftreten, wenn mehrere Fragen zu einem gemeinsamen Itemstamm (Testlet) gestellt werden. Dies tritt z. B. in den internationalen Vergleichsstudien des „Programme for International Student Assessment (PISA)“ (OECD 2014) auf, in denen zu einem vorgegebenen Problembereich jeweils mehrere Fragen gestellt werden. Die richtige oder falsche Beantwortung einer der Fragen kann somit die Beantwortung der verbliebenen Fragen beeinflussen (Rijmen 2010). Eine Kontrolle dieses durch das Design induzierten Effekts kann durch spezifische psychometrische Testlet-Modelle (Bradlow et al. 1999; Rijmen 2010; Wainer et al. 2007, vgl. ► Kap. 18) erfolgen, die berücksichtigen können, dass mehrere Antworten zu derselben Aufgabe gehören und somit von der jeweiligen Testperson wahrscheinlich ähnlicher beantwortet werden.

Nicht nur zwischen den Items, sondern auch innerhalb der einzelnen Items lassen sich *Reihenfolgeeffekte von Antwortalternativen* feststellen. Untersuchungen haben gezeigt, dass die Reihenfolge der Antwortalternativen einen Response-Bias induziert (Krosnick et al. 2014), und zwar insbesondere in Situationen, in denen Testpersonen Fragen verstärkt im Sinne des Satisficing beantworten. Werden die Antwortalternativen visuell präsentiert, so ist eine vermehrte Zustimmung zu den ersten Antwortalternativen zu beobachten (*Primacy Effect*; Galesic et al. 2008). Werden sie hingegen vorgelesen, so ist eine vermehrte Zustimmung zu den Antwortalternativen am Ende der Auswahl zu beobachten (*Recency Effect*; Holbrook et al. 2007).

Einen informativen Überblick über diese und weitere Arten von Antwortstilen, Antworttendenzen sowie über andere Fehlerquellen geben Jäger und Petermann (1999, S. 368 ff.) sowie Podsakoff et al. (2003, 2012; s. auch ► Studienbox 4.1 in ► Abschn. 4.6.1).

4.8 Computerunterstützte Itemkonstruktion

Über Möglichkeiten und Verfahren der computerunterstützten Itemkonstruktion informieren Goldhammer und Kröhne in ► Kap. 6. Eine ausführliche Darstellung von Reihenfolgeeffekten ist auch bei Eid und Schmidt (2014) zu finden.

4.9 Zusammenfassung

Die Itemgenerierung verfolgt das Ziel, repräsentative, inhaltsvalide Operationalisierungen des interessierenden Merkmals zu finden und diese in entsprechenden Aufgaben/Items abzubilden. Dazu wurde auf typische Vorgehensweisen eingegangen sowie auf wichtige Aspekte, die bei der Formulierung der Items beachtet werden müssen, vor allem auf die sprachliche Verständlichkeit, die Eindeutigkeit des Iteminhalts und die Vermeidung bestimmter Iteminhalte. Basierend auf der Erörterung von typischen kognitiven und motivationalen Prozessen bei der Itembeantwortung wurden verschiedene potentielle Störvariablen des Antwortverhaltens (Response-Bias, Antwortstil, Antworttendenz, Soziale Erwünschtheit, Akquieszenz, Tendenz zur Mitte und Effekte der Itemreihenfolge) näher erläutert. Diese Störvariablen sollten bereits bei der Itemgenerierung mitberücksichtigt wer-

4.10 · Kontrollfragen

den, da sie das Ergebnis von Tests und Fragebogen verfälschen können; Möglichkeiten zur Verringerung ihres Einflusses wurden diskutiert.

4.10 Kontrollfragen

- ?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Worin unterscheiden sich die intuitive und die rationale Strategie der Itemgenerierung?
 2. Worin unterscheiden sich die kriteriumsorientierte und die faktorenanalytische Strategie der Itemgenerierung?
 3. In welchen der Stadien der Aufgabenbeantwortung (nach Podsakoff et al. 2003) ist insbesondere mit Effekten der Selbstdäuschung (*Self-deceptive Enhancement*) zu rechnen, in welchen mit Effekten der Fremtdäuschung (*Impression Management*)?
 4. Erklären Sie die Begriffe Response-Bias, Antworttendenz und Antwortstil.
 5. Welche Aspekte sollte man bei der Reihenfolge von Items in einem Fragebogen berücksichtigen? Wie sollte man sie berücksichtigen?
 6. Sie erstellen einen Persönlichkeitstest, der sowohl bei unterdurchschnittlich als auch bei durchschnittlich begabten Testpersonen eingesetzt werden soll. Welche Störvariablen sind Ihrer Erwartung nach bei unterdurchschnittlich begabten Testpersonen stärker ausgeprägt als bei den durchschnittlich begabten? Wie könnten Sie diese Einflüsse verringern?

Literatur

- Angleitner, A., John, O. P. & Löhr, F.-J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. Wiggins (Eds.), *Personality assessment via questionnaires. Current issues in theory and measurement* (pp. 61–108). Berlin, Heidelberg: Springer.
- Baumgartner, H. & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, 38, 143–156.
- Ben-Porath, Y. S. & Tellegen, A. (2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form): Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Clark, H. H. & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.
- Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 60, 151–174.
- Critcher, C. R. & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, 21, 241–251.
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475–494.
- Cronbach L. J. (1950). Further evidence on response sets and test validity. *Educational and Psychological Measurement*, 10, 3–31.
- Crowne, D. & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York, NY: Wiley.
- Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50, 164–185.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eifermann, R. R. (1961). Negation: A linguistic variable. *Acta Psychologica*, 18, 258–273.
- Fahrenberg, J., Hampel, R. & Selg, H. (2010). *Freiburger Persönlichkeitsinventar (FPI-R)* (8. Aufl.). Göttingen: Hogrefe.

- Fahrenberg, J. & Selg, H. (1968). *Das Persönlichkeitssinventar ALNEV* (Unveröffentlichter Arbeitsbericht). Freiburg/Br.
- Galesic, M., Tourangeau, R., Couper, M. P. & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72, 892–913.
- Greenberger, E., Chuanheng, Ch., Dmitrieva, J. & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences*, 35, 1241–1254.
- Hardesty, F. P. & Priester, H. J. (1963). *Hamburg-Wechsler-Intelligenz-Test für Kinder. HAWIK* (2. Aufl.). Bern: Huber.
- Hathaway, S. R., McKinley, J. C. & Engel, R. (Hrsg.) (2000). *MMPI-2. Minnesota Multiphasic Personality Inventory 2*. Göttingen: Hogrefe.
- Hathaway, S. R. & McKinley, J. C. (1943). *Manual of the Minnesota Multiphasic Personality Inventory*. New York, NY: Psychological Corporation.
- Höfling, V., Moosbrugger, H., Schermelleh-Engel, K. & Heidenreich, T. (2011). Mindfulness or Mindlessness? *European Journal of Psychological Assessment*, 27, 1, 59–64.
- Holbrook, A. L., Green, M. C. & Krosnick, J. A. (2003). Telephone vs. face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67, 79–125.
- Holbrook A. L., Krosnick, J. A., Moore, D. & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of questions and respondent attributes. *Public Opinion Quarterly*, 71, 325–348.
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, 57, 253–270.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, 35, 21–35.
- Jäger, R. S. & Petermann, F. (Hrsg.) (1999). *Psychologische Diagnostik. Ein Lehrbuch* (4. Aufl.). Weinheim: Beltz PVU.
- Krebs, D. & Matschinger, H. (1993). *Richtungseffekte bei Itemformulierungen. Arbeitspapier*. Mannheim: ZUMA.
- Kreuter, F., Presser, S. & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847–865.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537–567.
- Krosnick, J. A. & Fabrigar, L. R. (1998). *Designing Good Questionnaires: Insights from Psychology*. New York, NY: Oxford University Press.
- Krosnick, J. A., Lavrakas, P. J. & Kim, N. (2014). Survey research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2nd ed.). New York, NY: Cambridge University Press.
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.
- McCrae, R. R. & Costa, P. T. (2010) *NEO inventories for the NEO Personality Inventory-3 (NEO-PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3), NEO Personality Inventory-Revised (NEO PI-R): professional manual*. Lutz, FL: PAR.
- Messick, S. (1991). Psychology and the methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 200–221). Hillsdale, NJ: Erlbaum.
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, 42, 779–794.
- Moosbrugger, H., Jonkisz, E. & Fucks, S. (2006). Studierendenauswahl durch die Hochschulen – Ansätze zur Prognostizierbarkeit des Studienerfolgs am Beispiel des Studiengangs Psychologie. *Report Psychologie*, 3, 114–123.
- Mussweiler, T. & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35, 136–164.
- Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Ergebnisse: Was Schülerinnen und Schüler wissen und können (Band I, überarbeitete Ausgabe): Schülerleistungen in Lesekompetenz, Mathematik und Naturwissenschaften*. Bielefeld: W. Bertelsmann.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. 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. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609.
- Podsakoff, P. M., MacKenzie, S. B., Lee J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88, 879–903.

Literatur

- Podsakoff, P. M., MacKenzie, S. B. & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569.
- Porst, R. (2008). *Fragebogen. Ein Arbeitsbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Rauch, W. Schweizer, K. & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, 42, 1597–1607.
- Reiß, S. & Moosbrugger, H. (2008) *Online Self Assessment Psychologie*. Institut für Psychologie der Goethe-Universität Frankfurt am Main. Verfügbar unter https://www.psychologie.uni-frankfurt.de/49829947/20_self-Assessment [20.12.2019]
- Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47, 361–372.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.
- Salancik, G. R. & Pfeffer, J. (1977). An examination of the need-satisfaction models of job attitudes. *Administrative Science Quarterly*, 22, 427–456.
- Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, CA: Academic Press.
- Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology*, 26, 249.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Thurstone, L. L. & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago, IL: University of Chicago Press.
- Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Van Herk, H., Poortinga, Y. H. & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35, 346–360.
- Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.
- Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, 52, 133–142.
- Weijters, B., Cabooter, E. & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236–247.
- Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535–548.
- Winkler, J. D., Kanouse, D. E. & Ware, Jr., J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67, 555–561.



Antwortformate und Itemtypen

Helfried Moosbrugger und Holger Brandt

Inhaltsverzeichnis

- 5.1 Antwortformate im Überblick – 93**
- 5.2 Aufgaben mit freiem Antwortformat – 94**
 - 5.2.1 Kurzaufsatzaufgaben – 94
 - 5.2.2 Ergänzungsaufgaben – 95
- 5.3 Aufgaben mit gebundenem Antwortformat – 96**
 - 5.3.1 Ordnungsaufgaben – 97
 - 5.3.2 Auswahlaufgaben – 99
 - 5.3.2.1 Konstruktion geeigneter Distraktoren – 99
 - 5.3.2.2 Disjunkttheit der Antwortalternativen – 101
 - 5.3.2.3 Exhaustivität der Antwortmöglichkeiten – 101
 - 5.3.2.4 Anzahl der Antwortalternativen – 102
 - 5.3.3 Beurteilungsaufgaben – 105
 - 5.3.3.1 Kontinuierliche vs. diskrete Beurteilungsskalen – 105
 - 5.3.3.2 Verwendung von Skalenstufen – 106
 - 5.3.3.3 Unipolare vs. bipolare Antwortskala – 106
 - 5.3.3.4 Bezeichnung der Skalenpunkte – 107
 - 5.3.3.5 Verwendung einer neutralen Mittelkategorie – 109
 - 5.3.3.6 „Weiß nicht“ als separate Antwortalternative – 110
 - 5.3.3.7 Festlegung der symptomatischen bzw. unsymptomatischen Antwortrichtung – 111
 - 5.3.3.8 Einsatz asymmetrischer Beurteilungsskalen und itemspezifischer Antwortformate – 111
 - 5.3.3.9 Zusammenfassende Bewertung – 112
- 5.4 Aufgaben mit atypischem Antwortformat – 112**
- 5.5 Entscheidungshilfen für die Wahl des Aufgabentyps – 114**
- 5.6 Computerunterstützte Antwortformate – 114**

5.7 Zusammenfassung – 115

5.8 Kontrollfragen – 115

Literatur – 115

5.1 · Antwortformate im Überblick

i Neben den Überlegungen bei der Formulierung des Aufgabenstamms muss bei der Konstruktion von Test- und Fragebogenitems auch eine Festlegung des Antwortformats und des Itemtyps erfolgen. Die Item-/Aufgabentypen unterscheiden sich vor allem darin, wie die Beantwortung der Aufgaben vorzunehmen ist. Daraus resultiert die Art und Weise, wie die einzelnen Aufgabenstellungen (d. h. Aufgabenstamm und Antwortformat) administriert werden. Diese Festlegungen sind für die Objektivität (Durchführung, Auswertung, Interpretation) sowie für die Ökonomie eines Fragebogens oder Tests von erheblicher Bedeutung.

5.1 Antwortformate im Überblick

Die Beantwortung einer Testaufgabe/-frage bzw. die Stellungnahme zu einem Statement kann auf sehr verschiedene Art in bestimmten Antwortformaten erfolgen, wobei die Klassifikation der Item-/Aufgabentypen in der Regel nach dem Strukturiertheitsgrad des jeweiligen Antwortformats vorgenommen wird (Abb. 5.1).

In grober Differenzierung unterscheidet man Aufgabentypen mit einem freien, einem gebundenen und einem atypischen Antwortformat (vgl. Lienert und Raatz 1998). Je nach Antwortlänge gliedern sich die Aufgaben mit dem freien Antwortformat in Kurzaufsatz- und Ergänzungsaufgaben. Die Aufgaben mit dem gebundenen Format gliedern sich in Ordnungsaufgaben, bei denen die einzelnen Bestandteile der Aufgabe umgeordnet oder einander so zugeordnet werden müssen, dass eine inhaltlich passende Ordnung entsteht (Zu- und Umordnungsaufgaben), in Auswahlaufgaben, bei denen die zutreffende Antwort aus mehreren Alternativen auszuwählen ist, sowie in Beurteilungsaufgaben, bei denen individuelle Einschätzungsurteile („Ratings“) zu den Inhalten im jeweiligen Itemstamm abgegeben werden müssen. Auswahlaufgaben mit zwei Antwortalternativen nennt man dichotome Aufgaben; im Falle mehrerer Antwortalternativen nennt man sie MehrfachwahlAufgaben (Multiple-Choice-Aufgaben). Bei Beurteilungsaufgaben unterscheidet man in Abhängigkeit vom Format der Beurteilungsskala zwischen kontinuierlichen Analogskala- und diskreten (geordnet kategorialen) Ratingskala-Aufgaben.

Bezüglich einer idealtypischen Kombination von Aufgabenstamm/Aufgabeninhalt und Aufgabentyp/Answerformat gibt es keine allgemeingültigen Regeln. Bei Leistungstests kann fast jeder Inhalt im Prinzip in jede Form gekleidet werden. Bei Persönlichkeitstests zählen dichotome Aufgaben und MehrfachwahlAufgaben, vor allem aber Beurteilungsaufgaben mit dem Ratingskalaformat zu den am häufigsten verwendeten Aufgabentypen. Zur Erfassung von Einstellungen und Meinungen wird meist nur das Ratingskalaformat herangezogen.

Im Folgenden werden die Charakteristika der vorgestellten Item-/Aufgabentypen detailliert beschrieben (Abb. 5.1).

Typen von Antwortformaten

Ordnungsaufgaben

AuswahlAufgaben

Beurteilungsaufgaben

Verwendung typischer

Antwortformate und Itemtypen in Leistungs- und Persönlichkeitstests

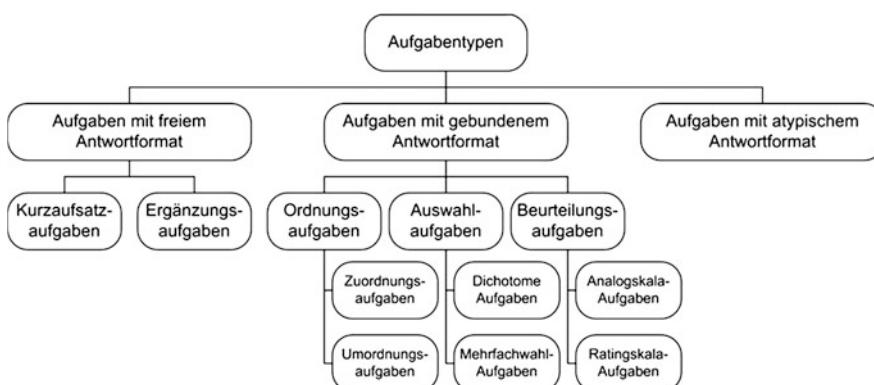


Abb. 5.1 Gliederungsschema der Aufgabentypen

Qualität der Kodierung beeinflusst Auswertungsobjektivität

Einsatzgebiete freier Antwortformate

Antwortlänge begrenzen

Kurzaufsatzaufgaben verlangen selbst erzeugte Antworten

5.2 Aufgaben mit freiem Antwortformat

Bei Aufgaben mit einem freien Antwortformat sind keine Antwortalternativen vorgegeben. Die Antworten werden von den Testpersonen selbst formuliert bzw. produziert. Dennoch sind die Antworten nicht völlig unstrukturiert, denn das Format, d. h. die Art, wie auf das Item geantwortet werden kann, ist in der Instruktion festgelegt. So muss z. B. eine Aussage gemacht, ein Text geschrieben oder eine Zeichnung angefertigt werden. Die besondere Herausforderung für die Testleitung bei der Auswertung besteht darin, dass die Antworten verschlüsselt werden müssen, indem man sie nach einem vorgefertigten Kategoriensystem „kodiert“, d. h. in einen numerischen Ausdruck übersetzt. Die Qualität dieses Kategoriensystems ist maßgebend für die Auswertungsobjektivität.

Das freie Aufgabenformat wird nicht nur für Erhebungen im schulisch-pädagogischen Bereich häufig verwendet, sondern z. B. auch für die Erfassung von Kreativität, bei der ein gebundener Antwortmodus per se keine kreativen Antworten zulassen würde (z. B. im Torrance Test of Creative Thinking; Torrance 1998; Torrance und Ball 1984; s. auch Pfiffer 2012). Des Weiteren findet das freie Antwortformat bei den „projektiven Verfahren“ Anwendung (s. ▶ Kap. 3, ▶ Abschn. 3.2.4). Das Format wird auch bevorzugt, wenn angenommen wird, dass die Reihenfolge der gegebenen Antwortteile eine wesentliche Informationsquelle darstellt, d. h. wenn anzunehmen ist, dass die Informationen, die von den Testpersonen zu Beginn genannt werden, wichtiger sind als die am Ende (z. B. bei der Frage: „Was ist Ihnen im Leben wichtig?“).

Die zumeist verwendeten Aufgabentypen mit freiem Antwortformat sind Kurzaufsatz- und Ergänzungsaufgaben.

5.2.1 Kurzaufsatzaufgaben

Bei Kurzaufsatzaufgaben werden die Testpersonen dazu angehalten, auf Fragen in Form von Kurzaufsätzen bzw. Essays zu antworten. Eine Antwort kann aber auch nur aus einem Satz oder aus einzelnen Wörtern bestehen. Sollen die Fragen mit mehreren Sätzen beantwortet werden, ist es sinnvoll, die Anzahl der Wörter auf höchstens 150 zu begrenzen (▶ Beispiel 5.1).

Beispiel 5.1: Kurzaufsatzaufgaben

- Geben Sie mit wenigen Worten die Bedingungen für die Entstehung einer affektiven Erkrankung an.*
- Geben Sie so viele kreative Ideen wie möglich an, was man mit einer Garnrolle und einem Nagel machen könnte.*
- „Was meinst du wohl, was der Junge oder das Mädchen auf dem Bild darauf antwortet? Schreib immer die erste Antwort, die Dir dazu einfällt, in das freigelassene Viereck.“*

■ Vorteile

Kurzaufsatzaufgaben erfordern in der Regel eine von der Testperson selbst erzeugte Antwort. Eine zufällig richtige Antwort wie bei Auswahlauflagen ist demnach fast nicht möglich. Bei der Erfassung von Merkmalen wie Kreativität oder stilistische Begabung und bei kognitiven Leistungen wie Leseverständnis oder Anwendung von Wissen stellt der Kurzaufsatz ein wichtiges Antwortformat dar. Im Persönlichkeitsbereich können Aufgaben dieses Typs vor allem zur Erfassung von qualitativen Motiven und Gründen verwendet werden. Am häufigsten sind sie in

5.2 · Aufgaben mit freiem Antwortformat

projektiven Testverfahren anzutreffen, bei denen es wichtig ist, nicht zwischen vorgefertigten Antworten wählen zu können.

■ Nachteile

Bei der Bearbeitung nehmen Kurzaufsatzaufgaben für die Testperson, vor allem aber für den Auswerter erheblich mehr Zeit in Anspruch als andere Aufgabentypen. Der hohe Zeitverbrauch bei der Bearbeitung, der große Auswertungsaufwand sowie die eingeschränkte Auswertungsobjektivität sind die Hauptnachteile dieses Antwortformats. Bei der Auswertung ist die Objektivität dabei umso eher beeinträchtigt, je länger und je komplexer eine Antwort sein darf. Testpersonen mit Formulierungsschwierigkeiten werden tendenziell benachteiligt. Problematisch an diesem Aufgabentyp ist auch die Mehrdeutigkeit der gegebenen Antworten, sodass sehr genaue Angaben des Testkonstrukteurs bezüglich einer angemessenen Kodierung der Antworten unerlässlich sind. (Ein Beispiel für ein solches, umfangreiches Kategoriensystem für den Rorschachtest findet man bei Exner 2010.) Fehlen relevante Antworten im Kategoriensystem zur Kodierung der Antworten, so ist die Auswertungsobjektivität beeinträchtigt.

**Hoher Aufwand in Durchführung und Auswertung;
Objektivitätsmängel**

5.2.2 Ergänzungsaufgaben

Ergänzungsaufgaben (*Completion Tests*, C-Tests) stellen den ältesten Aufgabentyp der Psychologie dar und werden bis heute im Leistungskontext häufig angewendet (s. Krampen 2015). Bei diesem Aufgabentyp kommt es darauf an, den Aufgabenstamm durch ein bestimmtes Wort („Schlüsselwort“) oder durch eine kurze Darstellung (ein Symbol, eine Zeichnung) sinnvoll zu ergänzen. Deshalb werden diese Aufgaben auch „Schlüsselwort-Ergänzungsaufgaben“ genannt. Man spricht von sog. „offenen Fragen“, wenn das Schlüsselwort am Ende von einzelnen Sätzen (Reihen) ausgelassen ist. Fehlt es in einem laufenden Text, spricht man von einem „Lückentext“ (► Beispiel 5.2).

Beispiel 5.2: Ergänzungsaufgaben

A. Offene Fragen:

- a. *Kenntnisse*: Kolumbus entdeckte Amerika im Jahr _____.

- b. *Oberbegriffe*: Specht und Ente sind _____.

- c. *Analogien*: Atheist verhält sich zu Religion wie Pazifist zu _____.

- d. *Folgen*: 2, 4, 8, _____.

(Lösung: a. 1492, b. Vögel, c. Krieg, d. 16)

B. Lückentexte:

- a. *Ergänzen Sie bei dem folgenden Text die fehlenden Wörter*:

Depression und _____ sind die dominierenden Emotionen bei _____. Störungen. Die meisten Menschen mit einer solchen Störung leiden ausschließlich an _____. Wenn beide Phasen mit jeweils einer dominierenden Emotion sich abwechseln, heißt dieses Muster _____.

(Lösung: Manie, affektiven, Depressionen, bipolar)

- b. *Ergänzen Sie bei dem folgenden Text die fehlenden Worthälften*:

Mit int_____ Validität ist die Eindeu_____ gemeint, mit der ein Untersuchungsergebnis inh_____ auf die Hy_____ bezogen werden kann.

(Lösung: -erner, -tigkeit, -altlich, -pothese)

Geringe Ratewahrscheinlichkeit**■ Vorteile**

Tests mit Ergänzungsaufgaben verlangen von Teilnehmerinnen und Teilnehmern eine Reproduktion gespeicherten Wissens und nicht nur eine Wiedererkennung. Somit ist die Wahrscheinlichkeit einer nur zufällig richtigen Beantwortung der Aufgaben sehr gering. Wenn eine standardisierte Instruktion und ein Auswertungsschema vorhanden sind, ist die Objektivität in der Regel gewährleistet, vor allem bei Leistungstests. Ergänzungsaufgaben können auch angezeigt sein, wenn nicht nur die Antwort selbst, sondern auch der Lösungsweg für komplexere Denkprobleme interessiert (z. B. das Lösen einer komplexen mathematischen Aufgabe).

Ergänzungsaufgaben erfassen nicht nur intendierte Inhalte**■ Nachteile**

Mit Ergänzungsaufgaben wird meist nur Faktenwissen geprüft. Die Auswertungsobjektivität kann eingeschränkt sein, wenn mehrere Begriffe als Antwort passen, aber nicht alle von der Testkonstrukteurin berücksichtigt wurden und als Richtigantworten zugelassen sind. Auch besteht die Gefahr von Suggestivwirkungen, wenn den Testpersonen durch die Art der Aufgabenformulierung eine bestimmte Antwort nahegelegt wird. Darüber hinaus besteht bei diesem Aufgabentyp die Gefahr, dass eher die allgemeine Intelligenz und die Lesefähigkeit erfasst werden und weniger die eigentlich interessierenden Merkmale. Des Weiteren erfordert sowohl die Bearbeitung als auch die Auswertung von Ergänzungsaufgaben mehr Zeit im Vergleich zu Aufgabentypen mit gebundenem Antwortformat.

5.3 Aufgaben mit gebundenem Antwortformat**Testperson muss aus vorgegebenen Antwortalternativen auswählen**

Aufgaben mit gebundenem Antwortformat sind dadurch gekennzeichnet, dass mehrere Alternativen für mögliche Antworten vorgefertigt sind. Die Testpersonen sind in ihren Antworten nicht frei, sondern an die vorgegebenen Antwortalternativen „gebunden“, aus denen die passende der möglichen Antwortalternativen auszuwählen ist. Hierbei ist darauf zu achten, dass die Antwortalternativen exhaustiv/erschöpfend sind, damit die Merkmalsausprägungen aller Personen abgebildet werden können.

Antworterfassung ökonomisch und eindeutig**■ Vorteile**

Aufgaben mit gebundenem Antwortformat erfordern keinen hohen Zeitaufwand aufseiten der Testpersonen und können sehr ökonomisch ausgewertet werden, da die Erfassung der gegebenen Antworten mit Schablonen bzw. Scanner erleichtert werden kann. Weiterhin ist eine unmittelbare Antwortheingabe in den Computer durch die Testperson möglich (s. computeradministrierte Tests in ► Kap. 3, ► Abschn. 3.5). Beim gebundenen Antwortformat ist die Kodierung und Beurteilung der Antworten eindeutig, da jeder vorgegebenen Antwortkategorie eine Zahl zugewiesen werden kann. Dadurch ist bei Aufgaben mit gebundenem Antwortformat die Auswertungsobjektivität zumeist deutlich höher als die von Aufgaben mit freiem Antwortformat.

■ Nachteile

Die Antwortalternativen können von den Testpersonen als zu „eng“ oder zu „knapp“ empfunden werden in dem Sinne, dass keine wirklich passende Antwortalternative gefunden wird. Dieser Nachteil kann durch sog. „Forced-Choice-Formate“ umgangen werden, bei denen die am ehesten passende Antwort gewählt werden soll (► Abschn. 5.3.2.4).

5.3.1 Ordnungsaufgaben

Ordnungsaufgaben werden bearbeitet, indem die einzelnen Bestandteile der Aufgabe umgeordnet oder einander so zugeordnet werden müssen, dass eine logisch-inhaltlich passende Ordnung entsteht. Es wird zwischen Zuordnungs- und Umordnungsaufgaben unterschieden.

Bei *Zuordnungsaufgaben* besteht die Anforderung an die Testteilnehmerinnen und Teilnehmer, eine richtige Zuordnung von jeweils zwei Elementen – Wörtern, Zahlen, Zeichnungen etc. – zueinander vorzunehmen. Da die Antwortalternativen voneinander nicht unabhängig sind, nimmt mit jeder richtigen Zuordnung die Anzahl der verbleibenden Antwortalternativen ab, wodurch die Wahrscheinlichkeit einer nur zufallsbedingt richtigen Lösung für die verbliebenen Zuordnungen steigt. Um dieses Problem zu umgehen, ist es daher grundsätzlich ratsam, die Anzahl der Antwortalternativen größer zu wählen als die Zahl der Aufgabestellungen (► Beispiel 5.3), damit auch bei der letzten Zuordnung noch mehreren Antwortalternativen zur Auswahl stehen und als Distraktoren fungieren können, also als Antwortalternativen, die zwar wie richtige Antworten aussehen, aber inhaltlich falsch sind (zu Distraktoren siehe auch ► Abschn. 5.3.2.1).

Zuordnungsaufgaben

Beispiel 5.3: Zuordnungsaufgaben

Ordnen Sie jedem Land die entsprechende Hauptstadt zu.

Land	Hauptstadt					
	a) Kuala Lumpur	b) Quito	c) Montevideo	d) Vientiane	e) Lima	f) Jakarta
1) Peru	a	b	c	d	<input checked="" type="radio"/> e	f
2) Laos	a	b	c	<input checked="" type="radio"/> d	e	f
3) Indonesien	a	b	c	d	e	<input checked="" type="radio"/> f
4) Ecuador	a	<input checked="" type="radio"/> b	c	d	e	f
5) Malaysia	<input checked="" type="radio"/> a	b	c	d	e	f

■ Vorteile

Zuordnungsaufgaben sind einfach, ökonomisch und objektiv. Eine große Anzahl von Aufgaben kann platzsparend auf einer kleinen Fläche untergebracht werden. Deshalb sind sie besonders für Wissens- und Kenntnisprüfungen gut geeignet. Eine nur zufällig richtige Beantwortung stellt bei Zuordnungsaufgaben ein geringes Problem dar, da die Ratewahrscheinlichkeit niedrig ist, sofern die Zahl der Antwortalternativen die Zahl der Fragen übersteigt.

**Ratewahrscheinlichkeit gering,
wenn mehrere Distraktoren
vorhanden sind**

■ Nachteile

Bei diesem Antworttyp ist keine Reproduktionsleistung, sondern lediglich eine Wiedererkennungsleistung (Rekognition) erforderlich. Bei der Konstruktion der unrichtigen Antwortalternativen (Distraktoren) ist ein erheblicher Konstruktionsaufwand notwendig, um (bei Unkenntnis der richtigen Antwort) für die zufällige Wahl der Richtigantwort bzw. der Distraktoren gleich hohe Auswahlwahrscheinlichkeiten zu gewährleisten (► Abschn. 5.3.2.1). Weiterhin kann die Wahrscheinlichkeit einer richtigen Zuordnung durch bereits getätigte Zuordnungen von Antwortalternativen beeinträchtigt sein, z. B. wenn eine der Antwortalternativen fälschlich zugeordnet wurde und somit als Richtigantwort nicht mehr zur Verfügung steht. Das Problem ist stärker ausgeprägt, wenn keine zusätzlichen Dis-

**Zuordnungsaufgaben erfordern nur
Wiedererkennungsleistung**

Umordnungsaufgaben erfordern das Auffinden einer sinnvollen Reihung

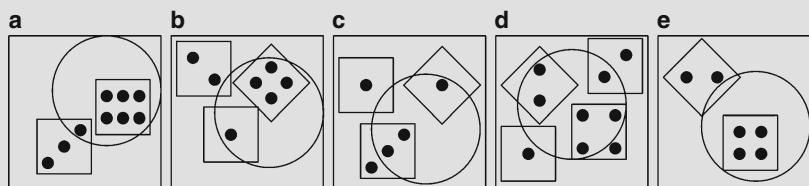
traktoren vorhanden sind. Jedenfalls sollte eine solche, durch vorherige Antworten veränderte Wahrscheinlichkeit einer richtigen Antwort bei der Auswertung bedacht werden. (Elaborierte Testmodelle der Item-Response-Theorie [IRT] sind in der Lage, die skizzierte Problematik geeignet zu berücksichtigen, s. z. B. ▶ Kap. 18.)

Umordnungsaufgaben verlangen als Bearbeitung das Umsortieren von Worten, Satzteilen, Zahlen, Bildern oder Gegenständen. Die Testpersonen sind angehalten, die einzelnen Teile der Aufgabe in eine logisch-sinnvolle Reihenfolge zu bringen. Die Schwierigkeit der Items hängt u. a. von der Anzahl der umzuordnenden Teile ab. Es sollte aber beachtet werden, dass mit einer steigenden Anzahl von Teilen auch ein erhöhter Konstruktionsaufwand einhergeht, um eine Eindeutigkeit der Reihenfolge der Teile zu gewährleisten. Sofern alternative richtige Reihungen existieren, die nicht vom Testkonstrukteur vorgesehen waren, kann die Auswertungsobjektivität erheblich beeinträchtigt sein.

Häufig wird dieser Aufgabentyp bei sog. „Materialbearbeitungstests“ angewendet. Typische Beispiele für dieses Antwortformat stellen Testaufgaben im Hamburg-Wechsler-Intelligenz-Test für Kinder (HAWIK; Hardesty und Priester 1963) bzw. in der aktuellen Version des Intelligenztests für Kinder (Wechsler Intelligence Scale for Children – Fourth Edition, WISC-IV; Petermann und Petermann 2011) dar, in dem Bildertafeln so angeordnet werden müssen, dass sie eine sinnvolle Geschichte ergeben; auch nach einer logischen Reihung kann gefragt werden (► Beispiel 5.4).

Beispiel 5.4: Umordnungsaufgaben

Ordne die Bilder so, dass sie eine logische Reihe ergeben.



(Lösung: c, e, b, d, a)

Postkorbaufgaben

Einen anderen Typus von Umordnungsaufgaben stellen sog. „Postkorbaufgaben“ in Assessment-Centern (AC) dar, die zur Personalauswahl dienen. In diesen Aufgaben werden die Testpersonen aufgefordert, eine Reihe von ungeordneten Schriftstücken verschiedenen Inhalts („Postkorb“) in eine sinnvolle Reihenfolge zu bringen, z. B. anhand der Wichtigkeit und Dringlichkeit. (Einschränkend ist zu erwähnen, dass bei solchen AC-Aufgaben die psychometrischen Eigenschaften nicht immer feststellbar sind und bisher nur selten überprüft wurden; Höft und Funke 2006.)

■ Vorteile

Die Umordnung von Bildmaterial ist besonders in den Fällen gut geeignet, in denen die Gefahr besteht, dass die Testergebnisse z. B. durch mangelnde Lesefähigkeit beeinträchtigt sein könnten (z. B. bei Kindern). Im Leistungsbereich sind Umordnungsaufgaben günstig, um z. B. schlussfolgerndes Denken, Ursache-Wirkungs-Zusammenhänge oder Abstraktionsfähigkeiten zu überprüfen – mit dem Vorteil, nicht auf verbales Material beschränkt zu sein.

■ Nachteile

Materialbearbeitungstests sind vor allem durch einen hohen Materialverbrauch gekennzeichnet und für Gruppentestungen nur bedingt anwendbar; bei einer Testadministration am Computer und in einem PC-Pool entfallen diese Einschränkungen.

Geringe Abhängigkeit von Lesefähigkeit

Hoher Materialverbrauch

5.3.2 Auswahlaufgaben

Im Gegensatz zu Ordnungsaufgaben werden die Testpersonen bei Auswahlaufgaben aufgefordert, aus mehreren vorgegebenen Antwortalternativen die richtige/-n bzw. zutreffende/-n Antwort/-en zu identifizieren. Im Folgenden werden wichtige Aspekte für die Konstruktion von Auswahlaufgaben vorgestellt: Bei Leistungstests sind die wichtigsten Aspekte die Wahl *geeigneter Distraktoren* sowie die *Disjunktivität* der Antwortalternativen; zudem muss eine Entscheidung über die Anzahl der Antwortalternativen sowie über die Anzahl der „richtigen“ Antwortalternativen getroffen werden. Bei Persönlichkeitstests kommt der *Exhaustivität* der Antwortmöglichkeiten eine besondere Bedeutung zu.

5.3.2.1 Konstruktion geeigneter Distraktoren

Im psychologischen und pädagogischen Leistungskontext gehören Auswahlaufgaben zu den am häufigsten verwendeten Aufgaben. Anders als beim freien Antwortformat ist das Auffinden der richtigen Lösung allein durch das Wiedererkennen (Rekognition) möglich. Um die Identifikation der richtigen Antwort zu erschweren, müssen weitere Antwortalternativen in der Weise konstruiert werden, dass sie zwar wie richtige Antworten aussehen, aber inhaltlich falsch sind. Solchermaßen falsche Antwortalternativen werden *Distraktoren* (lat. „*distrahere*“ = auseinanderziehen, ablenken) genannt und sollen unwissende Testpersonen auf falsche Fährten führen.

Bei der Suche nach der Richtigantwort werden von der Testperson in der Regel alle Antwortalternativen dahingehend geprüft, ob sie eine passende Antwort darstellen können. Je mehr Distraktoren vorgegeben werden, desto mehr wird das Auswahlverhalten auf die verschiedenen Antwortalternativen auseinandergesogen und desto geringer ist die Wahrscheinlichkeit für das rein zufällige Auffinden der richtigen Lösung. Doch nur, wenn die Distraktoren der Richtigantwort stark ähneln, weist der Auswahlprozess für Testpersonen, die nicht über die spezifischen Kenntnisse zur Auswahl der Richtigantwort verfügen, die notwendige Schwierigkeit auf; anderenfalls verfehlten die Distraktoren ihren Zweck, eine niedrige Ratewahrscheinlichkeit zu gewährleisten. Deshalb sind für die Konstruktion geeigneter Distraktoren die Auswahlwahrscheinlichkeiten („Attraktivität“), die Ähnlichkeit mit der richtigen Antwortalternative sowie die Plausibilität von größter Bedeutung. Um plausible Distraktoren zu finden, können z. B. die in Ergänzungsaufgaben (► Abschn. 5.2.2) häufig genannten Falschantworten herangezogen werden.

Eine sorgfältige Analyse der Distraktoren hinsichtlich ihrer Auswahlwahrscheinlichkeit stellt insoweit einen sehr wichtigen Schritt dar, als die Qualität der Distraktoren essentiell für die Qualität des gesamten (Leistungs-)Tests ist. Minderwertige Distraktoren, die zu geringe Auswahlwahrscheinlichkeiten aufweisen, verletzen nämlich die Annahmen der psychometrischen Testmodelle (s. z. B. IRT-Modelle, ► Kap. 16) und können u. a. zu Problemen bei der Testwertinterpretation führen.

Minderwertige Distraktoren können auch ohne spezifisches inhaltliches Wissen beim Auswahlprozess häufig daran erkannt werden, dass sie

- sprachlich weniger sorgfältig ausgearbeitet sind als die Richtigantwort,
- grammatisch nicht exakt zum Aufgabenstamm passen oder
- (etwa bei Matrizenaufgaben) Muster oder Formen verwenden, die in der Aufgabenstellung gar nicht vorkommen (Genaueres siehe Eid und Schmidt 2014, S. 107 f.).

Durch solche und ähnliche Strategien büßen minderwertige Distraktoren ihre Funktion ein und erhöhen die Wahrscheinlichkeit, dass eine Aufgabestellung von den Testpersonen nicht nur durch spezifisches inhaltliches Wissen, sondern auch durch geschicktes Ausnutzen anderer Hinweise richtig gelöst werden kann.

Wichtige Aspekte bei Auswahlaufgaben

Distraktoren sollen unwissende Testpersonen auf falsche Fährten führen

Gute Distraktoren bewirken eine niedrige Ratewahrscheinlichkeit

Wichtigkeit der Distraktorenanalyse

Das ► Beispiel 5.5 enthält Varianten von sorgfältig konstruierten Distraktoren. Die Anzahl der Antwortalternativen im Fall A beträgt 8, im Fall B und C jeweils 5.

Beispiel 5.5: Antwortalternativen in Auswahlaufgaben

- A. Wählen Sie aus den Antwortalternativen diejenige, die das Muster im Bild am besten ergänzt.

Übungsbeispiel 2

Welche Figur entspricht der geltenden Regel?

Zurück Ich weiß die Antwort nicht Weiter >

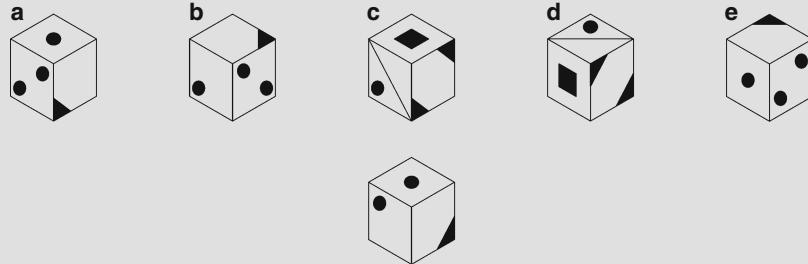
Beispielitem aus dem Adaptiven Matrizen Test (AMT). (Aus Hornke et al. 2005, mit freundlicher Genehmigung des Schuhfried Verlages)
(Lösung: 2. von rechts)

- B. Welche der folgenden Buchstabengruppen folgt nicht der Logik der übrigen?

- a. TUVW
- b. ABCD
- c. FGHJ
- d. PQRS
- e. JKLM

(Lösung: c)

- C. Der Würfel zeigt einen der vorgegebenen Würfel in veränderter Lage. Sie sollen herausfinden, um welchen der vorgegebenen Würfel es sich handelt. Der Würfel kann gedreht, gekippt oder gedreht und gekippt worden sein.



Beispielitem aus dem Intelligenz-Struktur-Test (I-S-T) 2000R. (Aus Amthauer et al. 2001, © by Hogrefe GmbH & Co. KG, Göttingen ● Nachdruck und jegliche Art der Vervielfältigung verboten. Bezugssquelle des Intelligenz-Struktur-Test 2000R (I-S-T 2000R): Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, www.testzentrale.de)
(Lösung: b)

5.3.2.2 Disjunktheit der Antwortalternativen

Bei der Erstellung der Antwortalternativen ist des Weiteren darauf zu achten, dass diese disjunkt sind, d. h. dass sie sich gegenseitig ausschließen und die Schnittmenge ihrer Aussagen leer ist. ► Beispiel 5.6 enthält Items, die das Gebot der Disjunktheit *nicht* erfüllen, was zur Folge hat, dass keine eindeutige Richtigantwort existiert: Im Fall A sind mehrere Antworten richtig, weil die Antworten b und c nicht disjunkt sind, sondern sich an den Kategoriengrenzen überlappen. Im Fall B weisen die Kategorien b und c eine nichtleere Schnittmenge auf, denn bei verschiedenen Sorten von Obst handelt es sich auch um Baumfrüchte.

Man muss also darauf achten, dass genau eine richtige Antwort existiert und nicht mehrere.

Eindeutigkeit der Richtigantwort

Beispiel 5.6: Nicht disjunkte Antwortalternativen

- A. Wie groß ist $1/5$ von 10 ?
 - a. kleiner als 1.5
 - b. 1.5 bis 2.0
 - c. 2.0 bis 2.5
 - d. größer als 2.5

- B. Welcher ist der Oberbegriff für Äpfel?
 - a. Gemüse
 - b. Obst
 - c. Baumfrüchte
 - d. Strauchfrüchte

Konsequenz: Bei beiden Aufgaben wären sowohl Antwort b als auch c richtig.

5.3.2.3 Exhaustivität der Antwortmöglichkeiten

Insbesondere für Persönlichkeits- und Einstellungstests ohne Richtigantwort muss bei der Konstruktion der Antwortalternativen auf die Exhaustivität der Antwortmöglichkeiten („Ausschöpfung“ aller möglichen Antworten, Vollständigkeit) geachtet werden. Exhaustivität ist gegeben, wenn alle Testpersonen unter den Antwortalternativen eine passende Antwort finden können. Ein nicht exhaustives Beispiel aus dem Persönlichkeitsbereich ist im Folgenden aufgeführt (► Beispiel 5.7).

Bei Persönlichkeitstests ist die Exhaustivität der Antwortalternativen wichtig

Beispiel 5.7: Nicht exhaustive Antwortalternativen in Persönlichkeitstests

Wählen Sie die Antwortalternative aus, die für Sie zutreffend ist:

- a. Ich bevorzuge harte, realistische Actionthriller.
- b. Ich bevorzuge gefühlvolle, feinsinnige Filme.

Konsequenz: Das Item ist ggf. nicht beantwortbar, wenn keine der beiden Antwortalternativen für die Testperson zutreffend ist. Eine tendenzielle Antwort kann allerdings mittels der „Forced-Choice-Instruktion“ erzielt werden (► Beispiel 5.10).

Bei Leistungstests ist Exhaustivität der Antwortalternativen nicht notwendig

Im Gegensatz zu Persönlichkeitstests ist bei Leistungstests (► Beispiel 5.8, Fall A und B) die Erfüllung der Exhaustivität nicht notwendig, sofern die Richtigantwort in der Menge der Antwortalternativen enthalten ist. (Die Exhaustivität wäre allerdings auch nicht erzielbar, weil die Menge falscher Antwortalternativen unendlich ist.)

Beispiel 5.8: Nicht exhaustive Antwortalternativen in Leistungstests

A. Zwischen dem ersten und zweiten Wort besteht eine ähnliche Beziehung wie zwischen dem dritten und einem der fünf zur Wahl stehenden Begriffe. Wählen Sie das Wort, das jeweils am besten passt:

Die Beziehung zwischen Hund und Welpe ist ähnlich wie zwischen Schwein und _____.

- a. Eber
- b. Ferkel
- c. Sau
- d. Lamm
- e. Nutztier

(Lösung: b)

B. Kreuzen Sie diejenige Lösung an, die Ihren Berechnungen nach richtig ist.

$$\frac{1}{3a} - \frac{1}{2a} + \frac{1}{a} = ?$$

- a. $\frac{1}{2a}$
- b. $\frac{5}{6a}$
- c. $\frac{1}{6a}$
- d. $\frac{5}{2a}$

(Lösung: b)

Konsequenz: Obwohl beliebig viele weitere Distraktoren denkbar sind, ist das Item beantwortbar, da die richtige Antwort in der Menge der Antwortalternativen enthalten ist.

5.3.2.4 Anzahl der Antwortalternativen

Je nach Anzahl der vorgesehenen Antwortalternativen spricht man von „dichotomen Aufgaben“, wenn die Testpersonen zwischen zwei Antwortalternativen wählen können; bei mehr als zwei Alternativen spricht man von „Mehrzahlwahlaufgaben“. Wie viele Antwortalternativen in einem Test sinnvoll sind, hängt vom gewählten Gegenstandsbereich ab. Neuere Forschungen weisen darauf hin, dass nicht eine möglichst hohe Anzahl von Antwortalternativen zu besonders günstigen Itemeigenschaften führt, sondern dass eine Begrenzung auf drei – allerdings sehr sorgfältig konstruierte – Antwortalternativen häufig ausreichend sein kann (Rodríguez 2005; s. auch Eid und Schmidt 2014, S. 107). Hierbei sollte darauf geachtet werden, dass die räumliche Positionierung der Richtigantwort und der jeweiligen Distraktoren auf dem Testbogen randomisiert erfolgt; man sollte die richtige Antwort also nicht immer z. B. an Position zwei platzieren.

Bei *dichotomen Aufgaben* werden zwei Antwortalternativen a und b angeboten (► Beispiel 5.9, Fall A). Oft wird anstelle der zwei Alternativen nur ein Statement angeboten, das dann mit „Ja/Nein“, „Stimmt/Stimmt nicht“ oder mit „Richtig/Falsch“ zu beantworten ist, weshalb auch die Bezeichnung Richtig-Falsch-Aufgaben gebräuchlich ist (► Beispiel 5.9, Fall B). Auch komplexere Problemstellungen können als dichotome Aufgaben formuliert werden, wie die Fälle C und D in ► Beispiel 5.9 zeigen; im Fall D bestehen die beiden Antwortalternativen aus

- a. Markierung des „Zielitems“ durch einen Zacken und
- b. Markierung der „Nichtzielitems“ durch eine Linie unter den Zeichen.

► Beispiel 5.9, Fall E zeigt zuletzt eine klassische Testlet-Aufgabe, bei der mehrere Items mit dichotomem Antwortformat auf einem gemeinsamen Bezugssatz aufgebaut. Hier beziehen sich alle vier dichotomen Items auf einen gemeinsamen Inhalt (hier: Satz D „Der Kant'sche Imperativ ist eine allgemeingültige sittliche Norm“). Die Wahrscheinlichkeiten, dass eine Testperson die vier Items korrekt löst, sind in diesem Fall nicht unabhängig voneinander, da sie allesamt ein Verständnis des Satzes D voraussetzen sowie auf das Konzept des Widerspruchs Bezug nehmen (man weiß entweder, was ein Widerspruch ist oder eben nicht). Testlets in diesem

Sinnvolle Begrenzung der Alternativen

Dichotome Aufgaben

5.3 · Aufgaben mit gebundenem Antwortformat

Stil werden z. B. häufig in der PISA-Studie verwendet (OECD 2014). Entsprechende testtheoretische Überlegungen und psychometrische Modelle finden sich in ► Kap. 18.

Beispiel 5.9: Dichotome Aufgaben

- A. Sir Karl Raimund Popper war
 - a. der Begründer des kritischen Rationalismus.
 - b. der Begründer des Neo-Positivismus.
 (Lösung: a)
- B. Fallschirmspringen würde ich gerne ausprobieren.
 - a. Stimmt
 - b. Stimmt nicht
- C. Das Verhältnis von Vegetariern zu Nichtvegetariern betrage in Deutschland 10 zu 90 %. In einer ökologisch orientierten Partei sind von 480 Mitgliedern 169 Vegetarier. Weicht die Häufigkeit der Vegetarier in der Partei von der Häufigkeit der Vegetarier in Deutschland ab?
 - a. Ja
 - b. Nein
 (Lösung: a)
- D. Ihre Aufgabe wird darin bestehen, in einer Liste von runden Zeichen jene zu finden, welche innen entweder einen

»Kreis mit 3 Punkten«  bzw. 
 oder ein »Quadrat mit 2 Punkten«  bzw.  zeigen.

Die Bearbeitung des Testbogens geschieht folgendermaßen: Sie beginnen am linken Blattrand bei dem angedeuteten Stift und ziehen eine Linie unter den Zeichen nach rechts. Immer wenn Sie einen „Kreis mit 3 Punkten“ oder ein „Quadrat mit 2 Punkten“ finden, ziehen Sie von unten einen Zacken in das Zeichen hinein, unter den anderen Zeichen ziehen Sie die Linie einfach vorbei. Die Linie ist genauso wichtig wie die Zacken.

Eine richtig bearbeitete Zeile sollte etwa so aussehen:



Instruktion aus dem Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2; Moosbrugger und Oehlschlägel 2011, mit freundlicher Genehmigung von Hogrefe)

- E. Welcher/welche der untergeordneten Sätze (1, 2, 3, 4) steht/stehen zu dem übergeordneten Satz D im Widerspruch?

Satz D: Der Kant'sche Imperativ ist eine allgemeingültige sittliche Norm.

1. Nicht alle Menschen halten sich an den Kant'schen Imperativ.
Widerspruch – Kein Widerspruch
2. Der Kant'sche Imperativ ist für den gläubigen Christen unmaßgeblich.
Widerspruch – Kein Widerspruch
3. Sittliche Normen kann es nicht geben, da jeder Mensch nur seinem Gewissen verantwortlich ist.
Widerspruch – Kein Widerspruch
4. Der Kant'sche Imperativ ist den meisten Menschen unbekannt.
Widerspruch – Kein Widerspruch

Beispiel nach Lienert und Raatz (1998, S. 23)

(Lösung: Widerspruch in 2 und 3; kein Widerspruch in 1 und 4)

**Einfach und ökonomisch
in Instruktion, Durchführung
und Auswertung**

**Ratewahrscheinlichkeit
beträgt 50 %**

**Mehrfachwahl-Aufgaben
(Multiple-Choice-Aufgaben)
mit einer Richtigantwort**

**Mehrfachwahlaufgaben
mit variabler Anzahl
von Richtigantworten**

**Kodierung bei mehreren
Richtigantworten**

■ **Vorteile**

Aufgaben mit einem dichotomen Antwortformat sind einfach und ökonomisch, sowohl in der Instruktion als auch in der Bearbeitung und Auswertung. Die Lösungszeit der einzelnen Items ist relativ kurz, da der Testperson nur zwei Antwortalternativen vorgegeben werden, aus denen sie wählen kann. Gleichzeitig erlauben dichotome Aufgaben eine große Variabilität der möglichen Aufgabenstellungen, wie das ► Beispiel 5.9 zeigt.

■ **Nachteile**

Der größte Nachteil dichotomer Items besteht bei Leistungstests in einer 50%igen Ratewahrscheinlichkeit für zufällig richtige Lösungen. Darüber hinaus wird lediglich die Wiedererkennungsleistung abgefragt. Die beiden Eigenschaften zusammenommen haben zur Folge, dass dichotome Aufgaben im Leistungskontext nicht sehr häufig zur Anwendung kommen.

Bei Persönlichkeitstests stellt sich die Problematik zufällig richtiger Lösungen naturgemäß nicht. Jedoch berichtet Krosnick (1999, S. 552), dass es bei Ja-Nein-Aufgaben Hinweise auf eine erhöhte Akquieszenz (Zustimmungstendenz, s. hierzu ► Kap. 4, ► Abschn. 4.7.3) gibt. Ein weiterer Nachteil besteht darin, dass es oft Schwierigkeiten bereitet, den Aufgabenstamm so zu formulieren, dass er mit „Ja“ oder „Nein“ beantwortet werden kann. Bei komplexen Aufgabeninhalten sollte deshalb eher ein anderes Antwortformat bevorzugt werden.

Für Aufgaben, bei denen mehr als zwei Antwortalternativen vorgegeben sind, wird die Bezeichnung *Mehrfachwahl- oder Multiple-Choice-Aufgabe* verwendet. Von den Alternativen ist in Leistungstests diejenige auszuwählen, die richtig ist, und in Persönlichkeitstests diejenige, die individuell zutrifft. Häufig hat sich gezeigt, dass schon drei Antwortalternativen in einem Test zu sehr guten psychometrischen Eigenschaften des Tests führen (Haladyna und Downing 1993; Lord 1944, 1977; Rodriguez 2005; Tversky 1964).

Über die „klassischen“ Mehrfachwahl-Aufgaben hinaus, bei denen lediglich eine Lösung richtig ist, sind in Leistungstests auch Antwortformate vorzufinden, bei denen mehrere (oder gar keine der) Antwortalternativen zutreffen; die zutreffenden Alternativen sollen herausgefunden und angekreuzt werden (*Multiple Mark Questions*; Cronbach 1941). Hierbei existieren zwei mögliche Instruktionen. In der einen Instruktion wird den Testpersonen die Anzahl der richtigen Antwortalternativen mitgeteilt, die Aufgabe besteht entsprechend in der Identifizierung der richtigen Antwortalternativen. In der anderen Instruktion muss die Testperson selbst entscheiden, wie viele Antworten sie für richtig hält (*Pick-any-out-of-n-Format*; vgl. Rost 2004), wodurch die Anforderung an die Testperson beträchtlich erhöht wird. Beide Instruktionen verringern die Ratewahrscheinlichkeit, d. h. die Wahrscheinlichkeit, durch Zufall die richtige(n) Lösung(en) zu wählen. Wenn im Test Items enthalten sind, bei denen keine der dargebotenen Antwortalternativen richtig ist, so sollte in der Instruktion explizit auf diesen Sachverhalt hingewiesen werden, und zwar möglichst auch mit einem Musteritem, damit die Variante „Keine Antwortalternative ist richtig“ auch ernst genommen wird (vgl. Eid und Schmidt 2014).

Ein wesentlicher Nachteil bei der Auswertung von Aufgaben, bei denen mehrere Antwortalternativen richtig sind, besteht darin, dass die Kodierung der Richtigantworten deutlich schwieriger ist. Der Testwert einer Person hängt dann nämlich davon ab, ob ausschließlich korrekte Antworten gezählt werden oder ob auch falsche Antworten Berücksichtigung finden (beispielsweise könnte eine Person immer alle Antwortalternativen ankreuzen und würde somit immer auch eine oder mehrere richtige Antwortalternativen korrekt ankreuzen). Untersuchungen hierzu haben gezeigt, dass eine Gewichtung der richtig und falsch angekreuzten Antwortalternativen zu psychometrisch günstigen Eigenschaften eines Tests führen (sog. *Partial Credit Scoring*; Bauer et al. 2011).

5.3 · Aufgaben mit gebundenem Antwortformat

Bei der Konstruktion der Antwortalternativen von Multiple-Choice-Aufgaben im Persönlichkeits- und Einstellungsbereich ist die Beachtung der Exhaustivität der Alternativen (► Abschn. 5.3.2.3) besonders wesentlich. Sofern keine Exhaustivität vorliegt, kann mit einer entsprechenden Instruktion veranlasst werden, dass sich die Testpersonen für die am ehesten auf sie zutreffende Antwortalternative entscheiden, auch wenn keine der Optionen für die Testperson genau passt. Solche Antwortformate werden „Forced-Choice-Formate“ genannt und z. B. bei Interessentests benutzt (► Beispiel 5.10).

Forced-Choice-Format

Beispiel 5.10: Forced-Choice-Fragetechnik

Was würden Sie am liebsten am kommenden Wochenende tun? Wählen Sie bitte die Antwortalternative, die am ehesten auf Sie zutrifft.

- a. Sich über die neuesten Ereignisse in der Wirtschaft informieren
- b. Einen Spaziergang in der Natur unternehmen
- c. Eine Kunstausstellung besuchen
- d. Sich in ein neues Computerprogramm einarbeiten

■ Vorteile

Mehrfachwahl-Aufgaben sind in Bezug auf die Durchführung und Auswertung fast ebenso einfach, ökonomisch und objektiv wie dichotome Aufgaben; bei Leistungstests ist die Ratewahrscheinlichkeit durch die erhöhte Anzahl der Antwortalternativen stark verringert. Sofern die Testpersonen mehrere Antwortalternativen als Richtigantworten identifizieren müssen, sinkt die Ratewahrscheinlichkeit nochmals beträchtlich.

**Geringere Ratewahrscheinlichkeit
als bei dichotomen Auswahlaufgaben**

■ Nachteile

Das erfolgreiche Bearbeiten von Aufgaben dieses Typs setzt lediglich eine Rekognitionsleistung (Wiedererkennen) voraus. Deshalb ist das Format nicht für alle Merkmale sinnvoll, beispielsweise für Kreativität. Wenn Mängel in den Distraktoren vorliegen (z. B. bei ungleich attraktiven Antwortalternativen) oder wenn die Aufgabe selbst Hinweise auf die Problemlösung beinhaltet, können Verzerrungen auftreten, die bei der Anwendung eines psychometrischen Modells beachtet werden müssen (► Kap. 18).

**Mängel bei der Distraktor-
konstruktion verursachen
Verzerrungen**

5.3.3 Beurteilungsaufgaben

In Persönlichkeitstests wird den Testpersonen häufig als Aufgabenstamm ein Statement/eine Aussage zur Beurteilung vorgelegt, wobei der Grad der Zustimmung oder Ablehnung als Indikator für die Ausprägung des interessierenden Persönlichkeitsmerkmals herangezogen wird. Solche Aufgaben werden als Beurteilungsaufgaben bezeichnet. Innerhalb der Gruppe der Beurteilungsaufgaben ist hinsichtlich des Antwortformats eine Reihe von Differenzierungsgesichtspunkten zu beachten. Bei der Konstruktion der Antwortskala sind jedenfalls acht Aspekte zu berücksichtigen.

**Aspekte bei der Konstruktion
von Beurteilungsaufgaben**

5.3.3.1 Kontinuierliche vs. diskrete Beurteilungsskalen

Als Antwortformat findet häufig eine kontinuierliche Analogskala oder eine diskret gestufte Ratingskala Verwendung, wobei Letztere verschiedentlich, wenn auch uneinheitlich und uneindeutig als „Likert-Skala“ bezeichnet wird (s. Eid und Schmidt 2014, S. 117 f.).

**Unterscheidung von Analogskala
und Ratingskala**

Testaufgaben, in denen als Antwortformat eine Ratingskala mit mehr als zwei graduell abgestuften Beurteilungskategorien benutzt werden, bezeichnet man auch

Visuelle Analogskala bietet Beurteilungskontinuum

Kein Informationsgewinn bei extrem vielen Skalenpunkten

Fünf bis sieben Skalenstufen sind optimal

Differenzierungsfähigkeit der Testpersonen berücksichtigen

Polarität der Antwortskala

als Stufenantwortaufgaben. Die Aufgabe der Testpersonen besteht darin, die für sie am besten passende oder zutreffende Antwortkategorie zu markieren; richtige oder falsche Antworten gibt es bei Persönlichkeitstests nicht. Jede Antwort wird gemäß einem zuvor festgelegten Punkteschlüssel gewichtet, sodass in der Regel für jede Antwort eine entsprechende Punktzahl vergeben wird. Charakteristisch ist, dass die Antwortkategorien meist nicht aufgabenspezifisch formuliert sind, sondern in einheitlicher Form für den gesamten Test gelten. Im Unterschied zur Ratingskala weisen kontinuierliche Analogskalen keine diskreten Abstufungen auf und ermöglichen somit zumindest theoretisch eine besonders feine Differenzierung der Beurteilung. Für welches Antwortformat (Rating- bzw. Analogskala) man sich entscheidet, hängt von der Zielsetzung der Messung und von der Art der Datenerfassung ab.

5.3.3.2 Verwendung von Skalenstufen

Dieser Aspekt betrifft die Differenziertheit einer Beurteilungsskala. Prinzipiell besteht die Möglichkeit, Skalenstufen zu verwenden oder nicht.

Ein Beispiel für eine *kontinuierliche Skala* ohne konkrete Skalenstufen stellt die visuelle Analogskala dar (► Beispiel 5.11, Fall A). Die Testpersonen werden angehalten, das Item zu beantworten, indem sie an der für sie zutreffenden Stelle der Skala eine Markierung setzen. Bei einer computerunterstützten (Online-)Implementierung kann die Beurteilung direkt am Bildschirm vorgenommen werden. Kontinuierliche Skalen sind im Vormarsch, da die seinerzeitigen Verrechnungsnachteile gegenüber ganzzahlig gestuften Skalenpunkten im Computerzeitalter nicht mehr zum Tragen kommen. Dennoch werden Analogskalen eher selten verwendet, da die Differenziertheit der Messung in der Regel nicht dem Differenzierungsvermögen der beurteilenden Testpersonen entspricht.

Sofern man sich für die konkrete Angabe von abgestuften Skalenpunkten entscheidet, spricht man von *diskret gestuften* oder *geordnet kategoriale Ratingskalen* (► Beispiel 5.11, Fall B). Bei der Entscheidung über die Anzahl der Skalenstufen ist der Grad der Differenziertheit des Urteils zu berücksichtigen, der von den Testpersonen erwartet wird. Eine Skala, die extrem viele numerische Stufen enthält (z. B. 100) führt zumeist zu keinem Informationsgewinn, da überwiegend jene Stufen ausgewählt werden, die durch 10 bzw. 5 teilbar sind (Döring und Bortz 2016; Henns 1989).

Allgemein kann davon ausgegangen werden, dass sich ab mehr als sieben Skalenstufen kein Informationsgewinn durch individuelle Urteilsdifferenzierungen ergibt (Cox 1980). In verschiedenen Studien konnte gezeigt werden, dass für fünf bis sieben Skalenstufen optimale Eigenschaften hinsichtlich der Validität und der Reliabilität erreicht werden (Alwin 1992; de Beuckelaer et al. 2013; Lozano et al. 2008; Preston und Colman 2000; Weijters et al. 2010).

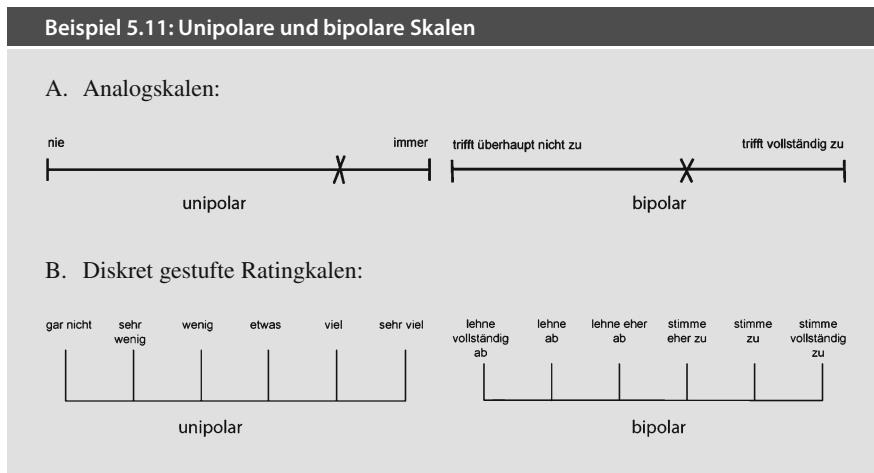
Bei der Entscheidung über die Anzahl der Skalenstufen sollten auch Überlegungen zur Differenzierungsfähigkeit der Testpersonen in der Zielpopulation einfließen (Alwin 1992; Döring und Bortz 2016; Rost 2004). Insbesondere bei Kindern kann eine zu große Anzahl von Skalenstufen zu einer Überforderung führen. Darüber hinaus sind sog. „Antworttendenzen“ oder „Response Sets“ zu berücksichtigen (► Exkurs 5.1 in ► Abschn. 5.3.3.5): So ist beispielsweise bei wenigen Antwortstufen die Tendenz zum extremen Urteil stärker ausgeprägt als bei mehr Stufen (Hui und Triandis 1989; Moors et al. 2014).

5.3.3.3 Unipolare vs. bipolare Antwortskala

Bei einer *bipolaren Skala* reicht der Zustimmungs-/Ablehnungsbereich zum jeweiligen Item von einem positiven Pol, der eine starke Zustimmung oder ein Zutreffen ausdrückt, über einen Indifferenzbereich zu einem negativen Pol, der eine starke Ablehnung oder ein Nichtzutreffen ausdrückt. Eine *unipolare Skala* hat hingegen einen „Nullpunkt“, womit jener Punkt als Bezugspunkt gemeint ist, der das geringste Ausmaß der Zustimmung (bzw. Ablehnung) kennzeichnet; der gegen-

5.3 · Aufgaben mit gebundenem Antwortformat

überliegende positive (bzw. negative) Pol markiert die stärkste Zustimmung (bzw. Ablehnung). Die Intensität, die Häufigkeit oder der Grad der Zustimmung (bzw. Ablehnung) steigt nur in eine Richtung (► Beispiel 5.11, Fall B). Die Entscheidung zugunsten einer unipolaren oder der bipolaren Skala ist von den Iteminhalten bzw. von der zu erfassenden Eigenschaft abhängig.



5.3.3.4 Bezeichnung der Skalenpunkte

Hat man sich für eine bestimmte Anzahl von Skalenpunkten (Stufen) entschieden, so stellt sich die Frage, wie diese bezeichnet werden sollen. Hierzu gibt es mehrere Möglichkeiten:

- Bei *numerischen Skalen* werden die Stufen häufig mit Zahlen markiert (► Beispiel 5.12). Diese erwecken den Anschein, dass es sich dabei um „präzise“ Messungen auf einer Intervallskala handelt. Die Anwendung einer numerischen Skala stellt jedoch nicht sicher, dass die Skalenpunkte gleichabständig sind und dass die Abstände zwischen den Skalenpunkten auch gleichen Ausprägungsunterschieden im interessierenden Merkmal entsprechen. Zudem erfolgt die Wahl eines bestimmten Zahlenformats oft recht willkürlich und ohne zu berücksichtigen, dass die Wahl bestimmte Folgen haben kann. So nehmen Testpersonen manchmal an, dass mit dem Zahlenformat die Polarität der Skala kommuniziert werden soll, indem mit der Beschriftung von -2 bis $+2$ ein bipolares, mit 1 bis 5 hingegen ein unipolares Merkmal gekennzeichnet ist. Die Wahl der Nummerierung kann somit eine Verschiebung der Antworten verursachen (Krebs und Hoffmeyer-Zlotnik 2010; Schwarz et al. 1991).

Ratingsskala mit numerischen Skalenpunkten



- Eine *verbale Ratingskala* liegt vor, wenn alle Skalenpunkte mit Worten bezeichnet werden. Sie hat den Vorteil, dass die Interpretation der Skalenpunkte intersubjektiv einheitlicher erfolgt – die Testpersonen brauchen sich nicht „vorzustellen“, was sich hinter den einzelnen Skalenpunkten verbirgt (► Beispiel 5.13). Zur verbalen Bezeichnung der Stufen werden u. a. Bewertungen, Häufigkeits- und Intensitätsangaben sowie Wahrscheinlichkeiten als Antwortoptionen herangezogen. Konkrete Häufigkeitsangaben (z. B. „mindestens einmal täglich“) zeichnen sich dabei durch die positive Eigenschaft aus,

Ratingsskala mit verbalen Skalenpunkten

dass sie einen verbindlichen, interpersonell vergleichbaren Maßstab darstellen. Die Testpersonen sind zufriedener, wenn nicht nur die zwei Extremwerte („Endpunktbenennung“), sondern auch weitere Skalenpunkte verbale Benennungen aufweisen (Dickinson und Zellinger 1980; Johnson et al. 2005; Weng 2004). Zudem verringert sich durch die Benennung aller Skalenstufen die Tendenz zum extremen Urteil (Moors et al. 2014). Insgesamt ist es jedoch schwierig, Benennungen zu finden, die äquidistante Abstände zwischen den Skalenstufen kennzeichnen.

Beispiel 5.13: Verbale Ratingskalen

A. Bipolare Skala mit Abstufungen des Zutreffens:

trifft voll und ganz zu	trifft über- wiegend zu	trifft gerade noch zu	trifft eher nicht zu	trifft über- wiegend nicht zu	trifft über- haupt nicht zu
-------------------------------	----------------------------------	--------------------------------	-------------------------------	--	--------------------------------------

B. Unipolare Skala mit Abstufungen der Häufigkeit (Intensität):

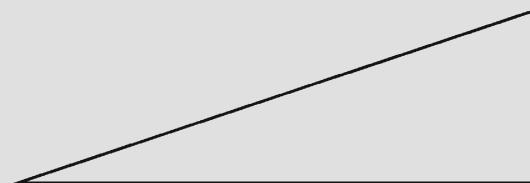
nie selten manchmal oft immer

Optische Skala und Symbolskala

- Optische Skalen und Symbolskalen unterliegen keinen subjektiven Schwankungen hinsichtlich der Bedeutung sprachlicher Bezeichnungen, wie dies bei rein sprachlichen Formaten der Fall ist (► Beispiel 5.14). Beide Skalentypen werden eingesetzt, um den Eindruck einer übertriebenen mathematischen Exaktheit zu vermeiden, die vom Testleiter de facto nicht sichergestellt werden kann.

Beispiel 5.14: Optische Skalen und Symbolskalen

A. Unipolare optische Analogskala:



B. Bipolare Symbolskala:



Kombinierte Skala

- Oft werden die verschiedenen Bezeichnungsarten miteinander kombiniert. Von einer Vermengung einer verbalen mit einer numerischen Skala erhofft man sich die Vorteile von beiden Formaten. Es ist dabei zu beachten, dass die verwendeten Bezeichnungen möglichst genau mit den Zahlen korrespondieren. So sollte man beispielsweise eine 5-stufige unipolare Intensitätsskala von „nie“ bis „immer“ nicht mit einem bipolar erscheinenden Zahlenschema von -2 bis +2 kombinieren, da dies die Eindeutigkeit der Interpretation absenkt (Hartley und Betts 2010; Lam und Kolic 2008; Rammstedt und Krebs 2007); angemessener wäre hingegen das Zahlenschema 0, 1, 2, 3, 4 (► Beispiel 5.15).

5.3 · Aufgaben mit gebundenem Antwortformat

Beispiel 5.15: Kombinierte Skalen

A. Verbal-numerische Skalen:

Unipolare Skala

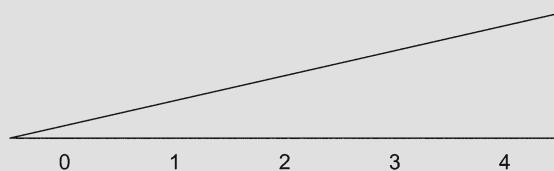


Bipolare Skala

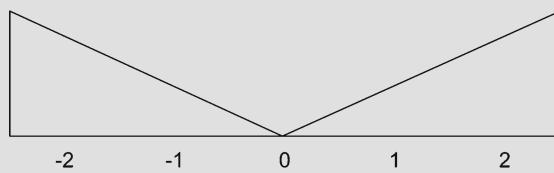


B. Optisch-numerische Skalen:

Unipolare Skala



Bipolare Skala



5.3.3.5 Verwendung einer neutralen Mittelkategorie

Bei der Entscheidung, ob die Skala eine neutrale Mittelkategorie enthalten soll oder nicht, sind verschiedene Aspekte in die Überlegung einzubziehen.

Verschiedentlich konnte empirisch gezeigt werden, dass eine neutrale mittlere Kategorie von den Testpersonen nicht nur instruktionsgemäß, d. h. im Sinne einer mittleren Merkmalsausprägung, benutzt wird (Bishop 1987; Hurley 1998; Katlon et al. 1980; Moors 2008; Presser und Schuman 1980). Vielmehr dient sie häufig auch als eine Ausweichoption, wenn die Testperson den angegebenen Wortlaut als unpassend beurteilt, die Frage nicht verstanden hat, die eigentliche Antwort verweigert oder diese nicht kennt. Im Antwortverhalten der Testpersonen resultiert daraus eine Konfundierung (Vermischung, Vermengung) des interessierenden Merkmals mit merkmalsfremden Aspekten, die zu erheblichen Validitätsproblemen und somit zu Verzerrungen bei der Interpretation der Befunde führt.

Manche Testpersonen nehmen auch an, dass die mittlere Kategorie von der „typischen“ oder „normalen“ Person angekreuzt wird, und platzieren deshalb ihre Antworten in dieser Kategorie, unabhängig davon, wie die Frage lautet. Diese sog. „Tendenz zur Mitte“ (s. ▶ Kap. 4) wird in ▶ Exkurs 5.1 näher beleuchtet.

Die der Tendenz zur Mitte entgegengesetzte „Tendenz zum extremen Urteil“ (s. z. B. Hurley 1998; Jäger und Petermann 1999) ist seltener zu beobachten; sie tritt

Merkmal konfundierung bei neutraler Mittelkategorie

Verstärkte Antworttendenzen bei neutraler Mittelkategorie

Exkurs 5.1**Tendenz zur Mitte**

Unter Tendenz zur Mitte wird die bewusste oder unbewusste Bevorzugung der mittleren (neutralen) Antwortkategorien unabhängig vom Iteminhalt verstanden (Paulhus 1991; van Herk et al. 2004). Die Bevorzugung kann entweder auf ein subjektiv unzureichendes Wissen zurückzuführen sein („Ich bin mir nicht ganz sicher in meiner Einschätzung, ich weiß zu wenig für ein sicheres Urteil – in der Mitte kann ich am wenigsten falsch machen!“) oder auf die Ansicht, dass sich die Antwortalternativen zur Beurteilung nicht eignen. Wenn Testpersonen dazu neigen, ihre Entscheidungen auf die mittleren Kategorien zu beschränken, führt dies zu einer verringerten Itemvarianz und zu Verzerrungen (vgl. ► Kap. 7).

Um der Tendenz zur Mitte entgegenzuwirken, sollte zum einen keine neutrale Mittelkategorie angeboten werden; zum anderen sollten möglichst keine zu extremen sprachlichen Bezeichnungen für die Pole der Beurteilungsskalen gewählt werden. Auch das Anbieten einer eigenen „Weiß-nicht“-Kategorie kann dieser Tendenz vorbeugen (► Abschn. 5.3.3.6).

vor allem dann auf, wenn die mittlere Kategorie z. B. von besonders motivierten Testpersonen gemieden wird. (Zu anderen Antworttendenzen s. ► Abschn. 4.7.1.) Zusammengenommen sprechen die Argumente eher gegen eine neutrale Mittelkategorie, obschon von manchen Autoren eine Mittelkategorie explizit empfohlen wird (s. z. B. Weijters et al. 2010).

5.3.3.6 „Weiß nicht“ als separate Antwortalternative

Die „Weiß-nicht“- oder „Kann-ich-nicht-beantworten“-Kategorie sollte als separate Antwortalternative dargeboten werden, wenn angenommen werden muss, dass es Testpersonen gibt, die zu dem jeweiligen Iteminhalt keine ausgeprägte Meinung haben, ihn nicht kennen, die Antwort nicht wissen oder die Frage sprachlich nicht verstanden haben. Gibt es diese Antwortoption nicht, sehen sich die Testpersonen veranlasst, auf der vorgegebenen Antwortskala z. B. die neutrale Mittelkategorie (► Exkurs 5.1) zu verwenden, was zu Antworten führt, die mit der Ausprägung des interessierenden Merkmals nur wenig oder gar nichts zu tun haben und für einen Rückschluss nicht zu gebrauchen sind.

Die „Weiß-nicht“-Kategorie vermindert das Problem der neutralen Mittelkategorie (► Abschn. 5.3.3.5), da den Testpersonen explizit die Möglichkeit einer Ausweichoption gegeben ist. Als Folge davon kann die neutrale Mittelkategorie ihre Funktion als Mitte der Beurteilungsskala besser erfüllen und muss nicht mehr als Sammelkategorie bei Verständnisschwierigkeiten, bei geringer Motivation, bei Erschöpfung wegen zu langer Tests oder als Ausweg bei Antwortverweigerung etc. dienen.

Der Nachteil der „Weiß-nicht“-Kategorie besteht darin, dass sie anstelle der intendierten merkmalsbezogenen Antworten „Missing Data“, also fehlende Daten erzeugt. Wird die „Weiß-nicht“-Kategorie von einer Testperson sehr häufig gewählt, so wird das Testergebnis ggf. unbrauchbar, weil nicht mehr genügend merkmalsbezogene Informationen vorliegen. Weiterhin ermöglicht die „Weiß-nicht“-Kategorie, schwierige Fragen zu umgehen; dies ist insbesondere dann problematisch, wenn Testpersonen unmotiviert sind oder kognitive Beeinträchtigungen haben (Krosnick et al. 2002). Zusammenfassend sollte die Aufnahme einer zusätzlichen „Weiß-nicht“-Kategorie sorgfältig abgewogen werden. Sie bietet sich insbesondere dann an, wenn eine berechtigte Vermutung besteht, dass einige Testpersonen nicht in der Lage sind, merkmalsbezogene Beurteilungen der Iteminhalte vorzunehmen.

Vorteile der „Weiß-nicht“-Kategorie**Nachteile der „Weiß-nicht“-Kategorie**

5.3.3.7 Festlegung der symptomatischen bzw. unsymptomatischen Antwortrichtung

Ein sehr wesentlicher weiterer Aspekt bei Ratingskalen besteht darin, dass die Beziehung zwischen der Antwortrichtung und der Ausprägung des interessierenden Merkmals eindeutig geklärt sein muss, d. h., es muss festgelegt sein, ob eine zustimmende bzw. ablehnende Antwort als symptomatisch bzw. als unsymptomatisch für eine hohe bzw. eine niedrige Ausprägung des interessierenden Merkmals zu bewerten ist.

Zumeist wird eine monotone Beziehung zwischen Itemantwort und Merkmalsausprägung angenommen: je größer die Zustimmung zum Item, desto höher die Ausprägung des Merkmals (► Kap. 16). Diese Annahme, die (im Sinne von Likert 1932) als „Dominanzansatz“ bezeichnet wird, ist z. B. für kognitive Tests angebracht, bei denen eine höhere Anzahl von korrekt gelösten Items auch eine höhere Merkmalsausprägung impliziert (Chernyshenko et al. 2001).

Zur Vermeidung von Akquieszenzeffekten (s. ► Kap. 4, ► Abschn. 4.7.3) wird verschiedentlich vorgeschlagen (z. B. Paulhus 1991), innerhalb eines Tests/Fragebogens zwei Arten von Items zu verwenden: Die Items sollen so konstruiert werden, dass bei einem Teil der Items eine Zustimmung als symptomatisch gilt; der andere Teil der Items soll so konstruiert werden, dass die Ablehnung als symptomatisch gilt. Neuere Untersuchungen haben allerdings gezeigt, dass diese Mischung von „positiven“ und „negativen“ Items zu anderen, zuvor nicht bekannten Problemen, vor allem zu Artefakten hinsichtlich der Dimensionalität des zu messenden Konstrukts, führen kann (s. z. B. Marsh 1996; Netemeyer et al. 2003; Rauch et al. 2007). Folglich ist die unkritische Anwendung des beschriebenen Verfahrens nicht uneingeschränkt empfehlenswert. Zur Überprüfung sollten statistische Verfahren zur Anwendung kommen, die eine Erfassung von Methodeneffekten erlauben (► Kap. 25).

Neben dem Dominanzansatz existiert bezüglich des Zusammenhangs zwischen Itemantwort und Merkmalsausprägung auch eine andere Überlegung, die (im Sinne von Thurstone 1927a, 1927b, 1928) als sog. „Idealpunktansatz“ bezeichnet werden kann (Stark et al. 2006): Bei Persönlichkeitsmerkmalen kann nämlich auch ein Idealpunkt vorliegen, d. h. eine bestimmte Ausprägung, bei der die Wahrscheinlichkeit einer zustimmenden Antwort besonders hoch ist. Beispielsweise würde einem Item „Ich habe kein Problem damit, soziale Events zu besuchen“ von durchschnittlich extravertierten Testpersonen mit einer höheren Wahrscheinlichkeit zugestimmt werden als von besonders Extravertierten oder von besonders Introvertierten (Stark et al. 2006). Als Konsequenz ist der Item-Merkmal-Zusammenhang dann nicht mehr monoton, sondern kurvilinear: Während die Wahrscheinlichkeit einer zustimmenden Antwort bis zu einer bestimmten Merkmalsausprägung ansteigt, fällt sie anschließend wieder ab. Ob der Dominanz- oder der Idealpunktansatz als zutreffend erachtet wird, sollte bereits bei der Itemkonstruktion und später bei der Itemanalyse und bei der psychometrischen Evaluation bedacht werden. Ist die Dominanzannahme verletzt, so können Standard-IRT-Modelle nicht mehr angewendet werden; vielmehr sollten Alternativmodelle herangezogen werden (s. nichtparametrische IRT-Modelle in ► Kap. 18).

5.3.3.8 Einsatz asymmetrischer Beurteilungsskalen und itemspezifischer Antwortformate

Asymmetrische Skalen werden vor allem dann eingesetzt, wenn damit zu rechnen ist, dass die Testpersonen kein vollständig symmetrisches Antwortspektrum nutzen werden. Psychologische Tests bedienen sich dieses Formats selten; bei Fragebogen in der Marktforschung und in der Kundenzufriedenheitsforschung findet sich ein asymmetrisches Antwortformat hingegen häufiger. Beispielsweise werden Schokolade- und Pralinenprodukte meist so positiv bewertet, dass symmetrische bipolare Beurteilungsskalen nur unzureichend in der Lage wären, Differenzen bei der Be-

Bedeutung von Antwortrichtung

Dominanzansatz: monotoner Item-Merkmal-Zusammenhang

Probleme bei der Mischung von „positiven“ und „negativen“ Items

Idealpunktansatz: kurvilinearer Item-Merkmal-Zusammenhang

Asymmetrische Beurteilungsskalen erzielen eine genauere Differenzierung im relevanten Merkmalsbereich

wertung unterschiedlicher Marken aufzudecken (Schuller und Keppler 1999). Eine asymmetrische Skala kann hier eine höhere Differenzierung in dem erwarteten positiven Bewertungsbereich erzielen und hat den Vorteil, dass beim Antwortverhalten kaum mit der Tendenz zur Mitte (► Exkurs 5.1) gerechnet werden muss (► Beispiel 5.16).

Beispiel 5.16: Asymmetrisches Antwortformat

Wie schätzen Sie den Geschmack dieser Schokoladensorte ein?

nicht lecker – lecker – sehr lecker – besonders lecker – exzeptionell lecker

Wechsel von Anzahl und Benennung der Antwortkategorien

Instruktion zum Hinweis von Formatwechseln

Obwohl es zugunsten einer einfacheren Handhabung wünschenswert ist, dass alle Items eines Tests oder Fragebogens möglichst mit einem einheitlichen Antwortformat beantwortet werden können, kommen auch *itemspezifische Antwortformate* zum Einsatz, d. h. Formate, bei denen sich die Anzahl und die Benennung der Antwortkategorien von Item zu Item unterscheiden. Itemspezifische Antwortformate finden eher in Fragebogen als in Tests Anwendung und vor allem bei der Erhebung demografischer Daten. Itemspezifische Antwortformate lassen sich auch mit asymmetrischen Beurteilungsskalen verbinden.

Sollte sich das Itemformat zwischen Blöcken von Items ändern (z. B. einige Items, die auf einer bipolaren Skala beurteilt werden sollen, gefolgt von Items, die auf einer unipolaren Skala beurteilt werden sollen), so sollte an dieser Stelle eine explizite Instruktion erfolgen, die auf den Wechsel des Antwortformats hinweist.

5.3.3.9 Zusammenfassende Bewertung

■ Vorteile von Beurteilungsaufgaben

In der Praxis sind Beurteilungsaufgaben leicht zu handhaben und ökonomisch bezüglich des Materialverbrauchs und der Auswertungszeit. Auch die Bearbeitungsdauer ist vergleichsweise kurz, da sich die Testpersonen auf einen Antwortmodus einstellen können und nicht bei jeder Aufgabe „umdenken“ müssen. Beurteilungsaufgaben werden sehr häufig eingesetzt.

■ Nachteile von Beurteilungsaufgaben

Häufig werden die Antwortkategorien mit Zahlen bezeichnet, um die Beurteilungsskala wie eine Intervallskala (Likert-Skala, ► Abschn. 5.3.3.1) benutzen zu können. Dies ist insofern etwas problematisch, als die Antworten streng genommen lediglich ordinalskaliert sind. Die Zuordnung von Zahlen zu den Skalenpunkten erleichtert aber die Anwendung von statistischen Auswertungsverfahren, die eine Intervallskalierung voraussetzen. An dieser Stelle sei jedoch angemerkt, dass dieses Vorgehen bei der Interpretation der Ergebnisse zu messtheoretischen Problemen führen kann. In den späteren Kapiteln wird dezidiert auf die verschiedenen Auswertungsmöglichkeiten von dichotomen, ordinal- und intervallskalierten Items eingegangen (► Kap. 12 und 16). Auch werden die Probleme dargestellt, die auftreten können, wenn falsche Annahmen über das Skalenniveau der Items getroffen werden.

5.4 Aufgaben mit atypischem Antwortformat

Nicht alle Antwortformate lassen sich in die oben aufgeführten Kategorien der freien und gebundenen Antwortformate einordnen. Durch Kombinationen der obigen Antworttypen lassen sich weitere Alternativen herstellen.

5.4 · Aufgaben mit atypischem Antwortformat

Unter das Prinzip atypischer Antwortformate lassen sich sehr verschiedene Aufgaben subsumieren. Im Zahlen-Verbindungs-Test (ZVT; Oswald 2016) beispielsweise werden solche Aufgaben verwendet, bei denen abgebildete Zahlen in einer aufsteigenden Reihenfolge auf dem Testbogen mit einer Linie verbunden werden müssen (► Beispiel 5.17, Fall A). Ein weiteres Beispiel stammt aus dem HAWIK (Hardesty und Priester 1963). Ähnliches findet auch in modernen Intelligenztests für Kinder Verwendung (z. B. WISC-IV; Petermann und Petermann 2011), wo beispielsweise aus einzelnen Teilen eine Figur gelegt werden soll. Zuletzt erlauben computerbasierte Tests neue Möglichkeiten für atypische Antwortformate, wie z. B. im Design a Matrix-Advanced (DESIGMA-Advanced; Becker und Spinath 2014), der einen distraktorfreien Matrizentest zur Erfassung der allgemeinen Intelligenz darstellt. Im Gegensatz zum klassischen Matrizentest (► Beispiel 5.5, Fall A), bei dem die richtige Antwort aus vorgegebenen Antwortalternativen ausgewählt werden muss, sollen die Testpersonen in diesem Test selbst die richtige Antwortalternative aus mehreren der angebotenen Konstruktionselementen zusammenstellen (► Beispiel 5.17, Fall B).

Arten von atypischen Antwortformaten

Beispiel 5.17: Atypische Antwortformate

A. Verbinden Sie die Zahlen in einer aufsteigenden Reihenfolge!

Übungsaufgabe 1:

Aufgabe: Verbinde die Zahlen in fortlaufender Folge:

1 – 2 – 3 – 4 – 5 – 6 usw. ...

<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6
ANFANG				
<input type="radio"/> 19	<input type="radio"/> 20	<input type="radio"/> 3	<input type="radio"/> 7	<input type="radio"/> 9
ENDE				
<input type="radio"/> 18	<input type="radio"/> 16	<input type="radio"/> 13	<input type="radio"/> 10	<input type="radio"/> 8
<input type="radio"/> 17	<input type="radio"/> 14	<input type="radio"/> 15	<input type="radio"/> 12	<input type="radio"/> 11

© Hogrefe Verlag Göttingen
Nachdruck und jegliche Art der Vervielfältigung verboten
Best. Nr. 01.066.07

Beispiel aus dem Zahlen-Verbindungs-Test (ZVT). (Aus Oswald 2016, © by Hogrefe Verlag GmbH & Co. KG, Göttingen ● Nachdruck und jegliche Art der Vervielfältigung verboten. Bezugsquelle des Zahlen-Verbindungs-Test (ZVT): Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, www.testzentrale.de)

B. „Ihre Aufgabe besteht darin zu erkennen, welchen Regeln die Muster der ersten 8 Kästchen folgen und das leere Kästchen so zu füllen, dass es die anderen logisch ergänzt. [...] Zum Füllen des leeren Kästchens stehen Ihnen die rot markierten Symbole in der unteren Hälfte des Bildschirms zur Verfügung. [...]“

Beim Anklicken eines Symbols wird das darin enthaltene Muster in dem leeren Kästchen angezeigt.“

Beispiel aus dem DESIGMA-Advanced. (Aus Becker und Spinath 2014, © by Hogrefe Verlag GmbH & Co. KG, Göttingen ● Nachdruck und jegliche Art der Vervielfältigung verboten. Bezugsquelle des Design a Matrix – Advanced (DESIGMA® – Advanced –): Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0551) 999-50-999, www.testzentrale.de)

5.5 Entscheidungshilfen für die Wahl des Aufgabentyps

Für die Test- und Fragebogenkonstrukteure lassen sich (in Anlehnung an Lienert und Raatz 1998, S. 24) allgemein relevante Gesichtspunkte für die Auswahl des Aufgabentyps bzw. des Antwortformats angeben.

Vorteilhafte Eigenschaften von guten Test- und Fragebogenaufgaben

Folgende Eigenschaften sind als Zielvorgabe für angemessene Aufgaben in Tests und Fragebogen sehr vorteilhaft:

- Leichte Verständlichkeit
- Einfache Durchführbarkeit
- Kurze Bearbeitungszeit
- Geringer Material- bzw. Papierverbrauch
- Leichte Auswertbarkeit
- Geringe Häufigkeit von Zufallslösungen

Neben der entsprechend dieser Zielvorgaben optimierten Auswahl angemessener Aufgabentypen müssen auch die in Moosbrugger und Brandt aufgeführten Gesichtspunkte der Itemformulierung Berücksichtigung finden (► Kap. 4).

5.6 Computerunterstützte Antwortformate

Über neuere Möglichkeiten computerunterstützter Antwortformate informieren Goldhammer und Kröhne in ► Kap. 6.

5.7 Zusammenfassung

Inhalt dieses Kapitels waren verschiedene Möglichkeiten, wie die Antworten der Testpersonen auf die Testaufgaben/-fragen erfasst und kodiert werden können („Antwortformate“). Daraus ergeben sich verschiedene Itemtypen. Unter Beachtung von Vor- und Nachteilen wurde das freie Antwortformat dem gebundenen Antwortformat gegenübergestellt. Bei Letzterem sind vor allem Ordnungs-, Auswahl- sowie kontinuierliche und diskrete Beurteilungsaufgaben als Itemtypen weitverbreitet, wobei Letztere auf „Ratingskalen“ beantwortet werden. Unter Heranziehung zahlreicher Beispiele wurden viele praxisrelevante Konstruktionsaspekte erörtert und unter Bezug auf verschiedene Zielvorgaben diskutiert. Mit Entscheidungshilfen für die Wahl des Aufgabentyps wurde das Kapitel abgerundet.

5.8 Kontrollfragen

- ?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Wozu dient eine sog. „Weiß-nicht“-Kategorie und wann wird sie eingesetzt?
 2. Was versteht man unter „Exhaustivität“ und unter „Disjunktheit“?
 3. Worauf muss bei der Generierung von Antwortmöglichkeiten im Rahmen von Auswahlaufgaben bei Leistungstests besonders geachtet werden, worauf bei Persönlichkeitstests?
 4. Welche Möglichkeiten zur Senkung der Ratewahrscheinlichkeit sollte man bei Zuordnungsaufgaben beachten?
 5. Was versteht man unter einem „Distraktor“ und was wird bei der Distraktorenanalyse genauer untersucht?
 6. Was versteht man unter der „Tendenz zur Mitte“?
 7. Worin unterscheiden sich „unipolare“ und „bipolare“ Antwortskalen?

Literatur

- Alwin, D. F. (1992). Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology*, 22, 83–118.
- Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *I-S-T 2000*. R. Göttingen: Hogrefe.
- Bauer, D., Holzer, M., Kopp, V. & Fischer, M. R. (2011). Pick-N multiple choice-exams: a comparison of scoring algorithms. *Advances in health sciences education: theory and practice*, 16, 211–221.
- Becker, N. & Spinath, F. (2014). *DESIGMA-Advanced – Design a Matrix-Advanced (Manual)*. Göttingen: Hogrefe.
- Bishop, G. F. (1987) Experiments with the Middle Response Alternatives in Survey Questions. *Public Opinion Quarterly*, 51, 220–232.
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F. & Williams, B. A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36, 523–562.
- Cronbach, L. J. (1941). An experimental comparison of the multiple true–false and multiple multiple-choice tests. *Journal of Educational Psychology*, 32, 533–543.
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, 17, 407–442.
- De Beuckelaer, A., Toonen, S. & Davidov, E. (2013). On the optimal number of scale points in graded paired comparisons. *Quality & Quantity*, 47, 2869–2882.
- Dickinson, T. L. & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating mixed standard scale formats. *Journal of Applied Psychology*, 65, 147–154.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Heidelberg: Springer.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.

- Exner, J. E. (2010). *Rorschach-Arbeitsbuch für das Comprehensive System: Deutschsprachige Fassung von A Rorschach Workbook for the Comprehensive System – Fifth Edition*. Göttingen: Hogrefe.
- Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53, 999–1010.
- Hardesty, F. P. & Priester, H. J. (1963). *Hamburg-Wechsler-Intelligenz-Test für Kinder: HAWIK* (2. Aufl.). Bern: Huber.
- Hartley, J. & Betts, L. R. (2010). Four Layouts and a Finding: The effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, 13, 17–27.
- Henss, R. (1989). Zur Vergleichbarkeit von Ratingskalen unterschiedlicher Kategorienzahl. *Psychologische Beiträge*, 31, 264–284.
- Höft, S. & Funke, U. (2006). Simulationsorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie*. (2. Aufl., S. 145–188). Göttingen: Hogrefe.
- Hornke, L. F., Etzel, S. & Rettig, K. (2005). *Adaptiver Matrizen Test. Version 24.00*. Mödling: Schuhfried.
- Hui, C. H. & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20, 296–309.
- Hurley, J. R. (1998). Timidity as a Response Style to Psychological Questionnaires. *Journal of Psychology*, 132, 202–210.
- Jäger, R. S. & Petermann, F. (Hrsg.) (1999). *Psychologische Diagnostik* (4. Aufl.). Weinheim: Beltz PVU.
- Johnson, T., Kulesa, R., Cho, Y. I. & Shavitt, S. (2005). The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, 36, 264–277.
- Katlon, G., Roberts, J. & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician*, 29, 65–78.
- Krampen, D. (2015). *Zur Bedeutung des Testformats für die Testauswertung. Aufgabenstamm- und Antwortabhängigkeiten im C-Test*. Frankfurt am Main: Lang.
- Krebs, D. & Hoffmeyer-Zlotnik, J. H. P. (2010). Positive first or negative first? Effects of the order of answering categories on response behavior. *Methodology*, 6, 118–127.
- Krosnick, J. A. (1999). Survey research. *Annual review of Psychology*, 50, 537–567.
- Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C. & Conaway, M. (2002). The impact of “no opinion” response options on data quality: Non-attitude reduction or an invitation to sacrifice? *Public Opinion Quarterly*, 66, 371–403.
- Lam, T. C. M. & Kolic, M. (2008). Effects of semantic incompatibility on rating response. *Applied Psychological Measurement*, 32, 248–260.
- Lienert, G. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz PVU.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5–53.
- Lord, F. M. (1944). Reliability of multiple choice tests as a function of number of choices per item. *Journal of Educational Psychology*, 35, 175–180.
- Lord, F. M. (1977). Optimal number of choices per item—a comparison of four approaches. *Journal of Educational Measurement*, 14, 33–38.
- Lozano, L. M., García-Cueto, E. & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4, 73–79.
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, 70, 810–819.
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, 42, 779–794.
- Moors, G., Kieruj, N. D. & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, 44, 369–399.
- Moosbrugger, H. & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2)*. Bern, Göttingen: Huber.
- Netemeyer, R. G., Bearden, W. O. & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications.
- Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Ergebnisse: Was Schülerinnen und Schüler wissen und können (Band I, überarbeitete Ausgabe): Schülerleistungen in Lesekompetenz, Mathematik und Naturwissenschaften*. Bielefeld: W. Bertelsmann.
- Oswald, W. D. (2016). *Zahlen-Verbindungs-Test ZVT* (3. Aufl.). Göttingen: Hogrefe.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman, (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
- Petermann, F. & Petermann, U. (Hrsg.) (2011). *WISC-IV. Wechsler Intelligence Scale for Children – Fourth Edition*. Frankfurt am Main: Pearson Assessment.
- Pfiffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, 7, 258–264.

Literatur

- Presser, S. & Schuman, H. (1980). The measurement of a middle position in attitude surveys. *Public Opinion Quarterly*, 44, 70–85.
- Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1–15.
- 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.
- Rauch, W. A., Schweizer, K. & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, 42, 1597–1607.
- Rodriguez, M. C. (2005). Three options for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24, 3–13.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. (2. Aufl.). Bern: Huber.
- Schuller, R. & Keppler, M. (1999). Anforderungen an Skalierungsverfahren in der Marktforschung/Ein Vorschlag zur Optimierung. *Planung & Analyse*, 2, 64–67.
- Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E. & Clark, L. (1991). Rating scales. Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570–582.
- Stark, S., Chernyshenko, O. S., Drasgow, F. & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91, 25–39.
- Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, 38, 368–389.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.
- Torrance, E. P. (1998). *The Torrance tests of creative thinking norms—technical manual figural (streamlined) forms A&B*. Bensenville, IL: Scholastic Testing Service.
- Torrance, E. P. & Ball, O. E. (1984). *Torrance test of creative thinking. Revised manual*. Bensenville, IL: Scholastic Testing Services.
- Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1, 386–391.
- van Herk, H., Poortinga, Y. H. & Verhallen, T. M. (2004). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35, 346–360.
- Weijters, B., Cabooter, E. & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236–247.
- Weng, L.-J. (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-retest Reliability. *Educational and Psychological Measurement*, 64, 956–972.



Computerbasiertes Assessment

Frank Goldhammer und Ulf Kröhne

Inhaltsverzeichnis

- 6.1 Computerbasiertes Assessment: Definition und Übersicht – 121**
 - 6.1.1 Definition – 121
 - 6.1.2 Assessment mit „Big Data“ – 122
 - 6.1.3 Etappen des Assessmentzyklus – 124
- 6.2 Itementwicklung: Antwortformat, Stimulus und Antwortbewertung – 124**
 - 6.2.1 Antwortformat – 125
 - 6.2.2 Komplexität – 125
 - 6.2.3 Wiedergabetreue – 126
 - 6.2.4 Interaktionsgrad – 127
 - 6.2.5 Medienverwendung – 128
 - 6.2.6 Antworthandlung – 129
 - 6.2.7 Antwortbewertung – 129
- 6.3 Testentwicklung: Testzusammenstellung und -sequenzierung – 130**
 - 6.3.1 Testzusammenstellung in computerisierten Tests – 130
 - 6.3.2 Navigation in computerbasierten Instrumenten – 131
 - 6.3.3 Sequenzierung computerisierter Fragebogen – 132
- 6.4 Testadministration – 132**
 - 6.4.1 Aktivitätsauswahl – 133
 - 6.4.2 Präsentation – 133
 - 6.4.3 Evidenzidentifikation – 134
 - 6.4.4 Evidenzakkumulation – 134
- 6.5 Datenanalyse und Rückmeldung – 135**
 - 6.5.1 Daten und Analysepotential – 135
 - 6.5.2 Rückmeldung von Testdaten – 136

6.6 Zusammenfassung – 137

6.7 EDV-Hinweise – 137

6.8 Kontrollfragen – 138

Literatur – 138

i Weshalb sollte ein Test oder Fragebogen computerbasiert durchgeführt bzw. beantwortet werden? Aus diagnostischer Sicht gibt es dafür eine Reihe gewichtiger Gründe: Mit computerbasierten Items lassen sich Personenmerkmale messen, die auf Papier schwer oder überhaupt nicht messbar sind (z. B. die Kompetenz, mit Computern umzugehen oder Probleme interaktiv durch die Manipulation einer simulierten Umgebung lösen zu können). Der Computer erlaubt außerdem, Items automatisch zu generieren und individuell für eine Testperson zu einem Test zusammenzustellen (adaptives Testen). Dazu werden Antworten automatisch bewertet und diese Information zur Steuerung des weiteren Testablaufs verwendet. Die automatische Antwortbewertung ermöglicht zudem eine zeitnahe Rückmeldung über das Bearbeitungsverhalten und den Bearbeitungserfolg, womit beispielsweise der Lernprozess unterstützt werden kann. Schließlich bieten computerbasierte Assessments die Möglichkeit, neben Ergebnis- auch Prozessdaten (z. B. Bearbeitungszeiten und Sequenzen von Bearbeitungsschritten) zu sammeln, die Einblicke in das Lösungsverhalten gewähren.

6.1 Computerbasiertes Assessment: Definition und Übersicht

Die Nutzung des Computers zur Messung von Individual- und Gruppenunterschieden zählt heute zu den methodischen Standards der psychologischen und sozialwissenschaftlichen Forschung. Computer werden beispielsweise eingesetzt, um allgemeine kognitive Fähigkeiten von Personen in beruflichen Auswahlverfahren zu messen (z. B. Funke 1995) oder auch die Lesekompetenz von Populationen in internationalen Bildungsvergleichsstudien (z. B. OECD 2011) oder das experimentell induzierte Ausmaß von Interferenzen in kognitionspsychologischen Studien zu erfassen (z. B. Steinwascher und Meiser 2016).

Gründe für computerbasiertes Assessment

6.1.1 Definition

Unter *computerbasiertem Assessment* verstehen wir eine Methode der technologiebasierten Messung von Individual- und Gruppenunterschieden in Bezug auf verschiedene Personenmerkmale wie Fähigkeiten, Kompetenzen, Persönlichkeitsdimensionen, Motivation oder Einstellungen.

Definition

Nach Scalise (2012, S. 134) kann Assessment als „collecting evidence designed to make an inference“ definiert werden. Das **computerbasierte Assessment** stellt damit ein Sammeln von empirischen Informationen unter Zuhilfenahme eines Computers im weiteren Sinne dar, wobei die Bedingungen für das Sammeln so gestaltet werden, dass auf Grundlage der gesammelten Informationen Schlussfolgerungen über Individual- und Gruppenunterschiede möglich sind. Der Computer wird dazu eingesetzt, die Items zu präsentieren (z. B. Text, Bild, Audio, Video, Simulation), ihre Abfolge zu steuern sowie die Interaktionen der Testperson mit der Aufgabe zu registrieren (z. B. über Mausklicks, Tastatur- und Touchdisplayeingaben) und ggf. automatisch auszuwerten.

Assessment als gezieltes Sammeln empirischer Belege für spezifische Schlussfolgerungen

Die Definition verweist zunächst darauf, dass im Rahmen eines Assessments Bedingungen gezielt hergestellt werden, um die interessierende empirische Information zum Vorschein zu bringen (z. B. die Vorgabe eines standardisierten Fragebogens oder von interaktiven Testaufgaben). Dies erfolgt in Abhängigkeit von den diagnostischen Zielsetzungen und der intendierten Interpretation des Testwertes (► Kap. 21) und stellt den Regelfall heutiger Assessmentpraxis dar.

Umstellung auf computerbasiertes Assessment in PISA

6

Für Assessments werden Computer im weitesten Sinne eingesetzt, nicht nur (vernetzte) Desktop-Computer und Laptops, sondern auch Tablets und Smartphones sowie spezifische Technologien zur Sammlung diagnostischer Daten wie digitale Stifte, Fitness-Tracker, Virtual-Reality-/Augmented-Reality-Brillen. Neben individuellen Unterschieden, die z. B. im Rahmen einer schulischen Förderdiagnostik erhoben werden, sind auch Gruppenunterschiede das Ziel von Assessments. Prominentes Beispiel dafür ist das „Programme for International Student Assessment“ (PISA; OECD 2014), in dem regelmäßig Kompetenzunterschiede (z. B. das Leseverständnis) zwischen Populationen (Bildungssystemen) gemessen werden. Seit PISA 2015 werden die Messungen in den Bereichen Lesen, Mathematik und Naturwissenschaften vollständig computerbasiert durchgeführt. Um die Vergleichbarkeit der PISA-2015-Messungen mit den früheren papierbasierten PISA-Messungen sicherzustellen, wurde eigens geprüft, ob bzw. wie der Wechsel des Modus von Papier zu Computer die psychometrischen Eigenschaften der Messinstrumente verändert („Moduseffekte“, s. Kroehne und Martens 2011; Mead und Drasgow 1993; Robitzsch et al. 2017).

6.1.2 Assessment mit „Big Data“

Im Unterschied zu dem vorgenannten Vorgehen gibt es jedoch aktuelle Entwicklungen, nach denen ein Assessment auch ohne die zielgerichtete Schaffung standar-

Exkurs 6.1

Big Data

In sozialen Netzwerken und Nachrichtendiensten informiert eine Vielzahl von Internetnutzern und -nutzerinnen täglich – wenn auch nicht immer bewusst – über aktuelle Aktivitäten, Gedanken und Gefühle. Es liegt nahe, diese große Menge anfallender Daten (*Big Data*) auch für Assessmentzwecke zu nutzen. Big Data unterscheiden sich von traditionellen Daten vor allem durch ihren erheblichen Umfang (Terabytes und mehr), ihre (Un-)Strukturiertheit, ihre vielfältigen Formate und ihre schnelle Verfügbarkeit (z. B. via Internet) sowie ihr rapides Wachstum (vgl. Deroos et al. 2012; Tien 2013). Daraus folgt, dass sie kaum mit gängigen Datenmanagement- und Analyseverfahren effizient genutzt werden können. Vielmehr kommen Verfahren des (*Educational*) *Data Mining* zum Einsatz, die beispielsweise sowohl Methoden des maschinellen Lernens zur Mustererkennung nutzen wie auch die natürliche Sprachverarbeitung zur Inhaltsanalyse von Texten. Quellen von Big Data sind vielfältig, beispielsweise Daten aus sozialen Medien und Netzwerken, persönliche Daten (z. B. das Bewegungsprofil), finanzielle Transaktionsdaten und administrative Daten (vgl. Chen et al. 2014).

Nutzt man Big Data für die Messung individueller Unterschiede, wird das Paradigma zur Datengewinnung grundsätzlich gewechselt: Es werden nicht mehr gezielt Bedingungen zur Datenerhebung hergestellt (z. B. durch Test oder Fragebogen), sondern ohnehin anfallende digitale Verhaltensdaten werden im Hinblick auf eine Assessmentfragestellung sekundär verwertet. Vorteilhaftweise können dadurch sowohl Versuchsleitereffekte wie auch verzerrnde Effekte bei Selbstberichten (z. B. soziale Erwünschtheit, ► Kap. 4) o. Ä. vermieden werden. Demgegenüber stehen – neben den oben genannten Herausforderungen der Datenstruktur – Probleme durch die Stichprobenverzerrung, die mehr oder weniger starke Tendenz zur attraktiven Selbstdarstellung, Fragen zur Sicherstellung der Anonymität von Nutzern sowie ethische Fragen zur Nutzung öffentlich verfügbarer Daten für Forschungszwecke (vgl. z. B. Shah et al. 2015).

6.1 · Computerbasiertes Assessment: Definition und Übersicht

disierter Bedingungen möglich erscheint, das Assessment mit „Big Data“, d. h. mit riesigen Datenmengen (► Exkurs 6.1). Nutzt man „Big Data“ für Assessmentfragestellungen, wird zur Datengewinnung ein Nebenprodukt von Nutzerprozessen im Internet herangezogen, deren hauptsächlicher Zweck nicht in der psychologischen oder sozialwissenschaftlichen Diagnostik besteht. Dieses Vorgehen scheint vor allem dann vielversprechend, wenn Indikatoren des interessierenden Konstrukts unmittelbar abgeleitet werden können. Statt beispielsweise selbstreguliertes Lernen und die Anwendung von Lesestrategien mit einem standardisierten Verfahren zu messen, kann auch das tatsächliche Lern- und Leseverhalten in einer aktiv genutzten Onlinelernumgebung (z. B. Navigation in digitalen Lernressourcen, Definition von Suchabfragen, Organisation von recherchiertem Wissen, Markieren wichtiger Inhalte und Zusammenfassung eines Textes) die nötigen Informationen liefern.

Zwei Beispiele sollen die Herangehensweise bei einem Big-Data-Assessment sowie dessen Potential illustrieren (► Beispiel 6.1, ► Beispiel 6.2).

Nutzung inzidenteller Daten für das Assessment

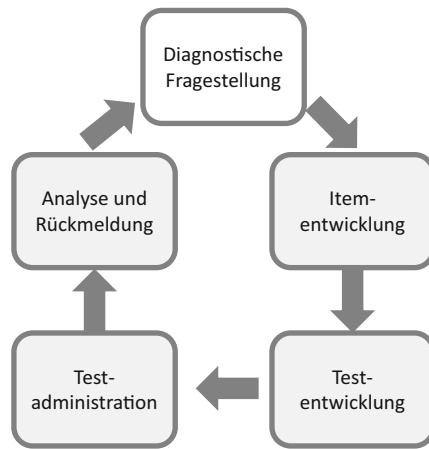
Aktivitäten in sozialen Netzwerken als Datenbasis für Assessment

Beispiel 6.1: Analyse von Facebook-Nutzern anhand ihres Like-Verhaltens

Kosinski et al. (2013) sind in ihrer Studie der Frage nachgegangen, ob sich unterschiedliche Merkmale von Facebook-Nutzern, u. a. Persönlichkeitsmerkmale, Intelligenz und politische Ansichten, auf der Grundlage ihrer *Likes*, d. h. der positiven Beurteilungen von Online-Inhalten erklären lassen. Datengrundlage war das Like-Verhalten einer Stichprobe von 58.466 freiwilligen Untersuchungsteilnehmern und -teilnehmerinnen sowie die in Teilstichproben zusätzlich per Test und Fragebogen erhobenen Personenmerkmale. Die *User-Like-Matrix* gab für die Nutzer an, ob eine Like-Assoziation zu einer der 55.814 berücksichtigten Online-Inhalte vorlag (1) oder nicht (0). Als Prädiktoren für die Personenmerkmale wurden die ersten 100 Hauptkomponenten (► Kap. 23) der User-Like-Matrix verwendet (es wurde eine Singulärwertzerlegung durchgeführt, in die die $(n \times m)$ -User-Like-Matrix einging). Die Klassifikation der politischen Einstellung (liberal vs. konservativ) gelang mit einer hohen Genauigkeit von 85 %. Die Korrelationen zwischen den durch das Like-Verhalten vorhergesagten und per Fragebogen gemessenen Big-Five-Persönlichkeitsdimensionen fielen moderat aus (zwischen .29 und .43), die Korrelation mit Intelligenz ebenfalls (.39).

Beispiel 6.2: Analyse von Facebook-Nutzern anhand ihrer Nachrichten

Schwartz et al. (2013) untersuchten inhaltsanalytisch u. a. die Persönlichkeit von Facebook-Nutzern auf Basis der Textinformationen in ihren Facebook-Nachrichten. Die Datenbasis bestand aus 700 Millionen Wörtern, Phrasen und Themen von 75.000 freiwilligen Untersuchungsteilnehmern und -teilnehmerinnen. Das Auftreten bestimmter sprachlicher Merkmale bzw. die Verwendung spezifischen Vokabulars korrelierte dabei signifikant mit der Ausprägung in den ebenfalls erfassten Big-Five-Persönlichkeitsdimensionen ($R = .31 - .42$). Beispielsweise war Extraversion gekennzeichnet durch eine häufigere Nennung des Wortes „Party“, Introversion hingegen durch „Internet“. Zudem konnten anhand linguistischer Merkmale mit sehr hoher Präzision das Alter ($R = .84$) und das Geschlecht (Genauigkeit: 91.9 %) der Nutzer und Nutzerinnen vorhergesagt werden.



■ Abb. 6.1 Assessmentzyklus

6.1.3 Etappen des Assessmentzyklus

Assessmentzyklus

Die Etappen des Assessmentzyklus (■ Abb. 6.1, vgl. auch „evidence-centered design layers“; Mislevy 2013) strukturieren das vorliegende Kapitel. Ausgehend von einer diagnostischen Fragestellung, die definiert, welche Schlussfolgerungen aufgrund des Testergebnisses gezogen werden sollen (z. B. Aussagen über die Fähigkeit einer Person), ist das fragliche Merkmal zu definieren und auf diesem Hintergrund zu klären, durch welche Verhaltensindikatoren es gemessen werden kann und wie diese Verhaltensweisen durch Items (z. B. Testaufgaben) entsprechend hervorgerufen werden können. Sind passende Items entwickelt (► Kap. 4), werden sie zu einem Test zusammengestellt. Der Test wird administriert (► Kap. 3) und schließlich das dabei beobachtete Verhalten zur Beantwortung der diagnostischen Fragestellung ausgewertet und aggregiert.

Aus diesem Zyklus ergibt sich der Aufbau dieses Kapitels. Mit Fokus auf das computerbasierte Assessment werden entlang des Assessmentzyklus die folgenden Themen behandelt:

- Itementwicklung (► Abschn. 6.2)
- Testentwicklung (► Abschn. 6.3)
- Testadministration (► Abschn. 6.4)
- Datenanalyse und Rückmeldung (► Abschn. 6.5)

6.2 Itementwicklung: Antwortformat, Stimulus und Antwortbewertung

Konstruktrepräsentation sicherstellen

Computerbasiertes Assessment ermöglicht neuartige Operationalisierungen von Konstrukten anhand von Verhaltensindikatoren, die nicht oder nur schwer mit Papier realisierbar wären. Solche Neuerungen auf Itemebene betreffen sowohl die Gestaltung des *Stimulus* als auch das *Antwortformat*. Das zentrale Motiv hierfür besteht darin, die Konstruktrepräsentation zu verbessern oder – vor allem in Hinblick auf komplexere kognitive Leistungen – überhaupt erst zu ermöglichen (Sireci und Zenisky 2006; s. auch Frey und Hartig 2013). Durch höhere Interaktivität bzw. Manipulierbarkeit des Stimulus und durch die mit multimedialer Anreicherung (z. B. Bild, Audio, Video, Animation) gesteigerte Authentizität sind die Aufgaben dazu geeignet, Fähigkeiten, Fertigkeiten und Wissen, die ein Konstrukt ausmachen, vollständig zu repräsentieren. Das bedeutet allerdings umgekehrt, dass beispielsweise multimediale Elemente, die nicht direkt aus dem zu messenden

6.2 · Itementwicklung: Antwortformat, Stimulus und Antwortbewertung

Konstrukt heraus begründet sind, potentiell die Validität der Testwertinterpretation beeinträchtigen (vgl. van der Linden 2002), weshalb für die Messung irrelevante Verarbeitungsschritte auf mentaler und Verhaltensebene zu vermeiden sind (vgl. Stout 2002). Gegebenenfalls kann eine zusätzliche multimediale Anreicherung die Attraktivität der Aufgabenstellung erhöhen, positive Effekte auf die Motivation der Testperson in die Aufgabenbearbeitung haben oder die Einstellung zum Assessment verbessern (z. B. Richman-Hirsch et al. 2000).

Parshall et al. (2010) schlagen eine Taxonomie vor, anhand derer sich computerbasierte Items auf sieben Dimensionen systematisieren lassen (s. auch Scalise und Gifford 2006). Entlang dieser Dimensionen (► Abschn. 6.2.1 bis 6.2.7) werden im Folgenden zentrale Aspekte der computerbasierten Itementwicklung behandelt.

6.2.1 Antwortformat

Die Dimension „Antwortformat“ bezieht sich darauf, ob eine Antwort durch die Auswahl einer oder mehrerer Antwortoptionen zustande kommt (z. B. Multiple-Choice-Item) oder ob die Antwort offen/frei konstruiert werden muss (Parshall et al. 2010; vgl. auch die Eingeschränktheit des Antwortformats bei Scalise und Gifford 2006; s. dazu auch ► Kap. 5).

Computerbasierte Formen von Auswahlaufgaben erlauben beispielsweise Sortier- und Zuordnungsaufgaben, in denen per Ziehen und Ablegen („Drag-and-drop“) eine Menge von Elementen nach einem bestimmten Kriterium sortiert, zugeordnet oder räumlich arrangiert werden muss, beispielsweise die vorgegebenen Schritte eines mathematischen Beweises. Ein weiteres Beispiel für eine innovative Form sind automatisch bewertbare „Hotspotaufgaben“, in denen ein Teil oder Teile einer Abbildung bzw. eines Diagramms ausgewählt werden müssen, oder auch Markieraufgaben, in denen für eine Antwort Textteile markiert werden (vgl. z. B. Programme for International Assessment of Adult Competencies, PIAAC; Upsing et al. 2013). Durch die höhere Zahl von Auswahlalternativen wird die Ratewahrscheinlichkeit gesenkt. Die direkte Interaktion in figuralen Auswahlaufgaben kann zudem zu einer besseren Konstruktrepräsentation beitragen, indem konstruktfremde verbale Repräsentationen und Verarbeitungsschritte vermieden werden.

Die offene/freie Konstruktion einer Antwort (*Constructed Response*) kann sehr unterschiedlich ausfallen, beispielsweise als Zeichnung, Formeln, Aufsatz (Essay), Kurztext (z. B. ein bis drei Sätze zur Begründung einer Auswahl), Wort oder Zahl. Textantworten werden in der Regel schriftlich gegeben, sind aber auch gesprochen von Bedeutung (z. B. beim Fremdsprachenlernen). Computerbasiert sind diese Konstruktionsaufgaben insbesondere innovativ, wenn neben der Antwortregistrierung auch die Auswertung computerbasiert erfolgen kann. Dies ist etwa im Fall eines einzelnen Wortes durch automatischen Abgleich mit einer Liste korrekter Lösungen (inklusive akzeptabler Rechtschreibfehler) vergleichsweise einfach, bei längeren Aufsätzen oder Kurztextantworten jedoch deutlich aufwendiger (► Abschn. 6.2.7).

Sieben Dimensionen zur Systematisierung computerbasierter Items

Geschlossene/gebundene Antwortformate

Offene/freie Antwortformate

6.2.2 Komplexität

Parshall et al. (2010) definieren die Komplexität von Items als die Anzahl und Vielfalt von Elementen, die eine Testperson für eine Antwortabgabe berücksichtigen muss. Diese betrifft sowohl inhaltliche (z. B. Informationen in Text und Bild auf unterschiedlichen Seiten platziert) als auch funktionale Aspekte (z. B. Hyperlinks, Geräte zum Abspielen von Medien). Eine klassische Leseverständnisaufgabe mit statischem Text und einem geschlossenen Antwortformat ist demnach deutlich

Schwierigkeitsbestimmende Merkmale

6

weniger komplex als eine computerbasierte Leseverständnisaufgabe mit digitalem Text auf simulierten Webseiten, die multimedial aufgebaut und miteinander verlinkt sind. Daraus wird ersichtlich, dass die Komplexität mit den beiden Dimensionen „Wiedergabetreue“ (► Abschn. 6.2.3) und „Interaktivität“ (► Abschn. 6.2.4) zusammenhängt.

Höhere Komplexität bedeutet oftmals, dass die Anforderungen beim Lösen des Items steigen, d. h. Komplexität kann ein mögliches schwierigkeitsbestimmendes Merkmal darstellen. Wichtig ist, dass solche Anforderungen als konstruktrelevant begründet sein müssen. Dies wäre beispielsweise nicht der Fall, wenn eine computerbasierte Mathematikaufgabe zusätzliche (für das Konstrukt irrelevante) Anforderungen allein durch den erforderlichen Umgang mit dem Computer stellen würde. Genauso wie Items auf Papier beispielsweise nicht durch die Formulierung unnötig verkompliziert werden sollen, ist bei computerbasierten Items darauf zu achten, dass eine sichere und benutzerfreundliche Handhabung des Assessment-systems gewährleistet ist und die Erfahrung im Umgang mit Computern keinen konfundierenden Faktor darstellt (Parshall et al. 2010).

6.2.3 Wiedergabetreue

Wiedergabetreue bezieht sich darauf, wie realistisch und genau das Item konstrukt-relevante Objekte, Situationen, Aufgaben und Umgebungen reproduzieren kann (Parshall et al. 2010). Diese Entsprechung bezieht sich sowohl auf die Präsentation als auch auf funktionale Eigenschaften von Itemelementen (► Beispiel 6.3).

Beispiel 6.3: Wiedergabetreue eingesetzter Items

Ein Beispiel dafür ist die Messung von Computerfertigkeiten (Goldhammer et al. 2014a). Wiedergabetreue bedeutet, dass die simulierte Software im Grundsatz so aussieht und funktioniert wie reale Software. Allerdings wäre es nicht zielführend, eine bestimmte Software identisch nachzubauen (z. B. einen bestimmten Browser), da Testpersonen benachteiligt werden könnten, die mit einer anderen Software vertraut sind. Das Ziel besteht also darin, das Design und die Funktionalität so zu gestalten, dass möglichst allgemeine, über verschiedene Softwareprodukte hinweg geltende Gestaltungsprinzipien umgesetzt werden (z. B. bedeutet das Symbolbild „Lupe“ in der Regel, dass mit dem Schalter eine Suchfunktion bedient werden kann).

Diagnostisches Ziel bestimmt das Maß der Wiedergabetreue

Eine höhere Wiedergabetreue bedeutet einen höheren Aufwand bei der Itementwicklung. Ob der Aufwand gerechtfertigt ist, hängt von der intendierten Testwertinterpretation ab. Für das PIAAC musste beispielsweise zur standardisierten Erfassung der Kompetenz, Probleme technologiebasiert zu lösen („Problem Solving in Technology-Rich Environments“, s. OECD 2013), ein Kompromiss zwischen realistischer Simulation von Softwareapplikationen (z. B. Webbrowser, Textverarbeitung, Tabellenkalkulation) und der Umsetzbarkeit des Assessments gefunden werden. Ein weiteres Beispiel geben Parshall et al. (2010): Während ein einfacher Flugsimulator am Computer für einen Auswahlprozess ausreichend sein mag, ist für die Pilotenausbildung ein voll funktionaler Flugsimulator erforderlich.

Höhere Wiedergabetreue führt nicht zwangsläufig zu einer Verbesserung der Konstruktrepräsentation. In manchen Fällen mag es ausreichen, eine Entsprechung wichtiger struktureller Merkmale herzustellen, wie im Beispiel zu „Problem Solving in Technology-Rich Environments“ des PIAAC bereits gezeigt. Unter Umständen kann es sogar sein, dass eine höhere Wiedergabetreue zu unerwünschten Stör-einflüssen führt. Soll etwa die Fehlerdiagnosekompetenz von Kfz-Mechatronikern per Simulation gemessen werden (Abele et al. 2014), ist es nicht konstruktrelevant,

umgebende Werkstattgeräusche, die für reale Settings typisch sein mögen (z. B. Motorgeräusche anderer Autos, Gespräche von Arbeitskollegen) zu simulieren.

6.2.4 Interaktionsgrad

Der Interaktionsgrad beschreibt das Ausmaß, in dem ein Item auf Aktionen der Testperson reagiert, indem sich Bestandteile des Stimulus ändern oder neue Informationen dargeboten werden. Einfache Auswahlaufgaben wie Multiple-Choice-Aufgaben sind wenig interaktiv, da mit einer Aktion die Aufgabe bearbeitet ist. Sowohl die in ► Abschn. 6.2.1 genannten Sortier- und Zuordnungsaufgaben wie auch figurale Auswahlaufgaben, die beispielsweise mit dem Antwortformat „Drag-and-drop“ (Ziehen und Ablegen mit gedrückter Maustaste) computerisiert wurden, sind dagegen etwas interaktiver.

Interaktive Simulationen können als komplexe Auswahlaufgaben verstanden werden, die in mehreren Schritten zu lösen sind, wobei je Schritt zahlreiche alternative Aktionen gewählt werden können und somit die Sequenzierung der Schritte unterschiedlich ausfallen kann. Nach Clark et al. (2009) stellen Simulationen Modelle realer oder hypothetischer Situationen dar, in denen Nutzer die Auswirkungen manipulierter oder modifizierter Modellparameter erfahren. Beispielsweise wird im PISA die Kompetenz, digitale Texte zu lesen, durch simulierte Internetseiten erfasst, d. h., Testpersonen navigieren zwischen Webseiten, um den zu lesenden Text auszuwählen, zu sequenzieren und zu integrieren (s. z. B. Hahnel et al. 2016).

Simulationsbasiertes Assessment

Aufgabe:

In deiner Abteilung hat eine neue Kollegin angefangen. Sie ist noch nicht im allgemeinen E-Mail-Verteiler der Abteilung aufgenommen. Du hast dich deshalb bereit erklärt, ihr wichtige E-Mails weiterzuleiten. Ihre E-Mail-Adresse ist caro.fischer@hfg.de.

Schau nun deine E-Mails durch und sende wichtige E-Mails an Caro weiter.

The screenshot shows a web-based email interface titled "MyMail". At the top, there are navigation icons for "Postfach" (Inbox), "Mail schreiben" (Compose), "Kalender" (Calendar), and "Suchen" (Search). The "Posteingang" (Inbox) is selected. On the left, there's a sidebar with links for "Entwürfe", "Gesendete Objekte", "Spam", and "Papierkorb". The main area displays a list of incoming emails:

Absender	Betreff	Größe
Peter Bär	Suche Datei	24 KB
Markus Höhnle	Mittagessen	8 KB
Marlene Schustc	Petition für den Regenwa	10 MB
Andrea Maur	Kollegin Jessika Beich	5 KB
Alexander Pfeife	Konzert Rosenpark	14 KB

Below the inbox, there are buttons for "Antworten..." (Reply) and "Weiterleiten..." (Forward), both with dropdown menus. The message preview shows:

Von: andrea.maur@hfg.de
An: alle@hfg.de
Betreff: Kollegin Jessika Beich

Liebe Kollegen und Kolleginnen,

kurz zur Info: Jessika Beich ist ab heute eine neue Kollegin bei uns. Ich werde sie später vorstellen.

Viele Grüße,
Andrea Maur

At the bottom, there are icons for "Weiter" (Continue), a monitor, a speech bubble, an '@' symbol, and a globe.

■ Abb. 6.2 Beispiel für eine interaktive Aufgabe zur Messung von Computerfertigkeiten

Herausforderungen interaktiver Items

Die Messung von „Problemlösen“ im PISA 2012 umfasst interaktive Aufgaben, in denen unbekannte Systeme zu manipulieren und zu beobachten sind, um daraus Regeln abzuleiten und sie am Ende zielgerichtet anzuwenden (Greiff et al. 2012). Die Antwortbewertung hängt davon ab, inwieweit es der Testperson gelungen ist, den anzustrebenden Zielzustand des Systems zu erreichen. □ Abb. 6.2 zeigt ein weiteres Beispiel für eine simulationsbasierte Aufgabe, die zur Messung von Computerfertigkeiten ein simuliertes E-Mail-Programm beinhaltet, in dem Funktionen eines realen E-Mail-Programms nachgebildet sind (s. Wenzel et al. 2016).

Ein höherer Interaktionsgrad stellt – vergleichbar der Komplexität – höhere Anforderungen an die Itementwicklung, insofern mehr (multimediale) Inhalte für die möglichen unterschiedlichen Zustände des Items produziert werden müssen. Mit zunehmender Interaktivität können zudem die Anforderungen an die automatische Antwortbewertung steigen. Eine Gefahr interaktiver Items besteht nach Parshall et al. (2010) darin, dass die Testpersonen bei fehlender Beschränkung möglicher Schritte sehr viele Aktionen durchführen und somit potentiell viel Zeit auf falsche Schritte verwenden.

6.2.5 Medienverwendung

Neben Text und Grafik erlauben computerbasierte Items innerhalb des Stimulus und der Antwortoptionen die Einbindung von Medien wie Audio, Video und Animation. Grafiken sind zwar auch auf Papier möglich, aber erst in computerbasierter Form erlauben sie Interaktivität, z. B. das Rotieren oder die Veränderung der Größe. Zudem kann die Interaktion mit computerbasierten grafischen Items direkt automatisch bewertet werden (s. Masters 2010).

Audio

Audiomaterial spielt vor allem bei der Erfassung von Hörverstehen in den Bereichen Sprache und Musik eine wichtige Rolle, kann aber auch zur Erhöhung der Wiedergabetreue simulierter Situationen oder zur Standardisierung der Vorgabe von Instruktionen von Bedeutung sein. Gegenüber separaten Abspielgeräten, die von der Testleitung bedient werden, hat die Testperson am Computer direkt die Möglichkeit, die Lautstärke und ggf. die Abspielzeitpunkte sowie die Abspielhäufigkeit individuell einzustellen. Die Verwendung von Audiomaterialien setzt natürlich voraus, dass das Hörvermögen der Testperson nicht beeinträchtigt ist.

Video und Animation

Videos sind prädestiniert zur Darstellung dynamischer Prozesse mit hohem Realitätsbezug. Ein Beispiel für die Anwendung von Videos ist die Erfassung der professionellen Wahrnehmung von Unterricht (Seidel et al. 2010). Lehramtsstudierende müssen hierzu eine in einem Video präsentierte Unterrichtssituation mit Schüler-Lehrer-Interaktion beurteilen. Videos geben dabei nicht nur die räumliche Situation und das gesprochene Wort wieder, sondern auch die nonverbalen Anteile der Kommunikation. Ein Problem von Videos kann darin bestehen, dass durch ablenkende Information (z. B. Merkmale der Darsteller) konstruktirrelevante Varianz entsteht (Parshall et al. 2010). Wie Videos eignen sich *Animationen* zur Darstellung dynamischer Abläufe. Ein möglicher Vorteil von Animationen besteht darin, dass die wesentlichen Merkmale eines Prozesses fokussiert werden können, d. h., ein weitergehender Realitätsbezug keinen Mehrwert ergibt bzw. unerwünscht ist.

Testerleichterung

Interaktive Multimediaelemente sind für die individuelle Testerleichterung („*Test Accommodation*“, Russell 2011; Sireci et al. 2005) von großer Relevanz. Testpersonen mit bestimmten Beeinträchtigungen (z. B. Seh- oder Hörschwäche) können durch kontrollierbare Medienelemente konstruktirrelevante Barrieren zu den Item- und Testinhalten abbauen. Beispiele dafür sind die Möglichkeit zur Vergrößerung von Bild und Text bei Sehschwäche, die auditive Darbietung von Textinhalten bei Dyslexie oder die videotragtztete Vorgabe von Textinhalten per

Gebärdensprache im Fall von Schwerhörigkeit. Die beiden zuletzt genannten Beispiele betreffen natürlich nur Tests, die kein Leseverständnis messen.

6.2.6 Antworthandlung

Die Dimension „Antworthandlung“ computerbasierter und innovativer Itemformate bezieht sich auf die erforderliche physische Aktion zur Abgabe einer Antwort. Computerbasierte Items verlangen hierbei den Umgang mit Tastatur und Maus sowie zunehmend auch mit Touchbildschirmen (Touchscreens). Die Tastatur spielt vor allem bei Konstruktionsaufgaben mit Text- und Zahleingaben eine Rolle. Ergänzend zur Tastatur kommt die Spracheingabe per Mikrofon infrage (z. B. zur Messung von sprachlichen Kompetenzen). Die Maus kommt vor allem bei Auswahlaufgaben zum Einsatz, bei denen ein grafisches Element anzuklicken oder ein textuelles Element durch Drag-and-drop zu markieren sind. Mit Drag-and-drop werden zudem Sortier- und Zuordnungsaufgaben gelöst. Alternativ zur Maus sind als Eingabegerät beispielsweise Touchpad, Touchscreen, Joystick oder Rollkugel (Trackball) möglich. Bei der Wahl ist die einfache und sichere Handhabbarkeit entscheidend, die wiederum von den Computerfertigkeiten der Testperson bzw. der Vertrautheit und Übung mit einem Eingabegerät abhängt. Jüngere Testpersonen könnten beispielsweise mit Touchscreens stärker vertraut sein als ältere. In jedem Fall ist im Rahmen einer Übung vor Beginn der Aufgaben sicherzustellen, dass sich die Testperson hinreichend mit dem Eingabegerät vertraut gemacht hat.

Manuelle und sprachliche Eingaben

6.2.7 Antwortbewertung

Die automatische Antwortbewertung ist ein zentraler Mehrwert computerbasierter Assessments (vgl. Williamson et al. 2006). Sie ermöglicht beispielsweise die adaptive Durchführung eines Testverfahrens, indem die Beurteilung der Korrektheit der abgegebenen Antworten bereits während der Testbearbeitung erfolgt und unmittelbar den weiteren Testablauf bzw. die Auswahl folgender Aufgaben bestimmt (► Kap. 20). Eine automatische Antwortbewertung (z. B. im Frankfurter Adaptiver Konzentrationsleistungs-Test II, FAKT-II; Moosbrugger und Goldhammer 2007) erlaubt zudem die Erstellung von Testberichten und Rückmeldungen unmittelbar nach der Testung. Wie das Antwortverhalten bewertet werden soll, ist bereits im Rahmen der Itementwicklung zu definieren und – wie die sonstige Gestaltung des Items auch – abhängig von den angestrebten Schlussfolgerungen auf der Basis des Testergebnisses (Mislevy et al. 2002).

Offene/freie vs. geschlossene/gebundene Antwortformate

In Aufgaben mit geschlossenem/gebundenem Antwortformat beschränkt sich die Antwortbewertung oft auf einen einzelnen Aspekt des Antwortverhaltens, z. B. auf die Auswahl einer Antwortalternative in einer Multiple-Choice-Aufgabe, und führt in der Regel zu einer dichotomen (richtig/falsch) oder polytomous (z. B. richtig/teilrichtig/falsch) Bewertung der Korrektheit der Antwort. Innovative Aufgaben, z. B. Simulationen, mit offenem/freiem Antwortformat erlauben es, weitere Informationen über den Antwortprozess (z. B. Zeit, Anzahl der Interaktionen, Interaktionssequenz) zu registrieren, die potentiell zusätzlich für die ggf. abgestufte Bewertung herangezogen werden können, etwa um die Effizienz einer korrekten Antwort zu gewichten. Für die Kombination vielfältiger und zahlreicher messbarer Aspekte des Antwortverhaltens zu einem Testwert wurden komplexe Bewertungsmodelle entwickelt.

Bewertungsmodelle

Regelbasierte Bewertungsmodelle funktionieren nach dem Prinzip von Expertensystemen, d. h., dass eine Menge relativ einfacher, miteinander verknüpfter Regeln auf einzelne Aspekte des Antwortverhaltens synthetisierend angewendet wird mit dem Ziel, das Urteilsverhalten von Experten über das gezeigte Antwortverhal-

Natürliche Sprachverarbeitung

6

ten nachzuahmen (s. Braun et al. 2006). Entsprechend wird ein solch komplexes Regelwerk unter Mitwirkung von Domänenexperten erstellt. In regressionsanalytischen Ansätzen (s. dazu Moosbrugger 2011) wird das mittlere Expertenurteil durch eine Menge von basalen Aspekten des Antwortverhaltens (z. B. in einer Patientensimulation die Anzahl von medizinischen Aktionen, die jeweils eine hohe bzw. eine niedrige Konstruktausprägung anzeigen) vorhergesagt. Auf der Basis eines empirisch begründeten, erklärangstarken Regressionsmodells kann in der Folge für neue Testpersonen aus ihrem Antwortverhalten auf das erwartete Expertenurteil geschlossen werden (vgl. Margolis und Clouser 2006).

Offene schriftsprachliche Antworten (d. h. Aufsätze oder Kurztextantworten), die auf Papier abgegeben und sodann transkribiert oder die gleich am Computer eingegeben werden, lassen sich anhand von Methoden der natürlichen Sprachverarbeitung automatisch bewerten (s. Shermis und Burstein 2003; Zehner et al. 2016). Eine verbreitete Methode ist die latente semantische Analyse (LSA; Landauer et al. 1998), die ähnlich einer Faktorenanalyse (► Kap. 23) einen mehrdimensionalen Raum aufspannt (z. B. basierend auf gemeinsam auftretenden Textinformationen in Wikipedia-Einträgen). In diesen Raum mit mehreren Hundert Dimensionen lassen sich Textantworten als Vektoren projizieren, wobei die räumliche Nähe einer Textantwort zu einer Musterantwort bzw. zu einem Cluster von korrekten Lösungen auf die Korrektheit dieser Antwort schließen lässt (vgl. Zehner et al. 2016). Beispielsweise würde das Wort „Ast“ eine hohe semantische Ähnlichkeit zu „Baum“ ausweisen, nicht jedoch das Wort „Computer“.

6.3 Testentwicklung: Testzusammenstellung und -sequenzierung

6.3.1 Testzusammenstellung in computerisierten Tests

Die Kombination computerbasierter Aufgaben zu Tests, die *Testzusammenstellung*, kann manuell oder automatisiert erfolgen. Automatisierte Verfahren der Testzusammenstellung bieten sich an, wenn aus einer Vielzahl geeigneter Aufgaben („Itempool“) ein oder mehrere Tests so zusammengestellt werden sollen, dass sie vorgegebene Eigenschaften erfüllen, oder wenn mehrere Testformen psychometrisch vergleichbar sein sollen.

Soll beispielsweise aus 100 Mathematikaufgaben ein Test mit 20 Aufgaben erstellt werden, der gleichzeitig eine Testinformationsfunktion besitzt (► Kap. 16), die eine hohe individuelle Testgenauigkeit für leistungsschwache Schüler und Schülerinnen einer bestimmten Jahrgangsstufe besitzt und darüber hinaus eine vorgegebene Anzahl von Aufgaben aus jedem Inhaltsbereich (z. B. Wahrscheinlichkeit, Algebra) enthält, so ist die optimale Lösung des Testzusammenstellungsproblems manuell nur mühsam zu finden (► Exkurs 6.2). Neben den Itemkennwerten werden beispielsweise fachbezogene Anforderungen, die Zuordnung von Aufgaben zu Inhaltsbereichen oder Aufgabeneigenschaften wie das Antwortformat als nicht statistische Kriterien für die Testzusammenstellung berücksichtigt. Die erwünschten Eigenschaften eines Tests werden dann als „Testspezifikation“ (*Blueprint*, vgl. Parshall et al. 2002) bezeichnet und die jeweiligen Zuordnungen der Aufgaben zu den Eigenschaften werden in einer Itembank gespeichert (Vale 2006).

Die Testzusammenstellung kann in Vorbereitung eines computerbasierten Tests oder während der Testdurchführung selbst (online oder „on-the-fly“) erfolgen. Wird die Testzusammenstellung vor der Testdurchführung festgelegt, hat der Testentwickler mehr Kontrolle über die Testzusammenstellung und kann beispielsweise die erstellten Tests einzeln prüfen. Wenn für die Testzusammenstellung die jeweils gegebenen Antworten der Testperson mitberücksichtigt werden, dann

Statistische und inhaltliche Kriterien definieren die Testspezifikation

Optimierung eines Zielkriteriums unter Berücksichtigung zusätzlicher Restriktionen

Exkurs 6.2

Automated Test Assembly (ATA)

Ein vielseitig einsetzbarer Ansatz zur Zusammenstellung von Aufgaben zu Tests wird als ATA bezeichnet (z.B. van der Linden 1998). Dazu wird die Zusammenstellung von Aufgaben für eine Testform als Optimierungsproblem formuliert. Für jede verfügbare Aufgabe i wird eine Entscheidungsvariable $y_i \in \{0, 1\}$ definiert, die mit dem Wert $y_i = 1$ anzeigt, dass eine Aufgabe in die Testform aufgenommen wird; andernfalls hat die Variable den Wert $y_i = 0$. Darauf aufbauend lässt sich die Testzusammenstellung als die Bestimmung der optimalen Werte für alle Entscheidungsvariablen y_i formulieren. Die Optimierung erfolgt im Hinblick auf ein Zielkriterium (*Objective Function*), beispielsweise die Maximierung der Testinformation für einen Fähigkeitswert η_p (oder auch für einen Fähigkeitsbereich, z.B. Leistungsschwäche), wodurch die Messgenauigkeit der Testform für Personen mit einer Fähigkeit von η_p maximiert wird. Dieses zu maximierende Zielkriterium lässt sich als Summe über die Iteminformationsfunktionen (► Kap. 16) für den Wert η_p über alle Aufgaben einer Testform ausdrücken, d.h., das zu optimierende Kriterium ist $k = \sum_i I_i(\eta_p) \cdot y_i$ (s. beispielsweise Kuhn und Kiefer 2013), wobei die Anzahl der Aufgaben festgelegt ist. Die Iteminformation $I_i(\eta_p)$ einer Aufgabe i wird dabei mit der Entscheidungsvariablen y_i multipliziert, d.h. es tragen nur Aufgaben zur Testinformation bei, für die $y_i = 1$ anzeigt, dass sie in der Testform enthalten sind. Als Ergebnis der Optimierung werden die Werte der Entscheidungsvariablen $y_i \in \{0, 1\}$ so bestimmt, dass das Zielkriterium maximiert ist. Im Fall leistungsschwacher Testpersonen führt dies zur Auswahl von eher leichten Aufgaben. Da jede Entscheidungsvariable nur die Werte 0 oder

1 annehmen kann, wird diese Optimierung als *Linear Integer Programming* bezeichnet.

Die Flexibilität dieses Ansatzes zur Testzusammenstellung entsteht u.a. durch die Möglichkeit, diese Optimierung durch weitere Nebenbedingungen einzuschränken. Eine typische Restriktion ist beispielsweise, dass die Testform eine vorgegebene Anzahl p an Aufgaben enthalten soll. Zusätzlich zum Zielkriterium wird dann die Restriktion formuliert, dass die Summe der Entscheidungsvariablen der Anzahl der Aufgaben p entsprechen soll: $\sum_i y_i = p$. Betrachtet werden dann nur Konstellationen der Entscheidungsvariablen y_i , die zu einer Testform mit p Aufgaben führen.

Eine Vielzahl weiterer Restriktionen können in diesem Rahmen formuliert und mithilfe der Optimierung gleichzeitig gelöst werden. Dazu zählt beispielsweise die Berücksichtigung von Aufgabenmerkmalen bei der Testzusammenstellung, die zur Erfüllung einer vorgegebenen Testspezifikation notwendig sind (z.B. Anzahl der Aufgaben je Inhaltsbereich entsprechend eines Blueprints). Aber auch wechselseitige Abhängigkeiten von Items, die nicht gemeinsam in einer Testform auftreten dürfen, weil sie zu derselben Itemfamilie gehören (*Enemies*), können bei der automatisierten Testzusammenstellung berücksichtigt werden (Kuhn und Kiefer 2013). Durch Anpassung des Zielkriteriums ist es schließlich möglich, auch Testformen zusammenzustellen, die eine bestimmte Form der Testinformationsfunktion aufweisen und somit optimal für die Messung einer Zielpopulation sind. Das Verfahren ist so flexibel, dass es auch für adaptives Testen unter Berücksichtigung von Restriktionen der Testzusammenstellung verwendet wird (*Shadow Testing*, s. van der Linden 2005).

spricht man allgemein vom adaptiven Testen (► Kap. 20). Die Anzahl der ausgewählten Items kann dabei von Item-by-Item-adaptiven Tests (jeweils nur ein Item wird ausgewählt) bis hin zu computerbasierten Multistage-Tests (MST) mit individuell angepassten Stufen (Han und Guo 2014) variieren.

6.3.2 Navigation in computerbasierten Instrumenten

Um in einem computerbasierten Instrument von Item zu Item zu wechseln, d.h. zu navigieren (vgl. Parshall et al. 2002), werden in der computerbasierten Benutzerschnittstelle spezielle Elemente benötigt. Die Umsetzung der Navigation richtet sich dabei nach den Eigenschaften der Fragen und Aufgaben. Instruktionsseiten zu Beginn eines Instruments sollen häufig nur genau einmal angesehen werden können, d.h., die Navigation zu einer bereits gezeigten Instruktionsseite wird nicht ermöglicht. Im Gegensatz dazu kann innerhalb von Items, die eine Unit-Struktur bilden, also beispielsweise mehrere Leseaufgaben zu einem gemeinsamen Textstimulus, in der Regel frei angesteuert werden. Dazu kann der Testteilnehmer beispielsweise mit Schaltflächen oder einer Navigationsleiste zu vorangegangenen Fragen innerhalb der Unit zurückkehren.

Freie vs. eingeschränkte Navigation**6**

Weitere Eigenschaften der Navigation betreffen beispielsweise die Möglichkeit, bereits bearbeitete Aufgaben für eine spätere Kontrolle zu markieren oder Aufgaben zu überspringen. Insbesondere bei computerisierten adaptiven Tests (► Kap. 20), bei denen die Antworten der Testperson für die Auswahl der folgenden Aufgaben berücksichtigt werden, wird die Navigation zu vorherigen Aufgaben häufig technisch unterbunden (s. z. B. Wise et al. 1999). Testteilnehmer bevorzugen die Möglichkeit, frei innerhalb von computerbasierten Instrumenten zu navigieren (Luecht und Sireci 2011); doch diese Möglichkeit wird vergleichsweise selten verwendet, wenn sie angeboten wird. Da die Navigation innerhalb computerbasierter Instrumente zusätzliche Testzeit benötigt, werden in der Praxis häufig nur eingeschränkte Navigationsmöglichkeiten angeboten. Wie Parshall et al. (2002) beschreiben, ist es dabei wichtig,

- a. dass die Navigationsmöglichkeiten konsistent sind, damit Testteilnehmer den Umgang mit dem computerbasierten Instrument anhand einer Instruktion leicht lernen können;
- b. darüber hinaus muss das computerbasierte Instrument Informationen bereitstellen, z. B. zur Orientierung innerhalb des Tests oder Fragebogens;
- c. schließlich empfehlen sie, dass Warnhinweise zur Verfügung gestellt werden, wenn Aktionen nicht rückgängig gemacht werden können (*Informative For-giveness*).

Derartige Warnhinweise für Aktionen, die nicht rückgängig gemacht werden können (die beispielsweise beim Verlassen einer Unit gegeben werden, wenn keine Rückwärtsnavigation mehr möglich ist), sind nur ein Beispiel für Feedback während der Aufgabenbearbeitung. Weitere Optionen für Feedback betreffen beispielsweise Informationen über die verbleibende Bearbeitungszeit, über fehlende Antworten oder die Anzahl noch zu bearbeitender Aufgaben (s. auch ► Abschn. 6.5.2).

6.3.3 Sequenzierung computerisierter Fragebogen

Sprung- und Filterregeln

Computerbasiert administrierte Fragebogen erlauben die Umsetzung von Sprung- und Filterregeln (vgl. z. B. Saris 1991). Nicht zutreffende Fragen können bei der computerbasierten Ablaufsteuerung (*Sequenzierung*) von Fragebogen in Abhängigkeit von vorausgehenden Antworten automatisch ausgeschlossen werden („filtering“). Ebenso ist die Verzweigung („branching“) in unterschiedliche Abschnitte eines computerisierten Instruments möglich, um unpassende Fragen zu vermeiden. So wäre es beispielsweise nicht zielführend, einer Person vertiefende Fragen zu ihren Kindern zu stellen, wenn sie angeben hat, dass sie keine hat. Aufwendige und potentiell fehleranfällige Instruktionen an die Zielpersonen zum Überspringen von Fragen oder Fragenblöcken (s. z. B. Brace 2008) können bei einer computerisierten Ablaufsteuerung vermieden werden. Je komplexer ein Fragebogen ist, desto aufwendiger ist entsprechend die Überprüfung des computerbasierten Instruments (vgl. z. B. Rölke 2012a). Wenn Sprung- und Filterregeln verwendet werden, werden computerbasierte Fragebogen, die von den Zielpersonen selbst ausgefüllt werden, häufig mit jeweils nur einer Frage pro Seite (One-Item-one-Screen, OIOS; Reips 2010) administriert.

6.4 Testadministration

Fragebogen und Tests können in Abhängigkeit von der diagnostischen Zielsetzung und den organisatorischen Rahmenbedingungen auf unterschiedliche Weise administriert werden. Nach Mislevy et al. (2012; s. auch Mislevy et al. 2003) lassen sich zur Strukturierung der Vielfalt in der computerbasierten Testadministration (auch

Testauslieferung, „delivery“) folgende vier Prozesse abstrahieren und entsprechende Infrastrukturkomponenten unterscheiden:

- Aktivitätsauswahl
- Präsentation
- Evidenzidentifikation
- Evidenzakkumulation

6.4.1 Aktivitätsauswahl

Der Prozess der Aktivitätsauswahl betrifft die Frage, wie der Test abläuft bzw. wie passende Items, die von der Testperson zu bearbeiten sind, nacheinander ausgewählt werden (► Abschn. 6.3), ggf. unter Berücksichtigung schon vorhandenen Wissens über die Person, also beispielsweise des bisherigen Antwortmusters beim adaptiven Testen. Auszuwählende Items können in einer Itemdatenbank fertig vorliegen oder anhand einer Aufgabenvorlage (*Template*) automatisch generiert werden (► Beispiel 6.4), womit in effizienter Weise die Produktion von zahlreichen Aufgabenvarianten ermöglicht wird (Gierl und Lai 2013).

Automatische Itemgenerierung

Beispiel 6.4: Aktivitätsauswahl über Templates

Für das Beispiel der quadratischen Gleichung $x^2 = 120 + 24$ definiert ein solches Template den Aufgabenstamm (z. B. „Bestimme den Wert von x in der folgenden Gleichung“), ein Modell ($x^2 = A$ Operator B), die variierten Elemente (Zahlen A und B, Operator + oder -) und die Restriktionen bezüglich der Variation (Lösungen nur ganzzahlig) sowie dynamische Text-, Zahl- und Bildteile, die im Zuge der (ggf. automatischen) Generierung für einzelne Aufgaben durch konkrete Werte der in Modellform beschriebenen Templates ersetzt werden. Ist durch ein empirisch geprüftes Schwierigkeitsmodell bekannt, wie die Werte dieser Aufgabenmerkmale die Schwierigkeit beeinflussen, können auf dieser Basis (neue) Items mit gewünschter Schwierigkeit generiert werden (s. z. B. Holling et al. 2009; Sinharay und Johnson 2013).

6.4.2 Präsentation

Der Prozess der Präsentation bezieht sich auf die Darbietung der Items. Bei computerbasierten Tests betrifft die Präsentation u. a. die Wahl des Geräts (Desktop-PC, Laptop, Tablet, Smartphone; s. dazu auch den ► Exkurs 6.3). Außerdem sind aus technischer Sicht Offline- und Onlineauslieferungen zu unterscheiden. Letztere setzen voraus, dass die Internetverbindung während der gesamten Testbearbeitung sicher und ausreichend leistungsstark zur Verfügung steht und der Test hinsichtlich des Layouts und der Funktionalität korrekt auf dem Browser des Endgeräts präsentiert wird (Browserkompatibilität). Bei Erfüllung der technischen Voraussetzungen bestehen Vorteile insbesondere darin, dass Tests räumlich und zeitlich flexibel bearbeitet werden können und das Datenmanagement erleichtert wird, in dem alle Antworten der Testpersonen zentral auf einem Server gesammelt und zur Weiterverarbeitung bereitgestellt werden können.

Offline- vs. Online-Anlieferung

Exkurs 6.3**Ambulantes Assessment**

Eine besondere Form der Itempräsentation zur Verlaufsmessung ist das *ambulante Assessment*, das üblicherweise computerbasiert mithilfe elektronischer Geräte, z. B. Smartphones durchgeführt wird (vgl. Ebner-Priemer et al. 2009). Messungen anhand von Tests, Selbstberichten, Biosensoren etc. werden „ambulant“ durchgeführt, um tatsächliches Verhalten und Erleben in alltäglichen Situationen und in Echtzeit erfassen zu können. Dies soll zu einem besseren Verständnis beitragen, wie sich psychologische, biologische und soziale Prozesse in einer natürlichen Umgebung zeitlich entwickeln (Ebner-Priemer et al. 2009).

Ein klassisches Beispiel ist die *Experience-Sampling-Methode* (Larson und Csikszentmihalyi 1983), nach der Personen an zufällig ausgewählten Gelegenheiten während der Wachzeiten einer Alltagswoche systematisch Selbstberichte abgeben. Der Hinweis auf die Messung („prompting“) kommt von einem elektronischen Gerät bzw. Zeitgeber nach einem experimentellen Design. Ein aktuelles Beispiel für die Anwendung eines ambulanten Assessments ist die Studie von Dirk und Schmiedek (2016), in der alltägliche Schwankungen der Arbeitsgedächtnisleistung und ihr Zusammenhang mit der Schulleistung über mehrere Wochen untersucht wurden. Die Schüler und Schülerinnen bearbeiteten dazu an drei vordefinierten Gelegenheiten in der Schule und zu Hause (morgens, mittags, nachmittags) auf den ihnen zur Verfügung gestellten Smartphones mehrere Blöcke mit Arbeitsgedächtnisaufgaben.

Das intensive längsschnittliche Design erlaubte es, die intraindividuelle Fluktuation der Arbeitsgedächtnisleistung in verschiedene Komponenten zu zerlegen, und zwar in Tag, Gelegenheit und Aufgabenblock, wobei individuelle Unterschiede aller drei Komponenten mit geringerer Schul- und Intelligenzleistung assoziiert waren.

Neben klassischen Verhaltensdaten aus der Fragebogen- und Testbearbeitung können durch Technologieeinsatz auch andere Arten von Daten kontinuierlich im Alltag gesammelt werden. Dazu zählen insbesondere (Verlaufs-)Messungen physiologischer und physischer Variablen (z. B. Herzfrequenzrate, Hautleitfähigkeit, Körpertemperatur, motorische Aktivität). In einer Studie von Kühnhausen et al. (2013) wurde beispielsweise untersucht, wie sich die körperliche Aktivität im Alltag von Kindern auf deren Affekt auswirkt. Dazu trugen die Grundschulkinder an der Taille einen Beschleunigungssensor (Akzelerometer), der die motorische Aktivität in Bezug auf alle drei Raumachsen mit einer Frequenz von 30 Hz erfasste. Auf diese Weise konnte das Bewegungsverhalten kontinuierlich erfasst und nach unterschiedlichen Aktivitäten, d. h. Liegen, Sitzen, Stehen, Gehen und Laufen, differenziert werden. Ähnlich dazu untersuchten Gawrilow et al. (2016) unter Verwendung eines Schrittmessers (Pedometer), inwieweit die körperliche Aktivität im Alltag den Affekt sowie exekutive Funktionen bei Kindern mit Symptomen einer Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (ADHS) beeinflusst.

6.4.3 Evidenzidentifikation

Der Prozess der Evidenzidentifikation bezieht sich auf die Verarbeitung des Antwortverhaltens bzw. der abgegebenen Antwort. Im Rahmen computerbasierter Tests kann die Antwortbewertung (d. h. die Beurteilung der Evidenz im Sinne des interessierenden Merkmals) je nach Antworttyp automatisch vorgenommen werden, indem eine Antwort mit einem Kodierschlüssel verglichen wird (► Abschn. 6.2.7). Das Itemergebnis kann zu Rückmeldezwecken auf Aufgabenebene verwendet werden (► Abschn. 6.5.2) und/oder in die Berechnung eines Testwertes eingehen.

6.4.4 Evidenzakkumulation

Der Prozess der Evidenzakkumulation bezieht sich auf die Aggregation der Bewertungen auf Itemebene, die von einer einfachen Aufsummierung (► Kap. 8, bzw. 9) bis hin zur Anwendung komplexer psychometrischer Modelle (► Kap. 12, 13, 16 und 18) reichen kann (s. auch ► Abschn. 6.5).

6.5 Datenanalyse und Rückmeldung

6.5.1 Daten und Analysepotential

Computerbasiertes Assessment kann das Analysepotential von Testdaten beträchtlich erweitern. Nach der Durchführung eines Tests stehen nicht nur die – manuell oder automatisch kodierten – Antworten (z. B. richtig vs. falsch) einer Person zur Verfügung, sondern auch alle Interaktionen der Person mit dem Assessmentsystem (d. h., wann welcher Schritt gemacht wurde, z. B. Klicks, Texteingaben, Markierungen etc.). Solche sog. „Prozessdaten“ werden automatisch generiert und in Protokolldateien („Logfiles“) abgespeichert (► Studienbox 6.1 mit Fallbeispiel).

Prozessdaten (Bearbeitungszeiten, Bearbeitungsschritte inklusive ihrer Sequenz) spiegeln also das Bearbeitungsverhalten wider und erlauben potentiell

Ergebnis- und Prozessdaten

Studienbox 6.1

Die □ Abb. 6.3 stellt exemplarisch dar, welche Prozessdaten bei der Bearbeitung der Beispieldaufgabe „My Mail“ (vgl. □ Abb. 6.2) in einem Logfile abgespeichert werden können. Die linke Spalte zeigt im XML-Format die durch die Testperson ausgelösten Ereignisse (*Events*) inklusive Zeitpunkt, Name, Typ und ereignisspezifische Werte. Die rechte Spalte zeigt, wie diese Information den Bearbeitungsprozess beschreibt. Das zweite Ereignis besteht beispielsweise im Öffnen des Posteingangs per Doppelklick. Aus dem Logfile lassen sich weitere Informationen ableiten, z. B. ob eine E-Mail geöffnet wurde und – wenn ja – wie lange sie offen war bzw. gelesen wurde (z. B. ergibt sich für die E-Mail von Marlene Schuster aus der Differenz der Zeitstempel 10:37:58.143 und 10:37:55.934 eine Zeit von etwas mehr als 2 Sekunden).

Log-Event	Prozessinformation
<pre><logEntry timeStamp="2017-02-06T10:37:53.276+0100" xsi:type="cbaloggingmodel:LogEntryTimeStamp"></logEntry> <xsi:type="cbaloggingmodel:TreeNodeLogEntry" nodeName="Posteingang" nodeType="Posteingang" nodePathId="xtr100_xtn100-0" treeUserDefinedId="I074422.xtr100" operation="Selection" id="I074422.cbaTree_330_23566721280649"/> </logEntry></pre>	Auswahl Ordner „Posteingang“
<pre><logEntry timeStamp="2017-02-06T10:37:53.550+0100" xsi:type="cbaloggingmodel:LogEntryTimeStamp"></logEntry> <xsi:type="cbaloggingmodel:TreeNodeLogEntry" nodeName="Posteingang" nodeType="Posteingang" nodePathId="xtr100_xtn100-0" treeUserDefinedId="I074422.xtr100" operation="DoubleClick" id="I074422.cbaTree_330_23566721280649"/> </logEntry></pre>	Doppelklick zum Öffnen des Ordners „Posteingang“
<pre><logEntry timeStamp="2017-02-06T10:37:55.934+0100" xsi:type="cbaloggingmodel:LogEntryTimeStamp"></logEntry> <xsi:type="cbaloggingmodel:TreeViewNodeLogEntry" nodeName="Marlene Schuster" nodeType="Mail" nodePathId="xtr100_xtn100-0_xtn103-2" operation="Selection" id="I074422.cbaTreeViewerExplore_353_23567883283263" treeViewUserDefinedId="I074422.xtv100"/> </logEntry></pre>	Anklicken/Öffnen der E-Mail von Marlene Schuster, E-Mail erscheint im Anzeigefenster
<pre><logEntry timeStamp="2017-02-06T10:37:58.143+0100" xsi:type="cbaloggingmodel:LogEntryTimeStamp"></logEntry> <xsi:type="cbaloggingmodel:TreeViewNodeLogEntry" nodeName="Andrea Maur" nodeType="Mail" nodePathId="xtr100_xtn100-0_xtn104-3" operation="Selection" id="I074422.cbaTreeViewerExplore_353_23567883283263" treeViewUserDefinedId="I074422.xtv100"/> </logEntry></pre>	Anklicken/Öffnen der E-Mail von Andrea Maur, E-Mail erscheint im Anzeigefenster

□ Abb. 6.3 Logfiledatenauszug der Bearbeitung der Beispieldaufgabe „My Mail“

Rückschlüsse auf zugrunde liegende kognitive Prozesse. Das heißt, man erfährt durch sie nicht nur, ob jemand bei einer Aufgabe Erfolg hatte oder nicht, sondern erhält auch Anhaltspunkte, wie der erfolgreiche Lösungsweg aussah bzw. an welcher Stelle eine Person wahrscheinlich gescheitert ist. Prozessdaten können somit genutzt werden, um sowohl allgemein das Verständnis des Antwort- und Lösungsprozesses zu verbessern als auch entsprechende individuelle Unterschiede zu identifizieren (z. B. Goldhammer et al. 2017; Greiff et al. 2016).

Ob kognitive Prozesse mittelbar über Logfiledaten beobachtbar sind, hängt u. a. vom Interaktionsgrad des Itemtyps bzw. der Art der geforderten Interaktionen ab (► Abschn. 6.2.4). Bei einem traditionellen geschlossenen Antwortformat, z. B. bei Multiple-Choice-Aufgaben, lassen sich nur Prozessdaten generischer Natur sammeln wie die Bearbeitungszeit, die Antwortänderung und der Wiederbesuch eines Items. Bei stärker interaktiven Items, die beispielsweise Simulationen präsentieren (z. B. einer Computerumgebung oder eines naturwissenschaftlichen Experiments), erhält man detailliertere Informationen über den individuellen Bearbeitungsprozess, beispielsweise über unterschiedliche Lösungsstrategien. Die Frage der Bedeutung von generischen Prozessindikatoren für den Bearbeitungserfolg haben Goldhammer et al. (2014b) untersucht. Sie konnten zeigen, dass der Effekt der Bearbeitungszeit auf den Aufgabenerfolg davon abhängt, wie schwer die Aufgabe und wie fähig die Person ist. Bei schweren Problemlöseaufgaben, die von schwachen Problemlösern bearbeitet wurden, zeigte sich ein positiver Effekt, während bei leichten Leseverständnisaufgaben, die von starken Lesern bearbeitet wurden, der Effekt negativ war. Bearbeitungszeiten erlauben zudem die Ableitung individueller Unterschiede in Bezug auf die Geschwindigkeit und können zu einer höheren Genauigkeit der Fähigkeitsschätzung beitragen (z. B. Klein Entink, Fox & van der Linden 2009). Generelle Bedeutung kommt Bearbeitungszeiten auch für die Qualitätssicherung zu, insofern sehr kurze Antwortzeiten als Indikator für unmotiviertes Antwortverhalten genutzt werden können (Kong et al. 2007, s. auch ► Kap. 4).

Im Bereich computergestützter Umfragen, z. B. im Computer Assisted Personal Interview (CAPI), werden Prozessdaten üblicherweise unter den Begriff der „Paradaten“ gefasst (Kreuter 2013). Darunter versteht man zusätzlich anfallende Daten, die den kompletten Prozess der Umfrage beschreiben. Dazu zählen Logfiledaten des Umfragesystems, aber auch darüber hinausgehende Daten, beispielsweise zum Kontakt mit der Befragungsperson oder auch Beobachtungsdaten des Interviewers. Aus Logfiles gewonnene Bearbeitungszeiten werden beispielsweise genutzt, um den kognitiven Antwortprozess in Einstellungsmessungen zu untersuchen (Bassili und Fletcher 1991).

6.5.2 Rückmeldung von Testdaten

Die automatische, zeitnahe und individuelle Rückmeldung von Testdaten an potentiell viele Personen gleichzeitig spielt vor allem im Bereich des Beratens und Lernens eine wichtige Rolle. Für Beratungszwecke werden oftmals sog. „Self-Assessments“ eingesetzt, die beispielsweise Orientierung bei der Entscheidung für einen Studiengang gemäß persönlicher Eignungen und Neigungen geben sollen und gleichzeitig über die Anforderungen der Universität informieren (Hornke et al. 2013). Self-Assessments werden üblicherweise online durchgeführt und bieten nach der Bearbeitung individuelle Rückmeldung über die eigenen Testwerte, die beispielsweise in Bezug zu einer Vergleichsgruppe interpretiert werden (s. z. B. Reiß und Moosbrugger 2008). Diese Art von Assessment kann als „summativ“ verstanden werden, insofern die Eignung als Ergebnis einer bisherigen Entwicklung festgestellt wird. Demgegenüber ist im Bereich des Lernens das „formative Assessment“ von großer Bedeutung, das darauf abzielt, durch die Rückmeldung von Testergebnissen an Lernende und/oder Lehrende die Lernentwicklung unmit-

Bearbeitungszeiten

6

Paradaten

Summatives und formatives Assessment

6.6 · Zusammenfassung

telbar positiv zu beeinflussen (z. B. Russel 2010; für Beispiele s. z. B. [Electronic]-Assessment Tools for Teaching and Learning, E-asTTle, Visible Learning Lab 2010; Lernverlaufsdiagnostik quop, Souvignier et al. 2014). Die Rückmeldung mit Informationen zu individuellen Stärken und Schwächen sowie lernförderlichen Hinweisen kann auf Grundlage des Antwortmusters im Test gegeben werden.

Andere Assessments weisen einen eher tutoriellen Charakter auf und erlauben (prozessorientierte) Rückmeldung auf der Ebene von Aufgaben. Eine Form des „computerbasierten Tutorings“ besteht darin, dass Lernende nach der Bearbeitung einer Aufgabe eine Rückmeldung und ggf. Hinweise zu ihrer Antwort erhalten (sog. „Computer Aided-Instruction“; VanLehn 2011; Waalkens et al. 2013). *Intelligent Tutorielle Systeme* (ITS) ermöglichen die Auswertung einzelner Schritte innerhalb des Lösungsprozesses, die Rückmeldung zu einzelnen Lösungsschritten, Hinweise für den folgenden Lösungsschritt sowie eine individuelle Aufgabenauswahl (z. B. Aleven et al. 2010; Koedinger und Aleven 2007; VanLehn 2011).

6.6 Zusammenfassung

Das Kapitel enthält einen Überblick, wie mithilfe von Computern im weiteren Sinne Tests und Fragebogen realisiert und dabei die Möglichkeiten von klassischen Paper-Pencil-Verfahren erweitert bzw. deutlich überschritten werden können. Dies betrifft beispielsweise die Entwicklung computerbasierter Items mit innovativen Antwortformaten und multimedialen Stimuli sowie die automatische Bewertung des gezeigten Antwortverhaltens. Des Weiteren ermöglicht der Computer eine flexiblere Testzusammenstellung, d. h., Items können automatisch unter Berücksichtigung inhaltlicher und statistischer Kriterien sequenziert werden. Außerdem behandelt wurde die Frage, wie durch Logfiledaten das Analysepotential gesteigert und durch die automatische und zeitnahe Rückmeldung von Testdaten beispielsweise das Lernen unterstützt werden kann. Das Kapitel schließt mit Hinweisen auf einschlägige und frei zugängliche Softwarelösungen für Assessmentzwecke.

6.7 EDV-Hinweise

Funktionale Itembeispiele zur Taxonomie computerbasierter Itemformate von Scalise und Gifford (2006) finden sich auf folgender Internetseite: ► <http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html>.

Der *CBA ItemBuilder* ist ein Autorenwerkzeug zur Erstellung von Aufgaben für computerbasierte Assessments (► <http://tba.dipf.de/de/>; s. Rölke 2012b). Die grafische Benutzeroberfläche ermöglicht es Nutzern, ohne Programmiererfahrung oder Kenntnis spezieller Beschreibungssprachen, komplexe, interaktive und multimediale Items umzusetzen. Die interaktive Beispielaufgabe in □ Abb. 6.2 wurde mit dem *CBA ItemBuilder* erstellt und auch der zugehörige Logfiledatenauszug (□ Abb. 6.3) stammt daraus.

Um im Rahmen einer automatischen Testzusammenstellung ein mithilfe von Entscheidungsvariablen formuliertes Optimierungsproblem zu lösen, werden sog. „Solver“ verwendet. Diese können beispielsweise mithilfe des R-Pakets *lp_Solve* eingesetzt werden (Diao und Linden 2011). Einfache Probleme lassen sich auch mit *Excel* lösen (Cor et al. 2009).

Das Softwaresystem *TAO* (*Testing Assisté par Ordinateur* = computerbasiertes Testen) stellt eine generische Open-Source-Plattform dar, die umfassende Funktionen zur kollaborativen und verteilten Entwicklung, Steuerung und Bereitstellung von computerbasierten Tests bietet (► <https://www.taotesting.com/>).

6.8 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Welche Vorteile haben computerbasierte Testverfahren im Vergleich zu Papier- und Bleistift-Verfahren?
2. Was ist bei der Wiedergabebreue von Testitems zu beachten?
3. Bei welchen Antwortformaten ist eine computerbasierte automatische Auswertung besonders hilfreich?
4. Was ist bei der Gestaltung von (eingeschränkten) Navigationsmöglichkeiten in computerbasierten Tests zu beachten?
5. Welche Vorteile bieten ambulante gegenüber klassischen Assessments?
6. Wofür können Logfiledaten aus computerbasierten Assessments genutzt werden?

Literatur

- Abele, S., Walker, F. & Nickolaus, R. (2014). Zeitökonomische und reliable Diagnostik beruflicher Problemlösekompetenzen bei Auszubildenden zum Kfz-Mechatroniker. *Zeitschrift für Pädagogische Psychologie*, 4, 167–179.
- Aleven, V., Roll, I., McLaren, B. & Koedinger, K. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educational Psychologist*, 45, 224–233.
- Bassili, J. N. & Fletcher, J. F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes, *Public Opinion Quarterly*, 55, 331–346.
- Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research* (2nd ed). London; Philadelphia: Kogan Page.
- Braun, H., Bejar, I. I. & Williamson, D. M. (2006). Rule-based methods for automated scoring: Application in a licensing context. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Mahwah, NJ: Lawrence Erlbaum Associates.
- Chen, M., Mao, S. & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19, 171–209.
- Clark, D., Nelson, B., Sengupta, P. & D’Angelo, C. (2009). *Rethinking science learning through digital games and simulations: Genres, examples, and evidence. An NAS commissioned paper*. Retrieved from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_080068.pdf [20.12.2019]
- Cor, K., Alves, C. & Gierl, M. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research & Evaluation*, 14. Retrieved from <http://pareonline.net/getvn.asp?v=14%26n=14> [20.12.2019]
- Deroos, D., Deutsch, T., Eaton, C., Lapis, G. & Zikopoulos, P. (2012). *Understanding big data: analytics for enterprise class Hadoop and streaming data*. New York, NY: McGraw-Hill.
- Diao, Q. & van der Linden, W. J. (2011). Automated Test Assembly Using lp_Solve Version 5.5 in R. *Applied Psychological Measurement*, 35, 398–409.
- Dirk, J. & Schmiedek, F. (2016). Fluctuations in elementary school children’s working memory performance in the school context. *Journal of Educational Psychology*, 108, 722–739.
- Ebner-Priemer, U. W., Kubiak, T. & Pawlik, K. (2009). Ambulatory Assessment. *European Psychologist*, 14, 95–97.
- Frey, A. & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von Papier- und Bleistift-basierten Verfahren eingesetzt werden? *Zeitschrift für Erziehungswissenschaft*, 16, 53–57.
- Funke, U. (1995). Using complex problem solving tasks in personnel selection and training. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 219–240). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gawrilow, C., Stadler, G., Langguth, N., Naumann, A. & Boeck, A. (2016). Physical activity, affect, and cognition in children with symptoms of ADHD. *Journal of Attention Disorders*, 20, 151–162.
- Gierl, M. J. & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32, 36–50.
- Goldhammer, F., Kröhne, U., Keßel, Y., Senkbeil, M. & Ihme, J. M. (2014a). Diagnostik von ICT-Literacy: Multiple-Choice- vs. simulationsbasierte Aufgaben. *Diagnostica*, 60, 10–21.

Literatur

- Goldhammer, F., Naumann, J., Rölke, H., Stelter, A. & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Eds.), *Competence Assessment in Education: Research, Models and Instruments* (pp. 407–425). Berlin, Heidelberg: Springer.
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H. & Klieme, E. (2014b). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106, 608–626.
- Greiff, S., Niepel, C., Scherer, R. & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46.
- Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement*, 36, 189–213.
- Hahnel, C., Goldhammer, F., Naumann, J. & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior*, 55, 486–500.
- Han, K. T. & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier & Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119–133). New York, NY: CRC Press.
- Holling, H., Bertling, J. P. & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, 35, 71–76.
- Hornke, L. F., Wosnitza, M. & Bürger, K. (2013). SelfAssessment: Ideen, Hintergründe, Praxis und Evaluation, *Wirtschaftspsychologie*, 15, 5–16.
- Klein Entink, R. H., Fox, J.-P. & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21–48.
- Koedinger, K. R. & Aleven V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19, 239–264.
- Kong, X., Wise, S. L. & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, 67, 606–619.
- 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.
- Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information* (Vol. 581). New York, NY: John Wiley & Sons.
- Kroehne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, 14, 169–186.
- Kuhn, J.-T. & Kiefer, T. (2013). Optimal test assembly in practice: The design of the Austrian Educational Standards Assessment in Mathematics. *Zeitschrift für Psychologie*, 221, 190–200.
- Kühnhausen, J., Leonhardt, A., Dirk, J. & Schmiedek, F. (2013). Physical activity and affect in elementary school children's daily lives. *Frontiers in Movement Science and Sport Psychology*, 4, 456.
- Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259–284.
- Larson, R. & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction. New directions for methodology of social and behavioral sciences* (pp. 41–56). San Francisco, CA: Jossey-Bass.
- Luecht, R. L. & Sireci (2011). *A review of models for computer-based testing*. Research report 2011–2012. New York, NY: The College Board.
- Margolis, M. J. & Claußer, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–168). Mahwah, NJ: Lawrence Erlbaum Associates.
- Masters, J. (2010). Automated Scoring of an Interactive Geometry Item: A Proof-of-Concept. *Journal of Technology, Learning, and Assessment*, 8. Retrieved from <https://ejournals.bc.edu/index.php/jtla/article/view/1626> [20.12.2019]
- Mead, A. D. & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*, 114, 449–458.
- Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, 178, 107–114.
- Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03-16). Princeton, NJ: Educational Testing Service.
- Mislevy, R. J., Behrens, J. T., Dicerbo, K. E. & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4, 11–48.
- Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G. & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363–389.
- Moosbrugger, H. (2011). *Lineare Modelle. Regressions- und Varianzanalysen* (4. Aufl.). Bern: Huber.

- Moosbrugger, H. & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test II. Computerprogramm. Grundlegend neu bearbeitete und neu normierte 2. Auflage des FAKT von Moosbrugger und Heyden (1997)*. Bern: Huber.
- Organisation for Economic Co-operation and Development (OECD). (2011). *PISA 2009 results Vol VI: Students on line – Digital technologies and performance*. Paris: PISA, OECD Publishing. Retrieved from <https://doi.org/10.1787/9789264112995-en> [20.12.2019]
- Organisation for Economic Co-operation and Development (OECD). (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Paris: OECD Publishing.
- Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014). Paris: PISA, OECD Publishing. Retrieved from <https://doi.org/10.1787/9789264201118-en> [20.12.2019]
- Parshall, C. G., Harmes, J. C., Davey, T. & Pashley, P. J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). New York, NY: Springer.
- Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.
- Richman-Hirsch, W. L., Olson-Buchanan, J. B. & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, 85, 880–887.
- Reiß, S. & Moosbrugger, H. (2008) *Online Self Assessment Psychologie*. Institut für Psychologie der Goethe-Universität Frankfurt am Main. https://www.psychologie.uni-frankfurt.de/49829947/20_self-Assessment [20.12.2019]
- Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. Gosling & J. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 29–43). Washington, DC: American Psychological Association.
- Robitzsch, A., Lüdtke, O., Kölner, O., Kröhne, U., Goldhammer, F. & Heine, J. H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica*, 63, 148–165.
- Rölk, H. (2012a). Automata and Petri Net Models for visualizing and analyzing complex questionnaires – A case study. In S. Donatelli, J. Kleijn, R. J. Machado & J. M. Fernandes (Eds.) *CEUR Workshop Proceedings* (Vol. 827, p. 317–329). Braga: CEUR-WS.org. Retrieved from https://www.researchgate.net/publication/220852463_Automata_and_Petri_Net_Models_for_Visualizing_and_Analyzing_Complex_Questionnaires_A_Case_Study [20.12.2019]
- Rölk, H. (2012b). The Item Builder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 344–353). Chesapeake, VA: AACE. Retrieved from <http://www.editlib.org/p/41614> [20.12.2019]
- Russel, M. (2010). Technology-aided formative assessment and learning: New developments and applications. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 125–138). New York, NY: Routledge.
- Russell, M. (2011). Computerized tests sensitive to individual needs. In S. N. Elliott, R. J. Kettler, P. A. Beddow & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students*. (pp. 255–273). New York, NY: Springer.
- Saris, W. E. (1991). *Computer-assisted interviewing*. Newbury Park, Calif: Sage Publications.
- Scalise, K. (2012). Creating innovative assessment items and test forms. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 134–156). Charlotte, NC: Information Age Publisher.
- Scalise, K. & Gifford, B. R. (2006). Computer-based assessment in E-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *Journal of Teaching, Learning and Assessment*, 4(6), 1–45.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8: e73791. <https://doi.org/10.1371/journal.pone.0073791>
- Seidel, T., Blomberg, G., Stürmer, K. (2010). „Observer“ – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. *Zeitschrift für Pädagogik*, 56, 296–306.
- Shah, D. V., Cappella, J. N. & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *Annals of the American Academy of Political and Social Science*, 659, 6–13.
- Shermis, M. D. & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Sinharay, S. & Johnson, M. S. (2013). Statistical modeling of automatically generated items. In M. J. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 183–195). New York, NY: Routledge.
- Sireci, S. G., Li, S. & Scarpati, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457–490.

Literatur

- Sireci, S. G. & Zenisky, A. L. (2006). *Innovative item formats in computer-based testing: In pursuit of improved construct representation*. In S. M. Downing & Th. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum Associates.
- Steinwascher, M. A. & Meiser, T. (2016). How a high working memory capacity can increase proactive interference. *Consciousness and Cognition*, 44, 130–145.
- Stout, W. (2002). Test Models for Traditional and Complex CBTs. In C. Mills, M. Potenza, J. Fremer & W. Ward (Eds.), *Computer-Based Testing* (pp. 103–118). Mahwah, NJ: Lawrence Erlbaum Associates.
- Souvignier, E., Förster, N. & Salaschek, M. (2014). quop: ein Ansatz internet-basierter Lernverlaufsdiagnostik und Testkonzepte für Mathematik und Lesen. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik: Test und Trends N. F.* (Bd. 12, S. 239–256). Göttingen: Hogrefe.
- Tien, J. M. (2013). Big data: Unleashing information. *Journal of Systems Science and Systems Engineering*, 22, 127–151.
- Upsing, B., Goldhammer, F., Schmitzler, M., Baumann, R., Johannes, R., Barkow, I., Rölke, H., Jars, I., Latour, T., Plichert P., Jadoul, R., Henry, C. &, Wagner, M. (2013). Chapter 5: Development of the Cognitive Items. In Organisation for Economic Co-Operation and Development (OECD) (Ed.), *Technical Report of the Survey of Adult Skills (PIAAC)* (p. 148–156). Paris: OECD.
- Vale, C. D. (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 261–285). Mahwah, NJ: Erlbaum.
- van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195–211.
- van der Linden, W. J. (2002). On complexity in CBT. In C. Mills, M. Potenza, J. Fremer & W. Ward (Eds.), *Computer-Based Testing* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum Associates.
- van der Linden, W. J. (2005). *Linear models of optimal test design*. New York, NY: Springer.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46, 197–221.
- Visible Learning Lab (2010). *Educator manual. E-asTTle fitness for national standards*. Retrieved from <http://e-astle.tki.org.nz/User-manuals> [20.12.2019]
- Waalkens, M., Aleven, V. & Taatgen, N. (2013). Does supporting multiple student strategies lead to greater learning and motivation? Investigating a source of complexity in the architecture of intelligent tutoring systems. *Computers and Education*, 60, 159–171.
- Wenzel, S. F. C., Engelhardt, L., Hartig, K., Kuchta, K., Frey, A., Goldhammer, F., Naumann, J. & Horz, H. (2016). Computergestützte, adaptive und verhaltensahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills) (CavE-ICT). In BMBF (Hrsg.). *Forschung in Ankopplung an Large-Scale Assessments* (S. 161–180). Bonn, Berlin: BMBF.
- Williamson, D. M., Mislevy, R. J. & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A. & Severance, D. D. (1999). Examinee Judgments of Changes in Item Difficulty: Implications for Item Review in Computerized Adaptive Testing. *Applied Measurement in Education*, 12, 185–198.
- Zehner, F., Sälzer, C. & Goldhammer, F. (2016). Automatic Coding of Short Text Responses via Clustering in Educational Assessment. *Educational and Psychological Measurement*, 76, 280–303.



Deskriptivstatistische Itemanalyse und Testwertbestimmung

Augustin Kelava und Helfried Moosbrugger

Inhaltsverzeichnis

- 7.1 Einleitung – 145
- 7.2 Erstellung der Datenmatrix – 145
- 7.3 Schwierigkeitsanalyse – 146
 - 7.3.1 Schwierigkeitsindex – 146
 - 7.3.2 Schwierigkeitsbestimmung bei Leistungstests – 147
 - 7.3.3 Schwierigkeitsbestimmung bei Persönlichkeitstests – 149
- 7.4 Itemvarianz – 151
 - 7.4.1 Differenzierungsfähigkeit eines Items – 151
 - 7.4.2 Berechnung der Itemvarianz – 152
- 7.5 Vorläufige Testwertermittlung – 153
- 7.6 Trennschärfe – 153
 - 7.6.1 Definition – 153
 - 7.6.2 Berechnung des Trennschärfeindex – 154
 - 7.6.3 Interpretation des Trennschärfeindex – 154
- 7.7 Itemselektion auf Basis von Itemschwierigkeit, Itemvarianz und Itemtrennschärfe – 155
- 7.8 Testwertbestimmung und Itemhomogenität – 156

7.9 Zusammenfassung – 157

7.10 EDV-Hinweise – 157

7.11 Kontrollfragen – 158

Literatur – 158

7.1 · Einleitung

i Im Folgenden wird der Frage nachgegangen, wie eine erste empirische deskriktivstatistische Evaluation der generierten Testitems vorgenommen werden kann. Die Items werden einer Erprobungsstichprobe von Testpersonen vorgelegt und das Antwortverhalten wird zur Gewinnung von Itemwerten numerisch kodiert. Im Anschluss können durch Aufsummierung der Itemwerte (vorläufige) Testwerte ermittelt werden, die zusammen mit den empirisch festgestellten Itemschwierigkeiten, Itemvarianzen und Itemtrennschärfen Auskunft darüber geben, ob die Items ihrer Aufgabe gerecht werden, die angezielten Differenzierungen zwischen den Testpersonen bezüglich des interessierenden Merkmals zu leisten. Basierend auf diesen Ergebnissen kann eine Itemselektion vorgenommen werden. Danach müssen die Kennwerte und Testwerte neu bestimmt werden.

7.1 Einleitung

Nachdem die gemäß ► Kap. 4 und 5 konstruierten Testitems/Fragen erfolgreich einer qualitativen Verständlichkeitsüberprüfung unterzogen wurden, kann eine vorläufige Test-/Fragebogenfassung zusammengestellt und an einer geeigneten Stichprobe einer ersten empirischen Erprobung („Pilotstudie“, vgl. ► Kap. 3) unterzogen werden. Dabei werden deskriktivstatistische Maße berechnet, die die Differenzierungsfähigkeit der Items quantifizieren. Das heißt, es werden Kennwerte berechnet, die eine Beurteilung darüber erlauben, ob und wie gut die Konstruktion geeigneter Items zur Abbildung der Unterschiedlichkeit der Merkmalsträger (Testpersonen) gelungen ist. Hierzu ist es notwendig, dass die Items Varianz erzeugen, indem sie bei unterschiedlichen Personen unterschiedliche Antworten hervorrufen. Darüber hinaus sollen die Items auch „trennscharf“ sein, d. h., sie sollen Personen mit höheren Merkmalsausprägungen von Personen mit niedrigeren Merkmalsausprägungen möglichst eindeutig unterscheiden. Items, die keine oder nur eine unzureichende Differenzierungsfähigkeit aufweisen, sind für den Einsatz in einem Test oder Fragebogen ungeeignet. Die verschiedenen Untersuchungsschritte der deskriktivstatistischen Erprobung werden unter der Bezeichnung „Itemanalyse“ zusammengefasst.

Die *Itemanalyse* besteht aus mehreren Schritten:

- Erstellung der Datenmatrix (► Abschn. 7.2)
- Analyse der Itemschwierigkeiten (► Abschn. 7.3)
- Bestimmung der Itemvarianzen (► Abschn. 7.4)
- Vorläufige Testwertermittlung (► Abschn. 7.5)
- Trennschärfeanalyse der Items (► Abschn. 7.6)
- Itemselektion auf Basis von Itemschwierigkeit, Itemvarianz und Trennschärfe (► Abschn. 7.7)
- Erneute Testwertbestimmung (► Abschn. 7.8)

Schritte der Itemanalyse

Die im Zuge der deskriktivstatistischen Itemanalyse gewonnenen Ergebnisse ermöglichen eine erste Qualitätsbeurteilung der Items/Fragen/Aufgaben des „neuen“ Messinstruments (Test oder Fragebogen). Dies geschieht mit dem Ziel, durch eine ggfs. erforderliche Itemselektion (s. dazu aber auch ► Kap. 13) eine möglichst qualitätsvolle Test-/Fragebogenfassung zu gewinnen.

7.2 Erstellung der Datenmatrix

Um die deskriktivstatistische Itemanalyse durchführen zu können, müssen die erhobenen Daten in systematischer Form aufbereitet werden. Dies geschieht am einfachsten durch das Anlegen einer *Datenmatrix*, in der die kodierten Antworten (Itemwerte y_{vi}) von n Testpersonen ($v = 1, \dots, n$) auf m Items ($i = 1, \dots, m$)

■ Tabelle 7.1 Datenmatrix der erhobenen Itemwerte (y_{vi}) von n Testpersonen (Tp) in m Items

Testperson	Item 1	Item 2	...	Item i	...	Item m	Zeilensumme
Tp 1	y_{11}	y_{12}	...	y_{1i}	...	y_{1m}	$\sum_{i=1}^m y_{1i} = Y_1$
Tp 2	y_{21}	y_{22}	...	y_{2i}	...	y_{2m}	$\sum_{i=1}^m y_{2i} = Y_2$
⋮	⋮	⋮		⋮		⋮	⋮
Tp v	y_{v1}	y_{v2}	...	y_{vi}	...	y_{vm}	$\sum_{i=1}^m y_{vi} = Y_v$
⋮	⋮	⋮		⋮		⋮	⋮
Tp n	y_{n1}	y_{n2}	...	y_{ni}	...	y_{nm}	$\sum_{i=1}^m y_{ni} = Y_n$
Spaltensumme	$\sum_{v=1}^n y_{v1}$	$\sum_{v=1}^n y_{v2}$...	$\sum_{v=1}^n y_{vi}$...	$\sum_{v=1}^n y_{vm}$	

7

eingetragen werden (■ Tab. 7.1). Bei einfachen Kodierungen werden in einem Leistungstest z. B. eine 0 für eine falsche Lösung und eine 1 für eine richtige Lösung als Itemwerte eingetragen; in einem Persönlichkeitstest können die Zustimmungsstufen zu Aussagen, die z. B. von 0 bis 5 reichen, als konkrete Itemwerte dienen. Es sind aber auch komplexere Kodierungen wie z. B. die Kehrwerte von Reaktionszeiten in Millisekunden als Itemwerte denkbar.

7.3 Schwierigkeitsanalyse

7.3.1 Schwierigkeitsindex

Der erste deskriktivstatistische Schritt besteht in der sog. „Schwierigkeitsanalyse“ der Items. Damit ein Test seiner Aufgabe gerecht werden kann, Informationen über die Unterschiedlichkeit der Testpersonen hinsichtlich des interessierenden Merkmals zu liefern, müssen die Items so konstruiert sein, dass nicht alle Testpersonen die gleiche Antwort auf ein Item geben. Das heißt, die Items dürfen weder zu „leicht“ sein, sodass alle Testpersonen das Item bejahen/lösen können, noch zu „schwierig“, sodass keine der Testpersonen das Item bejahen/lösen kann. Deshalb ist es notwendig, die Items hinsichtlich ihrer Schwierigkeit zu beurteilen. Als deskriptives Maß der Itemschwierigkeit verwendet man den sog. „Schwierigkeitsindex“ P_i (auch Itemschwierigkeitsindex genannt).

Definition

Der **Schwierigkeitsindex** P_i eines Items i ist der Quotient aus der bei diesem Item tatsächlich erreichten Punktsumme aller n Testpersonen und der maximal erreichbaren Punktsumme aller n Testpersonen (d. h. die Summe der höchstmöglichen Zustimmungsstufen, multipliziert mit dem Faktor 100).

Der Schwierigkeitsindex P_i eines Items i wird berechnet aus der Spaltensumme $\sum_{v=1}^n y_{vi}$ des Items i in der Datenmatrix geteilt durch die maximal mögliche Punktsumme bei n Testpersonen, nämlich $n \cdot \max(y_i)$; dieses Ergebnis wird mit 100 multipliziert:

$$P_i = \frac{\sum_{v=1}^n y_{vi}}{n \cdot \max(y_i)} \cdot 100 \quad (7.1)$$

Exkurs 7.1**Zur Unterscheidung von deskriptiven Schwierigkeitsindizes und Schwierigkeitsparametern in KTT und IRT**

An dieser Stelle sei angemerkt, dass der deskriptive Schwierigkeitsindex der Itemanalyse von den theoriebasierten Schwierigkeitsparametern in der Klassischen Testtheorie (KTT, ► Kap. 13) und in der Item-Response-Theorie (IRT, ► Kap. 16) unterschieden werden muss. Der Schwierigkeitsparameter in der KTT wird zumeist so definiert (technische Bezeichnung: parametrisiert) und geschätzt, dass seine numerische Höhe die Leichtigkeit des Items angibt, wohingegen der Schwierigkeitsparameter in der IRT tatsächlich die Schwierigkeit des Items kennzeichnet. Aber auch innerhalb der IRT gibt es die Möglichkeit, analog zu der hier beschriebenen Vorgehensweise die Itemschwierigkeit im Sinne einer Leichtigkeit zu parametrisieren.

Die Multiplikation des Quotienten mit dem Faktor 100 führt zu einem Wertebereich von P_i zwischen 0 und 100. Der Schwierigkeitsindex P_i wird umso größer, je mehr Testpersonen ein Item lösen konnten bzw. „symptomatisch“ im Sinne des interessierenden Merkmals beantwortet haben (d. h. bejahend bzw. verneinend, je nachdem, ob es sich um ein invertiertes Item handelt oder nicht, ► Abschn. 7.3.3). Die numerische Höhe des Schwierigkeitsindex P_i kennzeichnet somit in dieser Definition eigentlich die „Leichtigkeit“ des Items i und nicht die „Schwierigkeit“. Dieser etwas verwirrende Sachverhalt ist auf Konventionen zurückzuführen (► Exkurs 7.1).

Wenn der Wertebereich der Itemantworten auf einem Item i nicht bei 0 beginnt (sondern z. B. bei 1 aufgrund der Verwendung einer Ratingskala ohne 0), muss der potentiell erreichbare Minimalwert einer Itemantwort auf ein Item i , $\min(y_i)$, von jeder realisierten Itemantwort y_{vi} im Zähler abgezogen werden. Ebenso wird die minimal erreichbare Punktsumme der n Testpersonen auf Item i , $n \cdot \min(y_i)$, im Nenner abgezogen. Der Schwierigkeitsindex ergibt sich dann allgemeiner ausgedrückt als

$$P_i = \frac{\sum_{i=1}^n [y_{vi} - \min(y_i)]}{n [\max(y_i) - \min(y_i)]} \cdot 100 \quad (7.2)$$

Inhaltliche Interpretation des Schwierigkeitsindex**7.3.2 Schwierigkeitsbestimmung bei Leistungstests**

Im Folgenden soll die numerische Bestimmung des Schwierigkeitsindex für Leistungstests an einem vereinfachten Beispiel dargestellt werden.

■■ Klassifikation von Antworten

Zunächst empfiehlt es sich (in Anlehnung an Lienert und Raatz 1998), jede Messung y_{vi} unter den vier folgenden Gesichtspunkten zu klassifizieren, nämlich als

- richtig beantwortete Items (*R-Antworten*),
- falsch beantwortete Items (*F-Antworten*),
- ausgelassene (übersprungene) Items (*A-Antworten*),
- im Test unbearbeitete Items, weil z. B. die Zeit nicht ausgereicht hat (*U-Antworten*).

Anmerkung: Eine Unterscheidung zwischen A- und U-Antworten ist nur dann sinnvoll, wenn die Antworten der Testpersonen innerhalb einer vorgeschriebenen

	Item 1	Item 2	Item 3	Item 4	Item 5	Zeilensumme			
						m_R	m_F	m_A	m_U
Tp 1	R	R	R	R	R	5	0	0	0
Tp 2	R	F	A	F	F	1	3	1	0
Tp 3	R	R	R	F	U	3	1	0	1
Tp 4	R	R	F	F	U	2	2	0	1
Tp 5	R	R	F	F	U	2	2	0	1
n_R	5	4	2	1	1				
n_F	0	1	2	4	1				
n_A	0	0	1	0	0				
n_U	0	0	0	0	3				
P_i	100	80	40	20	50				

■ Abb. 7.1 Beispiel einer Datenmatrix mit klassifizierten Daten bei einem fiktiven Speedtest von $n = 5$ Testpersonen (Tp) in $m = 5$ Items mit Schwierigkeitsindizes

Zeitspanne erbracht werden müssen, d. h. bei Speedtests; bei Niveau-/Powertests können hingegen keine U-Antworten, sondern nur A-Antworten auftreten.

Bei solchermaßen klassifizierten Antworten in einem Leistungstest könnte die Datenmatrix mit n Zeilen für die (hier 5) Testpersonen und m Spalten für die (hier 5) Items folgendermaßen aussehen (■ Abb. 7.1).

Für jede Testperson v wird zeilenweise jeweils die Anzahl m_R, m_F, m_A und m_U ihrer R-, F-, A- bzw. U-Antworten bestimmt, wobei die Beziehung gilt:

$$m_R + m_F + m_A + m_U = m \quad (7.3)$$

Für jedes Item i wird spaltenweise jeweils die Anzahl n_R, n_F, n_A und n_U jener Testpersonen bestimmt, die eine R-, F-, A- bzw. U-Antwort gegeben haben, wobei die Beziehung gilt:

$$n_R + n_F + n_A + n_U = n \quad (7.4)$$

Zur Bestimmung der Schwierigkeitsindizes werden die Spaltensummen benötigt.

■ ■ Speedtests

Speedtests (Geschwindigkeitstests) erfordern eine Leistungserbringung unter zeitlichem Druck. Um bei Speedtests die Schwierigkeit P_i eines Items i nicht zu überschätzen, soll man die Anzahl n_R der auf dieses Item entfallenden R-Antworten nicht (wie bei Niveautests, s. u.) zur Gesamtzahl n aller Testpersonen in Beziehung setzen, sondern lediglich zur Anzahl $n_B = n_R + n_F + n_A$ der Testpersonen, die das Item i tatsächlich bearbeitet haben. Der Schwierigkeitsindex P_i eines Items i lautet also bei Speedtests:

$$P_i = \frac{n_R}{n_B} \cdot 100 \quad (7.5)$$

■ ■ Niveautests

Ein Niveautest (Powertest) ist ein Leistungstest, in dessen Durchführungs vorschrift entweder keine oder nur eine moderate Zeitbegrenzung, die von den Testpersonen nicht als Zeitdruck empfunden wird, vorgegeben ist. In einem Niveautest gibt es demzufolge Richtig- und Falschantworten sowie ausgelassene Aufgaben (A-Antworten). Hingegen gibt es keine Aufgaben, die unbearbeitet bleiben

7.3 · Schwierigkeitsanalyse

	Item 1	Item 2	Item 3	Item 4	Item 5	Zeilensumme		
						m_R	m_F	m_A
Tp 1	R	R	R	R	R	5	0	0
Tp 2	R	F	A	F	F	1	3	1
Tp 3	R	R	R	F	A	3	1	1
Tp 4	R	R	F	F	A	2	2	1
Tp 5	R	R	F	F	A	2	2	1
n_R	5	4	2	1	0			
n_F	0	1	2	4	2			
n_A	0	0	1	0	3			
P_i	100	80	40	20	20			

■ Abb. 7.2 Fiktives Beispiel einer Datenmatrix mit klassifizierten Daten bei Niveautests von $n = 5$ Testpersonen (Tp) in $m = 5$ Items mit Schwierigkeitsindizes. Wie man durch Vergleich mit ■ Abb. 7.1 leicht erkennen kann, sind n_A und n_U hier zu n_A zusammengefasst

(U-Antworten), da die Testpersonen genügend Zeit hatten, alle Aufgaben zu bearbeiten (■ Abb. 7.2).

Die tatsächlich erreichte Punktsumme aller Testpersonen bei einem Item ist durch die Anzahl n_R der Testpersonen gegeben, die bei diesem Item eine R-Antwort gegeben haben. Die Punktsumme bei Item i wird am größten, wenn alle Testpersonen eine R-Antwort geben, sodass die maximal erreichbare Punktsumme durch die Anzahl aller Testpersonen, d. h. n , gegeben ist.

Die Berechnung des Schwierigkeitsindex P_i für ein Item i vereinfacht sich bei Niveautests somit zu

$$P_i = \frac{n_R}{n} \cdot 100 \quad (7.6)$$

und entspricht der von Lienert und Raatz (1998, S. 73) vorgeschlagenen Definition:

Definition

„Der **Schwierigkeitsindex** einer Aufgabe ist gleich dem prozentualen Anteil der auf diese Aufgabe entfallenden richtigen Antworten in Beziehung zur Analysestichprobe von der Größe n ; der Schwierigkeitsindex liegt also bei schwierigen Aufgaben niedrig, bei leichten hoch.“

Bei Verwendung der Gl. (7.6) geht man davon aus, dass Einflüsse des Ratens auf die Beantwortung der Items vernachlässigbar und unbedeutend sind. Ist hingegen eine solche Annahme nicht gerechtfertigt, kann bei Auswahlaufgaben (also bei Items mit einer richtigen und mehreren falschen Antwortalternativen) eine *Ratekorrektur* angewendet werden. Hierzu muss die Gl. (7.6) angepasst werden, um einen korrigierten Schwierigkeitsindex zu erhalten (für Details sei an dieser Stelle auf Lienert und Raatz 1998, S. 75 ff., verwiesen).

7.3.3 Schwierigkeitsbestimmung bei Persönlichkeitstests

Bei Persönlichkeitstests erscheint es zunächst nicht ganz passend, von einer Item-schwierigkeit zu sprechen, da die Antwort auf ein Item nicht „richtig“ oder „falsch“

Symptomatische vs. unsymptomatische Antworten

Invertierte Items

k = 2 Antwortkategorien

Schwierigkeitsindex bei $k > 2$ Antwortkategorien

Interpretation des Schwierigkeitsindex

sein kann. Stattdessen legt man bei der Item- und Testkonstruktion fest, welche der Antwortmöglichkeiten als *symptomatische* und welche als *unsymptomatische Antwort* für das Vorhandensein bzw. für eine starke Ausprägung des untersuchten Merkmals (z. B. Extraversion) anzusehen ist.

Für gewöhnlich werden Items so konstruiert, dass eine *Bejahung/Zustimmung* als Hinweis für eine höhere Merkmalsausprägung spricht und eine *Verneinung/Ablehnung* für eine niedrigere. Eine Ausnahme bilden „*invertierte Items*“, die z. B. zur Abschwächung von Antworttendenzen, vor allem von Akquieszenz (vgl. ► Kap. 4) eingesetzt werden. Invertierte Items sind so konstruiert, dass die Bejahung/Zustimmung symptomatisch für eine niedrige Merkmalsausprägung ist. Von daher sind die Itemantworten *vor* der Itemanalyse wieder „umzupolen“, indem man z. B. eine fünfstufige Kodierung 0, 1, 2, 3, 4 *reinvertiert*, und zwar in 4, 3, 2, 1, 0.

■ ■ Numerische Bestimmung des Schwierigkeitsindex

Liegen für die Items lediglich $k = 2$ Antwortkategorien vor, bei denen die im Sinne des Merkmals „symptomatische“ Antwort mit $y_{vi} = 1$ und die „unsymptomatische“ Antwort mit $y_{vi} = 0$ kodiert wird, so kann man zur Schwierigkeitsbestimmung wie bei Leistungstests (Niveautests) verfahren (Gl. 7.6). Der Schwierigkeitsindex entspricht dann dem Anteil „symptomatischer“ Antworten an allen n Antworten (weil von jeder Testperson zu jedem Item i eine Antwort vorliegt).

Bei $k > 2$ Antwortkategorien kann man den „Schwierigkeitsindex“ für intervall-skalierte Stufen nach Dahl (1971) berechnen. Kodiert man hierzu die k Antwortstufen des Items i mit 0 bis $k - 1$, so ergibt sich der Schwierigkeitsindex wie in der allgemein gehaltenen Darstellung in Gl. (7.1) als

$$P_i = \frac{\sum_{v=1}^n y_{vi}}{n \cdot (k - 1)} \cdot 100, \quad (7.7)$$

d. h. als Quotient aus der i -ten Spaltensumme in □ Tab. 7.1 und der maximal möglichen Spaltensumme, multipliziert mit dem Faktor 100.

Diese Formel lässt sich mit dem Mittelwert der Itemwerte vereinfachen:

$$P_i = \frac{\bar{y}_i}{k - 1} \cdot 100 \quad (7.8)$$

Der Mittelwert \bar{y}_i im Zähler braucht nun nur noch durch die um eins verminderte Anzahl der Antwortstufen der Ratingskala geteilt werden. Weist die Ratingskala z. B. 5 Abstufungen auf (0 bis 4), so wird der Mittelwert durch 4 geteilt.

Hinweis: Eine verschiedentlich vorgeschlagene Dichotomisierung (d. h. die künstliche Zweiteilung k -stufiger Werte in „hohe“ und „niedrige“ Werte) kann nicht empfohlen werden, da hierdurch Informationsverluste oder sogar Verzerrungen entstehen würden (vgl. MacCallum et al. 2002).

Der deskriptiv festgestellte Schwierigkeitsindex P_i kann als durchschnittliches Ausmaß der Zustimmung auf der k -stufigen Antwortskala interpretiert werden (wobei noch mit 100 multipliziert wird). Der Schwierigkeitsindex weist einen Wertebereich von 0 bis 100 auf. Je höher der Wert P_i ist, desto leichter fällt es den Testpersonen im Durchschnitt, auf das Item i eine „symptomatische“, d. h. in der Regel zustimmende Antwort zu geben. Und umgekehrt: Je kleiner der Wert P_i ist, desto schwerer fällt die Zustimmung (vgl. dazu den Leichtigkeitsparameter Alpha in der KTT, ► Kap. 13).

7.4 Itemvarianz

Unter Itemvarianz versteht man ein Maß für die Differenzierungsfähigkeit eines Items i in der untersuchten Stichprobe.

7.4.1 Differenzierungsfähigkeit eines Items

Zur Veranschaulichung der Itemvarianz stelle man sich 10 Testpersonen vor, die z. B. vier fiktive Teilprüfungen (Items) eines Tests zu absolvieren haben. Dabei kodieren wir das Bestehen einer Testperson v in einer Teilprüfung i mit $y_{vi} = 1$ und das Scheitern mit $y_{vi} = 0$. Die fiktiven Ergebnisse werden der Datenmatrix in □ Tab. 7.2 dargestellt.

Die einfachen empirischen Lösungswahrscheinlichkeiten $p_i = P_i/100$ der vier Teilprüfungen (Items) sind $p_1 = 9/10 = .90$, $p_2 = 5/10 = .50$, $p_3 = 2/10 = .20$ und $p_4 = 0/10 = .00$.

Wie man an der Verteilung der Itemwerte (y_{vi}) erkennen kann, waren die Teilprüfungen nicht nur unterschiedlich schwer, sondern sie haben auch unterschiedliche Differenzierungen zwischen den Testpersonen, die bestanden haben, und jenen, die nicht bestanden haben, hervorgebracht:

1. *Niedrige Varianz*: In Teilprüfung 1 kann die Testperson 10 (d. h. diejenige, die durchgefallen ist), jeder der 9 anderen, die nicht durchgefallen sind, gratulieren. Teilprüfung 1 (Item 1) leistet hier also $1 \cdot 9 = 9$ Differenzierungen ($Var(y_1) = .09$).
2. *Hohe Varianz*: Nach Durchführung von Teilprüfung 2 kann jede der 5 Testpersonen, die bei Teilprüfung 2 durchgefallen sind, jeder der 5 Testpersonen, die bei der Teilprüfung 2 bestanden hat, gratulieren. Diese Teilprüfung (Item 2) leistet von daher $5 \cdot 5 = 25$ Differenzierungen ($Var(y_2) = .25$).

**Unterschiedliche
Differenzierungsfähigkeit**

□ **Tabelle 7.2** Beispiel einer Datenmatrix zur Veranschaulichung der Itemvarianz anhand der Itemwerte y_{vi} von $m = 4$ fiktiven Teilprüfungen (Items) und $n = 10$ Testpersonen (Tp)

	Item 1	Item 2	Item 3	Item 4	Zeilensumme	
					m_R	m_F
Tp 1	1	1	1	0	3	1
Tp 2	1	1	1	0	3	1
Tp 3	1	1	0	0	2	2
Tp 4	1	1	0	0	2	2
Tp 5	1	1	0	0	2	2
Tp 6	1	0	0	0	1	3
Tp 7	1	0	0	0	1	3
Tp 8	1	0	0	0	1	3
Tp 9	1	0	0	0	1	3
Tp 10	0	0	0	0	0	4
n_R	9	5	2	0		
n_F	1	5	8	10		
p_i	.90	.50	.20	.00		
$Var(y_i)$.09	.25	.16	.00		

3. *Mittlere Varianz*: Nach Teilprüfung 3 kann jede/r der 8 Durchgefallenen jeder/m der 2 Durchgekommenen gratulieren. Hier (Item 3) kommen $8 \cdot 2 = 16$ Differenzierungen zustande ($Var(y_3) = .16$).
4. *Keine Varianz*: Teilprüfung 4 (Item 4) leistet hingegen keinerlei Differenzierungen ($Var(y_4) = .00$). Hier kann keine Testperson aus der Gruppe der Durchgefallenen einer Testperson aus der Gruppe der Durchgekommenen gratulieren, weil letztere Gruppe leer ist. Alle sind durchgefallen.

Wie man sieht, ist die mögliche Ausprägung der Itemvarianz durch die Itemschwierigkeit begrenzt: Items mittlerer Schwierigkeit (z. B. Item 2) können viele Differenzierungen leisten, Items höherer oder niedrigerer Schwierigkeit (z. B. Item 1 oder 3) ermöglichen weniger Differenzierungen. Items mit einer extremen Schwierigkeit (Lösungswahrscheinlichkeit gleich 1 oder gleich 0, z. B. bei Item 4), sind hingegen nicht für eine Differenzierung geeignet.

7

7.4.2 Berechnung der Itemvarianz

Itemvarianz bei k -stufigen Items

Die Itemvarianz $Var(y_i)$ eines Items i informiert über das Ausmaß der Differenzierungsfähigkeit des Items. Sie wird numerisch berechnet als

$$Var(y_i) = \frac{\sum_{v=1}^n (y_{vi} - \bar{y}_i)^2}{n} \quad (7.9)$$

Da der Itemmittelwert \bar{y}_i (durchschnittliche Antwort aller Testpersonen auf das Item i) und die Itemschwierigkeit P_i (bzw. die Wahrscheinlichkeit $p_i = P_i/100$) bei Items mit k -stufigem Antwortmodus in funktionaler Abhängigkeit zueinander stehen ($\bar{y}_i = p_i \cdot (k - 1)$), lässt sich die Itemvarianz auch wie folgt berechnen:

$$Var(y_i) = \frac{\sum_{v=1}^n (y_{vi} - p_i \cdot (k - 1))^2}{n} \quad (7.10)$$

Itemvarianz bei zweistufigen Items

Für zweistufige Items lässt sich Gl. (7.10) zu

$$Var(y_i) = p_i \cdot (1 - p_i) \quad (7.11)$$

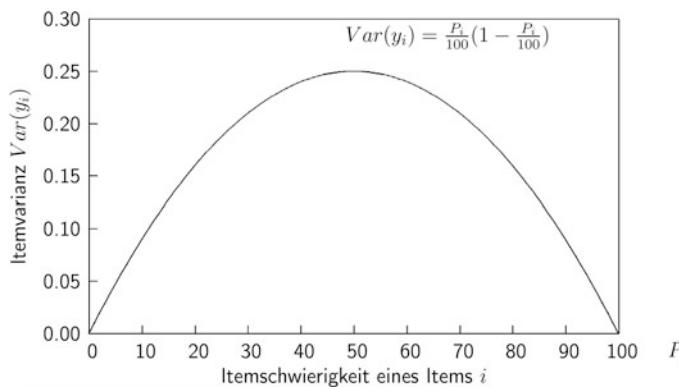
vereinfachen (vgl. z. B. Eid et al. 2017, S. 196; Kranz 1997, S. 52 ff.). Die Itemvarianz entspricht dann dem Produkt der Wahrscheinlichkeiten, das Item i zu lösen (p_i), und der Gegenwahrscheinlichkeit, das Item i nicht zu lösen ($1 - p_i$).

In der Abb. 7.3 wird der kurvilineare Zusammenhang zwischen der Itemvarianz und der Itemschwierigkeit für zweistufige (dichotome) Items veranschaulicht.

Wie man der Abb. 7.3 entnehmen kann, hat die Itemvarianz ihr Maximum ($Var(y_i) = .25$) bei mittlerer Itemschwierigkeit ($P_i = 50$). Das heißt, dass bei dichotomen Items die größte Differenzierung bei einer Itemschwierigkeit von $P_i = p_i \cdot 100 = 50$ erreicht wird, während sie zu den beiden extremen Ausprägungen hin (sehr niedrig, sehr hoch) stark abnimmt und bei $P_i = 0$ sowie bei $P_i = 100$ keine Differenzierung mehr vorliegt.

Kurvilinearer Zusammenhang von Itemvarianz und Itemschwierigkeit

7.5 · Vorläufige Testwertermittlung



■ Abb. 7.3 Zusammenhang von Itemvarianz $Var(y_i)$ und Itemschwierigkeit P_i bei zweistufigem Antwortmodus

7.5 Vorläufige Testwertermittlung

Wurden zur Messung desselben Merkmals (auf verschiedenen Schwierigkeitsniveaus) mehrere Items konstruiert, ist es sinnvoll, die einzelnen Itemwerte y_{vi} pro Person probeweise zu einem (vorläufigen) *Testwert* Y_v zusammenzufassen, der die Merkmalsausprägung der Person v widerspiegeln soll.

Die einfachste Möglichkeit, einen (vorläufigen, d. h. testtheoretisch noch nicht abgesicherten) Testwert Y_v für die Testperson v zu bestimmen, besteht darin, die einzelnen Itemwerte y_{vi} der m Items zu addieren. Den Testwert Y_v einer Testperson v erhält man somit als resultierenden Summenwert (Zeilensumme im Sinne der Datenmatrix).

$$Y_v = \sum_{i=1}^m y_{vi} \quad (7.12)$$

Die Sinnhaftigkeit der Bildung des Testwertes durch Addition der Itemwerte ist in diesem Stadium der Testentwicklung aber noch sehr ungewiss, denn weder die Itemvarianzen noch die Itemschwierigkeiten erlauben eine Beurteilung, ob die Items dasselbe Merkmal messen. Das Vorliegen von Eindimensionalität ist aber die wesentliche Grundvoraussetzung für die Sinnhaftigkeit der Summenbildung.

Erste deskriptive Hinweise, ob die einzelnen Items dasselbe Merkmal messen, können den *Itemtrennschärfen* (► Abschn. 7.6) entnommen werden, die anhand von Korrelationen der einzelnen Itemwerte y_{vi} mit den (vorläufigen) Testwerten Y_v berechnet werden. Mithilfe der Trennschärfen kann – in Verbindung mit den Ergebnissen von Itemvarianz und Itemschwierigkeit – eine *Itemselektion* (► Abschn. 7.7) vorgenommen werden, die dazu beitragen soll, dass nur diejenigen Items im Test verbleiben, die das interessierende Merkmal messen; die anderen Items sollen hingegen ausgesondert oder nachgebessert werden.

7.6 Trennschärfe

7.6.1 Definition

Wurden (vorläufige) Testwerte gemäß Gl. (7.12) ermittelt, so kann als weiteres deskriptivstatistisches Maß der Itemanalyse die *Itemtrennschärfe* bestimmt werden. Die Trennschärfe gibt an, ob und wie gut die Merkmalsdifferenzierung des jeweiligen Items i mit der Merkmalsdifferenzierung, die alle Items gemeinsam leisten,

übereinstimmt. Bei guter Übereinstimmung werden die Testpersonen von dem einzelnen Item in gleicher Richtung differenziert wie von dem Gesamttest. In diesem – im Sinne der angezielten Eindimensionalität wünschenswerten – Fall verfügt das Item über eine gute Trennschärfe, weil es die Testpersonen mit höheren Testwerten (gebildet aus den Ergebnissen aller Items gemeinsam, ▶ Abschn. 7.5) von jenen mit niedrigeren Testwerten deutlich („scharf“) trennen kann. Items mit hohen positiven Trennschärfen liefern für den Gesamttest einen wesentlichen Beitrag zur Merkmalsdifferenzierung.

Definition

Die Trennschärfe (eigentlich der Trennschärfeindex) r_{it} eines Items i drückt aus, wie groß der korrelative Zusammenhang zwischen der Variablen der Itemwerte y_i der Testpersonen und der Testwertvariablen Y ist. Die **deskriptive Trennschärfe** wird als Korrelation berechnet und kann Werte im Bereich zwischen -1 und 1 annehmen.

7

7.6.2 Berechnung des Trennschärfeindex

Ohne nähere testtheoretische Untermauerung, die zu wesentlich genaueren Ergebnissen führen kann (▶ Kap. 13), wird zur Berechnung der sog. „unkorrigierten Trennschärfe“ über alle n Testpersonen hinweg der korrelative Zusammenhang der Itemvariablen y_i mit der unkorrigierten Testwertvariablen Y bestimmt. Die Testwerte Y_v der einzelnen Testpersonen v werden dabei üblicherweise als Zeilensumme sämtlicher Itemwerte y_{vi} der Testperson v gebildet (▶ Abschn. 7.5).

■ ■ Unkorrigierte Trennschärfe

Itemwert-Testwert-Korrelation r_{it}

$$r_{it} = r_{(y_i, Y)} \quad (7.13)$$

Testwertkorrektur bei geringer Itemanzahl

Insbesondere dann, wenn der Test nur aus wenigen Items besteht, wird die unkorrigierte Itemtrennschärfe durch den Sachverhalt überschätzt, dass der jeweilige Itemwert auch selbst in den Testwert Eingang gefunden hat, was sich korrelationserhöhend auswirkt. Deshalb empfiehlt es sich, bei der Berechnung des Testwertes die sog. „part-whole correction“ vorzunehmen, indem man das betreffende Item i bei der Summenbildung auslässt, d. h. nicht in den Testwert Y_v aufnimmt, und stattdessen den korrigierten Testwert $Y_{v(i)} = Y_v - y_{vi}$ bildet.

Part-whole-korrigierte Trennschärfe

Die mit dem korrigierten Testwert vorgenommene Trennschärfebestimmung wird als *part-whole-korrigierte Trennschärfe* $r_{it(i)}$ bezeichnet. Sie wird als Korrelation zwischen der Itemvariablen y_i und der korrigierten Testwertvariablen $Y_{(i)}$ berechnet:

$$r_{it(i)} = r_{(y_i, Y_{(i)})} \quad (7.14)$$

7.6.3 Interpretation des Trennschärfeindex

Da der Trennschärfeindex eines Items i bezüglich Höhe und Vorzeichens unterschiedliche Ausprägung annehmen kann, ist es zweckmäßig, bei der Interpretation und den resultierenden Schlussfolgerungen die nachfolgenden *Fallunterscheidungen* zu beachten.

■ ■ Fallunterscheidungen

1. *Trennschärfe hoch positiv*: Das Item i wird von Testpersonen mit einem hohem Testwert (hoher Merkmalsausprägung) gelöst bzw. symptomatisch beantwortet

7.7 · Itemselektion auf Basis von Itemschwierigkeit, Itemvarianz und Itemtrennschärfe

und von Testpersonen mit niedrigem Testwert (niedriger Merkmalsausprägung) nicht. Liegen hohe positive Trennschärfen vor, d. h. hohe Korrelationen zwischen den Itemvariablen y_i und der Testwertvariablen Y , so kann man davon ausgehen, dass die einzelnen Items sehr ähnlich differenzieren wie der Gesamttest und einen guten Beitrag zu Messgenauigkeit und Validität des Tests liefern. Trennschärfen im Bereich von .4 bis .7 gelten als „gute“ Trennschärfen.

2. *Trennschärfe nahe null:* Die mit dem Item i erzielte Differenzierung weist keinen Zusammenhang mit der Differenzierung durch den Gesamttest auf. Das Item ist ungeeignet, zwischen Testpersonen mit hohem Testwert (hoher Merkmalsausprägung) und Testpersonen mit niedrigem Testwert (niedriger Merkmalsausprägung) zu unterscheiden. Was auch immer das Item i misst, es ist unabhängig von dem, was die zum Gesamttestwert aufsummierten weiteren Items des Tests messen.
3. *Trennschärfe hoch negativ:* Das Item i wird von Testpersonen mit niedriger Merkmalsausprägung gelöst bzw. symptomatisch beantwortet, von Testpersonen mit hoher Merkmalsausprägung hingegen nicht. Dies kann durch Mängel, z. B. bei der Instruktion oder bei der Itemformulierung, bedingt sein, denen nachgegangen werden muss. Sofern es sich nicht um einen Leistungs-, sondern um einen Persönlichkeitstest handelt, können Items mit hoher negativer Trennschärfe ggf. dennoch genutzt werden, wenn man sie begründet als invertierte Items auffassen kann und danach die Kodierungsrichtung ändert. Die Trennschärfe wird dadurch positiv. Aus inhaltlich-theoretischer Perspektive ist dieses Vorgehen jedoch nicht unproblematisch. Ein weiterer Grund dafür, dass negative Trennschärfen auftreten, könnte eine vergessene Rückinvertierung von invers konstruierten Items sein. Inverse Items müssen nämlich – am besten bereits vor der Erstellung der Datenmatrix – so umgepolt werden, dass ein höherer Itemwert mit einer höheren Merkmalsausprägung einhergeht (► Abschn. 7.3.3).

7.7 Itemselektion auf Basis von Itemschwierigkeit, Itemvarianz und Itemtrennschärfe

Als Konsequenz aus der Itemanalyse sollen bei der Zusammenstellung der Items zu einem Test oder Fragebogen diejenigen Items ausgewählt („selektiert“) werden, die zur Messung des interessierenden Merkmals psychometrisch gesehen am besten geeignet sind.

Dabei sind zunächst die Ergebnisse der deskriptivstatistischen Itemanalyse hinsichtlich der *Itemschwierigkeit* und der *Itemvarianz* zu berücksichtigen: Wie wir in ► Abschn. 7.3 und 7.4 gesehen haben, sind Items mit einer mittleren Schwierigkeit von $P_i = 50$ am besten in der Lage, deutliche Differenzierungen zwischen den Testpersonen mit höherer und Testpersonen mit niedriger Merkmalsausprägung zu erzeugen. Sofern die Absicht verfolgt wird, auch zwischen Testpersonen mit extremen Merkmalsausprägungen zu differenzieren, so sind nicht nur Items mit mittlerer Schwierigkeit, sondern auch solche mit Schwierigkeitsindizes von $5 \leq P_i \leq 20$ bzw. $80 \leq P_i \leq 95$ auszuwählen, obwohl sie nur eine geringere Itemvarianz aufweisen. Bei einem Test, der über das gesamte Merkmalsspektrum eine Differenzierung erlauben soll, sollten die Items typischerweise Schwierigkeitsindizes in allen Schwierigkeitsgraden aufweisen und gleichmäßig über den Bereich von $5 \leq P_i \leq 95$ verteilt sein.

Items mit einer Schwierigkeit von 0 oder 100 wären in jedem Fall aus dem Test zu entfernen, da sie keinerlei Differenzierung zwischen den Testpersonen liefern. In unserem Beispiel aus □ Tab. 7.2 wäre demzufolge das Item 4 zu entfernen, da es einen Schwierigkeitsindex von 0 aufweist.

Selektion nach Itemschwierigkeit

Selektion nach Itemvarianz und Trennschärfe

Um sich für die Aufnahme in einen Test, in dem mehrere Items zur Messung desselben Merkmals zusammengefasst werden, zu qualifizieren, müssen die Items aber nicht nur eine geeignete Itemschwierigkeit und eine hohe Itemvarianz, sondern auch eine ausreichend große *Itemtrennschärfe* (vgl. ► Abschn. 7.6) aufweisen. Eine hohe Trennschärfe wird im Allgemeinen durch eine hohe Itemvarianz begünstigt. Dies gilt sowohl bei intervallskalierten als auch bei dichotomen Items. Dennoch garantiert eine hohe Itemvarianz alleine nicht auch schon eine hohe Trennschärfe. Items mit Trennschärfen nahe null oder im negativen Bereich (► Abschn. 7.6.3) sollten nicht in den Test oder Fragebogen aufgenommen werden. Hat man bei der Selektion mehrere Items gleicher Schwierigkeit zur Verfügung, so ist jeweils das Item mit der höchsten Trennschärfe zu bevorzugen.

Bevor Items, die keine zufriedenstellenden Trennschärfen aufweisen, aus dem Test entfernt werden, sollte geprüft werden, ob die Items inhaltlich-theoretisch für das interessierende Merkmal bedeutsam erscheinen. Sollten die Trennschärfen trotz inhaltlicher Bedeutsamkeit sehr nahe bei null liegen, so könnte das interessierende Merkmal nicht eindimensional, sondern mehrdimensional sein, wobei faktorenanalytische Verfahren zwecks weiterer Analysen hilfreich sind (s. z. B. die exploratorische [EFA], ► Kap. 23, und konfirmatorische Faktorenanalyse [CFA], ► Kap. 24). Erweist sich die Vermutung, dass sich das interessierende Merkmal aus mehr Dimensionen als zunächst angenommen zusammensetzt, wäre eine entsprechende Konstruktion weiterer Items für jede der Merkmalsdimensionen angemessener.

Niedrige Trennschärfen können auf Mehrdimensionalität hinweisen

Voraussetzung für Summierung der Itemwerte: Itemhomogenität

Neubestimmung von Testwerten und Itemtrennschärfen

Wichtig: testtheoretische Überprüfung der Eindimensionalität

7.8 Testwertbestimmung und Itemhomogenität

Die Sinnhaftigkeit der Bildung eines Testwertes durch Addition der Itemwerte ist an die Voraussetzung gebunden, dass *Itemhomogenität/Eindimensionalität* der Items vorliegt; anderenfalls würden „Äpfel mit Birnen“ addiert. Hohe Trennschärfenfeindizes liefern grobe Hinweise dafür, dass die Items inhaltlich dasselbe Merkmal messen und Eindimensionalität vorliegt.

! Durch die Selektion der Items auf Basis von geeigneten Itemschwierigkeiten, Itemvarianzen und Trennschärfen und durch den Ausschluss ungeeigneter Items geht eine Veränderung der Datenmatrix einher, da jene Spalten entfernt werden müssen, in denen die Daten der nicht ausgewählten Items eingetragen sind. Aus diesem Grund müssen die Testwerte nach Gl. (7.12) neu bestimmt werden, was auch veränderte Itemtrennschärfen nach sich zieht. Je nach Datenlage kann es erforderlich sein, eine erneute Itemselektion vorzunehmen und den beschriebenen Prozess zu wiederholen.

Auch die Neubestimmung der Testwerte nach Gl. (7.12) kann nicht als endgültig angesehen werden. Vielmehr erlauben erst testtheoretisch begründete Modellierungen eine genaue Beurteilung der Homogenität/Eindimensionalität der Items. Zu nennen sind hier vor allem *faktorenanalytische Methoden*. Dazu gehören insbesondere Verfahren, die auf der KTT (► Kap. 13) aufbauen, z. B. die EFA (► Kap. 23), die CFA (► Kap. 24) bzw. Modelle der IRT (► Kap. 16, 17, 18 und 19) sowie die Latent-Class-Analyse (LCA, ► Kap. 22). Während sich die deskriptive Itemanalyse zur Beurteilung der Frage, ob die einzelnen Testitems dasselbe Merkmal messen, damit begnügen musste, dass die selektierten Items hinreichend hohe Trennschärfenfeindizes aufweisen, erlauben die testtheoretischen Verfahren darüber hinaus eine inferenzstatistische Überprüfung der (Ein-)dimensionalität, was zu wesentlich belastbareren Ergebnissen führt als das zuvor beschriebene deskriptive Vorgehen.

Um die besten Items auszuwählen, sollten darüber hinaus möglichst auch Überlegungen hinsichtlich der *Reliabilität* (vgl. ► Kap. 13, 14 und 15) und *Validität*

7.9 · Zusammenfassung

(► Kap. 21) einbezogen werden. Eine rein von Kennzahlen getriebene Itemselektion ohne theoretische Auseinandersetzung mit den Iteminhalten ist nicht zweckmäßig und auch nicht im Sinne der Psychometrie. Vielmehr ist die psychometrische Itemanalyse ein iterativer Prozess der gelingenden Auseinandersetzung/Verzahnung von Theorie und Empirie.

Verzahnung von Theorie und Empirie

7.9 Zusammenfassung

Die deskriptivstatistische Itemanalyse ist eine erste empirische Erprobung neu konstruierter Items an einer ersten Stichprobe („Pilotstudie“). Als deskriptive Maße zur Beurteilung der Itemqualität werden in der Regel die Itemschwierigkeit (► Abschn. 7.3), die Itemvarianz (► Abschn. 7.4) und nach erfolgter vorläufiger Testwertberechnung (► Abschn. 7.5) die Itemtrennschärfe (► Abschn. 7.6) berechnet.

Diese drei Maße werden verwendet, um zu einem (groben) psychometrischen Urteil zu gelangen, welche Items in einem neu konstruierten Test oder Fragebogen verbleiben können und welche Items nicht. Bei der Beurteilung stehen aus Sicht der Itemkonstruktion vor allem zwei Aspekte im Vordergrund: Erstens sollen die konstruierten Items differenzieren können. Das heißt, die Items sollten geeignet sein, die Unterschiedlichkeit der Testpersonen zu erfassen. Hierüber geben die Itemschwierigkeit und die Itemvarianz Auskunft. Zweitens sollen die Items, die zu einem Test(summen)wert zusammengefasst werden, nach Möglichkeit ein und dasselbe Merkmal erfassen. Als grobes Beurteilungsmaß der Ähnlichkeit zwischen Itemwerten und Testwert eignen sich die Itemtrennschärfen. Beide Aspekte werden anhand der drei Maße zu einem integrativen Urteil über die psychometrische Eignung der Items verbunden. Zusätzlich werden gleichermaßen inhaltlich-theoretische Überlegungen unternommen, um zu entscheiden, ob

- a. Items im Test oder Fragebogen verbleiben können oder
- b. Items entfernt bzw. nachgebessert werden müssen oder
- c. vielleicht weitere Items neu zu konstruieren sind, um einen zunächst nicht ausreichend genau erfassten Merkmalsbereich mit weiteren Items abzudecken.

Nachdem diese Analysen durchgeführt worden sind, werden die Items ausgewählt/ selektiert, die sich am besten für den Fragebogen oder den Test zur Erfassung des interessierenden Merkmals bzw. der interessierenden Merkmale eignen. Die Itemselektion soll sicherstellen, dass die empirisch-deskriptiv erprobte Test- oder Fragebogenfassung nur solche Items enthält, die eine geeignete Schwierigkeit, eine hohe Varianz und eine hinreichende Trennschärfe aufweisen. Items, deren Trennschärfe nahe bei null liegt, sind zur Erfassung eines eindimensionalen Merkmals ungeeignet. Die Durchführung der Itemselektion macht eine erneute Bestimmung der Testwerte notwendig. Doch auch die neu bestimmten Testwerte sind noch nicht endgültig, solange nicht durch Anwendung testtheoretischer Modelle (s. „Testtheorien“ in Teil II dieses Bandes) eine Dimensionalitätsüberprüfung vorgenommen wird, die eine wesentlich genauere Beurteilungen der Item- und Testqualität ermöglicht.

7.10 EDV-Hinweise

Mit einschlägiger Statistiksoftware (z. B. R, SPSS, Stata, SAS) können Itemschwierigkeiten, Itemmittelwerte, Itemvarianzen und Itemtrennschärfen berechnet werden. Ein weiteres Datenbeispiel findet sich im Bereich EDV-Hinweise unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

7.11 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Wie berechnet man den Schwierigkeitsindex P_i bei Persönlichkeitstests?
2. Welche Antworten lassen sich bei Speed- und Niveautests unterscheiden?
3. Gibt es einen Zusammenhang zwischen Itemvarianz und Itemschwierigkeit? Wenn ja, wie lässt sich dieser beschreiben und begründen?
4. Was sagt die Trennschärfe r_{it} eines Items i aus?
5. Können Items mit einer extremen Itemschwierigkeit (also sehr niedrig oder sehr hoch) extreme Trennschärfen haben? Falls ja, konstruieren Sie ein Beispiel. Falls nein, warum nicht?

7

Literatur

- Dahl, G. (1971). Zur Berechnung des Schwierigkeitsindex bei quantitativ abgestufter Aufgabenbewertung. *Diagnostica*, 17, 139–142.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
- Kranz, H. T. (1997). *Einführung in die klassische Testtheorie* (4. Aufl.). Frankfurt am Main: Verlag Dietmar Klotz GmbH.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.
- MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.



Testwertverteilung

Augustin Kelava und Helfried Moosbrugger

Inhaltsverzeichnis

- 8.1 Einleitung – 160**
- 8.2 Zentrale Tendenz der Testverteilung – 160**
- 8.3 Streuung der Testwertverteilung – 161**
- 8.4 Beurteilung der Verteilungsform – 161**
- 8.5 Ursachen für die Abweichung der Testwertverteilung von der Normalverteilung – 163**
- 8.6 Normalisierung der Testwertverteilung – 164**
 - 8.6.1 Nichtlineare Flächentransformation – 164
 - 8.6.2 Ablaufschritte – 165
- 8.7 Zusammenfassung und weiteres Vorgehen – 168**
- 8.8 EDV-Hinweise – 168**
- 8.9 Kontrollfragen – 168**
- Literatur – 168**

i Über die deskriptivstatistische Itemanalyse (► Kap. 7) hinausgehend liefert die Analyse der Testwertverteilung Informationen über die zu einem Test zusammengefassten Items. Mit den Maßen der zentralen Tendenz und der Streuung sowie von Schiefe und Exzess lassen sich die wesentlichen Eigenschaften der Testwertverteilung untersuchen. Abweichungen der Testwertverteilung von der Normalverteilung erlauben Rückschlüsse auf ungünstige Zusammensetzungen der Itemschwierigkeiten, die im Zuge von Testrevisionen ausgeglichen werden können. In begründeten Fällen kann eine Normalisierung der Testwerte vorgenommen werden.

8.1 Einleitung

Nach Abschluss der Itemselektion liegt das Augenmerk der Testentwicklung auf der Testwertverteilung, die über die Häufigkeit der verschiedenen Abstufungen der Testwertvariablen Y in einer Erprobungsstichprobe Aufschluss gibt und kann weitere Informationen über den Erfolg bei der Itemkonstruktion und -selektion liefern. Da die Ausprägungen psychologischer Merkmale häufig von vielen Einflussgrößen abhängen und deshalb als normalverteilt angenommen werden können, ist es von besonderem Interesse, ob die aus dem Konstruktionsprozess resultierende Testwertverteilung in etwa einer Normalverteilung folgt. Ist das der Fall, so ist die Verteilungsform eingipflig und symmetrisch. Kennwerte wie Schiefe und Exzess (► Abschn. 8.4) weichen dann nicht von den Kennwerten der Normalverteilung ab. Um eine geeignete Differenzierung zwischen den Testpersonen zu erzeugen, ist es auch wichtig, dass die Verteilung der Testwerte eine hinreichende Streuung aufweist.

8.2 Zentrale Tendenz der Testverteilung

Modalwert

Die zentrale Tendenz der Testwertverteilung kann durch verschiedene Maße angegeben werden. Zunächst empfiehlt es sich, als Maß der zentralen Tendenz den *Modalwert* der Testwertverteilung zu bestimmen. Der Modalwert ist der am häufigsten vorkommende Testwert in der Verteilung. Weist die Verteilung mehrere Modalwerte auf, kann das ein Hinweis für eine heterogen zusammengesetzte Stichprobe von Testpersonen, aber auch für eine mehrdeutige Instruktion oder für missverständliche Items sein.

Median

Ebenso wie der Modalwert ist auch die Berechnung des *Medians* bereits ab dem Ordinalskalenniveau der Testwerte möglich. Der Median (Mdn) der Testwertverteilung stellt jenen Testwert dar, der die Stichprobe bezüglich des interessierenden Merkmals in zwei gleich große Hälften zu je 50 % teilt. Das heißt mit anderen Worten, dass der Median derjenige Testwert ist, der von der Hälfte der n Testpersonen unterschritten oder erreicht und von der Hälfte der Testpersonen überschritten oder zumindest erreicht wurde.

Der Median einer geordneten Stichprobe (Y_1, Y_2, \dots, Y_n) von n Testwerten wird wie folgt errechnet:

$$Mdn = \begin{cases} \frac{Y_{\frac{n+1}{2}}}{2} & \text{für ungerade } n \\ \frac{1}{2}(Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}) & \text{für gerade } n \end{cases} \quad (8.1)$$

Arithmetischer Mittelwert

Der *arithmetische Mittelwert* \bar{Y} setzt intervallskalierte Testwerte voraus und kann bei n Testpersonen und m Items wie folgt berechnet werden:

$$\bar{Y} = \frac{\sum_{v=1}^n Y_v}{n} = \frac{\sum_{v=1}^n \sum_{i=1}^m y_{vi}}{n} \quad (8.2)$$

8.3 · Streuung der Testwertverteilung

Bei symmetrischen Verteilungen nehmen alle drei Maße denselben Wert an, bei nicht symmetrischen, also schiefen Verteilungen unterscheiden sie sich. Die Berechnung des Medians ist vor allem im Fall nicht-normalverteilter Testwerte sinnvoll, weil der Median robuster gegenüber Ausreißern ist als der arithmetische Mittelwert und die zentrale Tendenz besser charakterisiert.

8.3 Streuung der Testwertverteilung

Die Streuung der Testwertverteilung kann durch die Spannweite der Testwerte, den Interquartilabstand und durch die Varianz bzw. die Standardabweichung quantifiziert werden.

Die *Spannweite* („Range“) der Testwerte ist die Differenz aus dem höchsten beobachteten Testwert Y_{max} und dem niedrigsten beobachteten Testwert Y_{min} ($Range = Y_{max} - Y_{min}$).

Der *Interquartilabstand*, $IQR(Y)$, kann ebenso wie die Range bereits auf Ordinalskalenniveau der Testwerte bestimmt werden und bezeichnet die Differenz aus jenem Testwert, der von 25 % der Testpersonen überschritten wird, und jenem Testwert, der von 25 % der Testpersonen unterschritten wird (s. dazu auch ► Kap. 9). Der $IQR(Y)$ kann somit als Range zwischen den Quartilwerten $Q1$ und $Q3$ interpretiert werden.

Die *Testwertvarianz* setzt intervallskalierte Testwerte voraus und kann wie folgt berechnet werden:

$$Var(Y) = \frac{\sum_{v=1}^n (Y_v - \bar{Y})^2}{n - 1} \quad (8.3)$$

wobei \bar{Y} der Mittelwert der Testwerte ist. Die *Standardabweichung* $SD(Y)$ gewinnt man als positive Quadratwurzel aus der Varianz.

Sind die Testwerte normalverteilt, so lässt sich die Standardabweichung dahingehend interpretieren (vgl. ► Kap. 9), dass

- im Bereich des arithmetischen Mittelwertes ± 1 Standardabweichung ca. zwei Drittel der Testwerte vorzufinden sind; ein Drittel der Testpersonen weist extreme Testwerte auf, und zwar je zur Hälfte im oberen und im unteren Bereich;
- im Bereich des arithmetischen Mittelwertes ± 2 Standardabweichungen ca. 95 % der Testwerte vorzufinden sind; lediglich 5 % der Testpersonen weisen dann noch extremere Testwerte auf.

Für beliebig verteilte Testwerte lässt sich aus der sog. „Tschebyscheff’schen Ungleichung“ folgern, dass wesentlich höhere Prozentanteile außerhalb des Bereichs von ± 1 bzw. ± 2 Standardabweichungen zu erwarten sind (Näheres dazu findet sich z. B. in Eid et al. 2017).

Range/Spannweite

Interquartilabstand

Testwertvarianz und Standardabweichung

Interpretation der Standardabweichung

8.4 Beurteilung der Verteilungsform

Die Berechnung von *Schiefe* und *Exzess* erlaubt eine Beurteilung, ob die Form der Testwertverteilung von der Normalverteilung abweicht.

Dabei berechnet man die *Schiefe* („skewness“) der Testwertvariablen Y wie folgt:

$$Schiefe(Y) = \frac{\sum_{v=1}^n (Y_v - \bar{Y})^3}{n SD(Y)^3} \quad (8.4)$$

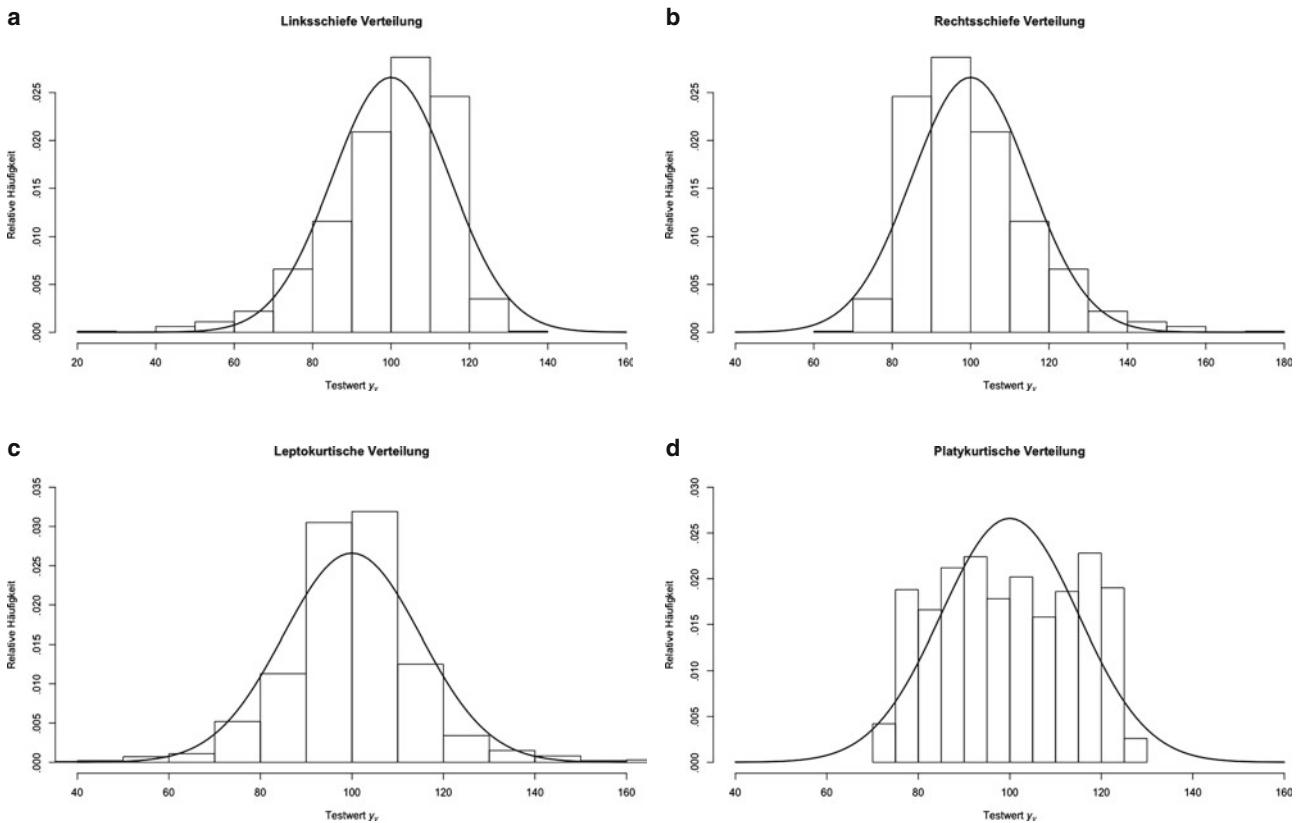


Abb. 8.1 Beispiele für die Schiefe der Verteilungen fiktiver Testwerte. Zum Vergleich ist jeweils eine Normalverteilung (schwarze Kurve) hinzugefügt. Beide Verteilungen haben jeweils einen Mittelwert von 100 und eine Standardabweichung von 15. **a** Linksschiefe (rechtssteile) Verteilung. **b** Rechtsschiefe (linkssteile) Verteilung. **c** Leptokurtische (steilgipflig, supergaußförmig) Verteilung. **d** Platykurtische (flachgipflig, subgaußförmig) Verteilung

Linksschiefe vs. rechtsschiefe Verteilung

Logarithmische Testwerttransformation zur Reduzierung der Schiefe

Ist die Schiefe ($Y < 0$), dann ist die Verteilung linksschief, d. h. rechtssteil (Abb. 8.1a); ist die Schiefe ($Y > 0$), so ist die Verteilung rechtsschief, d. h. linkssteil (Abb. 8.1b). Bei linksschiefen Verteilungen beispielsweise sind Werte, die größer als der Mittelwert sind, häufiger zu beobachten, so dass sich der Median rechts vom Mittelwert befindet; der linke Teil der Verteilung ist „flacher“ als der rechte.

Die Schiefe einer Verteilung kann durch geeignete Datentransformationen reduziert werden (s. dazu auch ► Abschn. 8.6). Bei einer rechtsschiefen Verteilung besteht die einfachste Transformationsform in der *Logarithmierung* der Testwerte. Die Logarithmierung beinhaltet, dass jeder Testwert Y_v in einen neuen Testwert Y'_v transformiert wird. Dies geschieht dadurch, dass man für jeden Testwert Y_v den logarithmierten Testwert $Y'_v = \ln Y_v$ berechnet. Dabei werden insbesondere die *Ausreißer* der rechtsschiefen Verteilung „näher“ an die Mitte der Verteilung verschoben (Achtung: Bei linksschiefen Verteilungen würde sich die Schiefe durch die Logarithmierung hingegen weiter verstärken). Spezialfälle von *Testwerttransformationen* anhand einer Logarithmusfunktion stellen z. B. das Box-Cox-Verfahren (Box und Cox 1964) oder die Yeo-Johnson-Transformation (Yeo und Johnson 2000) dar.

Den *Exzess* (Kurtosis, engl. „curtosis“) der Testwertvariablen Y berechnet man wie folgt:

$$\text{Exzess}(Y) = \frac{\sum_{v=1}^n (Y_v - \bar{Y})^4}{n SD(Y)^4} - 3 \quad (8.5)$$

8.5 · Ursachen für die Abweichung der Testwertverteilung von der Normalverteilung

Ist der Exzess (Y) = 0, so entspricht die Wölbung der Verteilung der Wölbung einer Normalverteilung („Gauß’sche Glockenkurve“). Diese Form der Verteilung heißt *mesokurtisch*. Ist hingegen der Exzess (Y) > 0, so handelt es sich um eine im Vergleich zur Normalverteilung schmalere, „spitzere“ Testwertverteilung, d. h. eine Verteilung mit einer stärkeren Wölbung; sie wird dann *leptokurtisch* (auch schmalgipflig, steilgipflig, supergaußförmig) genannt (► Abb. 8.1c). Ist der Exzess (Y) < 0, so ist die Verteilung vergleichsweise zu flach; sie wird dann *platykurtisch* (auch breitgipflig, flachgipflig, subgaußförmig) genannt (► Abb. 8.1d).

Mesokurtische, leptokurtische und plattykurtische Verteilungen

8.5 Ursachen für die Abweichung der Testwertverteilung von der Normalverteilung

Bei psychologischen Merkmalen kann eine normalverteilte Testwertverteilung häufig dahingehend interpretiert werden, dass der Test in Bezug auf die Schwierigkeit angemessene Anforderungen an die Testpersonen stellt. Weicht die Testwertverteilung hingegen deutlich von der Normalverteilung ab, so können verschiedene Ursachen unterschieden werden (vgl. Lienert und Raatz 1998).

Als *erste Ursache* kommt im Kontext der Testentwicklung eine *mangelhafte Konstruktion* des Tests in Frage. So ist z. B. einer linksschiefen, d. h. rechtssteilen Verteilung (► Abb. 8.1a) zu entnehmen, dass der Test insgesamt „zu leicht“ in dem Sinne ist, dass ein großer Teil der Testpersonen mehr als die Hälfte der Aufgaben beantworten konnte. Umgekehrt zeigt eine rechtsschiefe, d. h. linkssteile Verteilung (► Abb. 8.1b) an, dass der Test insgesamt „zu schwer“ ist, weil ein großer Teil der Testpersonen weniger als die Hälfte der Items beantworten konnte. Eine schmalgipflige, leptokurtische Verteilung (► Abb. 8.1c) zeigt an, dass der Test zu viele Items im mittleren Schwierigkeitsbereich enthält, aber zu wenige leichte bzw. schwere Items; bei breitgipfligen, plattykurtischen Verteilungen (► Abb. 8.1d) verhält es sich umgekehrt.

Erste Ursache: Konstruktionsmängel

Zu leichte bzw. zu schwere Tests wie auch Tests mit zu vielen bzw. zu wenigen Items im mittleren Schwierigkeitsbereich können im Zuge einer *Testrevision* durch Ergänzung von Items im unterrepräsentierten Schwierigkeitsbereich an das Niveau der Testpersonen angepasst werden. Eine Alternative stellt die *Normalisierung der Testwertverteilung* dar, bei der die Testwerte Y_v , z. B. mittels *Flächentransformation* so transformiert werden, dass die transformierten Testwerte der Normalverteilung folgen (► Abschn. 8.6). Dabei bleibt die geringere Differenzierungsfähigkeit des Tests im Bereich unterrepräsentierter Items allerdings erhalten (vgl. hierzu die Überlegungen zur „Informationsfunktion“ eines Tests in ► Kap. 16).

Zweite Ursache: heterogene Stichproben

Als *zweite Ursache* ist denkbar, dass es sich um eine *heterogene Stichprobe* handelt. Das bedeutet, dass sie sich aus Unterstichproben zusammensetzt, die für sich genommen durchaus normalverteilt sein können, aber zusammengekommen eine Mischverteilung bilden, die von der Normalverteilung abweicht. Dies kann daran liegen, dass die Unterstichproben unterschiedliche Mittelwerte und/oder unterschiedliche Varianzen aufweisen, sodass die resultierende Gesamtstichprobe von der Normalverteilung abweicht. So könnte es sich bei der plattykurtischen breitgipfligen Verteilung in ► Abb. 8.1d mit ihren drei Modalwerten z. B. um eine Mischverteilung aus drei heterogenen Unterstichproben handeln. Die Problematik unterschiedlicher Unterstichproben sollte bei der Testeichung ggf. in Form von differenzierten Testnormen berücksichtigt werden (vgl. ► Kap. 9).

Dritte Ursache: nicht normalverteilte Merkmale

Eine *dritte Ursache* könnte darin bestehen, dass das erhobene Merkmal auch in der Population *nicht normalverteilt* ist (z. B. Reaktionsfähigkeit). In einem solchen Fall hat der Testautor die Abweichung von der Normalverteilung nicht zu verantworten; vielmehr sollte er dann auch nicht daran interessiert sein, das Merkmal so zu erfassen, dass normalverteilte Testwerte resultieren.

8.6 Normalisierung der Testwertverteilung

Wenn sich die Annahme vertreten lässt, dass das gemessene Merkmal eigentlich normalverteilt ist und nur die Testwertverteilung in der Stichprobe eine Abweichung von der Normalverteilung aufweist, so kann eine nichtlineare Transformation der Testwerte vorgenommen werden, bei der die Testwertverteilung an die Normalverteilung angepasst wird. Diesen Vorgang bezeichnet man als *Normalisierung* (► Kap. 9).

Normalisierung vs. Normierung

! Von der nichtlinearen Transformation der Normalisierung zu unterscheiden ist die sog. *Normierung*, die eine lineare Transformation der Daten zwecks Interpretation vor dem Hintergrund eines Bezugsrahmens, der sog. Normverteilung, vornimmt.

8.6.1 Nichtlineare Flächentransformation

Eine an keine bestimmte Verteilungsform gebundene Möglichkeit der Normalisierung von nicht normalverteilten Testwerten besteht in der nichtlinearen Flächentransformation in Anlehnung an McCall (1939).

Definition

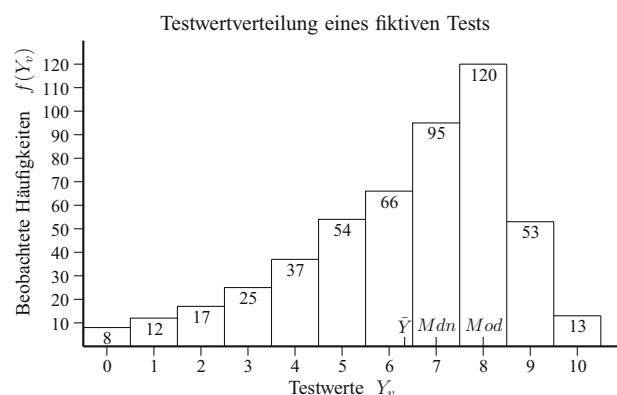
Bei der **Flächentransformation** bleiben die jeweiligen Flächen der einzelnen Histogrammsäulen – dem Prinzip der Flächentreue folgend – unverändert; hingegen werden die Breiten der Histogrammsäulen so verändert, dass die resultierenden Höhen der einzelnen Säulen mit denen einer Normalverteilung übereinstimmen.

Anpassung der Histogrammsäulen an die Normalverteilung

Im Folgenden wird die Flächentransformation an einem *Beispiel* genauer erläutert.

Gegeben sei eine Testwertverteilung von $n = 500$ Testpersonen in einem fiktiven Test, bei dem die Testwerte Y_v die Ausprägungen $0, 1, 2, \dots, 10$ annehmen können ($\bar{Y} = 6.334, SD(Y) = 2.234, Schiefe(Y) = -0.816, Exzess(Y) = 0.135$). Wie □ Abb. 8.2 zeigt, weist sie eine linksschiefe, d. h. rechtssteile und leptokurtische Form auf. Die Testwerte und die zugehörigen beobachteten Häufigkeiten der Testwerte sind in □ Tab. 8.1, Spalte 1 und 2, aufgeführt (► Abschn. 8.6.2).

Wenn sich die hier beobachtbare Anhäufung überdurchschnittlich hoher Testwerte z. B. darauf zurückführen lässt, dass die Items für die untersuchte Stichprobe zu leicht waren, so kann eine *Normalisierung der Testwerte* vorgenommen werden; sie ist vor allem dann rechtfertigbar, wenn für das mit dem Test erfasste Merkmal eigentlich eine Normalverteilung angenommen werden kann.



■ Abb. 8.2 Testwertverteilung eines fiktiven Tests für 500 Testpersonen. Abgetragen sind die beobachteten Häufigkeiten $f(Y_v)$ DER Testwerte Y_v , der Mittelwert \bar{Y} sowie der Median (Mdn) und der Modalwert (Mod)

8.6.2 Ablaufschritte

Der Ablauf der Flächentransformation erfolgt in drei Schritten.

■■ Schritt 1: Bildung von kumulierten relativen Häufigkeiten (bzw. Prozenträngen)

Im ersten Schritt werden für die Testwerte die relativen Häufigkeiten $p(Y_v)$ und die kumulierten relativen Häufigkeiten $p_{\text{cum}}(Y_v)$ gebildet, die mit dem Faktor 100 multipliziert Prozentränge ergeben (vgl. hierzu ► Abschn. 9.2.1). Die kumulierten relativen Häufigkeiten geben den Anteil der Testpersonen an, die einen Testwert aufweisen, der kleiner oder gleich Y_v ist. Dem Testwert $Y_v = 0$ entspricht z. B. ein $p_{\text{cum}}(Y_v = 0)$ von 0.016 (bzw. ein Prozentrang von 1.6) und dem Testwert $Y_v = 1$ ein $p_{\text{cum}}(Y_v = 1)$ von 0.040 (bzw. ein Prozentrang von 4.0); für alle weiteren Testwerte gilt Entsprechendes (► Tab. 8.1, Spalten 3 und 4). Bezogen auf die Gesamtfläche der relativen Häufigkeitsverteilungen von 1 (bzw. bezogen auf die Gesamtfläche der prozentualen Verteilung von 100) geben die kumulierten relativen Häufigkeiten (bzw. die Prozentränge) die Flächenanteile der einzelnen Histogrammsäulen in der beobachteten Häufigkeitsverteilung an, und zwar beginnend mit dem kleinsten Testwert bis einschließlich dem jeweiligen Testwert.

■■ Schritt 2: Bestimmung von z-Werten der Standardnormalverteilung gemäß den Flächenanteilen der beobachteten Häufigkeitsverteilung

Im zweiten Schritt werden die mit den kumulierten Flächenanteilen der beobachteten Histogrammsäulen korrespondierenden Klassengrenzen aus der Standardnormalverteilung (Gesamtfläche = 1) gesucht. Hierbei lässt man sich von dem Gedanken leiten, welche z -Werte der Standardnormalverteilung („ z -Verteilung“) als Klassengrenzen hätten auftreten müssen, damit die jeweiligen Histogrammsäulen unter Beibehaltung ihrer Fläche, aber unter Veränderung ihrer Breite und Höhe der Normalverteilung entsprechen. Man spricht hierbei vom *Prinzip der*

Schritt 1

Bildung von kumulierten relativen Häufigkeiten (bzw. Prozenträngen)

Schritt 2

Korrespondierende Klassengrenzen der Standardnormalverteilung

► **Tabelle 8.1** Häufigkeitstabelle der Testwerte (Y_v) von $n = 500$ Testpersonen; beobachtete ($f(Y_v)$), relative ($p(Y_v)$) und kumulierte relative Häufigkeiten ($p_{\text{cum}}(Y_v)$) der Testwerte, untere und obere Klassengrenzen von z_v , Klassenbreiten von z_v , normalisierte Testwerte z_v (Klassenmitte) und Höhe der Histogrammsäule

Y_v	$f(Y_v)$	$p(Y_v)$	$p_{\text{cum}}(Y_v)$	Untere Klassengrenze von z_v	Obere Klassengrenze von z_v	Klassenbreite von z_v	Normalisierter Testwert z_v	Höhe der Histogrammsäule von z_v
0	8	0.016	0.016	-3.00	-2.14	0.86	-2.572	0.019
1	12	0.024	0.040	-2.14	-1.75	0.39	-1.948	0.061
2	17	0.034	0.074	-1.75	-1.45	0.30	-1.599	0.112
3	25	0.050	0.124	-1.45	-1.16	0.29	-1.301	0.172
4	37	0.074	0.198	-1.16	-0.85	0.31	-1.002	0.241
5	54	0.108	0.306	-0.85	-0.51	0.34	-0.678	0.316
6	66	0.132	0.438	-0.51	-0.16	0.35	-0.332	0.376
7	95	0.190	0.628	-0.16	0.33	0.48	0.085	0.394
8	120	0.240	0.868	0.33	1.12	0.79	0.722	0.304
9	53	0.106	0.974	1.12	1.94	0.83	1.530	0.128
10	13	0.026	1.000	1.94	3.00	1.06	2.472	0.025

z-Tabelle**Konkrete Gewinnung von z -Werten**

Flächentreue. Basierend auf dieser Überlegung werden somit diejenigen z -Werte¹ bestimmt, welche die Fläche unter der Standardnormalverteilung entsprechend den vorgefundenen kumulierten relativen Häufigkeiten $p_{\text{cum}}(Y_v)$ der Testwerte Y_v so voneinander trennen, dass die resultierenden Histogrammsäulen hinsichtlich ihrer Breite und Höhe der Normalverteilung entsprechen.

Um die jeweiligen Klassengrenzen der normalisierten Testwerte z_v zu finden, wird die *Verteilungsfunktion der Standardnormalverteilung* („ z -Tabelle“) herangezogen, die im Anhang tabelliert ist. Die Tabelle zeigt, welche Flächenanteile unter der Standardnormalverteilung mit welchen z -Werten korrespondieren. Die mit den kumulierten Flächenanteilen der Histogrammsäulen korrespondierenden z -Werte fungieren dann als Klassengrenzen der gesuchten normalisierten z_v -Werte.

In unserem Beispiel geht man konkret wie folgt vor: Dem Testwert $Y_v = 0$ ordnet man auf Grundlage der kumulierten relativen Häufigkeit $p_{\text{cum}}(Y_v = 0)$ von 0.016 aus der Standardnormalverteilung als obere Klassengrenze jenen z -Wert zu, der von 1.6 % aller Fälle unterschritten wird, nämlich den z -Wert von -2.14 . Anschließend findet man für den Testwert $Y_v = 1$ mit der kumulierten relativen Häufigkeit $p_{\text{cum}}(Y_v = 1)$ von 0.024 als obere Klassengrenze den z -Wert -1.75 , der von 4 % aller Fälle in der Standardnormalverteilung unterschritten wird; gleichermaßen wird mit den oberen Klassengrenzen der weiteren Testwerte verfahren.

Somit werden als Klassengrenzen der normalisierten Testwerte anhand der in Schritt 1 berechneten kumulierten relativen Häufigkeit diejenigen z -Werte bestimmt, die – ausgehend vom kleinsten Testwert – jene kumulierten Flächenanteile unter der Standardnormalverteilung kennzeichnen, die den kumulierten relativen Häufigkeiten $p_{\text{cum}}(Y_v)$ entsprechen. Anstelle des z -Wertes von $-\infty$, der eigentlich als *untere Klassengrenze* des Testwertes $Y_v = 0$ fungieren müsste, verwendet man üblicherweise den konkreten z -Wert von -3 ; ebenso wird mit der *oberen Klassengrenze* des Testwertes $Y_v = 10$ verfahren, wobei anstelle von $+\infty$ der konkrete z -Wert von $+3$ verwendet wird (► Tab. 8.1, Spalten 5 und 6).

■ ■ **Schritt 3: Bestimmung der normalisierten Testwerte z_v**

Durch Subtraktion der jeweiligen Klassengrenzen lassen sich nun die aus der Flächentransformation resultierenden Klassenbreiten der neuen Histogrammsäulen berechnen. Abschließend werden durch Mittelung der Klassengrenzen die gesuchten normalisierten Testwerte z_v als Klassenmitten bestimmt.

Mit der Bestimmung der Klassenmitten z_v ist die Flächentransformation zur Gewinnung normalverteilter Testwerte abgeschlossen. Bei gegebenen Voraussetzungen können die normalisierten z_v -Werte anstelle der ursprünglichen Y_v -Werte für weitere anstehende Kalküle Verwendung finden. Die erhaltenen z_v -Werte können bei Bedarf so (linear) transformiert werden, dass sie auf einer anderen Skala interpretierbar sind. Wie man in ► Kap. 9 sehen wird, lassen sich z_v -Werte z. B. in IQ-Skalenwerte (Intelligenzquotienten-Skala) oder PISA-Skalenwerte (Skala der Studien des Programme for International Student Assessment [PISA] der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung [OECD]) transformieren.

Bezogen auf das Beispiel bildet man für den kleinsten Testwert die Differenz von -3.00 und -2.14 und erhält eine Klassenbreite von 0.86 , wobei die zugehörige Klassenmitte bei $z_v = -3.00 + 0.86/2 = -2.572$ liegt; für den zweiten Testwert bildet man die Differenz von -2.14 und -1.75 und erhält eine Klassenbreite von 0.39 , die zugehörige Klassenmitte liegt bei $z_v = -1.948$; für den dritten Testwert bildet man die Differenz von -1.75 und -1.45 und erhält eine Klassenbreite von 0.30 , die zugehörige Klassenmitte liegt bei $z_v = -1.599$; mit den übrigen Testwerten wird analog verfahren.

¹ Hinweis: Die hier beschriebene Gewinnung von z -Werten erfolgt *nicht* durch Anwendung der gebräuchlichen linearen z -Transformation („ z -Standardisierung“ oder auch „ z -Normierung“, ► Kap. 9), sondern anhand der nichtlinearen Flächentransformation.

8.6 · Normalisierung der Testwertverteilung

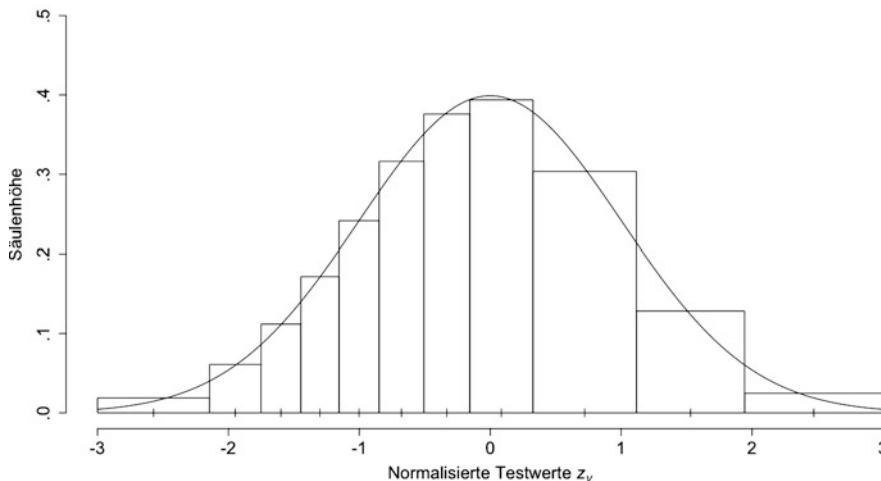


Abb. 8.3 Histogrammsäulen der an die Normalverteilung angepassten z_v -Werte nach Durchführung der nichtlinearen Flächentransformation. Zu Vergleichszwecken wurde die Dichte der Standardnormalverteilung eingefügt

In den Spalten 5 bis 8 der **Tab. 8.1** ist der Vorgang der Gewinnung von z -Werten für die jeweiligen Klassengrenzen und von normalisierten z_v -Werten (Klassenmittnen) für das Beispiel zusammengefasst. Wie man erkennt (s. auch **Abb. 8.2** und im Vergleich dazu **Abb. 8.3**), sind als Folge der Flächentransformation die Klassenbreiten der normalisierten Testwerte nicht mehr gleich, sondern sie variieren zwischen minimal 0.29 und maximal 1.06. Ebenso variieren die Abstände zwischen den normalisierten Testwerten z_v (Klassenmittten), und zwar in der Weise, dass die Abstände im linken Bereich der Verteilung verkleinert („gestaucht“), im rechten Bereich der Verteilung hingegen vergrößert („gespreizt“) sind, um die Nichtnormalität der beobachteten Testwerte auszugleichen.

Zur Beantwortung der Frage, wie die *Höhen der neuen Histogrammsäulen* in **Abb. 8.3** bestimmt wurden, kann man sich von folgender Überlegung leiten lassen:

- Die Flächen der jeweiligen neuen Histogrammsäulen sind wegen der Flächen-treue bekannt; sie sind – bezogen auf die Gesamtfläche unter der Standardnor-malverteilung von 1 – jeweils so groß wie die relative Häufigkeit $p(Y_v)$ des jeweiligen Testwertes;
- ebenfalls bekannt sind die Klassenbreiten der neuen Histogrammsäulen;
- dividiert man also die bekannten Flächen der Histogrammsäulen durch die neu-en Klassenbreiten, so erhält man die gesuchten neuen Säulen-höhen.

Höhen der neuen Histogrammsäulen

Bezogen auf das Beispiel hat der kleinste normalisierte Testwert von $z_v = -2.572$ eine Histogrammsäulenfläche von 0.016, die durch die Klassenbreite von 0.86 di-vidiert wird, woraus eine zugehörige Säulen-höhe von 0.019 resultiert; der zweite normalisierte Testwert von $z_v = -1.948$ hat eine Histogrammsäulenfläche von 0.024, die durch die Klassenbreite von 0.39 dividiert wird, woraus eine zugehörige Säulen-höhe von 0.061 resultiert; der dritte normalisierte Testwert von $z_v = -1.599$ hat eine Histogrammsäulenfläche von 0.034, die durch die Säulenbreite von 0.30 di-vidiert wird, woraus eine zugehörige Säulen-höhe von 0.112 resultiert; mit den weiteren Testwerten wird analog verfahren. Die Ergebnisse für die jeweiligen Säulen-höhen sind als letzte Spalte in **Tab. 8.1** aufgeführt.

8.7 Zusammenfassung und weiteres Vorgehen

Über die deskriptivstatistische Itemanalyse hinausgehend liefert die Analyse der Testwertverteilung Informationen über die zu einem Test zusammengefassten Items. Mit den Maßen der zentralen Tendenz und der Streuung sowie von Schiefe und Exzess lassen sich die wesentlichen Eigenschaften der Testwertverteilung untersuchen. Abweichungen der Testwertverteilung von der Normalverteilung erlauben Rückschlüsse auf ungünstige Zusammensetzungen der Itemschwierigkeiten, die im Zuge von Testrevisionen ausgeglichen werden können. In begründeten Fällen kann eine Normalisierung der Testwerte vorgenommen werden.

Weiteres Vorgehen: Nachdem die Testwertverteilung überprüft und ggf. eine Normalisierung vorgenommen wurde, gilt es, durch eine Testnormierung einen Bezugsrahmen für die gemessenen Testwerte herzustellen, um diese norm- oder auch kriteriumsorientiert interpretieren zu können. Die fachgerechte Interpretation von Testwerten ist Gegenstand von ► Kap. 9.

Genauere Qualitätsuntersuchungen der Items, die zu einem Test(summen)wert zusammengefasst werden, erfordern eine testtheoretische Untermauerung wie die der Klassischen Testtheorie (KTT, ► Kap. 13), auf deren Annahmen die Reliabilitätsanalysen vorgenommen werden können. Verschiedene Methoden der Reliabilitätsbestimmung (► Kap. 14), die auf unterschiedlichen Formen der Messäquivalenz basieren, erlauben eine Beurteilung der Messgenauigkeit des Messinstruments. Vor allem auf Basis der Annahmen der Item-Response-Theorie (IRT, ► Kap. 16) und der konfirmatorischen Faktorenanalyse (CFA, ► Kap. 24) sind auch genauere Beurteilungen der Itemhomogenität und der Dimensionalität von Tests und Fragebogen möglich, um belastbare Aussagen zur Validität der Verfahren zu gewinnen (► Kap. 21).

8.8 EDV-Hinweise

Mit einschlägiger Statistiksoftware (z. B. R, SPSS) können Itemschwierigkeiten, Itemmittelwerte, Itemvarianzen und Trennschärfen sowie die Kennwerte der Item- und Testwertverteilungen einfach berechnet werden. Ein Datenbeispiel finden Sie im Bereich EDV-Hinweise unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

8.9 Kontrollfragen

❓ Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Welche Maße würden Sie bestimmen, um zu beurteilen, ob eine Testwertverteilung von der Normalverteilung abweicht?
2. Welche Ursachen für die Abweichung der Testwertverteilung von der Normalverteilung kennen Sie?
3. In welchen Schritten erfolgt die Normalisierung einer Testwertverteilung?

Literatur

- Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–246.
 Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
 Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Literatur

- McCall, W. A. (1939). *Measurement*. New York: McMillan.
- Yeo, I.-K. & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.



Testwertinterpretation, Testnormen und Testeichung

Frank Goldhammer und Johannes Hartig

Inhaltsverzeichnis

- 9.1 Testwertbildung und Testwertinterpretation – 172
- 9.2 Normorientierte Testwertinterpretation – 173
 - 9.2.1 Bildung von Prozentrangnormen durch nichtlineare Testwerttransformation – 173
 - 9.2.2 Bildung von standardisierten z_v -Normwerten durch lineare Testwerttransformation – 176
- 9.3 Kriteriumsorientierte Testwertinterpretation – 179
 - 9.3.1 Bezug des Testwertes auf ein externes Kriterium – 180
 - 9.3.2 Bezug des Testwertes auf Aufgabeninhalte – 185
- 9.4 Integration von norm- und kriteriumsorientierter Testwertinterpretation – 187
- 9.5 Normdifferenzierung – 188
- 9.6 Testeichung – 189
 - 9.6.1 Definition der Zielpopulation – 190
 - 9.6.2 Erhebungsdesigns für Normierungsstichproben – 190
 - 9.6.3 Dokumentation der Normen im Testmanual – 192
- 9.7 Zusammenfassung mit Anwendungsempfehlungen – 193
- 9.8 EDV-Hinweise – 194
- 9.9 Kontrollfragen – 194
- Literatur – 195

i Wendet man einen psychologischen Test an, so erhält man in der Regel einen numerischen Testwert, der eine quantifizierte Auskunft über die Merkmalsausprägung der Testperson geben soll. Die Frage, wie dieser Testwert hinsichtlich der Merkmalsausprägung zu interpretieren ist, führt leicht zu Fehlinterpretationen. Erst die Hinzunahme weiterer Informationen erlaubt es, diese Frage in zweierlei Weise sinnvoll zu beantworten: einerseits dadurch, dass der Testwert in einen Normwert transformiert wird, der einen interpretativen Vergleich mit den Testwerten einer Bezugsguppe ermöglicht (*normorientierte Interpretation*), oder andererseits dadurch, dass eine genaue theoretische Vorstellung darüber besteht, wie der erzielte Testwert mit einem inhaltlich-psychologisch definierten Kriterium in Beziehung steht (*kriteriumsorientierte Interpretation*).

9.1 Testwertbildung und Testwertinterpretation

Testwert und Rohwert

9

Rohwerte sind uneindeutig

Bevor ein *Testwert*, d. h. das numerische Testresultat einer Testperson, interpretiert werden kann, muss er gemäß definierter Regeln gebildet werden (vgl. ► Kap. 8). In Abhängigkeit von dem im psychologischen Test gewählten Antwortformat sind die Regeln zur Testwertbildung unterschiedlich komplex. Im Falle von frei formulierten Verbalantworten ist eine ausführliche Anleitung nötig, um den Itemantworten in differenzierter Weise Punktwerte zuzuweisen, die zusammengekommen den Testwert ergeben. Bei Wahlaufgaben genügt dagegen ein Antwortschlüssel, anhand dessen einer gegebenen Itemantwort ein bestimmter Punktwert zugeordnet wird (z. B. Multiple-Choice-Persönlichkeitsfragebogen), oder aber es wird das Verhältnis von Arbeitsmenge und Bearbeitungszeit zur Ermittlung des Testwertes herangezogen (z. B. Leistungstests mit Zeitbegrenzung). Der Testwert wird auch als *Rohwert* bezeichnet, da er sich unmittelbar aus den registrierten Antworten ergibt und noch nicht weitergehend verarbeitet ist.

Obwohl also der Testwert das Antwortverhalten der Testperson widerspiegelt, führt die Interpretation des Testrohwertes ohne die Hinzunahme weiterer Informationen leicht zu Fehlinterpretationen, wie im nachfolgenden ► Beispiel 9.1 gezeigt wird.

Beispiel 9.1: Mangelnde Aussagekraft von Testrohwerten

Der Schüler Peter kann in einer Rechenprobe 18 von 20 Aufgaben richtig lösen, d. h. 90 % der Aufgaben. In der Vokabelprobe schafft er 21 von 30 Aufgaben, d. h. nur 70 %. Seine Eltern freuen sich über die Rechenleistung ihres Sohnes, mit der Vokabelliste sind sie weniger zufrieden. Zu Recht wendet Peter ein, dass die meisten seiner Mitschüler nur eine deutlich geringere Anzahl von Vokabellaufgaben richtig lösen konnten, d. h. im Vergleich zu den anderen Schülern habe er sehr gut abgeschnitten. Außerdem verteidigt er sich damit, dass in der Probe einige Vokabeln abgefragt wurden, die der Lehrer im Unterricht kaum vorbereitet hatte, d. h. die Vokabelprobe war sehr schwierig. Auch wenn Peter mit dieser Argumentation seine Leistung in der Vokabelprobe zu seinen Gunsten relativieren kann, stellt sich analog die Frage, ob Peter seine gute Rechenleistung der Auswahl einfacher Aufgaben verdankt, d. h., ob viele seiner Mitschüler ebenfalls mit Leistungen im oberen Punktbereich abgeschnitten haben.

Das Beispiel macht deutlich, dass der Testrohwert (hier z. B. der Anteil gelöster Aufgaben) per se nicht aussagekräftig ist, da die Höhe des Testwertes nicht nur von dem interessierenden Merkmal (z. B. Rechenfähigkeit) abhängt, sondern auch wesentlich durch die Aufgabenauswahl bzw. die Aufgabenschwierigkeit mitbestimmt wird.

9.2 · Normorientierte Testwertinterpretation

Um die aus der Aufgabenauswahl resultierende Uneindeutigkeit von Testwerten zu vermeiden, werden Testwerte anhand von Vergleichsmaßstäben interpretiert.

Vergleichsmaßstäbe zur Testwertinterpretation

Um eine aussagekräftige Entscheidung über die individuelle Merkmalsausprägung treffen zu können, wird zusätzlich zum Testwert ein *Vergleichsmaßstab* benötigt, anhand dessen der Testwert eingeordnet werden kann. Als Vergleichsmaßstäbe können entweder Merkmalsverteilungen von Bezugsgruppen (*normorientierte Testwertinterpretation*) oder genaue psychologisch-inhaltliche Beschreibungen, die für die Testwertausprägungen charakteristisch sind (*kriteriumsorientierte Testwertinterpretation*), herangezogen werden.

Bei der normorientierten Testwertinterpretation (► Abschn. 9.2) erhält man Informationen über die individuelle Merkmalsausprägung relativ zur Bezugsgruppe, wobei das Problem der Abhängigkeit des Testwertes von der Aufgabenauswahl insofern Berücksichtigung findet, als sich die Information aus der Testnorm nur mehr auf die relative Position der Testperson innerhalb der Bezugsgruppe bezieht und nicht mehr direkt von der Aufgabenauswahl abhängt. Eine solche Norm heißt auch *Realnorm*.

Bei der kriteriumsorientierten Testwertinterpretation (► Abschn. 9.3) wird die unkontrollierte Abhängigkeit des Testwertes von der Aufgabenauswahl von vornherein vermieden, indem eine genaue theoretische Vorstellung darüber besteht, wie das Beantworten bestimmter Testaufgaben und somit der jeweilige Testwert mit einem genau definierten psychologisch-inhaltlichen Kriterium in Beziehung steht. Eine solche Norm heißt auch *Idealnorm*.

Relation zu Bezugsgruppe

Relation zu inhaltlichem Kriterium

Testrohwert vs. Normwert

9.2 Normorientierte Testwertinterpretation

Die normorientierte Testwertinterpretation besteht darin, dass zu einem individuellen Testwert (Rohwert) ein *Normwert* bestimmt wird, anhand dessen die Testperson hinsichtlich der erfassten Merkmalsausprägung innerhalb der Bezugs- bzw. Referenzgruppe positioniert werden kann. Die Gewinnung von repräsentativen Normierungsstichproben für definierte Bezugsgruppen (Testeichung) wird später in ► Abschn. 9.6 beschrieben.

Im Rahmen normorientierter Testwertinterpretation lassen sich grundsätzlich zwei Vorgehensweisen unterscheiden, wie für einen bestimmten Testwert (Rohwert) ein Normwert ermittelt werden kann, und zwar zum einen die nichtlineare Transformation des Testwertes zur Gewinnung von *Prozentrangnormen* sowie zum anderen die lineare Transformation des Testwertes zur Gewinnung von *standardisierten z-Normwerten*.

9.2.1 Bildung von Prozentrangnormen durch nichtlineare Testwerttransformation

Die gebräuchlichste *nichtlineare Testwerttransformation* ist die Transformation des Testwertes in einen *Prozentrang*. Nichtlinearität bedeutet hierbei, dass der Prozentrang durch eine *Transformation der Testwertverteilung* der Bezugsgruppe gewonnen wird (und nicht durch eine lineare Transformation des Testwertes selbst, ► Abschn. 9.2.2). Dazu wird die relative Position des Testwertes Y_v in der aufsteigend geordneten Rangreihe der Testwerte in der Bezugsgruppe ermittelt.

**Nichtlineare
Testwerttransformation:
Prozentrang**

Definition

Ein **Prozentrang** gibt an, wie viel Prozent der Bezugsgruppe bzw. Normierungsstichprobe einen Testwert erzielen, der niedriger oder maximal ebenso hoch ist, wie der Testwert Y_v der Testperson v . Der Prozentrang entspricht somit dem prozentualen Flächenanteil der Testwertverteilung der Bezugsgruppe, der am unteren Skalenende beginnt und nach oben hin durch den Testwert Y_v begrenzt wird.

Um die Zuordnung eines Testwertes Y_v zu seinem korrespondierenden Prozentrang einfach vornehmen zu können, wird im Rahmen der Testkonstruktion aus den Testwerten Y_v der Normierungsstichprobe eine tabellarische Prozentrangnorm gebildet (vgl. ▶ Kap. 8).

Prozentrangnorm

Zur Erstellung einer *Prozentrangnorm* werden für alle Testwerte Y_v der Normierungsstichprobe die zugehörigen Prozentränge PR_v folgendermaßen bestimmt:

- Die Testwerte Y_v der Normierungsstichprobe vom Umfang n werden in eine aufsteigende Rangordnung gebracht.
- Die Häufigkeiten $freq(Y_v)$ der einzelnen Testwertausprägungen werden erfasst.
- Die kumulierten Häufigkeiten $freq_{cum}(Y_v)$ bis einschließlich des jeweiligen Testwertes Y_v werden bestimmt.
- Die kumulierten Häufigkeiten werden durch den Umfang n der Normierungsstichprobe dividiert und mit dem Faktor 100 multipliziert.

$$PR_v = 100 \cdot \frac{freq_{cum}(Y_v)}{n} \quad (9.1)$$

Prozentrang-Normtabelle

Aus den so gewonnenen Prozenträngen PR_v wird eine *Prozentrang-Normtabelle* gebildet, in der jedem Prozentrang zwischen 1 und 100 der jeweils zugehörige Testwert Y_v zugeordnet ist. Dabei ist es möglich, dass einem Intervall von unterschiedlichen Testwerten im Bereich geringer Testwertdichte derselbe Prozentrang zugeordnet wird; im Bereich hoher Testwertdichte können umgekehrt demselben Testwert mehrere aufeinander folgende Prozentränge zugeordnet werden (▶ Beispiel 9.2). Liegt für ein psychologisches Testverfahren eine Prozentrangnormierung vor, kann der Testwert Y_v einer Testperson dadurch interpretiert werden, dass anhand der Normtabelle der dem Testwert Y_v entsprechende Prozentrang abgelesen wird, der den relativen Grad der individuellen Merkmalsausprägung im Vergleich zu jenen in der Normierungsstichprobe angibt.

Perzentil, Quartile und Median

Während der Prozentrang die relative Position der Merkmalsausprägung einer Testperson in der Normierungsstichprobe beschreibt, bezeichnet das *Perzentil* jenen Testwert Y_v , der einem bestimmten Prozentrang in der Normierungsstichprobe zugeordnet ist. Das heißt z. B., dass derjenige Testwert, der von 30 % der Testpersonen unterschritten bzw. höchstens erreicht wird, 30. Perzentil genannt wird. Die grobstufigeren *Quartile* entsprechen dem 25., 50. bzw. 75. Perzentil. Als 1. Quartil (Q1) wird demnach derjenige Testwert bezeichnet, der von 25 % der Testpersonen unterschritten oder erreicht wird; das 2. Quartil (Q2) ist derjenige Testwert, den 50 % der Testpersonen unterschreiten oder erreichen, d. h. Q2 entspricht dem *Median*. Das 3. Quartil (Q3) schließlich ist derjenige Testwert, den 75 % der Testpersonen unterschreiten oder erreichen.

Beispiel 9.2: Prozentrangnorm

Für den computerbasierten Frankfurter Adaptiven Konzentrationsleistungs-Test (FAKT-II; Moosbrugger und Goldhammer 2007) liegen Prozentrangnormen vor, die in Bezug auf die Testform, die Durchführungsart und die Testungszahl differenziert sind (s. dazu auch ▶ Abschn. 9.5). Auf dem Ergebnisbogen wird u. a. für den Testwert „Konzentrations-Leistung KL1“ automatisch ein Prozentrang ausgegeben.

9.2 · Normorientierte Testwertinterpretation

Beispielsweise hat eine Testperson, die bei erstmaliger Bearbeitung des FAKT-II mit Testform E und der Durchführungsart „Standardtestzeit von 6 Minuten“ für die Konzentrations-Leistung KL1 einen Testwert von $Y_v = 146$ erzielt, den Prozentrang $PR_v = 76$. Der Prozentrang von 76 zeigt an, dass 76 % der Normierungsstichprobe eine geringere oder gleich hohe Leistung gezeigt haben; die Leistung der Testperson liegt oberhalb des 3. Quartils; sie zählt somit zum leistungsstärksten Viertel der Bezugsgruppe.

Zur Ermittlung eines Prozentrangs sucht das Computerprogramm in der entsprechenden Normtabelle automatisch denjenigen Prozentrang, der dem von der Testperson erzielten Testwert entspricht; bei fehlender Entsprechung wird der nächsthöhere Prozentrang ausgegeben.

Testwert KL1	Prozentrang
...	...
145	74
146	75
146	76
147	77
...	...

Zu beachten ist, dass in der Normierungsstichprobe der Testwert 146 relativ häufig auftrat (hohe Testwertdichte), sodass derselbe Testwert sowohl das 75. als auch das 76. Perzentil bezeichnet. In solch einem Fall wird gemäß obiger Prozentrangdefinition dem Probanden der jeweils höhere Prozentrang zugewiesen.

Prozentrangnormen stellen verteilungsunabhängige Normen dar, d.h., dass bei der Erstellung von Prozentrangnormen keine bestimmte Verteilungsform, z.B. die Normalverteilung der Testwerte in der Normierungsstichprobe, vorausgesetzt wird. Die Bedeutung von Prozenträgen und Quartilen ist anhand einer schießen Häufigkeitsverteilung von Testwerten Y_v in Abb. 9.1 veranschaulicht.

Für die Bildung von Prozentrangnormen ist das Ordinalskalenniveau der Testwerte ausreichend, da an den Testwerten keine Lineartransformation, sondern lediglich eine monotone Transformation vorgenommen wird. Falls für eine Test-

Für Prozentrangnormen sind ordinalskalierte Testwerte ausreichend

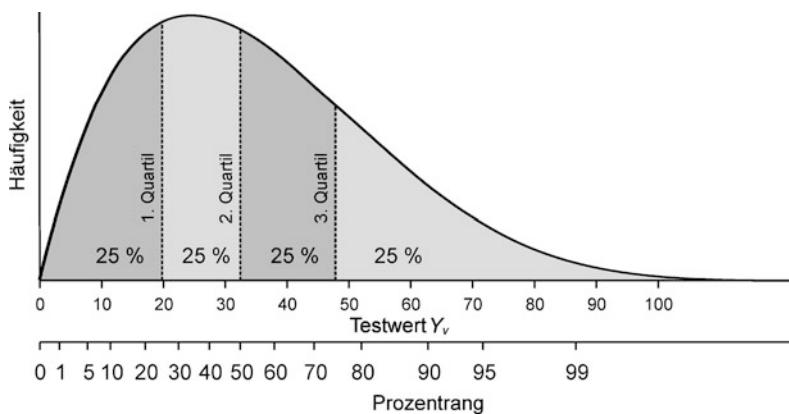


Abb. 9.1 Prozentränge und Quartile bei einer rechtsschiefen Testwertverteilung

Prozentrangnormen dürfen nicht intervallskaliert interpretiert werden

wertvariable nur ein Ordinalskalenniveau angenommen werden kann, kommt also lediglich die Bildung von Prozentrangnormen infrage (Rettler 1999).

Selbst bei intervallskalierten Testwerten können die durch nichtlineare Testwerttransformation gewonnenen Prozentrangnormen nicht als intervallskaliert bezüglich des gemessenen Merkmals aufgefasst werden, da durch die *Flächentransformation* die Differenzen zwischen je zwei Testwerten im Bereich geringer Testwertdichte verkleinert, im Bereich hoher Testwertdichte vergrößert werden (► Abb. 9.1). Dies bedeutet, dass Prozentränge im Bereich hoher Testwertdichte Unterschiede zwischen Merkmalsausprägungen in einer Weise hervortreten lassen, die empirisch gar nicht bestehen, wohingegen im Bereich geringer Testwertdichte tatsächlich bestehende Unterschiede weitgehend nivelliert werden (vgl. ► Kap. 8). Trotz der Verteilungsunabhängigkeit von Prozentrangnormen ist also die Kenntnis der Verteilungsform für den Testanwender beim Vergleich von Prozenträngen von Bedeutung. Liegt beispielsweise eine Normalverteilung vor wie für das Merkmal „Intelligenz“, dann werden durch die Anwendung von Prozentrangnormen kleine Testwertunterschiede im mittleren Skalenbereich überbetont, da bei normalverteilten Merkmalen im mittleren Skalenbereich eine hohe Testwertdichte besteht; liegt hingegen eine linkssteile Verteilung vor, z. B. für das reaktionszeitbasierte Merkmal „Alertness“, dann muss der Testanwender mit einer Überbetonung der Testwertunterschiede im unteren Skalenbereich rechnen, da hier eine hohe Testwertdichte besteht.

Aus der fehlenden Intervallskalierung der Prozentrangnormen folgt auch, dass Prozentrangdifferenzen nicht für Vergleiche herangezogen werden dürfen, denn eine Prozentrangdifferenz von $50 - 40 = 10$ hat – bezogen auf die untersuchte Merkmalsausprägung – eine ganz andere Bedeutung als eine numerisch gleiche Prozentrangdifferenz von $90 - 80 = 10$, wovon man sich in ► Abb. 9.1 leicht überzeugen kann. Vergleiche, die sich nicht auf die Merkmalsausprägung, sondern auf die Fläche der Merkmalsverteilung beziehen, sind hingegen zulässig, z. B. dass sich eine Person A mit dem Prozentrang von 90 von einer Person B mit Prozentrang 70 darin unterscheidet, dass in der Normierungsstichprobe 20 % der Personen einen höheren Testwert erzielt haben als Person B und zugleich den Testwert von Person A nicht überschritten haben.

9.2.2 Bildung von standardisierten z_v -Normwerten durch lineare Testwerttransformation

Lineare Testwerttransformation: z_v -Normwert

Wie die nichtlineare Prozentrang-Transformation dient auch die *lineare z_v -Transformation* von Testwerten dazu, die relative Position des Testwertes Y_v in der Verteilung der Bezugsgruppe anzugeben. Im Gegensatz zu Prozentrangnormen wird bei z_v -Normen für die Testwertvariable Y ein Intervallskalenniveau impliziert, da die Position des Testwertes Y_v einer Testperson als Abstand bzw. Differenz zum arithmetischen Mittelwert \bar{Y} der Verteilung der Bezugsgruppe ausgedrückt wird. Durch die Differenzbildung kann der Testwert Y_v als unter- oder überdurchschnittlich interpretiert werden. Um die Vergleichbarkeit von Testwerten aus Tests mit verschiedenen Testwertstreuungen und Skalenbereichen zu ermöglichen, wird bei der z_v -Transformation die Differenz $Y_v - \bar{Y}$ an der Standardabweichung $SD(Y)$ der Testwerte Y_v relativiert.

Definition

Der z_v -**Normwert** gibt an, wie stark der Testwert Y_v einer Testperson v vom Mittelwert \bar{Y} der Verteilung der Bezugsgruppe in Einheiten der Standardabweichung $SD(Y)$ der Testwerte Y_v abweicht. Der z_v -Normwert von Testperson v wird folgendermaßen berechnet:

$$z_v = \frac{Y_v - \bar{Y}}{SD(Y)}$$

z_v -Normwerte haben einen Mittelwert von $\bar{z} = 0$ und eine Standardabweichung von $SD(z) = 1$.

Im Gegensatz zu Prozenträngen können auf die z_v -Normwerte dieselben algebraischen Operationen angewendet werden wie auf die Testwerte. Die Differenz zwischen zwei z_v -Normwerten ist der Differenz der entsprechenden Testwerte proportional. Bei Statistiken, die invariant gegenüber Lineartransformationen sind (z. B. die Produkt-Moment-Korrelation), resultiert bei Verwendung von z_v -Normwerten das gleiche Ergebnis wie bei Verwendung der Testwerte. Die Berechnung von z_v -Normwerten ist prinzipiell verteilungsunabhängig, d. h. die z_v -Normwerte können bei beliebiger Verteilung der Testwerte Y_v (d. h. auch bei fehlender Normalverteilung) gebildet werden.

 z_v -Normen bei beliebig verteilten Testwerten

Die z_v -Normierung bewirkt keine Normalisierung der Testwerte.

Wenn jedoch die Testwerte Y_v normalverteilt sind, nimmt der interpretative Gehalt des z_v -Normwertes beträchtlich zu. Der z_v -Normwert heißt dann *Standardwert*, und es besteht die Möglichkeit, die prozentuale Häufigkeit der Standardwerte innerhalb beliebiger Wertebereiche über die Verteilungsfunktion der Standardnormalverteilung zu bestimmen (vgl. Rettler 1999). Eine z -Tabelle hierzu findet sich im ▶ Anhang Verteilungsfunktion der Standardnormalverteilung dieses Buches. Liegt dagegen keine normalverteilte Testwertvariable vor, ist diese Interpretation falsch und daher unzulässig¹.

 z_v -Normen bei normalverteilten Testwerten

Da z_v -Normen wegen negativer Vorzeichen und Dezimalstellen ziemlich unpraktisch sind, ist ihre Verwendung eher unüblich. Vielmehr wird der z_v -Normwert weiteren Lineartransformationen unterzogen, um Normwerte mit positivem Vorzeichen sowie mit möglichst ganzzahliger Abstufung zu erhalten. Auf diese Weise lassen sich Normen mit unterschiedlicher Metrik erzeugen (► Abb. 9.2). Beispielsweise wird für Intelligenzmessungen der Standardwert z_v oft noch durch Multiplikation mit dem Faktor 15 und Addition einer Konstante von 100 in den Intelligenzquotienten (IQ) umgeformt, dessen Mittelwert 100 und dessen Standardabweichung 15 beträgt². Für die im Rahmen der PISA-Studien (Programme for International Student Assessment) berichteten Schülerleistungen wurde eine normierte Skala gebildet, bei der der Leistungsmittelwert über alle teilnehmenden OECD-Staaten 500 und die Standardabweichung 100 Punkte beträgt (z. B. OECD 2001, 2004; vgl. auch ▶ Kap. 17). Am Prinzip der Testwertinterpretation ändert sich dadurch jedoch nichts (► Beispiel 9.3).

Weitere Transformationen des z_v -Normwertes

- 1 Bei fehlender Normalverteilung müsste für jeden spezifischen Verteilungsfall die Verteilungsfunktion bestimmt bzw. tabelliert werden. Im Prinzip wäre dies die Vorgehensweise wie bei der Bestimmung von Prozentrangnormen für beliebige Verteilungsformen (► Abschn. 9.3.1).
- 2 Die Wahl der additiven Konstante 100 und des Faktors 15 erfolgte mit dem Ziel, den auf dem z_v -Normwert basierenden IQ mit dem klassischen IQ vergleichen zu können bzw. die gebräuchliche Metrik beizubehalten. Der frühere IQ berechnete sich nämlich aus dem Verhältnis von Intelligenz- und Lebensalter multipliziert mit 100, d. h., er beträgt im Mittel 100, zudem ergab sich für ihn empirisch eine Standardabweichung von 15.

Beispiel 9.3: z_v -Normwert

Eine Testperson habe in einem Intelligenztest mit dem Mittelwert von $\bar{Y} = 31$ und der Standardabweichung von $SD(Y) = 12$ einen Testwert von $Y_v = 27$ erzielt. Der z_v -Normwert ergibt sich folgendermaßen:

$$z_v = \frac{27 - 31}{12} = -0.33$$

Der z_v -Normwert (und somit der Testwert) liegt also um ein Drittel der Standardabweichung unter der durchschnittlichen Testleistung.

An dieser Interpretation ändert sich nichts, wenn aus dem z_v -Normwert der Intelligenz-Quotient IQ_v wie folgt bestimmt wird:

$$IQ_v = 100 + 15 \cdot z_v = 100 + 15 \cdot (-0.33) = 95$$

Auch in der IQ-Metrik (Mittelwert $\bar{Y} = 100$; Standardabweichung $SD(Y) = 15$) liegt die Intelligenzleistung der Testperson mit 95 IQ-Punkten um 5 IQ-Punkte, d. h. um ein Drittel der Standardabweichung, unter dem Mittelwert von 100.

Ist die Testwertvariable Y normalverteilt, kann anhand von z_v der entsprechende Prozentrang aus der tabellarisierten Verteilungsfunktion der Standardnormalverteilung abgelesen werden (s. z -Tabelle im Anhang). Für das vorliegende Beispiel ist der $PR = 37$ (Flächenanteil unter der Standardnormalverteilung zwischen $z = -3.00$ und $z = -0.33$). Liegt dagegen keine Normalverteilung, sondern z. B. eine schiefe Verteilung vor, lassen sich mithilfe der Standardnormalverteilung *keine* Prozentränge ablesen und somit keine Aussagen darüber gewinnen, wie hoch die Wahrscheinlichkeit für einen Testwert ist, der niedriger oder maximal ebenso hoch ist wie z_v . Grobe Abschätzungen erlauben die Ungleichungen nach Tschebyscheff (s. z. B. Eid et al. 2017).

9 Zusammenhang zwischen Standardwerten und Prozenträngen

Standardnormen

Liegt eine normalverteilte Testwertvariable Y vor, dann wird die transformierte z -Norm als *Standardnorm* bezeichnet. Unter Annahme der Normalverteilung verschafft ▶ Abb. 9.2 einen Überblick über die *Standardnormen z-, IQ-, T-Werte und die PISA-Skala* mit der jeweils zugehörigen Vorschrift zur Transformation von z , ihren Mittelwerten und Standardabweichungen. Zusätzlich ist hier die sog. „Stanine-Norm“ („Standard Nine“, kurz: Stanine) dargestellt, bei der anhand der Werteverteilung neun Abschnitte mit Häufigkeiten von 4 %, 7 %, 12 %, 17 %, 20 %, 17 %, 12 %, 7 % und 4 % gebildet werden. Diese Unterteilung ergibt für normalverteilte Variablen gleiche Intervalle von der Breite einer halben Standardabweichung, wobei bezüglich Stanine 1 und 9 vereinfachend angenommen wird, dass Stanine 1 bei $z = -2.25$ beginnt und Stanine 9 bei $z = +2.25$ endet. Weiterhin sind in ▶ Abb. 9.2 die Prozentränge für eine normalverteilte Variable eingetragen.

Falls die Testwertvariable Y nicht normalverteilt ist, kann die Testwertverteilung über eine *Flächentransformation* in eine Normalverteilung umgewandelt werden (*Normalisierung*, ▶ Kap. 8). Allerdings sollte hierfür plausibel gemacht werden können, weshalb sich im konkreten Fall empirisch keine Normalverteilung gezeigt hat und warum die Annahme einer Normalverteilung für das jeweils untersuchte Merkmal dennoch begründbar ist.

Die normorientierte Interpretation von Testresultaten bezieht sich in der Regel auf psychologische Testverfahren, die nach der Klassischen Testtheorie (KTT, ▶ Kap. 13) konstruiert wurden. Doch auch bei probabilistischen Testmodellen der Item-Response-Theorie (IRT, ▶ Kap. 16) lassen sich die latenten Personenparameter η_v normorientiert interpretieren, wenn sie so normiert werden, dass ihre Summe gleich null ist (Rost 2004). In diesem Fall geben das Vorzeichen und der Betrag des

Normalisierung nicht normaler Verteilungen

Normorientierte Interpretation von Personenparametern

9.3 · Kriteriumsorientierte Testwertinterpretation

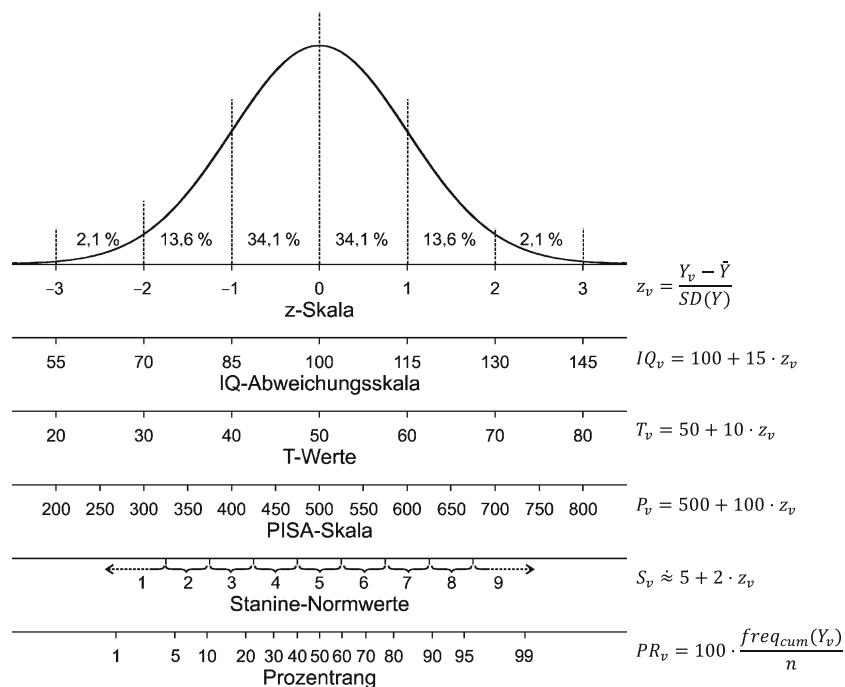


Abb. 9.2 Gebräuchliche Standardnormen, Stanine-Norm und Prozentrangnorm unter Annahme normalverteilter Testwerte

Personenparameters an, wie weit sich die Testperson über oder unter dem mittleren Personenparameter von null in der Bezugsgruppe befindet.

9.3 Kriteriumsorientierte Testwertinterpretation

Bei der kriteriumsorientierten Testwertinterpretation erfolgt die Interpretation des Testwertes nicht in Bezug zur Testwertverteilung einer Bezugsgruppe, sondern in Bezug auf ein spezifisches inhaltliches Kriterium. Ein derartiges Kriterium besteht in bestimmten diagnostischen Aussagen, die auf Basis des Testwertes über die getesteten Personen gemacht werden sollen. Bei der kriteriumsorientierten Testwertinterpretation interessiert nicht, wie viele Personen das Kriterium erfüllen. Theoretisch könnten alle getesteten Personen ein Kriterium erreichen, aber ebenso keine einzige (► Beispiel 9.4).

Beispiel 9.4: Diagnostische Aussagen bezogen auf ein spezifisches Kriterium

- Ein Patient in einer psychotherapeutischen Ambulanz wird mit einem Depressivitätsfragebogen untersucht. Aufgrund des Testresultats soll entschieden werden, ob eine genauere Diagnostik und Therapie hinsichtlich einer ausgeprägten Depression (*Major Depression*) angezeigt erscheint. Die Höhe des Testwertes soll also Auskunft darüber geben, ob der Patient zu jener Gruppe von Patienten gehört, die das Kriterium für eine Major Depression erfüllen oder nicht. Das Kriterium, hinsichtlich dessen eine Interpretation des Testergebnisses erfolgen soll, ist hier also das Vorliegen einer Major Depression.
- In einer Schulklasse werden die Schüler zum Ende des Schuljahres mit einem Vokabeltest in Englisch getestet. Es soll geprüft werden, wie viele Schüler in der Klasse das im Lehrplan gesetzte Leistungsniveau erreicht haben und das entsprechende Vokabular schriftlich beherrschen. Das Kriterium ist in diesem Fall also das Erreichen eines bestimmten, sachlich begründeten Leistungsniveaus.

Schwellenwerte zur Interpretation des Testergebnisses

Um eine kriteriumsorientierte Interpretation eines Testwertes vorzunehmen, werden in der Regel vorab bestimmte *Schwellenwerte* definiert. Wird der Schwellenwert erreicht oder überschritten, gilt das Kriterium (z. B. Schulreife) als erfüllt, bei Testwerten unterhalb des Schwellenwertes hingegen nicht. So könnte z. B. ab einem Wert von 19 Punkten im Depressivitätsfragebogen das Vorliegen einer Major Depression oder ab 30 Punkten im Vokabeltest das Erreichen des im Lehrplan definierten Leistungsziels als erfüllt angenommen werden. Solche Schwellenwerte können auf zwei unterschiedliche Weisen ermittelt werden. Zum einen kann der Testwert in eigens dafür durchgeführten Untersuchungen in Bezug zu einem *externen Kriterium* gesetzt werden (► Abschn. 9.3.1), wobei in der Untersuchungsphase das externe Kriterium, d. h. die auf einem anderen Weg bestimmte Gruppenzugehörigkeit, zusätzlich zu den individuellen Testwerten bekannt sein muss. Zum anderen können die *Inhalte der Testaufgaben* selbst herangezogen werden, um Schwellenwerte, z. B. in Form von Kompetenzniveaus, inhaltlich zu beschreiben (► Abschn. 9.3.2).

9.3.1 Bezug des Testwertes auf ein externes Kriterium

Grundvoraussetzung für eine sinnvolle Anwendung

Vom Grundsatz her kann die Interpretation eines Testwertes im Sinne der Zuordnung einer Testperson zu einer von zwei Gruppen nur dann gelingen, wenn sich die Testpersonen, die das Kriterium erfüllen, hinsichtlich ihrer Testwertausprägung deutlich von jenen Testpersonen, die das Kriterium nicht erfüllen, unterscheiden. Als Beispiel soll anhand eines Depressivitätsfragebogens beurteilt werden, ob ein Patient tatsächlich das Kriterium der Major Depression erfüllt oder nicht. Hierbei können die Klassifikationsentscheidungen mit einem grundsätzlich geeigneten Test in der diagnostischen Praxis in vier verschiedene Kategorien fallen:

- **Treffer:** Die Testperson erfüllt das Kriterium einer Depression und wird mit dem (vorläufigen) Schwellenwert korrekt als depressiv klassifiziert (*richtig positiv, RP*)
- **Verpasser:** Die Testperson erfüllt das Kriterium einer Depression, sie wird aber fälschlicherweise als nicht depressiv klassifiziert (*falsch negativ, FN*)
- **Falscher Alarm:** Die Testperson erfüllt das Kriterium einer Depression nicht, sie wird aber fälschlicherweise als depressiv klassifiziert (*falsch positiv, FP*)
- **Korrekte Ablehnung:** Die Testperson erfüllt das Kriterium einer Depression nicht und wird korrekt als nicht depressiv klassifiziert (*richtig negativ, RN*)

		Klassifikation		
		+	-	
Kriterium	+	RP	FN	
	-	FP	RN	

+	positiv
-	negativ
RP	richtig positiv
FN	falsch negativ
FP	falsch positiv
RN	richtig negativ

In Abhängigkeit davon, ob der Schwellenwert für die Klassifikation höher oder niedriger angesetzt wird, verändern sich die Wahrscheinlichkeiten für jede der vier Klassifikationskategorien.

Die Genauigkeit der Entscheidungen in Abhängigkeit vom Schwellenwert lässt sich anhand der Maße Sensitivität und Spezifität³ ausdrücken:

- *Sensitivität* oder *Trefferquote* bezeichnet die Wahrscheinlichkeit für die Klassifikationskategorie „RP“, d. h. dafür, dass eine Testperson, die das Kriterium

³ Der Begriff „Spezifität“, wie er in diesem Zusammenhang gebraucht wird, ist nicht zu verwechseln mit dem Spezifitätsbegriff in der Latent-State-Trait-Theorie (LST-Theorie, ► Kap. 26).

9.3 · Kriteriumsorientierte Testwertinterpretation

erfüllt, auch entsprechend als positiv klassifiziert wird. Aus dem Komplement ($1 - \text{Sensitivität}$) ergibt sich die *Verpasserquote*, die die Wahrscheinlichkeit für die Klassifikationskategorie „FN“ angibt, d. h. dafür, dass eine Testperson, die das Kriterium erfüllt, fälschlicherweise als negativ klassifiziert wird.

- *Spezifität* oder die *Quote korrekter Ablehnungen* bezeichnet hingegen die Wahrscheinlichkeit für die Klassifikationskategorie „RN“, d. h. dafür, dass eine Testperson, die das Kriterium nicht erfüllt, auch entsprechend als negativ klassifiziert wird. Aus dem Komplement ($1 - \text{Spezifität}$) ergibt sich die *Quote falscher Alarme*, die die Wahrscheinlichkeit für die Klassifikationskategorie „FP“ angibt, d. h. dafür, dass eine Testperson, die das Kriterium nicht erfüllt, fälschlicherweise als positiv klassifiziert wird.

Definition

Sensitivität und Spezifität

$$\text{Sensitivität} = \frac{\text{RP}}{\text{FN} + \text{RP}} \quad (\text{Trefferquote})$$

$$1 - \text{Sensitivität} = \frac{\text{FN}}{\text{FN} + \text{RP}} \quad (\text{Verpasserquote})$$

$$\text{Spezifität} = \frac{\text{RN}}{\text{FP} + \text{RN}} \quad (\text{Quote korrekter Ablehnungen})$$

$$1 - \text{Spezifität} = \frac{\text{FP}}{\text{FP} + \text{RN}} \quad (\text{Quote falscher Alarme})$$

In Abb. 9.3 ist für einen Depressivitätsfragebogen die grundsätzliche Voraussetzung für eine sinnvolle Klassifikation auf Basis des Testwertes dargestellt. An den verschiedenen Mittelwerten der Verteilungen ist klar erkennbar, dass sich die Testpersonen, die das Kriterium erfüllen („Positive“), deutlich von den übrigen Testpersonen („Negative“) hinsichtlich der Testwertausprägung unterscheiden. Die Gruppe von Personen mit Major Depression (rechte Verteilung, „Depressive“) weist im Durchschnitt höhere Testwerte (und zudem eine geringere Streuung) auf als die Vergleichsgruppe ohne Major Depression (linke Verteilung, „Gesunde“).

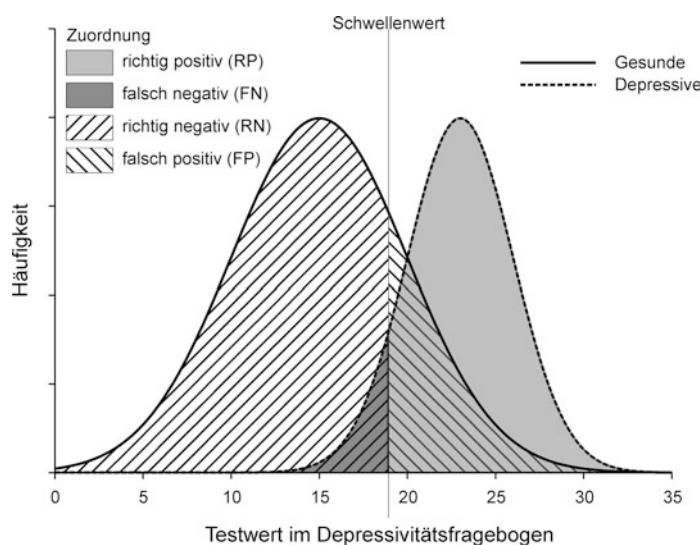


Abb. 9.3 Verteilung der Testwerte in der Gruppe mit Major Depression (rechts, „Depressive“) und in der Gruppe ohne Major Depression (links, „Gesunde“) in Abhängigkeit vom gewählten Schwellenwert (hier 18.5 Punkte). Die verschiedenen Schraffuren veranschaulichen die Zuordnung zu den vier Klassifikationskategorien

Wechselseitige Beziehung zwischen Sensitivität und Spezifität

Die □ Abb. 9.3 macht mittels der verschiedenen Schraffuren für die vier Klassifikationskategorien außerdem deutlich, dass die Sensitivität und die Spezifität vom Schwellenwert und voneinander abhängig sind.

Wenn der Schwellenwert im Depressivitätsfragebogen sehr hoch gesetzt würde (z. B. auf 25), wäre die Wahrscheinlichkeit für die Entscheidung „FP“ klein und es würde nur selten fälschlicherweise eine Major Depression angenommen werden; die Spezifität wäre also hoch. Gleichzeitig würden jedoch viele Patienten, die eine Major Depression haben, fälschlicherweise als nicht depressiv klassifiziert, obwohl sie einer Therapie bedürften; die Sensitivität wäre also niedrig. Ein niedriger Schwellenwert (z. B. 15) führt umgekehrt dazu, dass die Wahrscheinlichkeit für die Entscheidung „FN“ sehr klein ist und dass fast alle Patienten mit Major Depression als richtig positiv klassifiziert werden; allerdings würde auch fast die Hälfte der Nichtdepressiven fälschlicherweise als positiv klassifiziert. In diesem Fall lägen eine hohe Sensitivität und eine niedrige Spezifität vor. In □ Abb. 9.4 ist grafisch dargestellt, wie sich die Maße für die Genauigkeit der Klassifikation, d. h. Sensitivität und Spezifität, mit der Verschiebung des Schwellenwertes gegenläufig verändern. Bei dem höheren Schwellenwert (hier 25) sinkt die Sensitivität auf .27, während die Spezifität auf .98 ansteigt; mit dem niedrigeren Schwellenwert (hier 15) steigt die Sensitivität auf 1.00, die Spezifität fällt hingegen auf .50 ab.

9

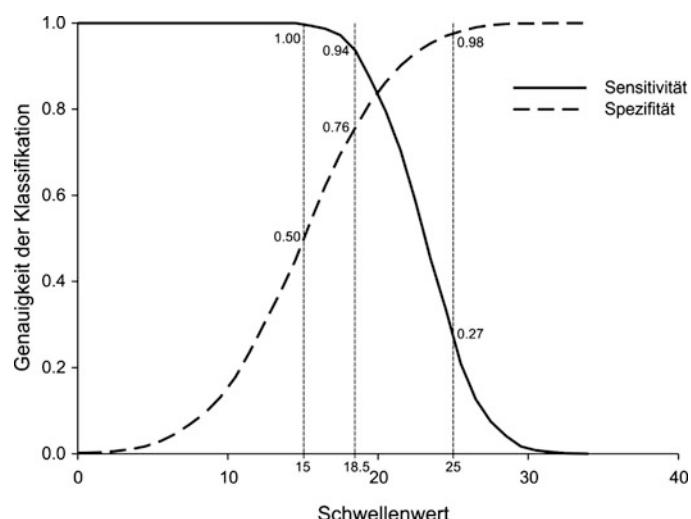
ROC-Analyse zur Bestimmung eines optimalen Schwellenwertes

Voraussetzung: Kriterium muss extern ermittelt sein

■ ■ Receiver-Operating-Characteristics-Analyse (ROC-Analyse)

Ein einfaches mögliches Verfahren, jene Höhe eines Testwertes zu ermitteln, der als optimaler Schwellenwert zur Unterscheidung von zwei Gruppen anhand eines externen Kriteriums herangezogen werden kann, ist die *Receiver-Operating-Characteristics-Analyse*, kurz: *ROC-Analyse*. Dieses Verfahren stammt aus der Signal-entdeckungstheorie der Psychophysik (Green und Swets 1966) und eignet sich für Situationen, in denen der eine Teil der Fälle das Kriterium erfüllt, der andere Teil hingegen nicht.

Bei der ROC-Analyse wird nach jenem Schwellenwert gesucht, der die Balance zwischen Sensitivität und Spezifität optimiert. *Voraussetzung* dazu ist eine bereits erfolgte Untersuchung, durch die sowohl das Testergebnis als auch das Kriterium (z. B. die Diagnose) der Personen bekannt sind. Das Vorliegen des Kriteriums „Major Depression“ im Beispiel kann etwa extern über ein ausführliches klinisches Interview ermittelt worden sein. In der ROC-Analyse wird nun für jeden der potentiellen Schwellenwerte (Testwerte) die Sensitivität und Spezifität berech-



□ Abb. 9.4 Sensitivität und Spezifität bei der Klassifikation depressiver und gesunder Patienten in Abhängigkeit vom Schwellenwert. Exemplarisch sind für das unten stehende ► Beispiel 9.5 die Schwellenwerte 15, 18.5 und 25 eingetragen

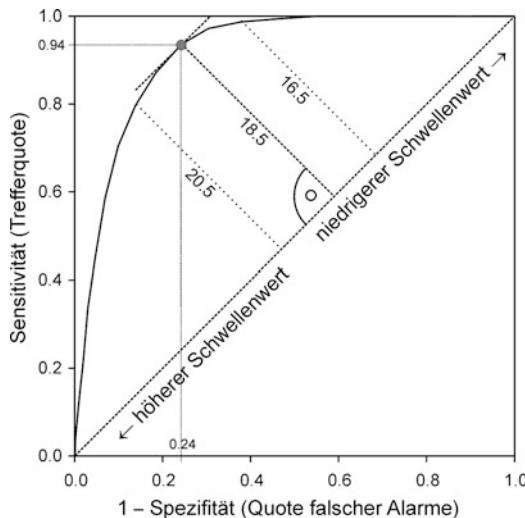


Abb. 9.5 ROC-Kurve. Der maximale Abstand zur Hauptdiagonalen wird bei einem optimalen Schwellenwert von 18.5 erreicht, der mit einer Sensitivität von 0.94 und einer Spezifität von 0.76 die Gruppe der Gesunden und die Gruppe der Depressiven optimal voneinander trennt, da hier die Summe aus Sensitivität und Spezifität am größten wird; zur Illustration sind zusätzlich die Projektionen eingezeichnet, die sich jeweils für einen niedrigeren (16.5) und einen höheren (20.5) Schwellenwert ergeben würden

net, die sich ergeben würde, wenn man diesen Wert als Schwellenwert verwenden würde. Dann werden die jeweils zusammengehörenden Werte für Sensitivität und 1 – Spezifität (d. h. Quote falscher Alarme) grafisch gegeneinander abgetragen. Die Beziehung ist kurvilinear und wird als *ROC-Kurve* bezeichnet (Abb. 9.5).

Die ROC-Kurve gibt Aufschluss darüber, ob und wie gut der Test geeignet ist, zwischen Testpersonen, die das Kriterium erfüllen, und den übrigen Testpersonen zu trennen. Wenn der Test nicht zwischen den beiden Gruppen trennen kann, verläuft die empirische ROC-Kurve nahe der Hauptdiagonalen, d. h. Sensitivität und 1 – Spezifität sind für alle Schwellenwerte gleich groß (in diesem Fall liegen die Verteilungen in Abb. 9.3 übereinander). Haben die Testpersonen, bei denen das Kriterium vorliegt, im Mittel höhere Testwerte als die übrigen, verläuft die Kurve oberhalb der Diagonalen. Je größer der Abstand zur Hauptdiagonalen ist, desto besser kann der Test mittels Schwellenwert zwischen den Gruppen trennen.

Als optimaler Schwellenwert wird derjenige bezeichnet, bei dem die Summe von Sensitivität und Spezifität am größten ist. Dies entspricht demjenigen Punkt in der ROC-Kurve, an dem das Lot auf die Hauptdiagonale den größten Abstand anzeigt; zugleich ist es genau derjenige Punkt, an dem die Tangente parallel zur Hauptdiagonalen verläuft (Abb. 9.5).

Rechnerisch lässt sich dieser Punkt auch über den *Youden-Index* (YI; Youden 1950) bestimmen, der so gebildet wird, dass er Werte zwischen 0 und 1 annimmt (YI = Sensitivität + Spezifität – 1).

Mit demjenigen Schwellenwert, für den die Summe von Sensitivität und Spezifität und somit der Youden-Index am größten wird, gelingt die Trennung der beiden Gruppen am besten (► Beispiel 9.5). Bis zu diesem Punkt nimmt der Gewinn an Sensitivität durch das Absenken des Schwellenwertes stark zu, während die Quote falscher Alarme (1 – Spezifität) vergleichsweise wenig ansteigt. Jenseits dieses Punktes steigt jedoch die Quote falscher Alarme schneller an, als die Sensitivität zunimmt; es lohnt sich also nicht, den Schwellenwert noch niedriger zu wählen.

ROC-Kurve zeigt die Beziehung zwischen Sensitivität und Spezifität

Rechnerische Bestimmung: Youden-Index

Beispiel 9.5: Bestimmung des optimalen Schwellenwertes

In der nachfolgenden Tabelle sind Sensitivität, Spezifität sowie der Youden-Index für potentielle Schwellenwerte des Beispiels in □ Abb. 9.3 aufgelistet (die aufgeführten Wahrscheinlichkeiten gewinnt man als prozentualen Anteil der Testpersonen, die in eine der vier Klassifikationskategorien fallen).

Tabellierte Koordinaten der ROC-Kurve (Ausschnitt)				
Schwellenwert	Sensitivität	1 – Spezifität	Spezifität	Youden-Index
14.5	1.00	0.55	0.45	0.45
15.5	0.99	0.46	0.54	0.53
16.5	0.99	0.38	0.62	0.61
17.5	0.97	0.30	0.70	0.67
18.5	0.94	0.24	0.76	0.70
19.5	0.87	0.18	0.82	0.69
20.5	0.80	0.14	0.86	0.66
21.5	0.70	0.10	0.90	0.60
22.5	0.59	0.07	0.93	0.52
23.5	0.45	0.05	0.95	0.40
24.5	0.34	0.03	0.97	0.31

Als optimaler Schwellenwert lässt sich im Beispiel ein Testwert von 18.5 ermitteln, der mit 0.70 den höchsten Youden-Index aufweist. Wie aus der Tabelle hervorgeht, wird im Beispiel mit dem Schwellenwert von 18.5 eine Sensitivität von 0.94 erreicht, d. h. 94 % der Patienten mit Major Depression werden korrekt klassifiziert. Zugleich werden 76 % der nicht depressiven Personen korrekt klassifiziert (Spezifität) bzw. 24 % der nicht depressiven Personen fälschlicherweise als depressiv diagnostiziert (1 – Spezifität). Würde man den Schwellenwert z. B. auf 20.5 Punkte heraufsetzen, würden nur noch 14 % der nicht depressiven Personen fälschlicherweise als depressiv diagnostiziert werden. Zugleich würden aber nur noch 80 % der Patienten mit Major Depression korrekt als solche erkannt.

Die entsprechende ROC-Kurve in □ Abb. 9.5 zeigt, dass der Test gut zwischen beiden Gruppen trennt, die Kurve liegt deutlich oberhalb der Hauptdiagonalen. Der dem optimalen Schwellenwert zugehörige Punkt auf der ROC-Kurve weist den maximalen Abstand zur Hauptdiagonalen auf (□ Abb. 9.5). Anhand des mit der ROC-Analyse ermittelten Schwellenwertes kann der Testwert im Depressivitätsfragebogen also kriteriumsorientiert interpretiert werden: *Personen mit einem Testwert von 18 oder weniger Punkten werden als nicht depressiv klassifiziert, Personen mit einem Wert von 19 oder mehr Punkten hingegen als depressiv.*

ROC-Analyse ist verteilungsfrei

Die grafische ROC-Analyse ist ein verteilungsfreies Verfahren, d. h., die Testwertverteilungen in den Gruppen („Positive“ und „Negative“) müssen keiner bestimmten Verteilungsfunktion folgen. Es ist für die Untersuchung auch unerheblich, wie groß die beiden anhand des Kriteriums gebildeten Gruppen im Verhältnis zueinander sind. Zur Bestimmung des Schwellenwertes können z. B. gleich große Gruppen von Depressiven und Nichtdepressiven untersucht werden, auch wenn die tatsächliche Quote Depressiver in der diagnostischen Praxis kleiner als 50 % ist. Die beiden Gruppen müssen allerdings möglichst repräsentativ für die Populationen derer sein (► Abschn. 9.6), die das Kriterium erfüllen bzw. nicht erfüllen, damit der gewonne-

9.3 · Kriteriumsorientierte Testwertinterpretation

ne Schwellenwert in der Praxis für die Klassifikation von neuen Fällen verwendet werden kann. Im Beispiel sollten also die Personen ohne Major Depression diejenige Population repräsentieren, aus der sich auch in der klinischen Praxis die interessierende „Vergleichsgruppe“ rekrutiert – z. B. die Gesamtheit der Patienten, die wegen Beschwerden in einer psychotherapeutischen Ambulanz untersucht werden, jedoch nicht das Kriterium der Major Depression erfüllen.

Anhand der ROC-Analyse können auch Schwellenwerte festgelegt werden, die ein anderes Optimierungsverhältnis ergeben, als das aus dem Youden-Index resultierende. Wenn etwa die Konsequenzen einer falschen negativen Diagnose („Verpasser“) schwerwiegender sind als diejenigen einer falschen positiven („falscher Alarm“), kann ein niedrigerer Schwellenwert mit einer höheren Sensitivität und niedrigeren Spezifität zweckmäßiger sein als der Wert mit der besten Balance zwischen beiden Kriterien. Wenn z. B. die Konsequenz aus dem Testergebnis des Depressivitätsfragebogens lediglich in einer intensiveren Diagnostik besteht (und nicht in einer therapeutischen Entscheidung), ist ein „falscher Alarm“ nicht sonderlich schwerwiegend. Einen Patienten mit Major Depression hingegen nicht als solchen zu erkennen („Verpasser“), könnte wegen einer bestehenden Suizidgefährdung wesentlich schwerwiegender Folgen haben. In diesem Fall könnte es angemessener sein, nicht den „optimalen“ Schwellenwert zu nehmen, sondern einen niedrigeren von z. B. 17.5, womit eine Sensitivität von 97 % erreicht werden würde.

An dieser Stelle sei darauf hingewiesen, dass das beschriebene Vorgehen zur Bestimmung eines Schwellenwertes auch als Validierung einer kriteriumsorientierten bzw. extrapolierenden Testwertinterpretation (► Kap. 21) verstanden werden kann. Validierung heißt in diesem Fall die Klärung der Frage, ob der Testwert eines neu entwickelten Depressivitätsfragebogens in der Lage ist, zwischen gesunden und depressiven Personen zu unterscheiden. Ist dies der Fall, weisen die Testwertverteilungen beider Gruppen nur eine geringe Überlappung auf und die ROC-Kurve verläuft weit von der Hauptdiagonalen entfernt. Wenn also die Interpretation des Testwertes dahingehend, ob eine Major Depression vorliegt oder nicht, empirisch gestützt wird, belegt dies die Kriteriumsvalidität der Testwertinterpretation.

9.3.2 Bezug des Testwertes auf Aufgabeninhalte

Eine zweite Möglichkeit, die Testwerte aus einem Test bezogen auf inhaltliche Kriterien zu interpretieren, ist der Bezug auf die Test- bzw. Aufgabeninhalte. Dieses Vorgehen ist dann möglich, wenn a priori eine genaue inhaltliche Vorstellung von der *theoretischen Grundgesamtheit* der für das interessierende Konstrukt relevanten Aufgaben existiert und die im Test verwendeten Aufgaben eine *Stichprobe* aus diesen möglichen Aufgaben darstellen. Beispielsweise werden die Fähigkeiten, die Schüler am Ende ihrer schulischen Ausbildung in der ersten Fremdsprache erreichen sollen, in den Bildungsstandards der Kultusministerkonferenz beschrieben und anhand von Beispielaufgaben illustriert (Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland 2004, 2005). Ein Vokabeltest kann so konstruiert werden, dass die Testaufgaben die in den Bildungsstandards definierten Lernziele repräsentieren. Die abgefragten Vokabeln können z. B. eine zufällige Stichprobe aus der Menge der Wörter darstellen, die bei einem bestimmten Fähigkeitsniveau in einer Fremdsprache beherrscht werden sollen. Für einen solchen Test ist es möglich, den Testwert im Sinne des Ausmaßes der *Erfüllung eines Lernziels* zu interpretieren. Dieser Testwert wird dann als der Anteil gelöster Aufgaben interpretiert, den ein Schüler gelöst hätte, wenn man ihm alle theoretisch möglichen Aufgaben des interessierenden Fähigkeitsbereiches vorgelegt hätte (Klauer 1987a, 1987b).

Andere Optimierungsverhältnisse

Validierung einer extrapolierenden Testwertinterpretation

Kriteriumsorientierte Interpretation bei Definition der Aufgabengrundgesamtheit

■ Tabelle 9.1 Zwei unterschiedlich schwierige Fragebogenitems zur Erfassung derselben depressiven Symptomatik mit einer vierstufigen Antwortskala

	trifft nicht zu	trifft eher nicht zu	trifft eher zu	trifft zu
Ich fühle mich manchmal grundlos traurig.	①	②	③	④
Mich überkommt oft ohne Anlass eine tiefe Traurigkeit.	①	②	③	④

Repräsentative Aufgabenstichprobe

9

Diese Form von kriteriumsorientierter Testwertinterpretation setzt voraus, dass eine Definition der theoretischen Grundgesamtheit von Aufgaben vorgenommen werden kann. Die wesentliche Anforderung an die Aufgabenzusammenstellung ist hierbei, dass die im Test verwendeten Aufgaben eine Schwierigkeitsverteilung aufweisen, die derjenigen in der Grundgesamtheit der Aufgaben entspricht. Im Falle von Lernzielen in pädagogisch-psychologischen Kontexten ist dies besonders gut möglich, wenn auch mit einem nicht unerheblichen Aufwand verbunden. So müssen Lernziele hinreichend (z. B. bildungspolitisch) legitimiert sein und es muss unter geeigneten Experten Einigkeit über die Definition der relevanten Aufgaben bestehen.

Während die kriteriumsorientierte Interpretation auf Basis der Aufgabeninhalte also vor allem bei Lernziel- oder Leistungstests möglich ist, bei denen die Grundgesamtheit möglicher Aufgaben und deren Schwierigkeitsverteilung definiert werden können, ist dieses Vorgehen im Falle von Fragebogen in der Regel nicht möglich. Dies liegt daran, dass die Schwierigkeiten der Items eines Fragebogens nicht nur durch die Inhalte, sondern auch durch die verbale Formulierung beeinflusst werden. Es ist daher ein häufig gemachter Fehler, den Testwert aus einem Fragebogen auf den „theoretischen Wertebereich“ zu beziehen, der sich aus der Antwortskala ergibt, und hieraus eine kriteriumsorientierte Interpretation abzuleiten.

Beispielsweise kann ein Fragebogen zur Erfassung depressiver Symptome einer der in **■ Tab. 9.1** aufgelisteten Items enthalten.

In beiden Items wird nach dem gleichen Inhalt gefragt, nämlich nach Traurigkeit ohne äußerer Anlass. Das erste Item ist jedoch so formuliert, dass ihm wahrscheinlich auch einige Personen ohne eine ausgeprägte Depression zustimmen können. Das zweite Item ist hingegen deutlich schwieriger, ihm würden wahrscheinlich deutlich weniger Personen zustimmen. Wenn nun die Antworten auf diese Items mit 1 bis 4 Punkten bewertet würden, würde sich beim ersten Item bei denselben Personen ein höherer Mittelwert ergeben als beim zweiten. Dieser Unterschied würde jedoch nichts über eine unterschiedliche Depressivität aussagen, sondern wäre allein auf die verbale Itemformulierung zurückzuführen.

Im Beispiel könnte sich etwa für das erste Item ein Mittelwert von 2.8 Punkten ergeben. Es wäre unzulässig, diesen Wert dahingehend zu interpretieren, dass die getesteten Personen „eher depressiv“ seien, da die Personen dem Item „eher zugestimmt haben“ und der Mittelwert über der „theoretischen Mitte“ des Wertebereichs von 2.5 Punkten liegt. Wenn dieselben Personen anstelle des ersten das zweite Item vorgelegt bekämen, würde der Mittelwert niedriger liegen (z. B. bei 1.6 Punkten), ohne dass die Personen deswegen „weniger depressiv“ wären. Eine kriteriumsorientierte Interpretation einer Fragebogenskala anhand des möglichen Wertebereichs ist also nicht sinnvoll, es müsste in diesem Fall ein externes Kriterium (**► Abschn. 9.3.1**) herangezogen werden, um zu beurteilen, welche Testwerte tatsächlich mit einer depressiven Störung einhergehen.

Weitaus differenziertere Möglichkeiten zur kriteriumsorientierten Testwertinterpretation auf Basis der Aufgabeninhalte bieten Auswertungen von Tests, die mit Modellen der IRT konstruiert wurden. Diese Modelle erlauben es, die Schwierig-

Häufiger Fehler: kriteriumsorientierte Interpretation anhand des theoretisch möglichen Wertebereichs der Antwortskala

Schwierigkeit eines Fragebogenitems wird durch verbale Formulierung beeinflusst

Differenziertere Interpretationsmöglichkeiten mit IRT-Modellen

keiten der Aufgaben und die Messwerte der Personen auf einer gemeinsamen Skala darzustellen. Die hierdurch mögliche kriteriumsorientierte Testwertinterpretation wird von Rauch und Hartig in ► Kap. 17 erörtert.

9.4 Integration von norm- und kriteriumsorientierter Testwertinterpretation

Norm- und kriteriumsorientierte Testwertinterpretation stellen grundsätzlich keine Gegensätze dar. Je nach diagnostischer Fragestellung kann einer der beiden Interpretationsansätze angemessener sein; sie können sich aber auch ergänzen, indem das Testresultat aus unterschiedlichen Perspektiven bewertet wird. Physikalische Maße, z. B. die Größe bzw. Länge eines Objekts, können hierbei als Analogie dienen. Beispielsweise wird bei einem Angelwettbewerb derjenige Angler gewinnen, der den größten oder schwersten Fisch fängt (normorientierte Diagnostik). In den gesetzlichen Regelungen, ab welcher Mindestgröße ein Fisch überhaupt von einem Angler aus dem Wasser entnommen werden darf („Schonmaß“), geht es hingegen um das theoretisch-inhaltlich definierte Kriterium, dass ein Fisch eine hinreichende Größe erreicht haben soll (kriteriumsorientierte Diagnostik), um sich wenigstens einmal im Leben fortpflanzen zu können, bevor er zum Angeln freigegeben wird.

In gleicher Weise ermöglicht die Beachtung sowohl des normorientierten als auch des kriteriumsorientierten Vergleichsmaßstabs eine differenziertere Bewertung von psychologischen Merkmalsausprägungen. Beispielsweise zeigen sich Peters Eltern erfreut über die Vokabellistung ihres Sohnes, nachdem sie erfahren haben, dass die meisten Klassenkameraden von Peter weniger als 70 % der Aufgaben richtig gelöst haben (normorientierte Interpretation ► Beispiel 9.1). Allerdings wurde ihre Freude wieder etwas getrübt, als sie erfuhren, dass das Lehrziel eine Lösungsrate von 90 % war (kriteriumsorientierte Interpretation).

Dieses Beispiel deutet bereits an, dass ein psychologisch-inhaltliches Kriterium oft nicht völlig unabhängig von der empirischen Verteilung der Merkmalsausprägungen erfolgen kann bzw. auch mit normorientierten Überlegungen in Zusammenhang steht. So wurden die Aufgaben für die Vokabelprobe derart zusammengestellt, dass eine Lösungsrate von 90 % in der entsprechenden Schülerpopulation hätte erzielt werden können.

Obwohl die Integration norm- und kriteriumsorientierter Testwertinterpretationen die angemessenste Form diagnostischer Informationsverarbeitung darstellt (Rettler 1999), können norm- und kriteriumsorientierte Testwertinterpretation bei konfligierenden Interessenlagen zu teilweise unvereinbaren Zielsetzungen führen. Cronbach (1990) unterscheidet im Zusammenhang mit der Festlegung von Standards für Auswahlprozesse empirische, politische und entscheidungsbezogene Aspekte:

- Aus empirischer Sicht sind solche Personen, z. B. für einen Studienplatz oder eine freie Stelle, auszuwählen, die mit ihrem Testwert ein psychologisch-inhaltlich gesetztes Kriterium erreichen (kriteriumsorientierter Standard) und somit wahrscheinlich einen hohen Studien- oder Berufserfolg haben werden.
- Hinzu kommen jedoch auch politische Einflüsse, die die Anhebung oder Absenkung von Kriterien zum Ziel haben können. Aus Arbeitgebersicht mag es z. B. von Interesse sein, dass strengere Standards (d. h. höhere Schwellenwerte) zur Erreichung eines Schulabschlusses gesetzt werden, um mit größerer Sicherheit geeignetes Personal einzustellen. Durch höhere Standards entstehen jedoch auch höhere soziale Kosten (z. B. durch Klassenwiederholung, Arbeitslosigkeit), sodass von Politikern ein normorientierter Standard eingefordert werden könnte, z. B. dass unabhängig vom Erreichen eines psychologisch-inhaltlich definierten Kriteriums nur 10 % der Schwächsten der Population in der Abschlussprüfung durchfallen dürfen, um gesellschaftliche Folgekosten zu begrenzen (normorientierter Standard).

Norm- und kriteriumsorientierte Testwertinterpretation ergänzen sich

Empirischer, politischer und entscheidungsbezogener Aspekt bei der Anwendung von Testnormen

- Dieses Spannungsfeld von unterschiedlichen Interessen und Einflüssen führt zu einem komplexen Entscheidungsprozess, in dem Kosten und Nutzen abgewogen werden müssen sowie politische Machbarkeit beachtet werden muss. Cronbach (1990, S. 98) stellt zusammenfassend fest, „standards must be set by negotiations rather than by technical analysis“.

9.5 Normdifferenzierung

Das Problem der Normdifferenzierung bezieht sich im Rahmen der normorientierten Testwertinterpretation auf die für Testentwickler und Testanwender gleichermaßen bedeutsame Frage, wie spezifisch eine Vergleichs- bzw. Referenzgruppe zusammengesetzt sein soll. Im Beispiel zur Prozentrangsnormalisierung in ► Abschn. 9.2.1 wurde deutlich gemacht, dass bei erstmaliger Erfassung der Konzentrationsleistung einer Testperson die Vergleichsgruppe nur solche Testpersonen beinhalten sollte, die ebenfalls nur einmal getestet wurden, d. h. die Vergleichsgruppe stimmt hinsichtlich des Übungsgrads mit dem der Testperson überein (vgl. Moosbrugger und Goldhammer 2007). Würde die Vergleichsgruppe aus Personen bestehen, die schon zweimal getestet wurden, würde die Testwertinterpretation aufgrund des übungsbedingt höheren Leistungsniveaus der Vergleichsgruppe zu einer Unterschätzung der Konzentrationsleistung der Testperson führen.

Eine Differenzierung von Normen ist dann angezeigt, wenn wesentliche, d. h. mit dem Untersuchungsmerkmal korrelierte Hintergrundfaktoren der Testpersonen zu anderen Testwerten als denen der Vergleichsgruppe führen. Wird für relevante Ausprägungen auf dem Hintergrundfaktor jeweils eine eigene Norm gebildet, kann der Einfluss des Faktors auf die Testwertinterpretation kontrolliert werden. Beispielsweise kann der Übungsgrad dadurch kontrolliert werden, dass eine Normdifferenzierung nach der Testungszahl erfolgt, d. h. danach, ob es sich um die erste, zweite, dritte etc. Testung handelt. Normdifferenzierungen werden nicht nur zum Ausgleich von Übungseffekten, sondern häufig auch zum Ausgleich von Alters-, Geschlechts- oder Bildungseffekten vorgenommen, d. h., es werden getrennte Normen nach Alter, Geschlecht, Bildung etc. gebildet (► Beispiel 9.6).

Beispiel 9.6: Vorteile/Nachteile einer geschlechtsbezogenen Normdifferenzierung

(nach Cronbach 1990)

Thomas erzielt in einem Test zur Erfassung von mechanischem Verständnis (*Mechanical Reasoning*) einen Testwert von 40. Nach der geschlechtsunspezifischen Prozentrangsnormaltabelle entspricht dem Testwert ein Prozentrang von 65. Den Erfolg von Thomas im Ausbildungsprogramm anhand dieser Information einzuschätzen, wäre allerdings verzerrend, da die meisten Auszubildenden in mechanischen Berufen männlich sind. Realistischer wird die Erfolgsaussicht also durch den Vergleich mit der männlichen Bezugsgruppe bestimmt. Es stellt sich heraus, dass der Prozentrang von 65 auf den weniger Erfolg versprechenden Prozentrang von 50 abfällt, weil Männer im Test zur Erfassung von Mechanical Reasoning im Allgemeinen höhere Testwerte erzielen als Frauen.

Clara erzielt in demselben Test ebenfalls einen Testwert von 40. Während der Vergleich mit der gemischtgeschlechtlichen Bezugsgruppe zu einem Prozentrang von 65 führt, steigt er bei einem Vergleich mit der weiblichen Bezugsgruppe auf 80. Steht Clara mit anderen Frauen um die Aufnahme in das Ausbildungsprogramm im Wettbewerb, hat sie somit sehr gute Erfolgsaussichten. Sofern ihre Mitbewerber im Ausbildungsprogramm hauptsächlich Männer sein sollten, ist zur Abschätzung

Auf Angemessenheit der Vergleichsgruppe achten

9

Kontrolle von Hintergrundfaktoren durch Normdifferenzierung

9.6 · Testeichung

ihres Ausbildungserfolgs aber ein Vergleich ihrer Testleistung mit der männlichen Bezugsgruppe angezeigt. Wie Thomas läge sie demnach nur noch im Durchschnittsbereich.

Am Beispiel nach Cronbach (1990) wird deutlich, dass in Wettbewerbssituativen nicht immer der Vergleich mit der Gruppe, die bestmöglich mit der Testperson übereinstimmt, diagnostisch am sinnvollsten ist. Vielmehr ist entscheidend, dass ein Vergleich mit der Gruppe der tatsächlichen Konkurrenten bzw. Mitbewerber vorgenommen wird, denn dadurch können Erfolgsaussichten realistisch eingeschätzt und negative Auswirkungen wie Frustrationen vermieden werden.

Während differenzierte Normen also auf der einen Seite zu diagnostisch sinnvollen Entscheidungen verhelfen können, besteht auf der anderen Seite die Gefahr, dass durch eine zu starke Anpassung der Testnorm an bestimmte Teilpopulationen die Bedeutung des normierten Testresultats an Aussagekraft verliert und Fehleinschätzungen vorgenommen werden (*Overadjustment*, Cronbach 1990). Da beispielsweise der Testwert in Intelligenztests auch mit dem soziokulturellen Hintergrund zusammenhängt, mag man es bei Selektionsprozessen als fairer ansehen, nur Bewerber mit vergleichbarem soziokulturellem Hintergrund miteinander zu vergleichen, d. h. Testnormen nach soziokulturellem Hintergrund zu differenzieren. Dies entspricht der Grundidee des Fairnessmodells der *proportionalen Repräsentation* (*Quotenmodell*, s. z. B. Amelang und Schmidt-Atzert 2012). Demnach gilt ein Selektionskriterium dann als fair, wenn in der Gruppe der ausgewählten Bewerber die Proportion unterschiedener Teilgruppen dieselbe ist wie in der Bewerberpopulation. Auf diese Weise werden jedoch tatsächlich vorhandene Unterschiede zwischen Bewerbern aus unterschiedlichen Teilpopulationen nivelliert. Im Extremfall kann die systematische Berücksichtigung von Unterscheidungsmerkmalen beispielsweise die äußerst bedenkliche Folge haben, dass ein 40-jähriger Alkoholkranker für eine verantwortungsvolle Tätigkeit (z. B. Überwachung eines Produktionsprozesses) eingesetzt wird, sofern er in Relation zur Teilpopulation der 40-jährigen Alkoholkranken sehr gute Testresultate erzielte.

Ein Kompromiss kann darin bestehen, dass durch gesellschaftspolitische Wertvorstellungen motivierte Normdifferenzierungen vorgenommen werden; gleichzeitig muss jedoch eine tatsächliche Bewältigung der Anforderungen erwartet werden können. Das obige Beispiel lässt sich in diesem Sinne auffassen, da Clara zwar auf der einen Seite durch Anwendung einer frauenspezifischen Norm leichteren Zugang zum Ausbildungsprogramm erhält als männliche Bewerber, auf der anderen Seite jedoch gleichzeitig ihre tatsächlichen Erfolgsaussichten im Ausbildungsprogramm anhand der männerspezifischen Norm abgeschätzt werden müssen.

Eine weitere Facette des *Overadjustment*-Problems besteht nach Cronbach (1990) darin, dass defizitäre Zustände in einer Teilpopulation als „normal“ bewertet werden. Das bedeutet beispielsweise, dass bei der mangelhaften Integration von Immigrantenkindern in das Schulsystem der Vergleich des Testwertes eines dieser Kinder mit der Teilpopulation der Immigrantenkinder zu der Annahme führt, dass ein „normales“ Leistungsniveau vorliegt. Dabei wird durch den inadäquaten Vergleichsmaßstab das eigentliche Problem, d. h. die sehr niedrige Schulleistung von Immigrantenkindern, unkenntlich gemacht und erforderliche Interventionen werden aufgrund des scheinbar „normalen“ Leistungsniveaus nicht initiiert.

**Vergleich mit Mitbewerbern
führt zu einer realistischen
Erfolgseinschätzung**

**Überanpassung der Normen kann
zu Fehleinschätzungen führen**

**Überanpassung von Normen kann
ein Zerrbild der „Normalität“
entstehen lassen**

9.6 Testeichung

Die Testeichung stellt den letzten Schritt einer Testkonstruktion dar und dient dazu, einen Vergleichsmaßstab für die normorientierte Testwertinterpretation (► Abschn. 9.2) zu erstellen. Um Normwerte gewinnen zu können, wird das

zu normierende psychologische Testverfahren an einer großen *Normierungsstichprobe* (auch: *Eichstichprobe*), die hinsichtlich einer definierten Bezugsgruppe repräsentativ ist, durchgeführt und die Verteilung der Testwerte (► Kap. 8) erfasst.

9.6.1 Definition der Zielpopulation

Anforderungen an Normwerte

In der DIN 33430 (s. DIN 2002, 2016; Westhoff et al. 2010) werden zur Qualitätssicherung von diagnostischen Beurteilungen hinsichtlich der Normwerte eine Reihe von Anforderungen formuliert:

- » [Norm] values must correspond to the research question and the reference group [...] of the candidates. The appropriateness of the norm values is to be evaluated at least every eight years. (Hornke 2005, S. 262)

In den International Guidelines for Test Use ist analog folgende Richtlinie enthalten (vgl. ► Kap. 10):

- » Ensure that invalid conclusions are not drawn from comparisons of scores with norms that are not relevant to the people being tested or are outdated. (ITC 2013, S. 20)

Die für die Testanwendung geforderte Entsprechung zwischen Normwerten und Forschungsfrage bedeutet, dass zu Beginn der Testeichung unter Berücksichtigung von Anwenderinteressen vom Testautor zu entscheiden ist, welche Fragen auf Basis der zu bildenden Normen beantwortet werden sollen. Beispielsweise kann die Frage nach der Rechenleistung eines Jungen in der dritten Klasse im Vergleich zu seinen Klassenkameraden nur dann beantwortet werden, wenn eine aktuelle schulspezifische Testnorm, die auf der Bezugsgruppe der Drittklässler und nicht etwa der Viertklässler basiert, zur Verfügung steht. Das heißt, es ist zu Beginn der Testeichung genau zu definieren und entsprechend im Testmanual zu dokumentieren, für welche Bezugsgruppe bzw. *Zielpopulation* die zu erstellenden Testnormen gelten sollen. Im Rahmen der Festlegung der Zielpopulation(en) ist auch die Frage zu klären, ob eine Normdifferenzierung (► Abschn. 9.5) vorgenommen werden soll.

Bezugsgruppe genau definieren

Repräsentativität der Stichprobe sicherstellen

9.6.2 Erhebungsdesigns für Normierungsstichproben

Um sicherzustellen, dass, wie in der DIN 33430 gefordert, die Testnorm der Merkmalsverteilung der Bezugsgruppe entspricht, ist bei der Erhebung der Normierungsstichprobe darauf zu achten, dass die Normierungsstichprobe bezüglich der Zielpopulation *repräsentativ* ist. Hierbei ist zwischen globaler und spezifischer Repräsentativität zu unterscheiden.

Globale Repräsentativität und spezifische Repräsentativität einer Stichprobe

Eine repräsentative Stichprobe liegt dann vor, wenn die Stichprobe hinsichtlich ihrer Zusammensetzung die jeweilige Zielpopulation möglichst genau abbildet. Dies bedeutet, dass sich Repräsentativität immer auf eine *bestimmte* Zielpopulation bezieht bzw. dass eine Stichprobe repräsentativ bezüglich einer vorher definierten und keiner beliebigen anderen Population ist.

Eine Stichprobe wird *global repräsentativ* genannt, wenn ihre Zusammensetzung hinsichtlich aller möglichen Faktoren mit der Populationszusammensetzung übereinstimmt – dies ist nur durch Ziehen einer echten Zufallsstichprobe aus einer definierten Population zu erreichen. Dagegen gilt eine Stichprobe als *spezifisch*

repräsentativ, wenn sie lediglich hinsichtlich derjenigen Faktoren der Populationszusammensetzung repräsentativ ist, die mit dem Untersuchungsmerkmal bzw. dem Testwert in irgendeiner Weise zusammenhängen, wobei für die Bildung von Testnormen insbesondere das Geschlecht, das Alter und der Bildungsgrad oder der Beruf von Bedeutung sind. Es sind also genau solche Faktoren, die auch Anlass zu einer Normdifferenzierung geben könnten (vgl. ► Beispiel 9.6 in ► Abschn. 9.5 zur geschlechtsspezifischen Normierung des Tests zum Mechanical Reasoning). Mangelnde Repräsentativität kann durch einen größeren Stichprobenumfang nicht kompensiert werden, d. h. eine kleinere repräsentative Stichprobe ist nützlicher als eine große, jedoch nicht repräsentative Stichprobe.

Liegen Kenntnisse vor, welche Faktoren mit dem Untersuchungsmerkmal zusammenhängen, bieten sich zur Testeichung einerseits spezifisch repräsentative *geschichtete Stichproben* und andererseits *Quotenstichproben* an (vgl. Döring und Bortz 2016). Durch diese Art der Stichprobenziehung wird erreicht, dass die prozentuale Verteilung der Ausprägungen auf merkmalsrelevanten Faktoren in der Stichprobe mit der Verteilung in der Population identisch ist. Hierfür muss jedoch in Erfahrung gebracht werden können, wie die Ausprägungen der Faktoren, die mit dem Untersuchungsmerkmal zusammenhängen, in der Population verteilt sind. Beispielsweise dürften für den Faktor „Geschlecht“ die Ausprägungen „weiblich“ und „männlich“ in vielen Zielpopulationen gleichverteilt sein bzw. mit einer Häufigkeit von jeweils 50 % auftreten. Wird eine Person aus der Menge von Personen mit gleichen Ausprägungen merkmalsrelevanter Faktoren zufällig ausgewählt, spricht man von einer *geschichteten (stratifizierten) Stichprobe*, bei einer nicht zufälligen Auswahl dagegen von einer *Quotenstichprobe*. Soll beispielsweise ein persönlichkeitspsychologisches Testverfahren für Jugendliche normiert werden, so ist davon auszugehen, dass der Testwert mit den Faktoren „familiäres Milieu“ und „Arbeitslosigkeit“ zusammenhängt. Wenn aus demografischen Erhebungen bekannt ist, dass von den Jugendlichen in der Population 30 % Arbeiterfamilien, 15 % Unternehmerfamilien, 15 % Beamtenfamilien und 40 % Angestelltenfamilien angehören und der Anteil der arbeitslosen Jugendlichen je nach familiärer Herkunft gemäß der angegebenen Reihenfolge 5 %, 1 %, 2 % und 3 % beträgt, kann ein entsprechender Erhebungsplan nach diesen Quoten erstellt werden. Beispielsweise beträgt die Quote Jugendlicher aus Arbeiterfamilien, die arbeitslos sind, 1.5 % (5 % von 30 %), wohingegen 28.5 % der Jugendlichen aus Arbeiterfamilien beschäftigt sind. Für die Normierungsstichprobe werden Personen so ausgewählt, dass sich für die kombinierten Ausprägungen merkmalsrelevanter Faktoren eine prozentuale Verteilung gemäß Erhebungsplan ergibt.

Liegen zur Gewinnung von geschichteten bzw. Quotenstichproben keine Informationen über die mit dem Untersuchungsmerkmal korrelierten Faktoren vor, ist eine *Zufallsstichprobe* zu ziehen, die zu globaler Repräsentativität führt. Ideal-typisch bedeutet dies für die Erhebung der Normierungsstichprobe, dass aus der Menge aller Personen der Zielpopulation nach dem Zufallsprinzip eine bestimmte Menge von Personen ausgewählt wird. In der Praxis wird dieses Vorgehen jedoch höchstens im Falle sehr spezifischer Populationen realisierbar sein, da mit echten Zufallsstichproben ein erheblicher finanzieller, organisatorischer und personeller Aufwand sowie datenschutzrechtliche Hürden verbunden wären.

Wird eine *anfallende Stichprobe* oder *Ad-hoc-Stichprobe* zur Bildung von Testnormen verwendet, bedeutet dies, dass nicht gezielt versucht wird, eine repräsentative Stichprobe hinsichtlich einer bestimmten Zielpopulation zu ziehen. Stattdessen werden sich bietende Gelegenheiten genutzt, um Testdaten für die Normierungsstichprobe zu sammeln. Anfallende Stichproben können an Wert gewinnen, wenn

Geschichtete Stichprobe und Quotenstichprobe

Zufallsstichprobe

Ad-hoc-Stichprobe

Umfang der Normierungsstichprobe

Feinstufige Normen erfordern hohe Reliabilität

9

Homogene vs. heterogene Zielpopulation

Verteilungseigenschaften der Normierungsstichprobe prüfen

sich aus ihnen in Anlehnung an die oben beschriebene Vorgehensweise nachträglich eine Quotenstichprobe für eine bestimmte Zielpopulation bilden lässt.

Der erforderliche *Umfang der Normierungsstichprobe* hängt von verschiedenen Faktoren ab. Allgemein gilt, dass eine gebildete Norm aus empirischer Sicht im Prinzip nur so fein zwischen Testpersonen differenzieren kann, wie vorher in der Normierungsstichprobe die unterschiedlichen Testwerte ausgefallen sind, wobei eine größere Anzahl unterschiedlicher Testwerte durch eine größere repräsentative Normierungsstichprobe gewonnen werden kann. Möchte man also fein abgestufte Normwerte bilden, z. B. in einer Prozentrang- oder Standardnorm, ist eine umfangreichere Stichprobe nötig, als wenn grobstufige Normen, z. B. Quartile, erstellt werden sollen.

Feinstufige Normen sollten jedoch nur dann angestrebt werden, wenn der Test eine hohe Reliabilität (► Kap. 13, 14 und 15) aufweist. Im Falle einer niedrigen Reliabilität würde eine feinstufige Norm zum Vergleich von Testpersonen lediglich eine hohe Genauigkeit vortäuschen, die sich aufgrund des hohen Standardmessfehlers aber als nicht gerechtfertigt erweisen würde. Es resultieren nämlich breite Konfidenzintervalle, die zahlreiche Normwerte ober- und unterhalb des individuellen Normwertes einer Testperson einschließen. Da jedoch Werte innerhalb eines Konfidenzintervalls nicht unterschieden werden können, wäre die feine Abstufung der Normwerte bei niedriger Reliabilität ohne Nutzen.

Bei der Planung des Stichprobenumfangs ist ausgehend von der oben angeführten Überlegung weiter zu beachten, dass die potentiell mögliche Anzahl unterschiedlicher Testwerte in einer Zielpopulation von deren jeweiliger Zusammensetzung abhängt. Handelt es sich um eine relativ homogene Zielpopulation (z. B. Gymnasialabsolventen in einem bestimmten Bundesland), ist die Spannweite der Testwerte, die vom Bildungsgrad abhängen, begrenzter als in einer sehr heterogen zusammengesetzten Zielpopulation (z. B. 18-jährige Jugendliche in Deutschland). Das bedeutet, dass zur Erreichung desselben Abstufungsgrads in der Testnorm bei einer heterogeneren Zielpopulation bzw. bei einem Test mit weitem Geltungsbereich eine größere Normierungsstichprobe gezogen werden muss als bei einer homogenen Zielpopulation bzw. bei einem Test mit engem Geltungsbereich.

Liegen die aus der Normierungsstichprobe gewonnenen Testwerte vor, sind zunächst deren *Verteilungseigenschaften* zu überprüfen. Insbesondere ist von Interesse, ob die Testwerte normalverteilt sind, sodass z_v -Normen bzw. Standardnormen (► Abschn. 9.2.2) berechnet und mittels Standardnormalverteilung interpretiert werden können. Prozentrangnormen (► Abschn. 9.2.1) können hingegen auch bei fehlender Normalverteilung erstellt werden. Unter bestimmten Bedingungen können die Daten auch normalisiert werden (► Kap. 8). Gegebenenfalls sind die Normen nach bestimmten Faktoren, z. B. Geschlecht, Alter, Bildungsgrad oder Beruf, zu differenzieren (► Abschn. 9.5). Die entsprechenden gruppenspezifischen Normen müssen erneut auf ihre Verteilungseigenschaften untersucht werden. Insbesondere sind die Mittelwertunterschiede hinsichtlich der Signifikanz und der Relevanz dahingehend zu prüfen, ob tatsächlich ein Anlass für die Normdifferenzierung besteht.

Die Darstellung der (differenzierten) Testnormen erfolgt in der Regel tabellarisch, sodass bei Papier-Bleistift-Tests vom Testanwender der dem jeweiligen Testwert zugeordnete Normwert in einer Tabelle aufgesucht werden muss. Bei computerbasierten Tests oder Auswertungsprogrammen (z. B. FAKT-II, Moosbrugger und Goldhammer 2007) wird der Normwert automatisch ausgegeben.

9.6.3 Dokumentation der Normen im Testmanual

Damit es dem Testanwender möglich ist, die Angemessenheit der Normen in Bezug auf seine Fragestellung beurteilen zu können, sind die erstellten Normen im

9.7 · Zusammenfassung mit Anwendungsempfehlungen

Testmanual hinsichtlich folgender Gesichtspunkte zu dokumentieren (s. z. B. Häcker et al. 1998; s. auch Cronbach 1990; ► Kap. 10):

- Geltungsbereich der Normen, d. h. Definition der Zielpopulation(en), die diejenige(n) sein sollte(n), mit der (denen) ein Anwender die Testpersonen in der Regel vergleichen will
- Erhebungsdesign bzw. Grad der Repräsentativität hinsichtlich der Zielpopulation
- Stichprobenumfang und -zusammensetzung
- Deskriptivstatistiken
- Jahr der Datenerhebung

Der letzte Dokumentationsgesichtspunkt hebt darauf ab, dass Normen über mehrere Jahre hinweg veralten können und somit Ihre Eignung als aktuell gültigen Vergleichsmaßstab verlieren (s. z. B. ► Beispiel 9.7 zum Flynn-Effekt; Flynn 1999). Die in der DIN 33430 geforderte *Überprüfung der Gültigkeit von Normen* nach spätestens acht Jahren (Hornke 2005) ist als Richtwert zu verstehen. Sofern empirische Evidenzen schon vor Ablauf der acht Jahre auf eine Änderung der Merkmalsverteilung in der Bezugsgruppe hinweisen, ist eine frühere *Normenaktualisierung* angezeigt.

Überprüfung der Gültigkeit von Normen nach spätestens acht Jahren

Beispiel 9.7: Flynn-Effekt

Der Flynn-Effekt ist ein eindrucksvolles Beispiel dafür, dass Testnormen über einen längeren Zeitraum ihre Gültigkeit verlieren können. Flynn (1999) hat gezeigt, dass in den westlichen Industrienationen der mittlere IQ über einen Zeitraum von mehreren Jahren hinweg ansteigt. Das Phänomen wurde von ihm zufällig entdeckt, als er Testmanuale studierte, in denen die Intelligenzausprägungen von Testpersonen mit der älteren, ersten und zudem mit der späteren, revidierten Version eines Intelligenztests bestimmt wurden. Unter Verwendung der für die jeweilige Version gebildeten Testnorm zeigte sich, dass Testpersonen im älteren Test einen deutlich besseren Normwert bzw. IQ erzielten als im zeitnah normierten Test. Die norm-orientierte Testwertinterpretation auf Basis von veralteten Normen führt also unter der Bedingung eines längsschnittlichen Intelligenzanstiegs zu einer Überschätzung der individuellen Intelligenzausprägung, da der Vergleich mit der früheren Population dem Vergleich mit einer insgesamt leistungsschwächeren Population entspricht. Nach Flynn stieg unter weißen Amerikanern zwischen 1932 und 1978 der IQ um 14 Punkte an, was einer Rate von etwa 1/3 IQ-Punkten pro Jahr entspricht.

Zur Vermeidung von Fehlern bei der Testwertinterpretation muss also die Gültigkeit von Normen regelmäßig überprüft werden. Bei einer geänderten Merkmalsverteilung in der Bezugsgruppe sollte eine Normenaktualisierung bzw. eine erneute Testeichung erfolgen.

9.7 Zusammenfassung mit Anwendungsempfehlungen

Ob ein Testwert norm- (► Abschn. 9.2) oder kriteriumsorientiert (► Abschn. 9.3) interpretiert werden kann, d. h., ob eine *Realnorm* in Form einer Bezugsgruppe (z. B. eine Prozentrangnorm) oder eine *Idealnorm* in Form eines Kriteriums (z. B. Lernziel) angelegt wird, hängt von den diagnostischen Zielsetzungen ab, für die ein Test geeignet sein soll.

Real- vs. Idealnorm

Vor der Bildung einer Bezugsgruppennorm (z. B. Prozentrangnorm, ► Abschn. 9.2) muss die Zielpopulation (► Abschn. 9.6.1) definiert werden, d. h. diejenige Population, mit der ein Testanwender den Testwert einer Testperson in der Regel

Zielgruppe und Eichstichprobe

Testnormen bei ordinal- vs. intervallskalierten Testwerten

Kriteriumsorientierte Interpretation anhand eines externen Kriteriums oder der Aufgabeninhalte

vergleichen will. Um eine repräsentative Stichprobe aus der Zielpopulation für die Testeichung zu gewinnen, kann eine Quoten- bzw. geschichtete Stichprobe oder eine Zufallsstichprobe gezogen werden. Liegt zunächst nur eine Ad-hoc-Stichprobe vor, sollte nachträglich die Bildung einer Quotenstichprobe für eine bestimmte Zielpopulation angestrebt werden (► Abschn. 9.6.2), damit geeignete Normen im Testmanual dokumentiert werden können (► Abschn. 9.6.3).

Falls die Testwertvariable nicht intervallskaliert ist, kommt für eine normorientierte Testwertinterpretation nur die Bildung einer Prozentrangnorm (► Abschn. 9.2.1) infrage, die die relative Position eines Testwertes in der aufsteigend geordneten Rangreihe der Testwerte in der Bezugsgruppe angibt. Falls hingegen eine intervallskalierte Testwertvariable vorliegt, ist auch die Bildung einer z_v -Norm (► Abschn. 9.2.2) möglich, die für einen Testwert seinen Abstand zum Mittelwert der Bezugsgruppe in Einheiten der Standardabweichung angibt. Wenn eine intervallskalierte Testwertvariable die Voraussetzung der Normalverteilung erfüllt, ist die z_v -Norm insbesondere von Vorteil, da anhand der tabellierten Standardnormalverteilung die prozentuale Häufigkeit der z-Werte innerhalb beliebiger Wertebereiche bestimmt werden kann.

Eine kriteriumsorientierte Interpretation eines Testwertes (► Abschn. 9.3) kann dadurch vorgenommen werden, dass anhand eines zusätzlich zu erhebenden externen Kriteriums auf der Testwertska ein Schwellenwert bestimmt wird, dessen Überschreitung anzeigen, dass das Kriterium erfüllt ist (► Abschn. 9.3.1). Die ROC-Analyse stellt eine Möglichkeit dar, einen Schwellenwert empirisch zu definieren. Alternativ kann eine kriteriumsorientierte Interpretation anhand der Aufgabeninhalte erfolgen (► Abschn. 9.3.2). Dieses Vorgehen stellt jedoch deutlich höhere Anforderungen an die Aufgabenkonstruktion, da eine genaue inhaltliche Vorstellung von der Grundgesamtheit der für das zu erfassende Merkmal relevanten Aufgaben bestehen muss und die Testaufgaben eine repräsentative Stichprobe aus der Aufgabengrundgesamtheit darstellen müssen. Der Testwert stellt in diesem Fall unmittelbar einen Indikator für die Merkmalsausprägung dar, da von der Leistung in der Aufgabenstichprobe auf die Leistung in der Aufgabengrundgesamtheit geschlossen werden darf. Verfahren zur Generierung repräsentativer Aufgabenstichproben werden von Klauer (1987a) beschrieben.

9.8 EDV-Hinweise

Mit einschlägiger Statistiksoftware (z. B. R, SPSS) können Perzentilwerte, z-Werte und ROC-Kurven, inklusive der Koordinatenpunkte Sensitivität und 1 – Spezifität, einfach berechnet werden.

Ein Datenbeispiel finden Sie im Bereich EDV-Hinweise unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

9.9 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Worin unterscheiden sich norm- und kriteriumsorientierte Testwertinterpretationen?
2. Welches Skalenniveau weisen Prozentränge auf, und was ist infolgedessen bei der Verwendung von Prozenträngen zu beachten?
3. Für eine Testperson mit dem Testwert $Y_v = 45$ soll ermittelt werden, wie groß der Personenanteil in der Bezugsgruppe ist, der einen geringeren oder maximal

- so hohen Testwert erzielt hat wie Y_v . Es ist bekannt, dass die Testwertvariable in der Bezugsgruppe normalverteilt ist ($\bar{Y} = 30$, $SD(Y) = 10$).
4. Ein Testentwickler hat mittels ROC-Analyse einen optimalen Schwellenwert definiert. Aus inhaltlichen Gründen hält er es für sinnvoll, den Schwellenwert so zu verschieben, dass die Rate falsch positiver Klassifikationen sinkt. In welche Richtung muss der Schwellenwert verschoben werden, wenn gilt, dass niedrige Testwerte auf das Vorliegen des Kriteriums hinweisen?
 5. Welche Rolle spielt die Normdifferenzierung bei der Testeichung?

Literatur

- Amelang, M. & Schmidt-Atzert, L. (2012). *Psychologische Diagnostik und Intervention* (5. Aufl.). Berlin, Heidelberg: Springer.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.
- Deutsches Institut für Normung e.V. (DIN). (2002). *DIN 33430:2002-06: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Deutsches Institut für Normung e.V. (DIN). (2016). *DIN 33430:2016-07: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage). Heidelberg: Springer.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
- Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist*, 54, 5–20.
- Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.
- Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe.
- Hornke, L. F. (2005). Die englische Fassung der DIN 33430. In K. Westhoff, L. J. Hellfritsch, L. F. Hornke, K. D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2. Aufl., S. 255–283). Lengerich: Pabst.
- International Test Commission (ITC). (2013). International Guidelines for Test Use. (Version 1.2). Verfügbar unter https://www.intestcom.org/files/guideline_test_use.pdf [20.12.2019]
- Klauer, K. J. (1987a). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.
- Klauer, K. C. (1987b). Kriteriumsorientiertes Testen: Der Schluß auf den Itempool. *Zeitschrift für differentielle und diagnostische Psychologie*, 8, 141–147.
- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2004). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss*. Neuwied: Luchterhand.
- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). (2005). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Hauptschulabschluss*. Neuwied: Luchterhand.
- Moosbrugger, H. & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test. Grundlegend neu bearbeitete und neu normierte 2. Auflage des FAKT von Moosbrugger und Heyden (1997)*. Testmanual. Bern: Huber.
- Organisation for Economic Co-operation and Development (OECD). (2001). *Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for Tomorrow's World – First Results from PISA 2003*. Paris: OECD.
- Rettler, H. (1999). Normorientierte Diagnostik. In R. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch* (4. Aufl., S. 221–226). Weinheim: Psychologie Verlags Union.
- Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.
- Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G. & Stemmler, G. (Hrsg.). (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Aufl.). Lengerich: Pabst.
- Youden, W. (1950). Index rating for diagnostic tests. *Cancer*, 3, 32–35.



Standards für psychologisches Testen

Volkmar Höfling und Helfried Moosbrugger

Inhaltsverzeichnis

- 10.1 Ziele von Teststandards – 198**
- 10.2 Standards für die Entwicklung und Evaluation psychologischer Tests – 198**
 - 10.2.1 Überblick – 198
 - 10.2.2 Standards zur Validität – 200
 - 10.2.3 Standards zur Reliabilität – 201
 - 10.2.4 Standards zu Itemgenerierung und Testentwicklung – 201
 - 10.2.5 Standards zu Normen und Testdokumentation – 202
- 10.3 Standards für die Übersetzung und Anpassung psychologischer Tests – 203**
- 10.4 Standards für die Anwendung psychologischer Tests – 204**
 - 10.4.1 Richtlinien für die Testanwendung und für die Kompetenzen der Testanwender – 205
 - 10.4.2 Testvorbereitungsphase – 206
 - 10.4.3 Testphase – 209
 - 10.4.4 Testauswertungsphase – 209
- 10.5 Standards für die Qualitätsbeurteilung psychologischer Tests – 210**
 - 10.5.1 Überblickswerke – 210
 - 10.5.2 Testbeurteilungssystem des Testkuratoriums (TBS-TK) – 210
- 10.6 Zusammenfassung – 213**
- 10.7 Kontrollfragen – 213**
- Literatur – 213**

i Standards für psychologisches Testen beziehen sich auf verschiedene Bereiche des Testens, z. B. auf die Entwicklung und Evaluation (*Testkonstruktion*), auf die Übersetzung und Anpassung (*Testadaptation*), auf die Durchführung, Auswertung und Interpretation (*Testanwendung*) sowie auf die Überprüfung der Einhaltung der Standards bei der Testentwicklung und -evaluation (*Qualitätsbeurteilung*) psychologischer Tests. Teststandards zielen in den genannten Phasen bzw. Bereichen auf größtmögliche Optimierung und wollen dazu beitragen, dass die im Rahmen psychologischen Testens getroffenen Aussagen mit hoher Wahrscheinlichkeit zutreffen.

10.1 Ziele von Teststandards

Definition

Teststandards sind vereinheitlichte Leitlinien, in denen sich allgemein anerkannte Zielsetzungen zur Entwicklung, Adaptation, Anwendung und Qualitätsbeurteilung psychologischer Tests widerspiegeln.

Bei der Anwendung von Standards für psychologisches Testen geht es nicht um deren buchstabengetreue Erfüllung, sondern um ihre souveräne Beachtung in den verschiedenen Bereichen bzw. Phasen psychologischen Testens. Teststandards vermögen ein auf verhaltenswissenschaftlicher, psychometrischer und anwendungspezifischer Kompetenz beruhendes Urteil nie zu ersetzen (vgl. Häcker et al. 1998, S. 3), aber zu erleichtern.

Verschiedene (nationale bzw. internationale) psychologische Organisationen (► Exkurs 10.1) haben Teststandardkompendien mit mehr oder weniger vergleichbarer Zielsetzung erarbeitet. Auf der Grundlage dieser Kompendien wird in diesem Kapitel ein Überblick über Teststandards gegeben, die die Entwicklung und Evaluation (*Testkonstruktion*; ► Abschn. 10.2), die Übersetzung und Anpassung (*Testadaptation*; ► Abschn. 10.3), die Durchführung, Auswertung und Interpretation (*Testanwendung*; ► Abschn. 10.4) und schließlich die Überprüfung der Einhaltung der Standards der Testentwicklung und -evaluation (*Qualitätsbeurteilung*; ► Abschn. 10.5) betreffen.

10.2 Standards für die Entwicklung und Evaluation psychologischer Tests

10.2.1 Überblick

Standards for Educational and Psychological Testing (SEPT)

Relevante Leitlinien für die Entwicklung und Evaluation psychologischer Testverfahren liegen vor allem in Form von zwei Teststandardkompendien vor: Das eine Werk ist die siebte Fassung der „Standards for Educational and Psychological Testing“ (SEPT), (AERA et al. 2014), deren vierte Fassung auch ins Deutsche übersetzt wurde („Standards für pädagogisches und psychologisches Testen“, SPPT; Häcker et al. 1998). Bei dem anderen Werk handelt es sich um die DIN 33430:2002-06: *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen* (DIN 2002; s. Westhoff et al. 2010, S. 232–259; vgl. Reimann 2009), die als DIN 33430: 2016-07: *Anforderungen an berufsbezogene Eignungsdiagnostik* (DIN 2016; vgl. Diagnostik- und Testkuratorium 2018) in ihrer aktuellsten Form vorliegt.

Die „SEPT“ können auf eine sehr lange Tradition zurückblicken. Die erstmals 1954 erschienenen Standards liegen seit 2014 in der siebten Fassung vor und tragen dem aktuellen Entwicklungsstand im Feld der psychologischen Diagnostik Rechnung. Sie thematisieren u. a. besonders die Konstruktvalidität, die

Exkurs 10.1**Organisationen, die Teststandards erarbeitet haben**— *American Educational Research Association (AERA)*:

Seit ihrer Gründung im Jahr 1916 hat es sich die AERA zur Aufgabe gemacht, für eine ständige Qualitätsverbesserung im Bildungsbereich zu sorgen. In diesem Zusammenhang fördert die AERA empirische Untersuchungen und die praktische Umsetzung von Forschungsergebnissen.

— *American Psychological Association (APA)*

Mit ca. 150000 Mitgliedern ist die APA die größte Psychologenvereinigung weltweit. Ihre Zielsetzungen sind u. a. die Förderung der psychologischen Forschung, die stetige Verbesserung von Forschungsmethoden und -bedingungen, die Verbesserung der Qualifikationen von Psychologen durch Standards in den Bereichen Ethik, Verhalten, Erziehung und Leistung, und die angemessene Verbreitung psychologischen Wissens durch Kongresse bzw. Veröffentlichungen.

— *National Council on Measurement in Education (NCME)*:

Das NCME fördert Projekte im Kontext der pädagogischen Psychologie. Hierbei geht es um die Optimierung psychometrischer Testverfahren und deren verbes-

serte Anwendung im Rahmen pädagogisch-psychologischer Diagnostik.

— *Diagnostik- und Testkuratorium (DTK) der Föderation Deutscher Psychologenvereinigungen*:

Das Diagnostik- und Testkuratorium (vormals: Testkuratorium) ist ein von der Föderation Deutscher Psychologenvereinigungen, d. h. von der Deutschen Gesellschaft für Psychologie (DGP) und dem Berufsverband Deutscher Psychologinnen und Psychologen (BDP) getragenes Gremium, dessen Aufgabe es ist, die Öffentlichkeit vor unzureichenden diagnostischen Verfahren und vor unqualifizierter Anwendung diagnostischer Verfahren zu schützen und verbindliche Qualitätsstandards zu formulieren (z. B. in Form der DIN 33430 zur berufsbezogenen Eignungsbeurteilung).

— *International Test Commission (ITC)*:

Die ITC setzt sich zusammen aus verschiedenen nationalen Psychologenvereinigungen und Testkommissionen, darunter auch Forscher zur Thematik des psychologischen Testens. Zielsetzung der ITC ist der Austausch und die Zusammenarbeit bezüglich der Konstruktion, der Verbreitung und der Anwendung psychologischer Tests.

Item-Response-Theorie (IRT), Kriterien für die Verwendung kritischer Trennwerte (Cut-off-Werte) und Überlegungen zur Anwendung technologisch gestützter Testsysteme.

Inhalte des ersten Teils der siebten Fassung der SEPT

Testkonstruktion, -evaluation und -dokumentation:

1. Validität
2. Reliabilität und Messfehler
3. Testentwicklung und -revision
4. Skalierung und Normierung
5. Testdurchführung, -auswertung und Ergebnisdarstellung
6. Testdokumentation

Die vom *Testkuratorium* der Föderation Deutscher Psychologenvereinigungen vorbereitete und vom Deutschen Institut für Normung e. V. herausgegebene „DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen“ (DIN 2002; Kersting 2016) enthält nicht nur Richtlinien, sondern festgelegte Qualitätsstandards für den Prozess des psychologischen Testens und bilden damit die erste Norm für wesentliche Aufgabenfelder des Personalwesens weltweit. Enthalten sind sowohl Anforderungen an diagnostische Verfahren an sich als auch Anforderungen hinsichtlich des Vorgehens bei der Anwendung dieser Verfahren für berufsbezogene Eignungsbeurteilungen. Die Norm wendet sich an *Auftraggeber*, die berufsbezogene Eignungsbeurteilungen in Auftrag geben, an *Auftragnehmer*, die für Auftraggeber berufsbezogene Eignungsbeurteilungen durchführen und an *Mitwirkende*, die unter Anleitung, Fachaufsicht und Verantwortung von Auftragnehmern Verfahren zur Eignungsbeurteilung durchführen.

DIN 33430

Qualitätskriterien und -standards für Verfahren zur berufsbezogenen Eignungsbeurteilung der DIN 33430 (DIN 2002; s. Westhoff et al. 2010, S. 237–239)

Abschn. 4.2: Auswahl und Zusammenstellung der Verfahren:

1. Verfahrenshinweise
2. Objektivität
3. Zuverlässigkeit
4. Gültigkeit
5. Normwerte; Referenzkennwerte

Die thematischen Überschneidungen der beiden Teststandardkompendien in Bezug auf die Entwicklung und Evaluation psychologischer Tests sind evident, beide thematisieren in Form von Handlungsanweisungen diejenigen Kriterien, die in der Phase der Entwicklung und Evaluation psychologischer Tests beachtet werden müssen. Sie bilden somit eine Schnittmenge zu dem, was in ► Kap. 2 zu den „Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“)“ behandelt wird, allerdings mit dem Unterschied, dass Teststandards stets in Form von Handlungsanweisungen formuliert werden.

10

10.2.2 Standards zur Validität

Empirische Validitätsbelege sollten alle acht Jahre auf ihre Gültigkeit überprüft werden

Zunächst werden Standards zur Validität (vgl. ► Kap. 21) aufgestellt. Testentwickler werden aufgefordert, für die verschiedenen Validitätsaspekte (Inhalts-, Kriteriums- und Konstruktvalidität) empirische Belege vorzulegen, die relativ aktuellen Datums und möglichst nicht älter als acht Jahre sein sollen. Im Kontext der Inhaltsvalidität muss beispielsweise der im Test abgebildete Inhaltsbereich definiert und in seiner Bedeutung für die vorgesehene Testanwendung beschrieben sein; bei etwaigen Expertenurteilen muss die Qualifikation der Experten dargelegt werden. Für den Nachweis von Kriteriumsvalidität fordern die Teststandardkompendien u. a. eine exakte Beschreibung etwaiger Kriteriumsmaße eines Tests und deren Erfassung. Zur differentiellen Vorhersagbarkeit müssen statistische Schätzungen Anwendung finden (z. B. multiple Regression), wobei Gruppenunterschiede zu berücksichtigen sind. Nachfolgend werden beispielhaft Auszüge zur „Konstruktvalidität“ aus den beiden Teststandardkompendien gegeben (► Beispiel 10.1).

Beispiel 10.1: Standards zur Konstruktvalidität laut SPPT und DIN 33430

SPPT, Standard 1.8 (Häcker et al. 1998, S. 18): „Ist ein Test zur Messung eines Konstrukts vorgesehen, sollte dieses Konstrukt von anderen abgegrenzt werden. Die vorgeschlagene Interpretation der Testwerte ist ausführlich darzustellen; konstruktbezogene Validitätsbelege sollten angeführt werden, um solche Schlussfolgerungen zu untermauern. Insbesondere sollten empirische Belege dafür erbracht werden, dass ein Test nicht in hohem Maße von anderen Konstrukten abhängt.“

DIN 33430, Abschn. A.7.1: Konstruktgültigkeit (DIN 2002; s. Westhoff et al. 2010, S. 248 f.): „Das interessierende Konstrukt muss von anderen Konstrukten klar abgrenzbar und in einen theoretischen Rahmen eingebettet sein. Das Konstrukt und die diesbezüglichen empirisch-psychologischen Forschungsergebnisse sind so zu beschreiben, dass sie ohne Sekundärliteratur verstehbar sind. Verfahrensrelevante

theoretische Alternativen sind ebenso darzustellen wie solche empirische Ergebnisse, die den zugrunde gelegten Annahmen widersprechen. Aufgrund von inhaltlichen Überlegungen und empirischen Ergebnissen ist darzulegen, wie sich das fragliche Konstrukt zu ähnlichen (konvergente Gültigkeit) und unähnlichen Konstrukten (diskriminante Gültigkeit) verhält.“

10.2.3 Standards zur Reliabilität

Weiterhin sind bei der Testentwicklung Standards zur Reliabilität/Zuverlässigkeit (vgl. ► Kap. 14 und 15) zu beachten. Auch im Kontext der Reliabilität wird in den Teststandardkompendien darauf hingewiesen, dass Kennwerte alle acht Jahre auf ihre Geltung hin empirisch überprüft werden müssen. Es wird ferner gefordert, dass Zuverlässigkeitsschätzungen sowohl für Gesamt- als auch für Subtestwerte vorzulegen sind, wobei auch hier die den Berechnungen zugrunde liegende Stichprobe möglichst genau beschrieben werden muss. Ferner ist es notwendig, zu den Zuverlässigkeitsskennwerten auch die Methode ihrer Bestimmung anzugeben. Die Bestimmung der internen Konsistenz zur Quantifizierung der Reliabilität allein ist nicht ausreichend, es sollten daher möglichst mehrere Reliabilitätskennziffern bzw. Standardmessfehlerangaben bereitgestellt werden (► Beispiel 10.2). Unterscheiden sich Zuverlässigkeitsskennwerte bzw. Standardmessfehler für verschiedene soziodemografische Gruppen, sollten die Kennwerte für alle Gruppen aufgeführt werden.

**Überprüfung
der Reliabilitätsangaben
alle acht Jahre und Angabe
der Bestimmungsmethoden**

Beispiel 10.2: Standards zur Reliabilität/Zuverlässigkeit laut SPPT und DIN 33430

SPPT, Standard 2.1 (Häcker et al. 1998, S. 24): „Für jeden angegebenen Gesamtwert, Subtestwert oder für jede Kombination von Testwerten sollen Schätzungen der relevanten Reliabilitäten und Standardmessfehler ausführlich und detailliert angegeben werden, damit der Testanwender einschätzen kann, ob die Testwerte für die von ihm vorgesehene Testanwendung ausreichend genau sind.“

DIN 33430, Abschn. A.6: Zuverlässigkeit (Reliabilität) (DIN 2002; s. Westhoff et al. 2010, S. 247): „Es ist anzugeben, nach welcher Methode die Zuverlässigkeit bestimmt wurde. Die Angemessenheit der herangezogenen Methode ist für verschiedene Typen von Eignungsbeurteilungen beispielhaft zu erläutern [...]. In den Verfahrenshinweisen muss beschrieben werden, wie die zur Zuverlässigkeitsbestimmung herangezogenen Untersuchungsgruppen zusammengesetzt waren.“

**Für jeden Subtest Inhaltsbereich
und Beispielitems angeben**

10.2.4 Standards zu Itemgenerierung und Testentwicklung

Im Kontext von Itemgenerierung und Testentwicklung verweisen insbesondere die SEPT bzw. SPPT auf eine Reihe von Standards. Beispielsweise müssen bei der Itemgenerierung der interessierende Inhaltsbereich beschrieben sowie repräsentative Items für jeden Subtest angegeben werden. Weiterhin sollten möglichst viele fundierte empirische Nachweise gesammelt werden, die mit dem neuen Test zusammenhängen. Weisen neue Forschungsergebnisse auf bedeutsame Veränderungen des Inhaltsbereichs, der Testanwendung oder der Testinterpretation hin, muss

Zugrunde liegendes IRT-Modell beschreiben

eine entsprechende Testrevision durchgeführt werden (► Beispiel 10.3). Empirische Belege und theoretische Begründungen sind auch für etwaige Kurzformen eines neu konstruierten Tests vorzulegen.

Werden im Rahmen probabilistischer Testtheorien sog. „Itemcharakteristische Funktionen“ (IC-Funktionen) grafisch dargestellt, muss auch das diesen Grafiken zugrunde liegende probabilistische Modell (IRT-Modell) beschrieben werden. Für die hierbei verwendete Stichprobe gilt, dass sie ausreichend groß sein und möglichst differenziert beschrieben werden muss. Sind bei der Testkonstruktion Subtests gebildet worden, so müssen diese gemäß der beiden Teststandardkompendien bezüglich ihrer theoretischen Konzeption beschrieben werden, um zu zeigen, dass sie mit dem beabsichtigten Zweck des Tests in Einklang stehen und angemessen interpretiert werden können.

Beispiel 10.3: Standards zur Testentwicklung laut SPPT

(Häcker et al. 1998, S. 65)

Standard 3.1: „Tests [...] sollten auf einer fundierten wissenschaftlichen Basis entwickelt werden. Testentwickler sollten alle empirischen Nachweise sammeln, die mit einem Test zusammenhängen. Sie sollten entscheiden, welche Informationen schon vor der Testveröffentlichung oder -distribution benötigt werden bzw. welche Informationen später vorgelegt werden können, und sie sollten die erforderliche Forschung durchführen.“

10

10.2.5 Standards zu Normen und Testdokumentation

Repräsentative Normierungsstichprobe

Testmanual muss alle relevanten Informationen enthalten

Testnormen sollten in Bezug auf diejenigen Gruppen erhoben werden, für die die Anwendung des psychologischen Tests von besonderer Bedeutung sind. Die durchgeführte Normierungsstudie (einschließlich des Stichprobenplans zur Gewinnung der repräsentativen Normierungsstichprobe, s. ► Kap. 9, ► Abschn. 9.6) sollte ausführlich dargestellt werden. Normwerte sind alle acht Jahre auf ihre Angemessenheit bzw. Gültigkeit hin zu überprüfen.

Für die Dokumentation psychologischer Tests muss vor allem das Testmanual (Testhandbuch bzw. Verfahrenshinweise) bestimmte Informationen enthalten. Die theoretische Grundkonzeption eines Tests sollte darin ohne Sekundärliteratur verständlich sein, seine Zielsetzung bzw. sein Anwendungsbereich deutlich werden (► Beispiel 10.4). Zudem muss aufgeführt sein, welche Qualifikationen für die Testanwendung erforderlich sind.

Beispiel 10.4: Standards zu Verfahrenshinweisen laut DIN 33430

(DIN 2002; s. Westhoff et al. 2010, S. 237)

Abschn. 4.2.1: Verfahrenshinweise: „In den Verfahrenshinweisen für standardisierte Verfahren zur Eignungsbeurteilung müssen die Zielsetzungen und Anwendungsbereiche benannt, relevante empirische Untersuchungen nachvollziehbar beschrieben, Konstruktionsschritte in angemessener, ausführlicher und verständlicher Weise dargestellt und alle Gütekriterien und eingesetzten Analysemethoden nachvollziehbar dokumentiert werden.“

10.3 Standards für die Übersetzung und Anpassung psychologischer Tests

Die „International Test Commission Guidelines for Translating and Adapting Tests“ (ITC-G-TA) wurden von der International Test Commission entwickelt (Hambleton 2001; International Test Commission 2005b), um eine qualitativ hochwertige Adaptation/Anpassung psychologischer Tests über Sprachen und Kulturen hinweg zu gewährleisten. Im Verständnis der ITC-G-TA erschöpft sich die Adaptation psychologischer Tests nicht einfach in der Übersetzung von Items, sondern erfordert darüber hinaus die Berücksichtigung weiterer Anpassungsaspekte.

Beispiel 10.5: Standards zur Anpassung psychologischer Tests laut ITC-G-TA

(International Test Commission 2017)

D.7: Testentwickler sollten geeignete statistische Techniken anwenden, um

1. die Äquivalenz der verschiedenen Testversionen sicherzustellen und
2. Testkomponenten oder -aspekte zu identifizieren, die für eine oder mehrere der interessierenden Populationen ungeeignet sind.

ITC-Guidelines for Translating and Adapting Tests

In vier Sektionen werden Richtlinien vorgestellt, die eine optimale Adaptation psychologischer Tests gewährleisten sollen:

- In Sektion 1 geht es um die Frage der Konstruktäquivalenz in Bezug auf eine Population mit anderem sprachlichen und kulturellen Hintergrund. Mit Psychologen der betreffenden Kulturen bzw. Sprachen sollte die Frage erörtert werden, ob davon ausgegangen werden kann, dass es sich bei dem psychologischen Konstrukt um ein sprach- und kulturgebundenes Konstrukt handelt oder nicht (► Beispiel 10.5).
- In Sektion 2 werden die Vorgänge der Übersetzung, der Datenerhebung und der statistischen Überprüfung thematisiert. Durch mindestens zwei Übersetzer, die im Idealfall sowohl über ausgewiesene Kenntnisse der beteiligten Kulturen als auch der zu messenden Konstrukte verfügen, soll eine optimale Übertragung der Operationalisierungen sichergestellt werden. Durch die Erhebung geeigneter Stichproben mit anschließenden statistischen Analysen sollen empirische Belege für die Konstruktäquivalenz bzw. Reliabilität und Validität der adaptierten Testversion bereitgestellt werden.
- In Sektion 3 werden Fragen zur Testdurchführung bei sprachlich und kulturell unterschiedlichen Gruppen geklärt, wobei auf die Auswahl von Testanwendern, die Wahl der Aufgabenstellungen und auf Zeitbeschränkungen eingegangen wird.
- Sektion 4 schließlich betont die Notwendigkeit einer ausführlichen Testdokumentation (z. B. mittels Testhandbuch) zur Gewährleistung einer zufriedenstellenden Validität und sorgfältigen Testwertinterpretation (z. B. mittels Normierung), um diagnostische Fehlentscheidungen aufgrund sprachlicher und kultureller Unterschiede zu vermeiden (► Studienbox 10.1; vgl. Van de Vijver und Watkins 2006).

Optimale Adaptation psychologischer Tests

Die ITC-G-TA wurden u. a. bereits erfolgreich bei der Adaptation von Messinstrumenten für international sehr beachtete Studien wie die dritte „Trends in International Mathematics and Science Study“ (TIMSS; vgl. Baumert et al. 2000) und das „Programme for International Student Assessment“ (PISA; vgl. Haider und Reiter 2004; Prenzel et al. 2007) eingesetzt.

Bedeutende erfolgreiche Adaptationen

Studienbox 10.1**Überprüfung der Übertragbarkeit der amerikanischen Normen des „Youth Self Report“ (YSR/11-18) an einer nicht klinischen deutschen Stichprobe (Roth 2000)**— **Fragestellung:**

Sind die amerikanischen Normen des YSR/11-18 (Achenbach 1991) auf deutsche Jugendliche übertragbar?

— **Methode:**

An einer Stichprobe deutscher Jugendlicher zwischen 12 und 16 Jahren ($N = 352$) wurde die deutsche Bearbeitung des YSR/11-18 (vgl. Arbeitsgruppe Deutsche Child Behavior Checklist 1998) überprüft. Anhand der Cut-off-Werte von Achenbach (1991) wurden die Jugendlichen in die Kategorien „Klinisch auffällig“ ($T > 63$), „Übergangsbereich“ ($63 \geq T \geq 60$) und „Klinisch unauffällig“ ($T < 60$) eingeteilt (beobachtete Häufigkeiten). Gemäß der amerikanischen Normierungsstichprobe wurden erwartete Häufigkeiten für die Kategorien „Übergangsbereich“ und „Klinisch auffällig“ ermittelt. Die erwarteten und beobachteten Häufigkeiten wurden mittels χ^2 -Tests inferenzstatistisch verglichen.

— **Ergebnisse:**

Die für die jeweiligen Kategorien ermittelten Häufigkeiten wichen deutlich von den anhand der amerikanischen Normierungsstichprobe erwarteten Häufigkeiten ab (Tab. 10.1). Mehr als doppelt so viele deutsche Jugendliche (im Vergleich zu amerikanischen Jugendlichen) wurden anhand der amerikanischen Normen den Kategorien „Klinisch auffällig“ und „Übergangsbereich“ zugeordnet.

— **Diskussion:**

Die Ergebnisse zeigen, dass die amerikanische Normierung nicht auf deutsche Jugendliche übertragen werden kann und eine deutsche Normierung des Verfahrens unumgänglich ist. Als mögliche Erklärung der Unterschiede wird sensitiveres Antwortverhalten deutscher Jugendlicher diskutiert.

■ **Tabelle 10.1** Beobachtete und erwartete Häufigkeiten in den drei Auffälligkeitskategorien bezüglich des Gesamtwertes

	$f_{\text{beob.}}$	$f_{\text{erwart.}}$	$\chi^2 (df = 2)$
Klinisch auffällig	79	28.44	
Übergangsbereich	59	27.42	149.00*
Klinisch unauffällig	214	296.14	

$f_{\text{beob.}}$ = beobachtete Häufigkeit aufgrund der Einteilung der Stichprobe nach den Cut-off-Werten von Achenbach (1991); $f_{\text{erwart.}}$ = erwartete Häufigkeit aufgrund der Verteilung in der amerikanischen Normierungsstichprobe; * = $p \leq .001$.

10.4 Standards für die Anwendung psychologischer Tests

Testanwendungstandards beinhalten viele Aspekte

Die Standards für die Anwendung (Vorbereitung, Durchführung, Auswertung und Interpretation) psychologischer Tests stellen eine bedeutsame Bedingung für objektive, reliable und valide Testergebnisse dar und wollen ethische Aspekte der Testanwendung angemessen berücksichtigen. Eine Testanwendung beinhaltet stets in umfassender Weise die Fragestellung, die Anforderungsanalyse, die Untersuchungsplanung und die Verschränkung der in der Testanwendung gewonnenen Informationen in diagnostischen Urteilen bzw. Entscheidungen (vgl. Westhoff et al. 2003).

10.4.1 Richtlinien für die Testanwendung und für die Kompetenzen der Testanwender

Drei Teststandardkompendien geben vor allem relevante Leitlinien für die Durchführung, Auswertung und Interpretation psychologischer Tests, und zwar die SEPT, die DIN 33430 (► Abschn. 10.2) sowie die „International Test Commission Guidelines on Test Use“ (ITC-G-TU; ITC 2001).

Die ITC-G-TU repräsentieren – wie die meisten anderen Teststandardkompendien auch – die Arbeit von Experten für psychologisches und pädagogisches Testen aus verschiedenen Nationen über einen langen Zeitraum. Insofern stellen sie keine neuen Standards dar, sondern fassen die übereinstimmenden Stränge bestehender Richtlinien zur Durchführung, Auswertung und Interpretation psychologischer Tests zusammen und bieten dem Testanwender somit eine nachvollziehbare Struktur.

Zweck und Zielsetzung der ITC-G-TU (ITC 2001, S. 7 f.)

„Das langfristige Ziel dieses Projektes ist u. a. die Erstellung von Richtlinien, die die für einen Testanwender erforderlichen fachlichen Kompetenzen (Fachwissen, Fertigkeiten, Fähigkeiten und andere persönliche Merkmale) betreffen. Diese Kompetenzen werden in Form von nachvollziehbaren Handlungskriterien spezifiziert. Auf Grundlage dieser Kriterien kann genau beschrieben werden, welche Kompetenzen von einem qualifizierten Testanwender erwartet werden können:

- Fachliche und ethische Standards beim Testen,
- Rechte des Probanden und anderer am Testprozess Beteiligter,
- Auswahl und Evaluation alternativer Tests,
- Testvorgabe, Bewertung und Interpretation und
- Anfertigung von Testberichten und Rückmeldung der Ergebnisse.“

Internationale Richtlinien für die Testanwendung

Kompetenzen der Testanwender

Die Umsetzung dieser Zielsetzung führte im Rahmen der ITC-G-TU zur Formulierung von Richtlinien zu verschiedenen Fertigkeitsebenen, über die der Testanwender verfügen sollte:

- Persönliche und handlungsorientierte Fertigkeiten: z. B. hinreichende mündliche und schriftliche Kommunikationsfähigkeiten
- Kontextbezogene Kenntnisse und Fertigkeiten: z. B. notwendiges fachliches und ethisches Wissen für die Testauswahl
- Fertigkeiten für die Aufgabenhandhabung: z. B. fachliche und ethische Verhaltensgrundsätze für den Umgang mit Tests bzw. Testdaten von der Erhebung über die Auswertung und Interpretation bis zur Sicherung
- Fertigkeiten zur Bewältigung unvorhergesehener Situationen: z. B. kompetente Bewältigung von Störungen im Routineablauf oder kompetenter Umgang mit Fragen von Testpersonen während der Testvorgabe

Auf Basis der drei relevanten Teststandardkompendien folgt nun ein Überblick über bedeutsame Standards zur Durchführung und Auswertung bzw. Interpretation psychologischer Tests, wobei ein *Phasenmodell* den übergeordneten Bezugsrahmen bildet. Psychologische Testungen werden demzufolge in zwei Hauptabschnitte unterteilt (► Abb. 10.1):

- Der erste Hauptabschnitt (*Testdurchführung*), wird in eine Vorbereitungs- (s. ► Abschn. 10.4.2) und in eine Testphase untergliedert, in der die eigentliche Testung stattfindet (s. ► Abschn. 10.4.3).
- Der zweite Hauptabschnitt (*Testauswertungsphase*) wird in die Bereiche Testergebnisse, Interpretation und Sicherung gegliedert (► Abschn. 10.4.4).

Phasenmodell psychologischer Testungen

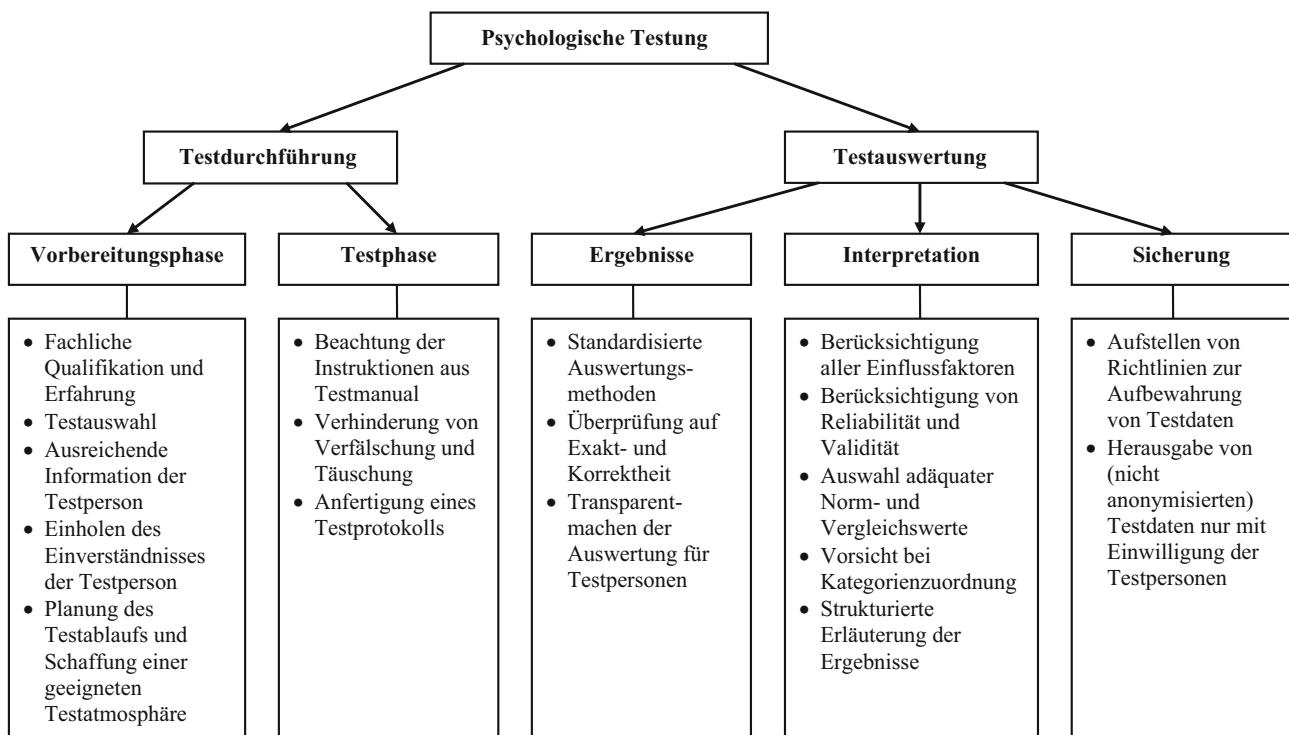


Abb. 10.1 Phasenmodell für die Durchführung und Auswertung psychologischer Testungen mit zugehörigen Standards (vgl. Moosbrugger und Höfling 2006, mit freundlicher Genehmigung von Hogrefe)

Anhand dieser Gliederung werden im Folgenden wesentliche Standards zur Testanwendung vorgestellt.

10.4.2 Testvorbereitungsphase

Vor der Testanwendung sollten die individuellen Testanwender zunächst verifizieren, ob sie bezüglich der Fragestellung, der durchzuführenden Tests (einschließlich der ihnen zugrunde liegenden Theorien und Konzepte), der technischen Handhabung und der zu testenden Personengruppen über ausreichende fachliche Qualifikation und Erfahrung verfügen (► Kritisch nachgefragt 10.1). Andernfalls erscheint eine Ablehnung bzw. Weiterleitung des Auftrags an kompetentere Testanwender angemessen. Verschiedentlich wird darauf hingewiesen, dass Probleme im Bereich der Testanwendung in nicht zu vernachlässigender Weise mit einer mangelhaften Qualifikation von individuellen Testanwendern assoziiert sind (vgl. Turner et al. 2001). Hinzuweisen ist in diesem Zusammenhang auch auf die Tatsache, dass bei der Auswahl von geeignetem Testmaterial Urheber- und Lizenzrechte zu beachten sind. Dies soll u. a. dazu beitragen, dass ausschließlich zur Testdurchführung autorisierte und qualifizierte Personen psychologische Tests anwenden.

Prüfung der Qualifikation von Testanwendern

Nur adäquate Tests einsetzen

Bei der Testauswahl sollte Hypothesengeleitet vorgegangen werden, d. h., zur Überprüfung apriorisch aufgestellter Hypothesen sind ausschließlich adäquate Tests auszuwählen, die sich durch gute psychometrische Eigenschaften, ausführliche und plausible theoretische Konzeptualisierung und stets aktuelle und repräsentative Normdaten auszeichnen (► Beispiel 10.6).

Kritisch nachgefragt 10.1

Wer ist zur Anwendung psychologischer Tests autorisiert?

Der Einsatz nicht qualifizierter Testanwender (ohne Master bzw. Diplom in Psychologie, sog. „testing technicians“) im Rahmen psychologisch-diagnostischer Urteils- und Entscheidungsprozesse gilt als umstritten. Hall et al. (2005) kommen in ihrem Review zu dem Schluss, dass folgende Argumente letztlich gegen den Einsatz von „testing technicians“ sprechen:

- Mangelnde Kenntnisse und Fähigkeiten in Bezug auf die theoretische und methodische Konzeption und Anwendung psychologischer Tests
- Mangelnde Professionalität im Kontext diagnostischer Urteils- und Entscheidungsfindung
- Unzulässige Veränderung standardisierter Testbedingungen, da in der Testentwicklungsphase die Tests in der Regel nur von lizenzierten Fachkräften durchgeführt würden
- Empirische Hinweise auf geringere Reliabilität und Validität von Testdaten, wenn sie von „testing technicians“ erhoben werden
- Mangelnde Fähigkeit zum Aufbau testförderlichen Verhaltens bei den Testpersonen
- Mangelnde Fähigkeit zur Verarbeitung qualitativer Testinformationen (Verhalten der Testpersonen während der Testung)

Gemäß der DIN 33430 (DIN 2002, 2016) sind nur *autorisierte Inhaber von DIN-Lizenzen* zur Testdurchführung berechtigt. Die DIN-Lizenzen sind an den Nachweis einschlägiger Kenntnisse gebunden und können im Rahmen einer Prüfung gemäß der „Fortsbildungs- und Prüfungsordnung der Föderation Deutscher Psychologenvereinigungen zur Personenlizenzierung für berufsbezogene Eignungsdiagnostik nach DIN 33430“ (Föderation Deutscher Psychologenvereinigun-

gen 2017) erworben werden. Die für die Lizenz erforderlichen Kenntnisse und Fertigkeiten sind in sechs Module gegliedert, deren Inhalte den sechs Kapiteln in *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (Westhoff et al. 2010) folgen:

- Modul 1: Einführung in die DIN 33430
- Modul 2: Anforderungsprofil, Verhaltensbeobachtung und Verhaltensbeurteilung
- Modul 3: Eignungsinterviews, direkte mündliche Befragungen
- Modul 4: Anforderungsanalyse, Verfahren der Eignungsbeurteilung sowie rechtliche Rahmenbedingungen
- Modul 5: Statistisch-methodische Grundlagen der Eignungsbeurteilung
- Modul 6: Evaluation der Eignungsbeurteilung

Für den Erwerb der *Lizenz A für Auftragnehmer* von Eignungsbeurteilungen sind alle sechs Module Gegenstand der Modulprüfung. Anstelle der Volllizenz kann aber auch eine *Lizenz MV zur Mitwirkung an Verhaltensbeobachtungen* bzw. eine *Lizenz ME zur Mitwirkung an Eignungsinterviews* erworben werden. Für die Lizenz MV sind die Module 1 und 2, für die Lizenz ME die Module 1, 2 und 3 Gegenstand der Modulprüfung. Inhaber der Lizenzen MV und ME können unter Anleitung, Fachaufsicht und Verantwortung von Lizenz-A-Inhabern an Eignungsbeurteilungen mitwirken.

Die Inhalte der Modulprüfungen können im Selbststudium oder in Seminaren erarbeitet werden, die von der Deutschen Psychologen Akademie (DPA) veranstaltet werden. Die DPA (2005) führt auch ein aktuelles Register der Lizenzinhaber.

Im Rahmen des Psychologiestudiums können die Lizenzprüfungen ebenfalls direkt an Universitäten abgelegt werden (s. Dormann et al. 2009).

Beispiel 10.6: Standards zur Auswahl angemessener Tests

(ITC-G-TU; ITC 2001, S. 15)

Standard 2.2.2: „Fachkompetente Testanwender 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;
- 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 vorgesehenen Testpersonengruppen;
- e. Annehmbarkeit für die an der Testanwendung Beteiligten, unter anderem die von diesen wahrgenommene Fairness und Bedeutsamkeit;
- f. Praktikabilität, unter anderem hinsichtlich des notwendigen Zeit-, Kosten- und anderen Ressourcenaufwands.“

Exkurs 10.2**Die „ITC Computer-Based and Internet Delivered Testing Guidelines“ (ITC-G-CB; ITC, 2005a)**— **Zielsetzung:**

Entwicklung und Publikation international anerkannter Richtlinien zur Verbesserung von computerbasierten und internetgestützten psychologischen Testungen

— **Zielgruppen:**

Testentwickler, Testherausgeber und Testanwender

— **Inhalte:**

1. Berücksichtigung technologischer Aspekte:

- Erfordernisse in Bezug auf Hardware und Software
- Technische Robustheit gegenüber Systemfehlern oder Störungen
- Angemessene optische und akustische Darstellungsweise
- Hilfestellungen für Testpersonen mit Behinderungen
- Ausreichende technische Unterstützung und Information

2. Qualifizierte Testanwendung:

- Ausreichende Informationen zu theoretischen Konzeptualisierungen und kompetenter Testanwendung
- Berücksichtigung der psychometrischen Qualität des jeweiligen Tests
- Evaluation der Äquivalenz im Falle der Übertragung von Tests aus Pa-pierversionen
- Standardisierte Auswertungsmethoden
- Angemessene Interpretation und Darstellung der Ergebnisse
- Gewährleistung von Testfairness

3. Gewährleistung ausreichender Kontrollebenen:

- Ausführliche und detaillierte Darstellung des Testablaufs unter Berück-sichtigung aller technischen Erfordernisse
- Dokumentation des jeweiligen Ausmaßes von Überwachung der Tes-tung
- Exakte Authentifizierung der Testpersonen

4. Gewährleistung von Sicherheit und Privatsphäre bei Datenübertragungen:

- Sicherung des Testmaterials
- Sicherung der Testdaten, die via Internet transferiert werden
- Gewährleistung der Vertraulichkeit von Testdaten

Zur Vorbereitungsphase gehört weiterhin eine ausreichende Aufklärung und Information der Testpersonen bzw. ihrer gesetzlichen Vertreter über die Ziele der psychologischen Testung. Im Regelfall ist bei der Testperson die Zustimmung zum Test einzuholen, wobei die Notwendigkeit einer Zustimmung unter gewissen Umständen entfallen kann (z. B. bei Anordnung durch gesetzliche Regelung, landesweite Testprogramme bzw. Berufseignungstests).

Für objektive, reliable und valide Testdaten entscheidend sind schließlich auch die sorgfältige Planung des Ablaufs der psychologischen Testung und die optimale Gestaltung der äußeren Bedingungen für die eigentliche Testphase. Hierbei sollten mögliche Störquellen (z. B. Mobiltelefone) ausgeschaltet, eine angenehme Umgebung (z. B. ausreichendes Licht, angemessene Temperatur) geschaffen und gut leserliches und verständliches Testmaterial vorgelegt werden. Von Bedeutung sind auch die besonderen Richtlinien für computerbasierte Testungen und für Test-administrationen via Internet (► Exkurs 10.2).

10.4.3 Testphase

Die eigentliche Testung erfolgt in der Testphase. Hierbei gilt der Grundsatz, dass Testanwender/Testleiter zu Beginn und während des gesamten Verlaufs der psychologischen Testung mit ihrem Verhalten zur Verbesserung der Motivation der Testpersonen und zur Reduktion der Testängstlichkeit beitragen sollten.

Zentral für die Testphase ist weiterhin die genaue Beachtung der Instruktion, die im Testmanual vorgegeben ist, die exakte Einhaltung der Bearbeitungszeit und das Vorliegen des entsprechenden Testmaterials. Abweichungen von der Instruktion sind im Testprotokoll zu vermerken und hinsichtlich etwaiger Beeinflussungen der Validität der Testdaten zu diskutieren (► Beispiel 10.7). Besondere ethische Verantwortung kommt Testanwendern in den Fällen zu, in denen Testpersonen getestet werden, die aufgrund spezifischer Besonderheiten nicht mit der typischen Zielpopulation übereinstimmen. Hierzu gehören Testpersonen mit körperlicher oder geistiger Behinderung bzw. Testpersonen, deren Muttersprache nicht Testsprache ist. Ihnen gegenüber haben Testanwender besonders auf die Wahrung der Testfairness zu achten.

Der Vollständigkeit halber sei noch darauf hingewiesen, dass die Identität der Testpersonen (insbesondere im Rahmen von Gruppentestungen) zweifelsfrei gesichert sein muss und dass Testanwender während der Testphase dafür zu sorgen haben, dass eine Verfälschung oder Täuschung vermieden werden.

Grundsätze der Testphase

Beispiel 10.7: Standards zur Testdurchführung laut SPPT

(Häcker et al. 1998, S. 94)

Standard 15.1: „Bei der Vorgabe gängiger Tests sollte der Testleiter den vom Testherausgeber und -verleger spezifizierten standardisierten Verfahren der Testdurchführung und -auswertung gewissenhaft Folge leisten. Die Ausführungen bezüglich der Instruktionen für die Probanden, der Bearbeitungszeiten, der Form der Itemvorgabe oder -antwort und des Testmaterials oder -zubehörs sollten genauestens beachtet werden. Ausnahmen sollten nur auf der Basis sorgfältiger fachlicher Beurteilung gemacht werden, vornehmlich im klinischen Bereich.“

10.4.4 Testauswertungsphase

In der Phase der Testauswertung sind zum Zwecke größtmöglicher Exakt- und Korrektheit von Testergebnissen standardisierte Auswertungsmethoden anzuwenden, was entsprechende statistisch-methodische Kenntnisse erfordert. Testanwender sollten Test- und Subtestwerte auf unwahrscheinliche bzw. unsinnige Werte überprüfen, die Verwendung verschiedener (z. B. grafischer) Darstellungsformen beherrschen und den Testpersonen das Zustandekommen der Testergebnisse und deren Bedeutung für etwaige psychologisch-diagnostische Entscheidungen transparent machen.

Überprüfung der Testergebnisse

Bei der Interpretation von Testergebnissen ist auf Angemessenheit zu achten, d.h., Testanwender sollten Testergebnisse nicht überbewerten und alle zur Verfügung stehenden Informationsquellen über die Testpersonen (z. B. Alter, Geschlecht, Bildungsniveau, kulturelle Faktoren, Vorerfahrungen mit bestimmten psychologischen Tests) einbeziehen. Weiterhin müssen bei der Interpretation von Testergebnissen die Reliabilität (z. B. durch die Bildung von Vertrauensintervallen) und Validität der Test- oder Subtestwerte angemessen berücksichtigt werden.

Sorgfältige Interpretation von Testergebnissen

Kommen Norm- oder Vergleichswerte zum Einsatz, ist darauf zu achten, dass diese aktuell und für die jeweiligen Testpersonen relevant sind. Im Kontext des kriteriumsorientierten Testens ist zu beachten, dass für die verwendeten *kritischen Trennwerte (Cut-off-Scores)* Validitätsbelege vorzulegen sind (► Kap. 9). Etwaiige Erläuterungen und Empfehlungen sollen entweder in mündlicher Form oder schriftlich in Form eines Testberichts den Testpersonen konstruktiv, sprachlich angemessen und inhaltlich verständlich übermittelt werden (► Beispiel 10.8).

Beispiel 10.8: Standards zum Schutz von Probanden laut SPPT

(Häcker et al. 1998, S. 98)

Standard 16.6: „Werden Personen aufgrund ihrer Testwerte Kategorien zugeordnet, sollten diese auf der Grundlage sorgfältig ausgewählter Kriterien bestimmt werden. In Übereinstimmung mit einer präzisen Berichterstellung sollten immer die am wenigsten stigmatisierenden Kategorien gewählt werden.“

Sicherung von Testergebnissen

10

Bedarf nach vereinheitlichten Qualitätsbeurteilungssystemen

Alle relevanten Testdaten einer Person einschließlich des Testprotokolls und alle schriftlichen Belege müssen gesichert und aufbewahrt werden. Testanwender sollten klare Richtlinien über die Verfügbarkeit, Aufbewahrungsdauer und weitere Verwendung der Testdaten erstellen. Testergebnisse, die im Zusammenhang mit computerbasierten und internetgestützten Testungen in Datenbanken gespeichert sind, müssen hinreichend geschützt werden (vgl. British Psychological Society 2002). Testdaten, die einer Person namentlich zugeordnet werden können, dürfen nur nach vorheriger Einwilligung der Testperson bzw. ihres gesetzlichen Vertreters anderen Personen oder Forschungseinrichtungen zugänglich gemacht werden.

10.5 Standards für die Qualitätsbeurteilung psychologischer Tests

10.5.1 Überblickswerke

Da die Zahl der auf dem Markt befindlichen psychologischen Tests und ihr Einsatz stets zunehmen, ist der Testanwender auf übersichtliche Informationen angewiesen. Hierzu stehen mehrere Überblickswerke zur Verfügung (► Exkurs 10.3). Neben den Übersichtswerken besteht aber auch vermehrter Bedarf nach Qualitätsbeurteilungen. Die Rezensionen psychologischer Tests erfolgten in Deutschland häufig weitgehend frei und unstandardisiert (vgl. Kersting 2006), wobei ein Vorteil unstandardisierter Testrezensionen in der großen Gestaltungsfreiheit bezüglich der auf jeden einzelnen Test individuell zugeschnittenen Kriterien gesehen werden könnte. Diesem Vorteil steht allerdings ein entscheidender Mangel gegenüber, da das Fehlen eines verbindlichen Kriteriensystems bei der Testbeurteilung dazu führt, dass die Tests untereinander nur schwierig zu vergleichen sind.

10.5.2 Testbeurteilungssystem des Testkuratoriums (TBS-TK)

Um diesem Mangel abzuheften, hat das Testkuratorium der Föderation Deutscher Psychologenvereinigungen unter Beachtung internationaler Testrezensionssysteme (Commissie Testaangelegenheden Nederland, COTAN; Evers 2001; European Federation of Psychologists' Associations, EFPA; Bartram 2001) und unter Berücksichtigung der DIN 33430 das für Deutschland maßgebliche „Testbeur-

Exkurs 10.3**Übersichtswerke für psychologische Tests in Deutschland**

Sonderheft der <i>Zeitschrift für Differentielle und Diagnostische Psychologie</i> (Kubinger 1997)	Testrezensionen von 25 gängigen Tests und eine Übersicht über bis dato erschienene Testrezensionen
<i>Brickenkamp Handbuch psychologischer und pädagogischer Tests</i> (Brähler et al. 2002)	Die in Kurzbeiträgen vorgestellten Tests sind in drei Hauptgruppen unterteilt: Leistungstests, psychometrische Persönlichkeitstests und Persönlichkeitsentwicklungsverfahren.
<i>Tests unter der Lupe</i> (Fay 1996, 1999, 2000, 2003, 2005)	Ausführliche Rezensionen aktueller psychologischer Tests
<i>Diagnostische Verfahren in der Psychotherapie</i> (Brähler et al. 2003)	Beschreibung von 94 Testverfahren für die psychotherapeutische Forschung und Praxis
<i>Handbuch personaldiagnostischer Instrumente</i> (Kanning und Holling 2002)	Vorstellung und Beurteilung von 50 personaldiagnostischen Testverfahren
<i>Handbuch wirtschaftspsychologischer Testverfahren</i> (Sarges und Wottawa 2005)	Zusammenstellung und Beschreibung von 140 wirtschaftspsychologischen Testverfahren

teilungssystem des Testkuratoriums“ (TBS-TK; Testkuratorium 2009, 2010)

entwickelt, das eine Reihe von Vorteilen bietet:

Vorteile des TBS-TK

- Höhere Transparenz und Objektivität
- Standardisierte Bewertung aufgrund vorgegebener Beurteilungskriterien
- Größere Vollständigkeit in Bezug auf relevante Aspekte
- Testübergreifende Vergleichsmöglichkeiten verschiedener Verfahren

Um die Qualität der Testbeurteilung sicherzustellen, beauftragt das Testkuratorium jeweils zwei fachkundige Rezessenten. Diese prüfen in *Schritt 1* zunächst anhand der „DIN Screen-Checkliste 1“ (Kersting 2008) das Testmanual bzw. Testhandbuch („die Verfahrenshinweise“) auf grundsätzliche Erfüllung der Anforderungen gemäß DIN 33430. Bei Erfüllung der Anforderungen wird der Test als prüffähig beurteilt; in *Schritt 2* erfolgt eine Testkategorisierung nach formalen Merkmalen und Inhalten gemäß ZPID (Zentrum für psychologische Information und Dokumentation) und in Ausschnitten gemäß EFPA. Die eigentliche Bewertung des Tests erfolgt schließlich in *Schritt 3* anhand der Besprechungs- und Beurteilungskategorien des Testkuratoriums. Die gemeinsame Abschlussbewertung bzw. Empfehlung der Rezessenten („Der Test erfüllt die Anforderungen voll, weitgehend, teilweise bzw. nicht“) erfolgt als Würdigung der Gesamtheit aller Aspekte. Vor Veröffentlichung wird die jeweilige Rezession den Testautoren zur Stellungnahme zugesandt. Das komplette TBS-TK ist in folgender Übersicht dargestellt.

Testbeurteilungssystem des Testkuratoriums (TBS-TK)

Das TBS-TK der Föderation Deutscher Psychologenvereinigungen (Testkuratorium 2009, 2010) dient zur Qualitätssicherung psychologischer Tests. Hierbei wird ein Beurteilungsprozess vorgenommen, in dem Testrezessenten in Bezug auf einen psychologischen Test in drei Schritten

- die Verfahrenshinweise auf grundsätzliche Erfüllung der in der DIN 33430 formulierten Anforderungen prüfen, und wenn ja,
- eine Testkategorisierung nach ZPID (► <https://www.zpid.de>) und Merkmalen aus EFPA (► <http://www.efpa.be>) vornehmen und
- den Test anhand der Besprechungs- und Beurteilungskategorien des Testkuratoriums bewerten.

Die *Besprechungs- und Beurteilungskategorien des Testkuratoriums* sind folgende:

1. Allgemeine Informationen über den Test, Beschreibung des Tests und seiner diagnostischen Zielsetzung
2. Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion
3. Objektivität
4. Normierung (Eichung)
5. Zuverlässigkeit (Reliabilität, Messgenauigkeit)
6. Gültigkeit (Validität)
7. Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)
8. Abschlussbewertung/Empfehlung

Das TBS-TK hat sich von Anfang an (s. Moosbrugger et al. 2008) sehr gut bewährt. Die Rezensionen werden in *Report Psychologie* und *Psychologische Rundschau* veröffentlicht, wobei in den besonders relevanten Besprechungs- und Beurteilungskategorien 1, 3, 5 und 6 standardisierte Bewertungen nach einem vorgegebenen Schema vorgenommen werden. □ Tab. 10.2 zeigt exemplarisch die standardisierte Bewertung gemäß TBS-TK (Schmidt-Atzert und Rauch 2008) für den „Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)“ von Liepmann et al. (2007). Allerdings schneiden keineswegs alle Verfahren in ähnlich positiver Weise ab. □ Tab. 10.3 zeigt als Beispiel die deutlich negativeren Bewertung (Baumgärtel und Thomas-Langel 2014) für das projektive Verfahren „Familie in Tieren“ von Brehm-Gräser (2011).

Alle bisher veröffentlichten Testrezensionen gemäß TBS-TK stehen auf der Homepage ► <https://www.zpid.de/index.php?wahl=Testkuratorium> des Leibniz-Zentrums für Psychologische Information und Dokumentation an der Universität Trier zur Verfügung.

□ **Tabelle 10.2** Testrezension gemäß TBS-TK zum I-S-T 2000 R (Liepmann et al. 2007) von Schmidt-Atzert und Rauch (2008)

I-S-T 2000 R	Die TBS-TK-Anforderungen sind erfüllt			
	voll	weitgehend	teilweise	nicht
Allgemeine Informationen, Beschreibung und diagnostische Zielsetzung	●			
Objektivität	●			
Zuverlässigkeit		●		
Validität	●			

□ **Tabelle 10.3** Testrezension gemäß TBS-TK zu Familie in Tieren (Brehm-Gräser 2011) von Baumgärtel und Thomas-Langel (2014)

Familie in Tieren	Die TBS-TK-Anforderungen sind erfüllt			
	voll	weitgehend	teilweise	nicht
Allgemeine Informationen, Beschreibung und diagnostische Zielsetzung			●	
Objektivität				●
Zuverlässigkeit				●
Validität				●

10.6 Zusammenfassung

Teststandards sind vereinheitlichte Leitlinien, in denen sich allgemein anerkannte Zielsetzungen zur Entwicklung und Evaluation (Testkonstruktion), Übersetzung und Anpassung (Testadaptation) sowie Durchführung, Auswertung und Interpretation (Testanwendung) psychologischer Tests widerspiegeln. Verschiedene nationale und internationale Teststandardkompendien haben mit unterschiedlicher Schwerpunktsetzung solche Teststandards zusammengetragen (SEPT, DIN 33430, ITC-G-TA, ITC-G-TU, ITC-G-CB).

Die Überprüfung der Einhaltung der Standards bei der Testentwicklung und -evaluation (Qualitätsbeurteilung psychologischer Tests) erfolgt in Deutschland unter Berücksichtigung der DIN 33430 mit dem TBS-TK, das die standardisierte Erstellung und Publikation von Testrezensionen anhand eines vorgegebenen Kriterienkatalogs vorsieht. Um die Standards bei der Testanwendung sicherzustellen, wurden vom Testkuratorium im Auftrag der Föderation Deutscher Psychologenvereinigungen Personenlizenzierungen nach DIN 33430 eingeführt.

10.7 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Nennen und erläutern Sie kurz, für welche Aspekte innerhalb der Testentwicklung und -evaluation Teststandards beachtet werden sollen.
2. Welche wesentlichen Standards existieren gemäß der SEPT und der DIN 33430 für die Validität eines Tests?
3. Beschreiben Sie kurz die Richtlinien in den vier Sektionen der Test-Adaption Guidelines (TAG).
4. Worauf sollte innerhalb der Testauswertung beim Ermitteln der Ergebnisse geachtet werden?
5. Welche Qualifikationen sollten Testanwender nach Möglichkeit aufweisen?

Literatur

- Achenbach, T. M. (1991). *Manual for the Youth Self Report and 1991 Profile*. Burlington: University of Vermont, Department of Psychiatry.
- 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: AERA.
- Arbeitsgruppe Deutsche Child Behavior Checklist (1998). *Fragebogen für Jugendliche; deutsche Bearbeitung der Youth Self-Report Form der Child Behavior Checklist (YSR). Einführung und Anleitung zur Handauswertung*. (2. Aufl., mit deutschen Normen, bearbeitet von M. Döpfner, J. Plück, S. Bölte, P. Melchers & K. Heim). Köln: Arbeitsgruppe Kinder-, Jugend- und Familiendiagnostik (KJFD).
- Bartram, D. (2001). Guidelines for test users: A review of national and international initiatives. *European Journal of Psychological Assessment*, 17, 173–186.
- Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (2000). *TIMSS III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Opladen: Leske + Budrich.
- Baumgärtel, F. & Thomas-Langel, R. (2014). TBS-TK Rezension: „Familie in Tieren“. *Psychologische Rundschau*, 66, 152–154.
- Brähler, E., Holling, H., Leutner, D. & Petermann, F. (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests*. Göttingen: Hogrefe.
- Brähler, E., Schumacher, J. & Strauß, B. (Hrsg.). (2003). *Diagnostische Verfahren in der Psychotherapie* (Bd. 1, 2. Aufl.). Göttingen: Hogrefe.
- Brehm-Gräser, L. (2011). *Familie in Tieren. Die Familiensituation im Spiegel der Kinderzeichnung*. München: Reinhardt.

- British Psychological Society (2002). *Guidelines for the Development and Use of Computer-Based Assessment*. Leicester, UK: Psychological Testing Centre.
- Deutsche Psychologen Akademie (DPA). (2005). *Inhaber(innen) Lizenz A für berufsbezogene Eignungsbeurteilungen nach DIN 33430*. Verfügbar unter <https://www.din33430portal.de/> [20.12.2019]
- Deutsches Institut für Normung e.V. (DIN). (2002). *DIN 33430:2002-06. Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Deutsches Institut für Normung e.V. (DIN). (2016). *DIN 33430:2016-07: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.
- Diagnostik- und Testkuratorium (Hrsg.) (2018). *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430*. Berlin, Heidelberg: Springer.
- Dormann, C., Moosbrugger, H., Stemmler, G. & Maier, G. A. (2009). Erwerb von Personenlizenzen zur DIN 33430 im Rahmen des Psychologiestudiums. Ein Modellversuch an der Johann Gutenberg-Universität Mainz. *Psychologische Rundschau*, 60, 23–27.
- Evers, A. (2001). The Revised Dutch Rating System for Test Quality. *International Journal of Testing*, 1, 155–182.
- Fay, E. (1996). *Tests unter der Lupe, Band 1*. Heidelberg: Asanger.
- Fay, E. (1999). *Tests unter der Lupe, Band 2*. Lengerich: Pabst.
- Fay, E. (2000). *Tests unter der Lupe, Band 3*. Lengerich: Pabst.
- Fay, E. (2003). *Tests unter der Lupe, Band 4*. Göttingen: Vandenhoeck & Ruprecht.
- Fay, E. (2005). *Tests unter der Lupe, Band 5*. Göttingen: Vandenhoeck & Ruprecht.
- Föderation Deutscher Psychologenvereinigungen. (2017). Fortbildungs- und Prüfungsordnung der Föderation Deutscher Psychologenvereinigungen zur Personenlizenzierung für berufsbezogene Eignungsdiagnostik nach DIN 33430. Verfügbar unter <https://www.din33430portal.de/din33430/din33430> [20.01.2020]
- Häcker, H., Leutner, D. & Amelang, M. (1998). Standards für pädagogisches und psychologisches Testen. *Zeitschrift für Differentielle und Diagnostische Psychologie, Supplementum*. Göttingen: Hogrefe.
- Haider, G. & Reiter, C. (Hrsg.). (2004). *PISA 2003. Nationaler Bericht. Mathematik, Lese-Kompetenz, Naturwissenschaft, Problemlösen*. Graz: Leycam.
- Hall, J. D., Howerton, D. L. & Bolin, A. U. (2005). The use of testing technicians: critical issues for professional psychology. *International Journal of Testing*, 5, 357–375.
- Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17, 164–172.
- International Test Commission (ITC). (2001). *International Guidelines on Test Use*. Retrieved from <https://www.intestcom.org/> [20.12.2019]
- International Test Commission (ITC). (2005a). *International Guidelines on Computer-Based and Internet Delivered Testing*. Retrieved from <https://www.intestcom.org/> [20.12.2019]
- International Test Commission (ITC). (2005b). *International Guidelines on Test Adaptation*. Retrieved from <https://www.intestcom.org/> [20.12.2019]
- International Test Commission (ITC). (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. Retrieved from <https://www.intestcom.org/> [20.12.2019]
- Kanning, U. P. & Holling, H. (Hrsg.). (2002). *Handbuch personaldiagnostischer Instrumente*. Göttingen: Hogrefe.
- Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau*, 57, 243–253.
- Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. In M. Kersting (Hrsg.), *Qualität in der Diagnostik und Personalauswahl – der DIN Ansatz* (S.141–210). Göttingen: Hogrefe.
- Kersting, M. (2016). DIN 33430 reloaded. Mit Qualität die Zukunft der Personalauswahl gestalten. *Report Psychologie*, 41, 291–295.
- Kubinger, K. D. (1997). Editorial zum Themenheft „Testrezensionen: 25 einschlägige Verfahren“. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 18, 13.
- Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID). (2001). *Internationale Richtlinien für die Testanwendung, Version 2000. Erstellt in Zusammenarbeit mit der International Test Commission*. Verfügbar unter https://www.psystdex.de/pub/tests/itc_richtlinien.pdf [20.12.2019]
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)* (2. Aufl.). Göttingen: Hogrefe.
- Moosbrugger, H. & Höfling, V. (2006). Teststandards. In F. Petermann & M. Eid (Hrsg.), *Handbuch der psychologischen Diagnostik* (S. 407–419). Göttingen: Hogrefe.
- Moosbrugger, H., Stemmler, G. & Kersting, M. (2008). Qualitätssicherung und -optimierung im Aufbruch. Die ersten Testrezensionen nach dem neuen TBS-TK System. *Psychologische Rundschau*, 59, 182–184.
- Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Literatur

- Reimann, G. (2009). *Moderne Eignungsbeurteilung mit der DIN 33430*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Roth, M. (2000). Überprüfung des Youth Self-Report an einer nichtklinischen Stichprobe. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 21, 105–110.
- Sarges, W. & Wottawa, H. (2005). *Handbuch wirtschaftspsychologischer Testverfahren*. (2. Aufl.). Lengerich: Pabst Science Publishers.
- Schmidt-Atzert, L. & Rauch, W. (2008). Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). Testrezensionen nach dem TBS-TK-System. *Report Psychologie*, 33, 303–304.
- Testkuratorium (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Report Psychologie*, 34, 470–478.
- Testkuratorium (2010). TBS-TK – Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau*, 61, 52–56.
- Turner, S. M., DeMers, S. T., Fox, H. R. & Reed, G. M. (2001). APA's Guidelines for Test User Qualifications. An Executive Summary. *American Psychologist*, 56, 1099–1113.
- Van de Vijver, F. J. R. & Watkins, D. (2006). Assessing similarity of meaning at the individual and country level. *European Journal of Psychological Assessment*, 22, 69–77.
- Westhoff, K., Horne, L. F. & Westmeyer, H. (2003). Richtlinien für den diagnostischen Prozess. *Report Psychologie*, 28, 504–517.
- Westhoff, K., Hagemeyer, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G. & Stemmler, G. (Hrsg.). (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Aufl.). Lengerich: Pabst.



Standards für pädagogisches Testen

Sebastian Brückner, Olga Zlatkin-Troitschanskaia und Hans Anand Pant

Inhaltsverzeichnis

- 11.1 Die „Standards for Educational and Psychological Testing“ im Überblick – 219**
- 11.2 Domänen, Ziele und Designs pädagogischen Testens – 220**
 - 11.2.1 Domänen pädagogischen Testens – 220
 - 11.2.2 Ziele pädagogischen Testens – 221
 - 11.2.3 Designs pädagogischen Testens – 225
- 11.3 Validitätsstandards und pädagogisches Testen (Standards 1.0–1.25) – 229**
- 11.4 Standards zur Reliabilität (Standards 2.10–2.20) – 234**
- 11.5 Schwellenwerte und ihre Bedeutung für die Testwertinterpretation – 234**
 - 11.5.1 Standards zur Definition von Schwellenwerten (Standards 5.21–5.23) – 235
 - 11.5.2 Standardsettings zur Bestimmung von Schwellenwerten – 235
- 11.6 Weitere Implikationen der Standards für pädagogisches Testen – 238**
 - 11.6.1 Fairness – 239
 - 11.6.2 Transparenz des Untersuchungs- und Interpretationsgegenstands (Konstrukte, Anforderungen und Inhalte) – 240
 - 11.6.3 Variation der Prüfungsformen – 240
 - 11.6.4 Feedback – 241
- 11.7 Standards zum Management und zur Archivierung von Daten pädagogischen Testens – 242**
- 11.8 Standards für Forschungsethik – 244**

11.9 Zusammenfassung – 245

11.10 Kontrollfragen – 245

Literatur – 245

i Die „Standards for Educational and Psychological Testing“ bieten eine umfangreiche Darstellung von über 240 Standards, die zur Entwicklung, Durchführung und Evaluation pädagogischer und psychologischer Tests praktische Handlungsempfehlungen geben. Zur Berücksichtigung der Besonderheiten der Zielsetzung und Entwicklung pädagogischer Tests können die Standards insbesondere zu Fragen der Validität einen Beitrag leisten. Weitere Implikationen aus den Standards für Anforderungen pädagogischen Testens lassen sich u. a. zum Standardsetting, zur Fairness, zur Transparenz des Untersuchungsgegenstands und Interpretation, zu Formen der Diagnostik, zum Feedback sowie zum Datenmanagement finden. Zusätzlich zu den Standards gibt die American Educational Research Association (AERA) mit dem „Code of Ethics“ Richtlinien zu Fragen der Forschungsethik heraus, denen auch beim pädagogischen Testen aktuell eine immer größere Bedeutung zukommt.

11.1 Die „Standards for Educational and Psychological Testing“ im Überblick

Die „Standards for Educational and Psychological Testing“¹ der drei amerikanischen erziehungswissenschaftlich und psychologisch orientierten Organisationen: *American Educational Research Association* (AERA), *American Psychological Association* (APA) und *National Council on Measurement in Education* (NCME), welche die Testentwicklung und -nutzung übergreifend an einheitlichen Kriterien ausrichten, werden seit dem Jahr 1954 veröffentlicht. Sie sind im Jahr 2014 in der siebten Version als eines von mehreren Kompendien (zu den Standards für psychologisches Testen vgl. ► Kap. 10) erschienen. Bereits in der ersten Version der Standards aus dem Jahr 1954 zeigte sich ein erster übergreifender Systematisierungsansatz der Testentwicklung. Als Ziel der Standards wird in der aktuellen Ausgabe von 2014 beschrieben:

» The purpose of the Standards is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses. (AERA et al. 2014, S. 1)

Ziel der „Standards“ ist es demzufolge, Kriterien und Maßnahmen für die Konzeptionierung und Konstruktion (vgl. ► Kap. 2, 3, 4, 5 und 6), den Einsatz und die Evaluation von Tests unter Berücksichtigung grundlegender Gütekriterien wie der Objektivität (vgl. ► Kap. 2), der Reliabilität (vgl. ► Kap. 14 und 15) und insbesondere der Validität (vgl. ► Kap. 21) bereitzustellen.

Die aktuelle Version der Standards umfasst 13 sog. „Chapters“, die den folgenden drei inhaltlichen Bereichen zugeordnet sind sowie insgesamt 249 Standards enthält (in Klammern ist die englischsprachige Bezeichnung angegeben; ► Tab. 11.1; AERA et al. 2014):

1. „Grundlagen“ (Foundations)
2. „Merkmale der Testdurchführung“ (Operations)
3. „Testanwendungen“ (Testing Applications)

Ziele der „Standards for Educational and Psychological Testing“

Als grundlegendes Kompendium zur Qualitätsbeurteilung von Testverfahren beziehen sich die Standards auf verschiedene Aspekte der Testentwicklung und -bewertung und weisen Bezüge zu mehreren Kapiteln dieses Lehrbuchs auf (► Tab. 11.1). In diesem Kapitel werden exemplarische Bezüge einzelner Standards zum pädagogischen Testen herausgestellt, um zu verdeutlichen, wie Standards verwendet und interpretiert werden können, sodass sie auf pädagogische Fragestellungen anwendbar sind. Da sich insbesondere der Abschnitt der Grundlagen auf zentrale Themenbereiche dieses Lehrbuchs bezieht, wird diesen Standards im Folgenden

1 Im weiteren Verlauf dieses Textes wird nicht mehr von „Standards for Educational and Psychological Testing“ gesprochen, sondern der Lesbarkeit halber nur noch von den „Standards“.

Tabelle 11.1 Beziehungen zwischen den Chapters der Standards und den Kapiteln dieses Buches

Chapter der Standards	Bezeichnungen in den Chapters (Anzahl der Standards)	Inhaltliche Beziehungen in diesem Buch
1	Validität (26)	► Kap. 2, 10, 11, ► Kap. 21, 25, 27
2	Reliabilität und Messfehler (21)	► Kap. 2, 10, 13, 14, 15, 19, 26, 27
3	Testfairness (21)	► Kap. 2, 10, 11
4	Testdesign und Testentwicklung (26)	► Kap. 3, 10
5	Testwerte, Skalen, Normierung, Testwert-Linking und Schwellenwerte (24)	► Kap. 2, 7, 8, 9, 10, 13, 16
6	Testdurchführung, Testauswertung, Ergebnisdarstellung und Interpretation (17)	► Kap. 2, 7, 8, 9, 10, 16
7	Dokumentation (15)	► Kap. 2, 9, 10
8	Rechte und Verantwortlichkeiten von Testteilnehmern (13)	Nicht enthalten
9	Rechte und Verantwortlichkeiten von Testanwendern (24)	► Kap. 2, 10
10	Psychologisches Testen und Diagnostik (18)	► Kap. 9, 10, 21
11	Berufliches Testen und Zertifikation (16)	► Kap. 10, 21
12	Pädagogisches Testen und Diagnostik (19)	► Kap. 11, 21
13	Testanwendungen für Programmevaluationen, politische Studien und Rechenschaftsberichte (9)	Nicht enthalten

eine detaillierte Analyse zuteil. Zudem werden in Chapter 12 der Standards die Besonderheiten beim pädagogischen Testen als eine von mehreren Testanwendungen deutlich und sollen im Folgenden herausgestellt werden.

11.2 Domänen, Ziele und Designs pädagogischen Testens

11.2.1 Domänen pädagogischen Testens

Nicht nur für das psychologische Testen (► Kap. 10), sondern auch für die praktische Entwicklung und Nutzung von Tests, die sich an pädagogischen Fragestellungen orientieren, stellen die Standards eine hilfreiche Informationsquelle dar. Analog zu den Standards des psychologischen Testens enthalten die Standards des pädagogischen Testens keine verbindlichen Kriterien, eine schablonenartige Nutzung wird explizit nicht empfohlen:

» The standards should not be used as a checklist. (AERA et al. 2014, S. 2)

Geltungsrahmen der Standards für pädagogisches Testen

Vielmehr handelt es sich bei den Standards um grundsätzliche Empfehlungen und Richtlinien, zur Entwicklung und Anwendung von Tests, die in verschiedenen Inhaltsbereichen oder Domänen flexibel eingesetzt werden können. Dabei sind auch weitere rechtliche und ethische Richtlinien zu beachten. Als exemplarische Domänen können im schulfachlichen Kontext z. B. Mathematik oder Physik, im akademischen Kontext z. B. die Wirtschaftswissenschaften, Ingenieurwissenschaften oder die Lehrerbildung genannt werden.

Definition

Zum Begriff der **Domäne** finden sich in der einschlägigen Forschungsliteratur unterschiedliche Definitionen. Klieme et al. (2007) heben in Rekurs auf die Expertiseforschung die fachinhaltliche Spezifität eines Gegenstands- oder Leistungsbereichs hervor. Achtenhagen (2004, S. 22 ff.) hingegen geht bei der Bestimmung von Domänen in der Berufsbildung von einem „übergeordneten sinnstiftenden, thematischen Handlungskontext“ aus.

So unterschiedlich die Eingrenzungen von Domänen auch sein können, wird doch überwiegend ein relativ breites Verständnis angelegt, das ähnliche Inhalte und Themen in einem funktionalen und zielorientierten Verständnis zu voneinander abgrenzenden Inhaltbereichen zusammenfasst.

Während beim psychologischen Testen vor allem psychologisch relevante Konstrukte operationalisiert und Items auf dieser Basis entwickelt werden (z. B. fluide Intelligenz), werden beim pädagogischen Testen vor allem erlernbare, pädagogisch beeinflussbare Konstrukte in bestimmten inhaltlichen Domänen fokussiert und in Aufgaben operationalisiert (z. B. das erlernte Wissen im Fach Mathematik). Eine Domäne beschreibt damit den Geltungsbereich eines pädagogischen Tests und muss bei der Testentwicklung präzise spezifiziert werden (vgl. ► Kap. 3). Weiterführend ist für das pädagogische Testen auch von Interesse, inwiefern diese erlernbaren Konstrukte domänenspezifisch oder -übergreifend von Bedeutung sind. So könnte mathematisches Wissen domänenübergreifend auch hilfreich sein, um pädagogische Phänomene sozialwissenschaftlich/statistisch zu analysieren.

11.2.2 Ziele pädagogischen Testens

Die Ziele pädagogischen Testens differenzieren sich je nach der (pädagogischen) Intention der Testung und unterscheiden sich von Zielen psychologischen Testens dahingehend, dass Testwerte hinsichtlich pädagogisch und erziehungswissenschaftlich relevanter Fragestellungen auf *individuelle, institutionelle und systemische Kontexte einer Domäne* bezogen werden können.² Die Optimierung des Lernens bzw. von Lehr-Lern-Umwelten bleibt dabei stets das perspektivische Ziel des pädagogischen Testens. Je nach der verfolgten Zielstellung – z. B. Individualdiagnostik vs. Bildungsmonitoring – können mittels pädagogischen Testens verschiedene Informationen gewonnen werden, die eine empirische Grundlage für evidenzbasiertes Handeln bilden.

- Auf *individueller Ebene* werden Informationen über den Lernenden³ generiert (z. B. was und wie erfolgreich gelernt wurde) und – wie auch die Informationen auf institutioneller Ebene – verschiedenen Bezugsgruppen adressatengerecht zur Verfügung gestellt, damit pädagogische Implikationen für die Domäne abgeleitet werden können. Dies stellt einen wesentlichen Unterschied zum psychologischen Assessment dar, bei dem der domänenspezifische Bezug zu Lehr-Lern-Inhalten und instruktionalen Lehr-Lern-Angeboten in der Regel nicht gegeben ist, z. B. bei einem Intelligenztest (► Beispiel 11.1).
- Auf *institutioneller Ebene* können die Tests z. B. zum Zwecke der Qualitätskontrolle eingesetzt werden, um formelles Lernen in Schulen, Hochschulen oder anderen Bildungseinrichtungen zu analysieren und domänenspezifische Hinweise für die Optimierung der institutionellen Lehr-Lern-Angebote zu geben.

Differenzierte Zielsetzungen

Beschaffung von Informationen für evidenzbasiertes Handeln

² Die Vorgehensweisen beim Testen sind dabei ähnlich und werden in ► Tab. 11.2 strukturiert dargestellt.

³ Die in dem vorliegenden Text verwendeten Personenbezeichnungen sind grundsätzlich geschlechtsunspezifisch zu verstehen und umfassen alle Geschlechtsausprägungen. Die Verwendung der männlichen Form dient lediglich der leichteren Lesbarkeit.

Als Beispiel sei das „Bildungsmonitoring“ benannt, das auch auf der *systemischen Ebene* erfolgen kann (► Beispiel 11.2).

Beispiel 11.1: Intelligenz

Einige Modelle der Intelligenzforschung unterscheiden fluide und kristalline Intelligenz. Häufig wird dabei unterstellt, dass fluide Intelligenz biologisch bedingt oder angeboren und nicht durch Lerngelegenheiten veränderbar sei. So gehört insbesondere visuelle Wahrnehmungsfähigkeit, figurales Schlussfolgern und der Umgang mit neuartigen Informationen zu solchen Intelligenzfacetten: „Fluid ability represents novel or abstract problem solving capability and is believed to have a physiological basis. In contrast, crystallized ability [...] is associated with learned or acculturated knowledge“ (Postlethwaite 2011, S. IV). Beispielsweise kann eine Lehrerin erste Hinweise über die Ausprägung der kristallinen Intelligenz einzelner Schüler gewinnen, wenn die Schüler z. B. Aufgaben zum Allgemeinwissen korrekt oder falsch beantworten. Auf Basis dieser Individualdiagnostik kann die Lehrerin eine pädagogische Entscheidung treffen, wie der einzelne Schüler zu fördern ist (z. B. mit weiteren Übungsaufgaben).

Beispiel 11.2: Bildungsmonitoring

Das Bildungsmonitoring befasst sich mit Fragen der evidenzbasierten Steuerung von Bildungsinstitutionen und -systemen auf Basis systematisch generierter, umfassender Daten, die z. B. im Rahmen von Schulleistungstests erhoben werden können oder durch das statistische Bundesamt oder andere Institutionen, die Daten über Schulen und Hochschulen archivieren, zu diesem Zweck bereitgestellt werden. Sowohl die Bildungspolitik als auch die Bildungspraxis erhalten damit Informationen über die Leistungsfähigkeit des gesamten Bildungssystems und seiner Institutionen und können Veränderungen vornehmen wie die Anpassung von Bildungsstandards, Lehrplänen oder die Ressourcenallokation im Bildungssystem. Als wichtige Datenquelle für das Bildungsmonitoring sind exemplarisch die Ergebnisse international vergleichender Schulleistungsstudien („Large-Scale-Assessments“) wie das „Programme for International Student Assessment“ (PISA) zu nennen, die länderübergreifend die Leistungen der Bildungssysteme in den Blick nehmen, dabei jedoch keine individualdiagnostischen Informationen bereitstellen (s. z. B. Böttcher et al. 2008; OECD 2009).

Zu beachtende Aspekte beim pädagogischen Testen

Ausgehend von der beim pädagogischen Testen verfolgten Zielstellung (wie individuelle Erfassung relevanter Informationen vs. Bildungsmonitoring) spielen gemäß den Standards insbesondere die folgenden Aspekte eine zentrale Rolle, die stets zu beachten sind, weil sie die Validität der Testung sowie die Aussagereichweite der dabei gewonnenen Informationen entscheidend bestimmen:

1. *Was soll getestet werden?*
 - Gegenstand der Testung: exakte Beschreibung der zu messenden Fertigkeiten, Fähigkeiten bzw. Kompetenzen (z. B. hinsichtlich der inhaltlich-dimensionalen und kognitiv-graduellen Struktur und der Zielpopulation)
2. *Wann und wie soll getestet werden?*
 - Zeitpunkt der Testung: Eingangsdiagnostik (z. B. vor Studienbeginn im Rahmen von Vorkursen), Prozessdiagnostik (z. B. zu mehreren Zeitpunkten während des Studiums), Abschlussdiagnostik (z. B. am Ende des Studiums)
 - Design der Testung: Format und Sequenzierung von Tests und Aufgaben (z. B. wiederholte oder adaptive Testung), Durchführung der Testung (z. B.

- unter experimentellen Laborbedingungen), Konstellation der Testpersonen (z. B. Einzel- oder Gruppentestung)
3. *Wo, d. h. in welchem Bezugssystem soll getestet werden?*
 - Lehr-Lern-Bezug der Testung: bezogen auf die formellen Lehr-Lern-Angebote an einer Bildungsinstitution (z. B. Studium an einer Hochschule) oder auf informelles Lernen (z. B. Bildungserwerb durch ein Hobby)
 4. *Wozu sollen die Testergebnisse verwendet werden?*
 - Konsequenzen aus der Testung: Definition von Implikationen des Testens (z. B. Ist der Test als eine Übung gedacht oder folgt daraus die Zulassung oder Ablehnung zu einem Studiengang?)

Diese Aspekte erfordern verschiedene Überlegungen, die z. B. den Lernort (*formelles vs. informelles Lernen*), den Zeitpunkt und die Implikationen für das Lernen (*formative vs. summative Testung*), die Konsequenzen aus der Testung (*Low-Stakes- vs. High-Stakes-Testung*) oder auch die fokussierten Komponenten des Lernens (z. B. *Motivation, Selbstregulation, Kognition*) für verschiedene Bezugsgruppen bzw. Interessenvertreter (sog. *Stakeholder*) betreffen. Im Folgenden werden diese Aspekte definitorisch näher ausgeführt.

Definition

Formelles Lernen (formal learning) findet üblicherweise in einer Bildungs- oder Ausbildungseinrichtung statt, ist (in Bezug auf Lernziele, -zeit oder -förderung) strukturiert und führt in der Regel zur Zertifizierung. Formelles Lernen ist aus der Sicht des Lernenden zielgerichtet (Overwien 2005, S. 346). Ein Beispiel dafür ist das Lernen in Schulen.

Formelles Lernen

Definition

Informelles Lernen (informal learning) umfasst Lernprozesse, die außerhalb formeller Bildungsinstitutionen (wie der Schule) erfolgen. Dabei wird meist zwischen non-formellen (z. B. Nachhilfe durch Institute) und informellen Bildungssettings (z. B. Erfahrungen bei einem Museumbesuch) unterschieden (Wild und Möller 2009, S. 464).

Informelles Lernen

Definition

Formatives Testen bezeichnet Leistungsbeurteilungen, die dem Lernenden oder dem Lehrenden eine Auskunft über den aktuell erreichten Stand in einem Lernprozess geben, indem ein Abgleich mit den jeweiligen Lernzielen vorgenommen und Implikationen für die Steuerung des Lernprozesses gegeben werden. Dem Feedback kommt hierbei eine besondere Bedeutung zu, da es Lehrenden und Lernenden handlungsleitende Hinweise zur Verbesserung des Lernens liefert (z. B. indem Maßnahmen zur Optimierung ergriffen werden können; Klieme et al. 2010, S. 64 ff.). Typische Aufgaben sind z. B. das Kompetenzportfolio oder auch klassische Übungsaufgaben, die zurückgemeldet werden (► Beispiel 11.3).

Formatives Testen

Definition

Summatives Testen ist als Bericht über den abschließenden Lernstand zu verstehen (z. B. die Note in einem Zeugnis) und erfolgt meist mit Bezug auf formale nationale oder internationale Vorgaben, wie sie u. a. in Bildungsstandards oder curricularen Modulbeschreibungen zu finden sind. Das summative Testen ist – anders als das formative Testen – auch eher zum Vergleich und zu Rechenschaftslegungszwecken geeignet (Harlen und James 1997, S. 372 f.).

Summatives Testen

Low-Stakes-Testung**Definition**

Low-Stakes-Testungen bezeichnen Leistungsmessungen, deren Ergebnisse mit keinen bedeutsamen oder unmittelbaren Konsequenzen für die Testteilnehmer verbunden sind. Die Teilnahme daran ist zumeist freiwillig (z. B. Teilnahme an Online-Befragungen von Forschungsinstituten) und das gesellschaftliche, private oder berufliche Leben der Testpersonen wird durch die Ergebnisse nicht tangiert. Allerdings können diese Tests Konsequenzen für einzelne Bildungseinrichtungen haben, was u. a. auch die öffentlichen und bildungspolitischen Diskussionen infolge der Veröffentlichungen von Ergebnissen mehrerer empirischer Large-Scale-Studien (z. B. PISA) zeigen (für den nationalen Bildungsbericht PISA 2015 s. Reiss et al. 2016). Low-Stakes-Testungen sind die gängigste Form des Testens zu Forschungszwecken. Kritisch wird mitunter die Gültigkeit der erzielten Testwerte diskutiert, da die Teilnahme(bereitschaft) der Testpersonen erheblich variieren und dies die Validität der Testwertinterpretationen (► Kap. 21) einschränken kann.

High-Stakes-Testung**Definition**

High-Stakes-Testungen bezeichnen Leistungsmessungen, deren Ergebnisse mit mittelbaren oder unmittelbaren Konsequenzen für die einzelnen Testteilnehmer verbunden sind (z. B. Tests für die Zulassung zum Studium oder Abschlussprüfungen; Heubert und Hauser 1999, S. 36 ff.). Die Teilnahme an diesen Tests ist meist obligatorisch (Haertel 1999). Aufgrund der erwarteten Konsequenzen ist anzunehmen, dass die Testteilnehmer stets versuchen, die bestmöglichen Testergebnisse zu erzielen.

Stakeholder**Definition**

Stakeholder (Interessenvertreter bzw. Bezugsgruppen) sind in aller Regel an unterschiedlichen Informationen der Testung interessiert und können einen Beitrag zur Konzeption der Testung leisten. Die Standards (vgl. AERA et al. 2014, S. 3) unterscheiden Personengruppen mit Verantwortlichkeiten bei der

- Entwicklung des Tests (z. B. Forschungsinstitute),
- Veröffentlichung des Tests (z. B. Verlage),
- Durchführung und Auswertung des Tests (z. B. Testanwender),
- Interpretation der Testergebnisse (z. B. statistische Institute),
- Entscheidung zu bildungspolitischen oder sozialen Zwecken (z. B. Politiker),
- Form der Testteilnahme (einschl. Testpersonen),
- Finanzierung der Testentwicklung oder des Testeinsatzes (z. B. Stiftungen oder Unternehmen),
- Überprüfung und Auswahl von Tests (z. B. für Large-Scale-Assessments oder internationale Vergleiche).

Beispiel 11.3: Formatives Testen in der Hochschullehre

In der Universität werden formative Tests z. B. über webbasierte Lernplattformen eingesetzt, die sowohl das Lernen als auch das Testen computerbasiert ermöglichen. Bei geschlossenen oder halboffenen Fragetypen kann die Auswertung automatisch erfolgen und an die Studierenden zurückgemeldet werden. So verfügen die Studierenden unmittelbar über das Feedback und können es für eine Veränderung ihres Lernverhaltens nutzen. Gleichzeitig stehen dem Lehrenden Statistiken zum absolvierten Test zur Verfügung, mittels derer z. B. Probleme bei bestimmten Aufgabeninhalten oder -formaten erkannt werden können oder die Passung zur Lerngruppe analysiert werden kann.

Bei pädagogischen Tests im Schulbereich stellen insbesondere Schulbezirke, Schulen, schulisches Personal und Lehrkräfte sowie Eltern und die Schüler selbst jene Bezugsgruppen dar, für die die Testergebnisse von unmittelbarer Bedeutung sind (Hattie et al. 1999, S. 396).

11.2.3 Designs pädagogischen Testens

Beim pädagogischen Testen werden Instrumente konstruiert und eingesetzt, die die Beantwortung von pädagogischen Fragestellungen bzw. die Realisierung von pädagogischen Zielen ermöglichen sollen. Die Standards bieten verschiedene Kriterien an, um die Erreichung der Ziele des pädagogischen Testens überprüfen zu können. Mit den Testinstrumenten sollen Erkenntnisse bereitgestellt werden, die über das Lehren und Lernen sowie die Lernergebnisse informieren. Die Testergebnisse sollen die Beantwortung von pädagogischen Fragestellungen sowie angemessene Interpretationen vor dem Hintergrund elaborierter Lehr-Lern-Theorien erlauben. In diesem Kontext ist insbesondere der Bezug zu Lehr-Lern-Inhalten (z. B. Curricula) sowie Lehr-Lern-Prozessen (z. B. Instruktionen) bedeutsam.

Die Ziele des pädagogischen Testens sollen bei der Entwicklung und Anwendung der Instrumente erfüllt und ihre Erreichung anhand der Standards überprüft werden. Hinsichtlich curricularer oder instruktionaler Vorgaben und Eigenschaften sind dazu Erkenntnisse bereitzustellen, die über das Lehren und Lernen sowie die Lernergebnisse informieren und die pädagogische Bedeutung von Testwerten auf Basis elaborierter Lehr-Lern-Theorien herausstellen. In der Triade von Assessment, Curriculum und Instruktion von Pellegrino (2010) kommt das sog. „Constructive Alignment“ (Biggs 1999, S. 64) zum Ausdruck, das bei der Entwicklung pädagogischer Tests berücksichtigt werden sollte.

Definition

Das **Constructive Alignment** nach Biggs (1999) beschreibt die Kohärenz von Lernergebnissen, Lehr-Lern-Methoden und Prüfungs- bzw. Testverfahren. Das Zusammenspiel der drei Komponenten der Constructive-Alignment-Triade setzt die Übereinstimmung aller Lehr-, Lern- und Prüfungsaktivitäten voraus, die damit in Zusammenhang stehend auf das Ziel einer optimalen Erreichung des Lernergebnisses ausgerichtet werden (Abb. 11.1).

Testinstrumente sollen Erkenntnisse des Lehrens und Lernens bereitstellen

Constructive Alignment

Um die Lehr- und Lernziele mit den konkreten Assessments (Test- bzw. Prüfungsinstrumenten) zu verbinden, werden spezielle Assessmentdesigns verwendet, bei denen die Ziele in die Konstruktformulierung und Testdefinitionen mit aufgenommen und dann in geeigneten Testaufgaben operationalisiert werden. In Tab. 11.2 sind fünf einschlägige Assessmentdesigns abgebildet, die zum pädagogischen Testen in den letzten Jahren entwickelt und eingesetzt wurden.

Assessmentdesigns

Die in Tab. 11.2 abgebildeten Assessmentdesigns weisen einige systematische Unterschiede auf. Das von Crooks et al. (1996) entwickelte Design legt einen besonderen Fokus auf die Festlegung und Verarbeitung von Testwerten. Die Autoren beschreiben detailliert die acht Komponenten des Designs von der Administration und Auswertung der Aufgabenlösungen bis hin zur Entscheidung und zu den Konsequenzen der Testwerte (z. B. die Zulassung zum Studium).

Umfassender und differenzierter gehen Wilson (2005) und Hattie et al. (1999) auf die Konzeption und Gestaltung von Tests ein. Nach der Beschreibung der jeweiligen Konstrukte werden diese in entsprechenden Aufgaben und Items operationalisiert und evaluiert; abschließend werden die formulierten Hypothesen zu den Konstrukten und dem Zweck des Testeinsatzes geprüft.

Die Assessment-Triade von Pellegrino et al. (2001) stellt die drei zentralen Komponenten eines Assessments in aggregierter Form dar, wobei alle drei Kom-

Lernergebnisse:

Welche Fähigkeiten und Fertigkeiten sollen die Studierenden am Ende der Veranstaltung bzw. des Studiums erreichen?



Abb. 11.1 Constructive Alignment (eigene Übersetzung aus Zlatkin-Troitschanskaia et al. 2017, S. 4; nach Biggs und Tang 2011)

11

ponenten aufeinander bezogen werden. Sie unterscheiden zwischen Kognition (dem intendierten Konstrukt und den zu formulierenden Hypothesen, z. B. Fachwissen in Mathematik), der Beobachtung (der Messbarmachung des Konstrukt und der Quantifizierung; z. B. ein Test in Algebra in einer Lehrveranstaltung) sowie der Interpretation (der Auswertung der Daten und die Prüfung der formulierten Hypothesen; z. B. Analyse der Testergebnisse am Ende der Lehrveranstaltung; **Abb. 11.2**).

In den letzten Jahren findet auch das Design von Mislevy und Haertel (2006) eine immer häufigere Verwendung, wobei die Autoren u. a. auf die Bedeutung tech-

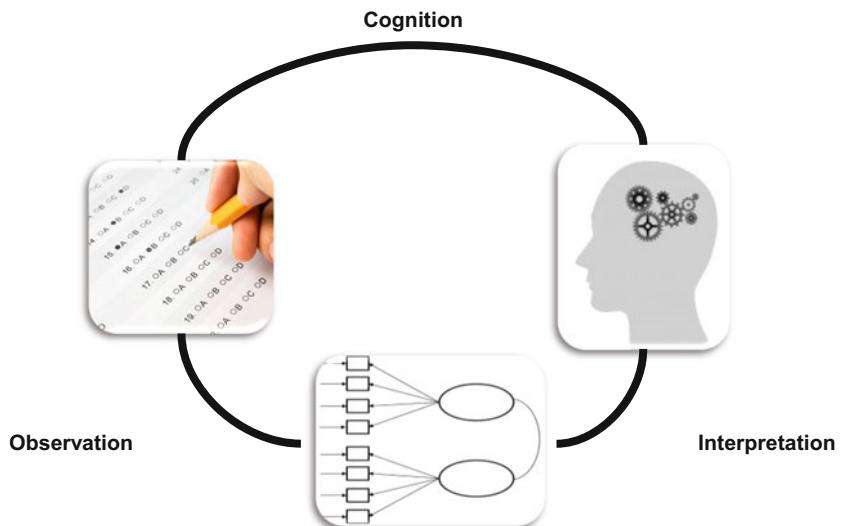


Abb. 11.2 Assessment-Triade nach Pellegrino et al. (2001). Republished with permission of The National Academies Press, © 2001; permission conveyed through Copyright Clearance Center, Inc.)

■ Tabelle 11.2 Assessmentdesigns zum pädagogischen Testen (Brückner et al. 2014, S. 138).

Republised with permission of Peter Lang GmbH, © 2014; permission conveyed through Copyright Clearance Center, Inc.)

Bezeichnung	Autoren	Komponenten pädagogischen Testens
<i>Modell des pädagogischen Testens</i> (Model of educational assessment)	Crooks et al. (1996)	<ul style="list-style-type: none"> – Administration (Administration) – Wertung (Scoring) – Aggregation (Aggregation) – Generalisierung (Generalization) – Extrapolation (Extrapolation) – Evaluation (Evaluation) – Entscheidung (Decision) – Auswirkung (Impact)
<i>Modell des pädagogischen Testens</i> (Model of educational testing)	Hattie et al. (1999)	<ul style="list-style-type: none"> – Konzeptionelles Messmodell (Conceptual Models of Measurement) – Test- und Aufgabenentwicklung (Test and Item Development) – Testadministration (Test Administration) – Testeinsatz (Test Use) – Testevalution (Test Evaluation)
<i>Assessment-Triade</i> (Assessment Triangle)	Pellegrino et al. (2001)	<ul style="list-style-type: none"> – Kognition (Cognition) – Beobachtung (Observation) – Interpretation (Interpretation)
<i>Vier Basiskomponenten</i> (Four Building Blocks)	Wilson (2005)	<ul style="list-style-type: none"> – Konstruktnetz (Construct Map) – Aufgabendesign (Item Design) – Ergebnisraum (Outcome Space) – Messmodell (Measurement Model)
<i>Evidenzbasiertes Testen</i> (Evidence-centered Assessment)	Mislevy und Haertel (2006)	<ul style="list-style-type: none"> – Domänenanalyse (Domain Analysis) – Domänenmodellierung (Domain Modeling) – Testkonstruktionsrahmen (Assessment Framework) – Testimplementierung (Assessment Implementation) – Testeinsatz (Assessment Delivery)

nologischer Veränderungen für das pädagogische Testen (z. B. höhere Leistungsfähigkeit von Computern, flexiblerer Einsatz von mobilen Endgeräten, mediale Unterstützung bei der Testentwicklung) besonders eingehen und eine Anwendbarkeit des Designs auf vielfältige Testentwicklungen gegeben ist (Mislevy 2016).

In ► Beispiel 11.4 wird das evidenzbasierte Assessmentdesign von Mislevy und Haertel (2006) anhand des Projekts „Wirtschaftswissenschaftliche Fachkompetenz“ exemplarisch erläutert.

Beispiel 11.4: Design des Projekts „Wirtschaftswissenschaftliche Fachkompetenz“

(WiwiKom; Zlatkin-Troitschanskaia, Förster, Brückner & Happ, 2014, S. 175)

Das Projekt WiwiKom verfolgt das Ziel, wirtschaftswissenschaftliche Fachkompetenz bei Studierenden bzw. Hochschulabsolventen zu modellieren und zu messen. Aufgrund des Defizits an geeigneten deutschsprachigen Testinstrumenten wurden hierzu zwei internationale Tests übersetzt und zu einem deutschsprachigen Test zusammengeführt. Dabei wurde überwiegend dem evidenzbasierten Testen gefolgt (Mislevy und Haertel 2006; ■ Tab. 11.2).

Domänenanalyse (Domain Analysis) zur Beschreibung der Domäne „Wirtschaftswissenschaften“: Sie beinhaltet die curriculare Analyse der Modulhandbü-

cher von 96 Studiengängen aus 64 Fakultäten für Wirtschaftswissenschaften deutscher Universitäten sowie Fachhochschulen. Die Befunde dieser Analyse wurden mit Ergebnissen einer Lehrbuchanalyse verglichen und durch Experteninterviews mit 78 Dozenten ergänzt. Durch die verschiedenen Blickwinkel konnte die Domänenanalyse breit abgesichert werden.

Domänenmodellierung (Domain Modeling): Sie beinhaltet die Definition des zu testenden Zielkonstrukts (wirtschaftswissenschaftliche Fachkompetenz) unter Berücksichtigung der Erkenntnisse der Domänenanalyse. Hierbei wird u. a. dargelegt, welches Fachwissen im Studium erworben werden soll und wie es in unterschiedliche Teilgebiete (Rechnungswesen, Marketing etc.) unterteilt werden kann.

Testkonstruktionsrahmen (Assessment Framework): Er umfasst die Modellierung des definierten wirtschaftswissenschaftlichen Fachwissens für verschiedene Anforderungsniveaus und Teilgebiete der Wirtschaftswissenschaften, sodass die Aufgaben des Tests diesen verschiedenen Bereichen (Aufgaben zu Grundlagen der Betriebswirtschaftslehre, Aufgaben zum Marketing etc.) zugeordnet werden können.

Testimplementierung (Assessment Implementation): Sie umfasst alle operativen Fragen bezüglich der Aufgabenkonstruktion, -übersetzung und -adaptation, des Untersuchungsdesigns und der Testwerte. Ausgehend von den Erkenntnissen der vorhergehenden Analysen wird aus dem Pool mit über 400 übersetzten und adaptierten Aufgaben ein systematischer Aufgabenauswahlprozess durchgeführt. Dieser umfasst neben einer Online-Bewertung der Items (z. B. im Bezug zur Praxisrelevanz) durch 78 Lehrende auch Interviews mit Studierenden, in denen sie sich mittels der Methode des lauten Denkens zu formalen Fehlern in den Aufgaben oder zu Verständnisschwierigkeiten bei den Aufgabenstellungen äußern (vgl. ▶ Kap. 4). Des Weiteren können Bewertungskriterien definiert werden, die eine Quantifizierung von richtigen, teilweise richtigen oder falschen Aufgabenlösungen gewährleisten (vgl. ▶ Kap. 5).

Testeinsatz (Assessment Delivery): Dieser beschreibt die Anforderungen, die unmittelbar mit dem Einsatz des Tests verbunden sind. Aufgrund der großen Zahl an Aufgaben ist z. B. zu entscheiden, wie der Test ökonomisch an Hochschulen und Universitäten, bei nur begrenzter Befragungszeit (z. B. 45 Minuten), eingesetzt werden kann (z. B. durch die Verwendung von Testheften, in denen die Studierenden dann nur einen Teil der Aufgaben bearbeiten müssen). In WiwiKom werden 43 Testhefte, die über ein Testheftdesign miteinander kombiniert sind, in mehreren Erhebungszyklen bei über 10.000 Studierenden an 57 Universitäten und Hochschulen eingesetzt.

Gemeinsamkeiten der Assessmentdesigns

Die Strukturen der Assessmentdesigns mit ihren einzelnen Komponenten sind unterschiedlich, weisen jedoch inhaltliche Gemeinsamkeiten auf, indem sie der Entwicklung von Aufgaben und der Durchführung des Testens generell die pädagogische Zielsetzung und Modellierung der Konstrukte voranstellen.

Diesen Designs können die Standards auf verschiedenen Ebenen zugeordnet werden. Beispielsweise spielen im Rahmen des Testkonstruktionsrahmens (Mislevy und Haertel 2006), des Konstruktnetzes (Wilson 2005), der Interpretation (Pellegrino et al. 2001), des konzeptionellen Messmodells (Hattie et al. 1999) oder der Extrapolation (Crooks et al. 1996) Validitätsstandards eine größere Rolle als Standards, die sich auf die Testadministration, die Wertung oder die Verantwortlichkeiten der Testentwickler beziehen. Ferner können sich mehrere Standards auf einzelne Aspekte der Designs zum pädagogischen Testen beziehen. Beispielsweise sind für die Testevaluation (Hattie et al. 1999) mehrere Standards aus dem Bereich der Testanwendungen besonders relevant (▶ Abschn. 11.1).

Zudem können manche Standards für einzelne Bezugsgruppen mehr oder weniger bedeutsam sein. Um für die einzelnen Bezugsgruppen (z. B. Schüler/Studierende, Lehrer/Dozierende, Eltern, Bildungspolitiker) Informationen über das Lernen bereitzustellen, ist es notwendig, dass die mit dem Test ermittelten Testwerte den an sie gestellten Qualitätsanforderungen hinsichtlich ihrer Objektivität (vgl. ► Kap. 2), Reliabilität (vgl. ► Kap. 14 und 15) und Validität (vgl. ► Kap. 21) genügen. Die Gütekriterien der Objektivität und Reliabilität sowie die mit ihnen verbundenen psychometrischen Anforderungen lassen sich sowohl auf psychologische als auch auf pädagogische Tests gut anwenden. Hingegen stellt die Validität als bedeutsamstes Qualitätskriterium das am meisten mit einer inhaltlichen Bedeutung auszufüllende Kriterium dar und muss im Rahmen des pädagogischen Testens besonderen Anforderungen entsprechen, wie im nächsten Abschnitt dargelegt wird.

11.3 Validitätsstandards und pädagogisches Testen (Standards 1.0–1.25)

Die Validitätsstandards bilden das erste Chapter der Standards for Educational and Psychological Testing (vgl. AERA et al. 2014). Die Standards unterscheiden drei Cluster von Standards zur Validität, von denen das dritte Cluster auf fünf grundlegende Ansätze von empirischen Validitätsnachweisen ausgerichtet ist (AERA et al. 2014, S. 23 ff.). Insgesamt werden 26 Validitätsstandards herausgearbeitet, von denen ein Großteil auf die Zusammenhänge der Testergebnisse mit anderen Variablen entfällt. Diese Fokussierung spiegelt zugleich die hohe Bedeutung wider, die dem klassischen Validierungsansatz im nomologischen Netz der Konstruktvalidität (vgl. ► Kap. 21) beigemessen wird.

Die drei Cluster haben folgende Inhalte:

1. Festlegung der Verwendungszwecke und Interpretationen (Establishing intended uses and interpretations) (Standards 1.1–1.7)
2. Aspekte der Stichprobe und des Settings der Validierung (Issues regarding samples and settings used in validation) (Standards 1.8–1.10)
3. Spezifische Formen von Validitätsevidenzen (Specific forms of validity evidence):
 - Analyse des Testinhalts (Standard 1.11)
 - Analyse der Aufgaben- und Testbearbeitungsprozesse (Standard 1.12)
 - Analyse der inneren Struktur eines Tests (Standards 1.13–1.15)
 - Analyse der Zusammenhänge mit anderen Variablen (Standards 1.16–1.24)
 - Analyse der Konsequenzen einer Testung (Standard 1.25)

Drei Cluster mit insgesamt
26 Standards zur Validität

Im ersten und zweiten Cluster werden die Zwecke, Interpretationen und Entscheidungen, die auf Basis der Testwerte für bestimmte Zielgruppen getroffen werden, plausibel beschrieben und in Bezug zum pädagogischen Testen gesetzt. Im dritten Cluster werden Evidenzen generiert, die dem Testentwickler oder -anwender helfen, eine Entscheidung über die Robustheit seiner Interpretationen und Entscheidungen zu treffen. Die Analyse der inneren Struktur eines Tests (z. B. ob sich in einem Test zur Erfassung mathematischen Fachwissens von Schülern das Wissen über Geometrie von dem Wissen über Analysis unterscheidet) sowie die Zusammenhänge der Testwerte zu anderen, externen Variablen (z. B. der verbalen Intelligenz von Schülern) erfordert eine genaue Kenntnis der Eigenschaften des zu analysierenden Konstrukts. Geeignete Informationen für diesen Cluster liefern die im Teil III dieses Bandes aufgeführten Verfahren, insbesondere die exploratorische Faktorenanalyse (EFA, ► Kap. 23), die konfirmatorische Faktorenanalyse (CFA, ► Kap. 24), die Multitrait-Multimethod-Analyse (MTMM-Analyse, ► Kap. 25) und die Modelle der Latent-State-Trait-Theorie (LST-Theorie, ► Kap. 26).

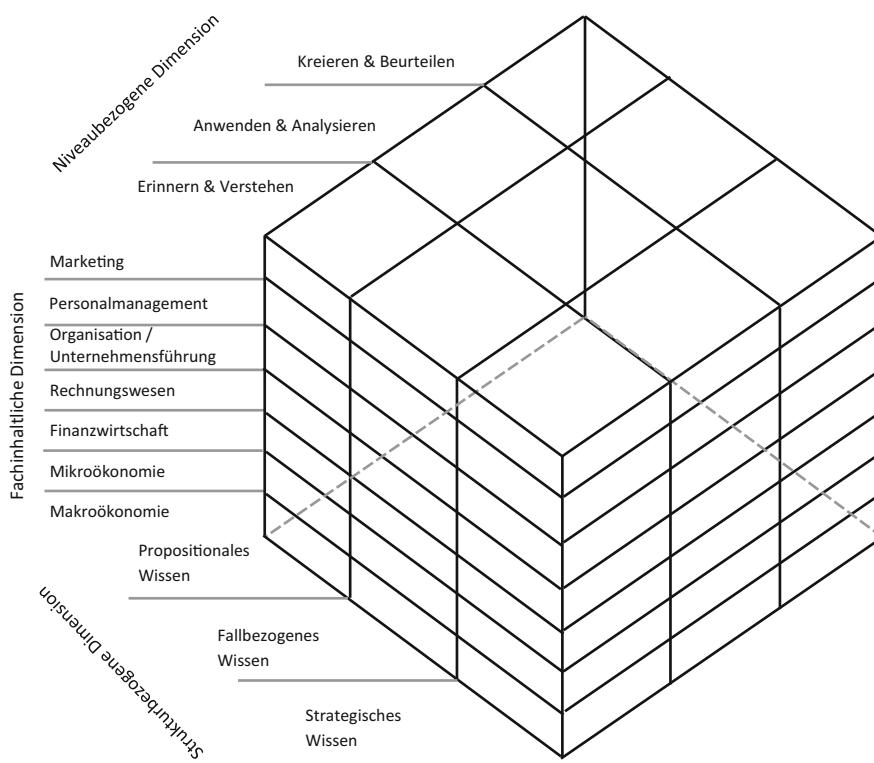
Vorgehen zur Konstruktdefinition

Das etablierte Vorgehen zur Konstruktdefinition basiert auf einem mehrdimensionalen Modellierungsansatz, in dem neben den Inhalten (z. B. Wissen in der Geometrie) einzelne (z. B. kognitionsbezogene) Lernziele oder -niveaus definiert (z. B. ob das Wissen nur memoriert oder auf einen Fall angewendet werden soll) und gegenübergestellt werden (Hattie et al. 1999).

Neben den einzelnen Inhaltsbereichen sind pädagogische Tests häufig darauf ausgerichtet, eine Graduierung des Lernfortschritts zu erfassen und zurückzumelden. Lernziele und -fortschritte können dabei in sog. „lehr-/lernzielbezogenen Taxonomien“ formuliert sein, die für eine standardbasierte Validierung verwendet werden können (☞ Tab. 11.3).

☞ Tabelle 11.3 Lehr-/lernzielbezogene Taxonomien

Bezeichnung	Autoren	Komponenten pädagogischer Taxonomien
<i>Lehr-Lernziel-Taxonomie</i> (Taxonomy of Educational Objectives)	Bloom et al. (1956)	<ul style="list-style-type: none"> – Wissen (Knowledge) – Verstehen (Comprehension) – Anwendung (Application) – Analyse (Analysis) – Synthese (Synthesis) – Bewertung (Evaluation)
<i>SOLO-Taxonomie (Struktur beobachteter Lernergebnisse)</i> (Structure of Observed Learning Outcome [SOLO] taxonomy)	Biggs und Collis (1982)	<ul style="list-style-type: none"> – Prästrukturell (Prestructural) – Unistrukturrell (Unistructural-quantitative) – Multistrukturrell (Multistructural-quantitative) – Relational (Relational-qualitative) – Abstrahierend (Extended Abstract-qualitative)
<i>Taxonomie des Lernens, Unterrichtens und Testens</i> (A Taxonomy for Learning, Teaching and Assessing)	Anderson und Krathwohl (2001)	<p><i>Kognitive Prozesse (Cognitive Processes)</i></p> <ul style="list-style-type: none"> – Erinnern (Remember) – Verstehen (Understand) – Anwenden (Apply) – Analysieren (Analyze) – Bewerten (Evaluate) – (Er)Schaffen (Create)
		<p><i>Wissen (Knowledge)</i></p> <ul style="list-style-type: none"> – Faktisch (Factual) – Konzeptuell (Conceptual) – Prozedural (Procedural) – Metakognitiv (Metacognitive)
<i>Taxonomie des nachhaltigen Lernens</i> (Taxonomy of Significant Learning)	Fink (2003)	<ul style="list-style-type: none"> – Grundwissen (Foundational knowledge) – Anwendung (Application) – Verknüpfung (Integration) – Menschliche Dimension (Human dimension) – Werte (Caring) – Lernen zu Lernen (Learning How to Learn)
<i>Die neue Lehr-Lern-Zieltaxonomie</i> (The new taxonomy of educational objectives)	Marzano und Kendall (2007)	<ul style="list-style-type: none"> – Abrufprozesse (Retrieval processes) – Verstehensprozesse (Comprehension processes) – Analyseprozesse (Analysis processes) – Prozesse der Wissensverwendung (Knowledge utilization processes) – Metakognitive Prozesse (Metacognitive processes) – Prozesse des Selbst-Systems (Self-system processes)



■ Abb. 11.3 Konzeptuelles Modell für die innere Struktur des Fachwissens über Wirtschaftswissenschaften. (Aus Zlatkin-Troitschanska et al. 2013, S. 118, mit freundlicher Genehmigung des Verlags Empirische Pädagogik)

Viele pädagogische Tests greifen auf diese Taxonomien zurück, um bestimmte Lernergebnisse mit einzelnen Aufgaben zu erfassen, und richten daran auch ihre Validierungsmaßnahmen aus. Die älteste und trotz einer Vielzahl an Modifikationen und Subklassifizierungen (z. B. nach Dubs 1978 oder Metzger et al. 1993) in ihrer ursprünglichen Form am häufigsten verwendete Taxonomie ist die nach Bloom et al. (1956). In einer klassischen Hierarchieform werden sechs Arten von Lernzielen beschrieben, die auf unterschiedliche Inhalte bezogen werden können – z. B. zur Erfassung des biologischen Wissens (Crowe et al. 2008), zur Erfassung des geschichtlichen Wissens (Moosbrugger 1985) oder zur Erfassung des ökonomischen Wissens von Schülern über verschiedene Schulformen hinweg (Walstad et al. 2013). Vielfach finden sich hierzu eine Reihe von Formulierungshilfen, die die Testentwickler unterstützen, Lernziele oder Aufgaben für die einzelnen Komponenten der Taxonomien zu entwickeln, z. B. für die Taxonomie von Bloom et al. (1956) durch die Angabe von Verben für die Komponenten (Wissen: benennen, [wieder]-erkennen, ...; Verstehen: erläutern, begründen, ...; Anwendung: identifizieren, ermitteln). In ■ Abb. 11.3 ist als Beispiel das Rahmenmodell zum Fachwissen in Wirtschaftswissenschaften dargestellt.

Theoriebasierte Kritik (z. B. mangelnde kognitionspsychologische Fundierung der Taxonomie) und empiriebasierte Kritik (z. B. mangelnde empirische Replikation der Taxonomie in Schwierigkeitsparametern von Testaufgaben) führten neben den Variationen der Bloom'schen Taxonomie zur Entwicklung weiterer Taxonomien, die ebenfalls bei der Entwicklung und Evaluation von Testaufgaben verwendet werden können (■ Tab. 11.3). Anderson und Krathwohl (2001) lösen in ihrer Überarbeitung der Taxonomie von Bloom et al. (1956), in der Taxonomie des Lernens, Unterrichtens und Testens, die vorgegebene Hierarchie durch eine Matrixstruktur auf, indem sie bei der Zielkonstruktion des Lehrens und Lernens explizit zwischen Arten von „Wissen“ und darauf gerichtete „kognitive Prozesse“

Taxonomien als mehrdimensionale Modellierungsansätze

Lernziele und Aufgaben formulieren

unterscheiden. Somit können jeder Wissensart alle kognitiven Prozesse zugeordnet werden und vice versa. So wurde z. B. im Rahmen des Projekts zur „Untersuchung von Leistungen, Motivation, und Einstellungen in der beruflichen Bildung“ (UL-ME) mit der Taxonomie von Anderson und Krathwohl (2001) berufsrelevantes Wissen in Wirtschaft und Technik erfasst (s. Brand et al. 2005).

Eine explizite Einbeziehung von metakognitiven Prozessen findet sich in der Taxonomie von Marzano und Kendall (2007), die somit eine Erweiterung der kognitiven Taxonomien von Anderson und Krathwohl (2001) resp. Bloom et al. (1956) darstellt. Eine ähnliche Ergänzung findet sich auch bei Fink (2003). Kognitionen werden in dieser Taxonomie um den ganzen Menschen betreffende Prozesse erweitert (Human dimension), die zusätzlich eine ethisch-moralische (Caring) sowie eine affektive Ebene adressieren.

Eine weitere von Biggs und Collins (1982) entwickelte (SOLO-)Taxonomie, die auf eine übergeordnete Struktur der Lehr-Lern-Ergebnisse fokussiert, wurde auf die Hochschullehre in Mathematik angewendet (s. Heinisch et al. 2016). Im Unterschied zu den anderen Taxonomien beziehen sich Biggs und Collins (1982) explizit auf die kumulativen Eigenschaften von Lernprozesszuständen. Diese zeigen sich in der Beschreibung von quantitativen und qualitativen Phasen des Lernens (z. B. muss der Lerner sich zunächst einen Zugang zu einem bestimmten Lerngegenstand verschaffen [quantitativ] und ihn mit anderen Lerninhalten in Verbindung bringen, bevor er analytisch und metatheoretisch über diesen nachdenken kann [qualitativ]).

Trotz der Unterschiedlichkeit – insbesondere im Hinblick auf die Integration non-kognitiver Elemente – zeigen alle in der ▶ Tab. 11.3 dargestellten Taxonomien eine Entwicklungsperspektive auf, die einen Lernfortschritt explizit machen und sowohl für die unterrichtlich-didaktische als auch für die psychometrische Testung bedeutsam sein können (um z. B. einen Test oder eine Aufgabe auf ein Lernziel zu beziehen).

Im Rahmen der standardbezogenen Validierung sind taxonomische Graduierungen nicht nur zu erfassen, sondern als pädagogisches Zielsystem an die entsprechenden Bezugsgruppen zu kommunizieren (z. B. um Schulen zu evaluieren; s. Hattie et al. 1999) und mit Validitätsevidenzen zu stützen, die sich nach den Standards vor allem auf die folgenden fünf Aspekte beziehen können (Näheres dazu s. ▶ Beispiel 11.5):

- Testinhalt
- Testaufgabenbearbeitungsprozesse
- Innere Struktur des Tests
- Zusammenhänge mit anderen Variablen
- Konsequenzen einer Testung

! Standards zur Validität sind auf die spezifischen Anforderungen pädagogischen Testens anzuwenden.

Beispiel 11.5: Validitätsanalyse für einen pädagogischen Test

(WiWiKom-Test; Zlatkin-Troitschanskaia et al. 2014)

Wird ein Test zur Erfassung des ökonomischen Wissens im Studium eingesetzt, sollten nach den Standards folgende Validitätsevidenzen erbracht werden:

— *Testinhalt:*

Es sollten Evidenzen erbracht werden, die darauf hinweisen, dass die in den Aufgaben abgebildeten Inhalte zugleich Inhalte des hochschulichen, ökonomischen Curriculums sind und entsprechende Lernziele widerspiegeln. Verschiedene Verfahren, die von einer Dokumentenanalyse, über Befragungen, Interviews

oder (Online-)Ratings mit Experten reichen, können zu diesem Zweck eingesetzt werden. Im Studiengang Wirtschaftswissenschaften ist z. B. im ersten Semester der Besuch der Einführungsveranstaltung „Einführung in die Volkswirtschaftslehre (VWL)“ vorgesehen. Der konstruierte Test prüft, ob die Studierenden grundlegende Analysemethoden der modernen VWL beherrschen sowie mit den Grundkonzepten und Modellen der VWL vertraut sind.

— *Testaufgabenbearbeitungsprozesse:*

Es sollten Evidenzen erbracht werden, die darauf hinweisen, dass die mit den Aufgaben erfassten Bearbeitungsprozesse typischerweise Lern- und Verstehensprozesse, die mit dem Erwerb ökonomischen Wissens assoziiert werden, widerspiegeln. Diese Prozesse sind von konstruktirrelevanten Prozessen, die nicht durch das Konstrukt definiert sind, abzugrenzen (z. B. Raten als Aufgabenbearbeitungsstrategie). Verschiedene Verfahren, beispielsweise kognitive Interviews, Analysen der Beantwortungszeiten, neuropsychologische Verfahren (z. B. anhand biometrischer Daten) oder Analysen von Mustern in den Aufgabenlösungen, können hierzu verwendet werden. Beim Test in der VWL-Vorlesung kann z. B. die Bearbeitungszeit der Studierenden kontrolliert werden. Studierende, die den Test auffällig schnell gelöst haben, könnten auf Raten als Aufgabenbearbeitungsstrategie zurückgegriffen haben.

— *Innere Struktur eines Tests:*

Es sollten Evidenzen erbracht werden, die darauf hinweisen, dass die mit den Aufgaben erfassten Inhalte und Anforderungsniveaus – typischerweise die intendierten Lernbereiche ökonomischen Wissens – auch in entsprechenden Testwerten zum Ausdruck kommen. Hierbei wird insbesondere die Dimensionalität des Konstrukts ökonomischen Wissens in den Testwerten geprüft. Neben Testmodellen der Klassischen Testtheorie (KTT, ► Kap. 13) in Verbindung mit einer EFA (► Kap. 23) oder CFA (► Kap. 24) und Strukturgleichungsmodellierungen können auch Testmodelle der Item-Response-Theorie (IRT, ► Kap. 16 und 18) für eine Analyse der inneren Struktur eines Tests verwendet werden. Bei der Analyse der Dimensionalität des ökonomischen Fachwissens könnte z. B. in konfirmatorischen Faktormodellen geprüft werden, ob die Aufgabenlösungen von ökonomischen Testinhalten zum Export und Import von Gütern auf andere Weise korrelieren als Aufgabenlösungen zu Testinhalten, die sich mit Markteinträgen der europäischen Zentralbank (z. B. Veränderungen des Leitzinssatzes) befassen.

— *Zusammenhänge mit anderen Variablen:*

Es sollten Evidenzen erbracht werden, die darauf hinweisen, dass die mit den ökonomischen Aufgaben erfassten Konstrukte mit relevanten externen Variablen in Beziehung stehen (z. B. mit erfolgreich absolvierten Lehrveranstaltungen zum ökonomischen Wissen). Hierzu sind mehrere Verfahren korrelativer Analysen, die die Testwerte mit quantifizierten Variablen in Verbindung bringen, geeignet. Wird der Test bei Studierenden durchgeführt, so kann über Korrelationsanalysen (vgl. z. B. ► Kap. 21) untersucht werden, ob Studierende, die z. B. die Einführungsveranstaltung in VWL absolviert haben, die Aufgaben besser lösen als Studierende, die die Veranstaltung nicht besucht haben.

— *Konsequenzen einer Testung:*

Die Konsequenzen, die mit den Ergebnissen eines ökonomischen Tests einhergehen, sollten umfassend dokumentiert und kommuniziert werden, bevor Entscheidungen auf Basis der Testwerte getroffen werden. Treten unerwartete Konsequenzen auf, sollten Analysen durchgeführt werden, ob diese Ergebnisse möglicherweise auf eine konstruktirrelevante Varianz der Testwerte oder eine Konstruktunterrepräsentativität (vgl. z. B. ► Kap. 21) zurückgeführt werden können. Ob der Test summativ das ökonomische Wissen in einem Studienein-

gangstest abbildet oder ob dieser formativ im Rahmen einer hochschulischen Übung eingesetzt wird, ist hierbei in Testdokumentationen festzuhalten. In diesem Kontext ist insbesondere vor einer sukzessiven Funktionenzunahme zu warnen, indem Tests zu anderen als bei der Testentwicklung vorgesehenen Zwecken in der Praxis eingesetzt werden und zu nicht angemessenen Testwertinterpretationen führen. Hängt die Zulassung zu einer weiterführenden Prüfung in VWL, z. B. einer Bachelorarbeit, von dem Ergebnis der Studierenden im Test der Einführungsveranstaltung ab, so ist dies vorher bekannt zu geben und entsprechend zu dokumentieren.

11.4 Standards zur Reliabilität (Standards 2.10–2.20)

Wie auch für das psychologische Testen (vgl. ► Kap. 10) stellt die Reliabilität (vgl. ► Kap. 14 und 15) für das pädagogische Testen, neben der Validität, ein weiteres zentrales Gütekriterium dar. In den Standards for Educational and Psychological Testing werden 20 Standards bezüglich der Reliabilität beschrieben, die sowohl an psychologisches als auch an pädagogisches Testen identische Anforderungen stellen. Die Bewertung der Ausprägung von Reliabilitätskennziffern ist für pädagogisches und psychologisches Testen identisch. Neben einer angemessenen Begründung der Methoden sowie der Kennzahlen zur Untersuchung der Reliabilität gilt es auch hier, die innere Struktur eines Tests angemessen zu berücksichtigen. Eine Kombination mehrerer Kennzahlen der Reliabilitätsbestimmung ist dabei stets einer einzelnen Kennzahl vorzuziehen.

Die Standards unterscheiden acht Aspekte, die im Rahmen der Reliabilitätsanalyse beachtet werden sollten:

- Vorgaben für die Wiederholung der Testdurchführung (Specifications for Replications of the Testing Procedure) (Standards 2.1–2.2)
- Schätzung der Reliabilität/Genauigkeit (Evaluating Reliability/Precision) (Standards 2.3–2.5)
- Reliabilität/Generalisierbarkeitskoeffizienten (Reliability/Generalizability Coefficients) (Standards 2.6–2.7)
- Einflussfaktoren der Reliabilität/Genauigkeit (Factors Affecting Reliability/Precision) (Standards 2.8–2.12)
- Standardfehler (Standard Errors of Measurement) (Standards 2.13–2.15)
- Konsistenz von Entscheidungen (Decision Consistency) (Standard 2.16)
- Reliabilität/Genauigkeit von Gruppenmittelwerten (Reliability/Precision of Group Means) (Standards 2.17–2.18)
- Dokumentation (Documenting Reliability/Precision) (Standards 2.19–2.20)

! Standards zur Reliabilität sind sowohl für pädagogisches wie auch für psychologisches Testen in gleichem Maße zu beachten.

11.5 Schwellenwerte und ihre Bedeutung für die Testwertinterpretation

Um die Testwerte von pädagogischen Tests angemessen interpretieren, vergleichen und den an der Testung beteiligten Bezugsgruppen angemessen zurückmelden zu können, ist es erforderlich, Testwerte derart zu transformieren, dass ein Verständnis und Implikationen zum pädagogischen Handeln für mehrere Bezugsgruppen bereitgestellt werden können (beispielsweise durch Testwertnormierung, ► Kap. 9).

Das Festlegen von Schwellenwerten (sog. „Cut-Scores“ bzw. „Cut-off-Scores“) mit voneinander abgrenzbaren Niveaus kann theoriebasiert im Vorhinein (a priori) oder empiriebasiert nach einer Testung (post hoc) erfolgen. Hierbei werden die Testwerte (KTT: Summenwerte über alle Items; IRT: Fähigkeitsparameter, ► Kap. 13 bzw. 16) diesen Niveaus zugeordnet und den Bezugsgruppen wird eine Rückmeldung über vorhandene Fähigkeiten der Lernenden sowie zu ihren Möglichkeiten zur Leistungsverbesserung gegeben (Zumbo 2016, S. 75). Neben dem Einsatz zur formativen Unterstützung der Lernprozesse sind sie auch zu summativen Zwecken geeignet, indem sie die Evaluation einer Zulassung oder eines erfolgreichen Abschlusses eines Bildungsgangs unterstützen können.

11.5.1 Standards zur Definition von Schwellenwerten (Standards 5.21–5.23)

In den Standards for Educational and Psychological Testing werden im Unterkapitel 4 von Chapter 5 (AERA et al. 2014 S. 107 ff.) drei Standards zur Bestimmung von Schwellenwerten vorgestellt:

- Standard 5.21: Hier nach sind sowohl die Verwendung von Schwellenwerten als auch die zu ihrer Bestimmung („Setting“) eingesetzten Verfahren zu begründen.
- Standard 5.22: Das Beurteilungsverfahren, in dem die Items oder die Testleistungen durch Experten hinsichtlich ihrer Niveaus eingeschätzt werden, sollte transparent, nachvollziehbar und derart gestaltet sein, dass die Experten ihre Erfahrungen und ihr Wissen angemessen einbringen können.
- Standard 5.23: Die Schwellenwerte sollten, sofern sie strikt trennbare Testleistungen und Interpretationen definieren, mit umfassenden empirischen Daten belegt und auf entscheidungsrelevante Kriterien bezogen werden können.

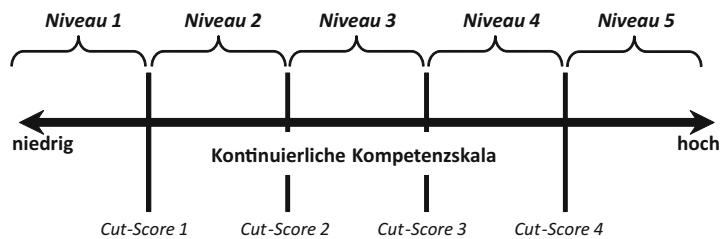
Insgesamt wird in den Standards das konkrete Setting von Schwellenwerten eher wenig thematisiert, was mitunter auf den noch mangelnden Forschungsstand und die Komplexität der Entscheidungs- und Interpretationsverfahren sowie der Beschreibung von Schwellenwerten zurückgeführt werden kann (► Kap. 17). Das im Folgenden beschriebene Verfahren berücksichtigt alle Gesichtspunkte der AERA und kann im Bereich des pädagogischen Assessments als Standardverfahren zur Bestimmung von Schwellenwerten angesehen werden; es wird deshalb als „Standardsetting“ bezeichnet.

11.5.2 Standardsettings zur Bestimmung von Schwellenwerten

Als „Standardsetting“⁴ zur Bestimmung von Kompetenzniveaus auf einer metrischen Kompetenzskala (Pant et al. 2010, S. 185) wird folgendes Verfahren bezeichnet: Zunächst eruieren Experten die Schwierigkeit von Testaufgaben, mit denen bestimmte Kompetenzniveaus (z. B. in Mathematik) festgelegt werden soll. Dieser Prozess involviert die Festlegung von Schwellenwerten, die die Übergänge zwischen Aufgabengruppen definieren, die (empirisch) abgrenzbare Anforderungen beinhalten und zu deren sicherer Lösung zunehmend komplexere kognitive Fähigkeiten notwendig sind. Dadurch werden die Lernenden in zwei oder mehr Gruppen eingeteilt. Mithilfe der daraus resultierenden Kompetenzniveaumodelle (s. z. B. Hartig et al. 2008) können die Fähigkeiten von Lernenden beschrieben und die Testitems hinsichtlich ihrer Inhalte und Qualität interpretiert werden (vgl. Cizek und Bunch 2007; Kaftandjieva 2010; Pant et al. 2013).

Schwellenwerte zur Bestimmung von Kompetenzniveaus

⁴ In diesem Kontext sind nicht die Standards for Educational and Psychological Testing gemeint, sondern das sog. „Standardsetting“ als Methode zur Definition von Schwellenwerten.



■ Abb. 11.4 Schematische Darstellung eines Standardsettings. (Aus Pant et al. 2013, S. 57, mit freundlicher Genehmigung von Waxmann)

Ein Beispiel, wie eine kontinuierliche Kompetenzskala durch Festlegung von Schwellenwerten in verschiedene Kompetenzniveaus (graduelle Abstufungen) zerlegt werden kann, ist in ■ Abb. 11.4 schematisch wiedergegeben (ein reales Beispiel findet sich bei Rauch und Hartig in ▶ Kap. 17).

Die Verfahren des Standardsettings können dahingehend unterschieden werden, ob sie eher personen- oder testzentriert sind (Pant et al. 2010).

Bei den *personenzentrierten Verfahren* werden Einschätzungen über reale Testpersonen bzw. deren Leistungen getroffen. Dabei verorten die Experten die Testpersonen unter Berücksichtigung der Kompetenzniveaubeschreibung auf den bereits definierten Kompetenzniveaus. Ein solches Verfahren wird im Allgemeinen dann präferiert, wenn die Leistungen der Lernenden den Experten bzw. Beurteilern hinreichend bekannt sind (Pant et al. 2010, S. 176). Ein personenzentriertes Verfahren stellt die sog. „Contrasting-Groups-Methode“ dar, bei der die Testpersonen direkt auf der Kompetenzskala klassifiziert werden, je nachdem, ob sie ein bestimmtes Niveau erreicht haben oder nicht. Der Schwellenwert wird dann an der Stelle gesetzt, an der die Diskriminierung zwischen den beiden Gruppen maximal wird (Pant et al. 2010; vgl. dazu auch ▶ Kap. 9).

Nach Kane (1994) sollte bei personenzentrierten Verfahren des Standardsettings zwischen Schwellenwerten und dem Leistungsstandard differenziert werden. Der Schwellenwert wird dabei als ein Punkt auf der Kompetenzskala definiert, wohingegen der Leistungsstandard als wünschenswerte Kompetenzausprägung, die den für einen bestimmten Zweck minimal akzeptablen Testwert beschreibt, bezeichnet wird (Zumbo 2016, S. 75).

Bei den *testzentrierten Verfahren* liegt der Fokus auf der Beurteilung von Testaufgaben bzw. Items, die in der Regel durch Expertinnen und Experten erfolgt. Zwei etablierte testzentrierte Verfahren sind das Angoff-Verfahren und die Bookmark-Methode (Pant et al. 2010, S. 176):

- Im *Angoff-Verfahren* erhalten die Experten die Aufforderung, sich eine hypothetische Testperson vorzustellen, die sich an der Schwelle zweier benachbarter Kompetenzniveaus befindet („Borderline-Person“). Zu jedem Testitem ist anschließend von jedem Experten die Wahrscheinlichkeit anzugeben, mit der diese Borderline-Person das Item löst. Auf diese Weise können die zu erwartenden Mindestanforderungen an die Lösung einer Aufgabe beschrieben werden. Nach einer Diskussion der Experteneinschätzungen werden die Wahrscheinlichkeitsratings pro Experten und über alle beteiligten Experten aggregiert, um den Schwellenwert zu ermitteln. Eine Herausforderung stellt hierbei die exakte Beschreibung der zu erwartenden Leistungen dar, die auf den beiden benachbarten Niveaus liegen (vgl. Angoff 1971; Brandon 2004).
- Bei der *Bookmark-Methode* erhalten die Experten ein Buch, das alle Items geordnet nach der empirischen Schwierigkeit auflistet. Die Beurteiler werden anschließend aufgefordert, unter Berücksichtigung der Kompetenzniveaubeschreibung alle Items zu markieren, die eine hypothetische Testperson mit einer spezifizierten Antwortwahrscheinlichkeit lösen kann.

Personenzentrierte Verfahren des Standardsettings

Testzentrierte Verfahren des Standardsettings

Angoff-Verfahren und Bookmark-Methode

11.5 · Schwellenwerte und ihre Bedeutung für die Testwertinterpretation

Obwohl bislang kein Verfahren generell bevorzugt wird, gilt die Bookmark-Methode als das derzeit etablierte Verfahren (Pant et al. 2010, S. 177). So zeichnet sich das Angoff-Verfahren zwar durch Transparenz und Einfachheit aus, stellt zugleich aber auch hohe Anforderungen an die Bewertung durch die Experten. Denn zum Festlegen sinnvoller Niveaus ist eine wesentliche Übereinstimmung der Experteneinschätzungen ebenso erforderlich wie eine Präzision bei der Prognose der Lösungswahrscheinlichkeiten (Grotjahn 2010). Hingegen bietet die Bookmark-Methode (► Beispiel 11.6) eine relativ einfache und effiziente Auswertung, die auf bereits generierten Daten eine evidenzbasierte Einschätzung von Niveaus zulässt, mehrere Schwellenwerte bei einem Test erlaubt und somit weniger hohe Ansprüche an die Experten stellt (Grotjahn 2010). Als problematisch erweist sich allerdings, dass die theoretisch erwarteten Schwierigkeitsparameter oftmals von den empirisch festgestellten Schwierigkeitsparametern abweichen (Freunberger und Yanagida 2012).

Vor und Nachteile des Angoff-Verfahrens und der Bookmark-Methode

Beispiel 11.6: Standardsetting im Fach Chemie

Zur Beschreibung der Kompetenzniveaus von Schülern im Fach Chemie werden in den Bildungsstandards Kompetenzen ausgewiesen, die jeweils als Kriterien zur Beschreibung niedriger und hoher Kompetenzausprägungen verwendet werden. Zur Überprüfung des Erreichens dieser Bildungsstandards werden zunächst Testaufgaben von Fachlehrern entwickelt, die die Grundlage für die Erarbeitung der Kompetenzniveaus bilden. Unter Verwendung der Bookmark-Methode identifizieren Fachexperten schwierigkeitserzeugende Merkmale der Aufgaben (von der leichtesten bis zur schwersten Testaufgabe). Dadurch können Schwellenwerte an den Stellen gesetzt werden, an denen ein qualitativer Sprung im kognitiven Anforderungsniveau sichtbar wird. Durch vier gesetzte Schwellenwerte (Ziffern in Quadranten: 435, 505, 605 und 680) werden fünf Kompetenzniveaus (römische Ziffern I–V) festgelegt (► Abb. 11.5), für die jeweils Kompetenzbeschreibungen entwickelt werden. Diese reichen von dem niedrigsten Niveau I „Unterer Mindeststandard“ bis zum höchsten Niveau V „Optimal- bzw. Maximalstandard“. Um die Testaufgaben zu normieren, erhalten diese einen Punktwert als Maß für ihre Schwierigkeit (Ziffern in Kreisen: 320, 495, 600, 650 und 825) und können so den Kompetenzniveaus (► Abb. 11.5, links: Niveau I–III; rechts: Niveau IV–V) zugeordnet werden (IQB 2013, S. 19 ff.). In ► Abb. 11.5 werden etwa auf dem Niveau IV Aufgaben zum Aufstellen von Reaktionsgleichungen zugeordnet, während zwei Niveaus darunter, auf Niveau II, eine Aufgabe mit einer Schlussfolgerung zur Bedeutung von Energie für chemische Reaktionen genannt ist.

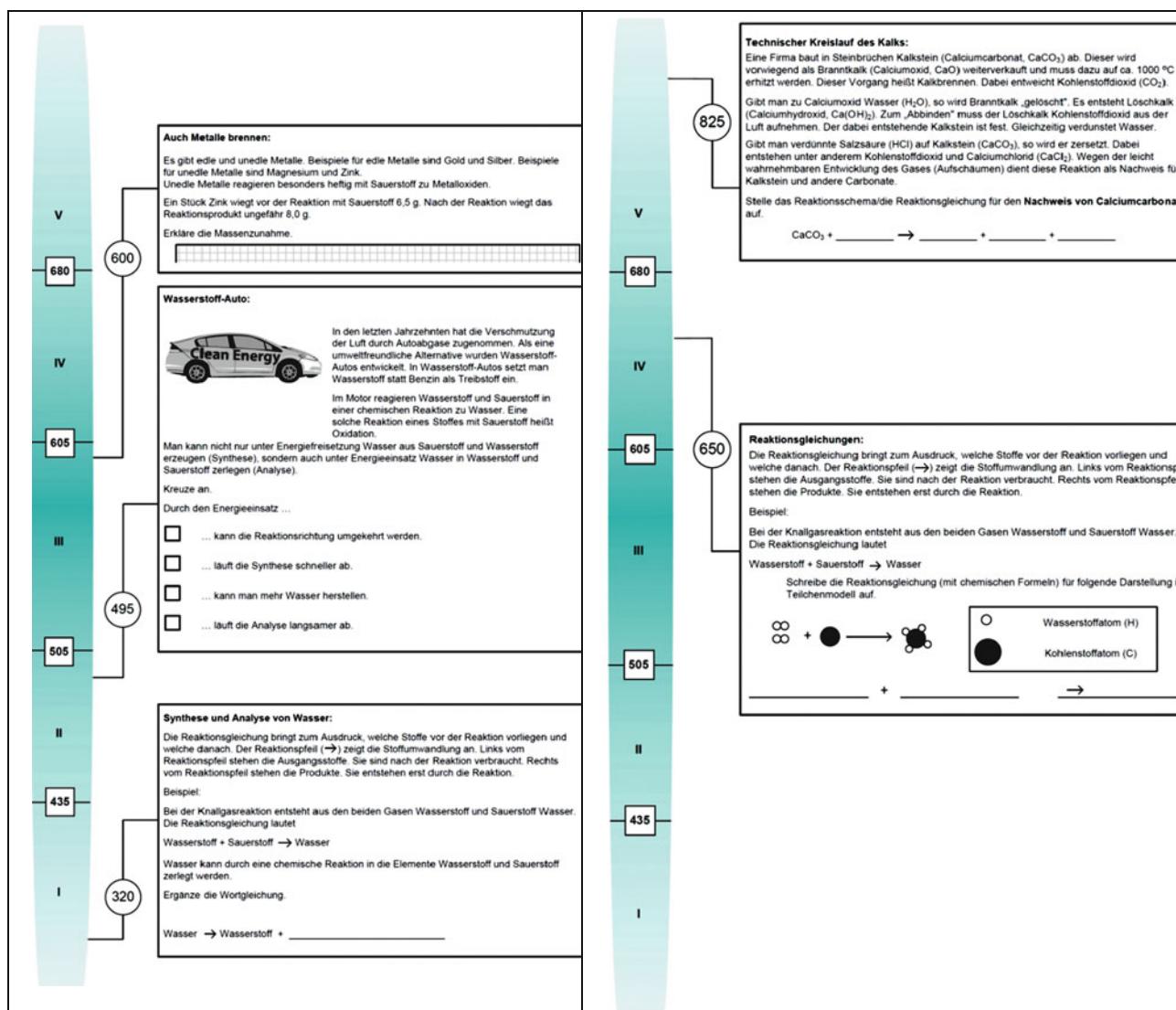


Abb. 11.5 Kompetenzniveaus in Chemie (Fachwissen). (Aus IQB 2013, S. 28 f., mit freundlicher Genehmigung des IQB)

11.6 Weitere Implikationen der Standards für pädagogisches Testen

In Chapter 12 der Standards for Educational and Psychological Testing werden 19 Standards für pädagogisches Testen und die Diagnostik in drei Cluster aufgeteilt, die sich

- auf das Design und die Entwicklung pädagogischer Tests (6 Standards),
- auf die Verwendungsweise und Interpretation von pädagogischen Tests (9 Standards) sowie
- auf die Durchführung, Bewertung und Berichterstattung von pädagogischen Tests (4 Standards) beziehen (AERA et al. 2014).

Die besondere Rolle des Testens im angloamerikanischen Raum, in dem standardisierte Schulleistungstests, die von kommerziellen und nicht kommerziellen Instituten regelmäßig zur Verfügung gestellt und von Ministerien, Schulen oder Firmen in Auftrag gegeben werden, spiegelt sich auch in den Standards wider. So

11.6 · Weitere Implikationen der Standards für pädagogisches Testen

sollten die Verwendung der Testergebnisse immer klar kommuniziert und negative Konsequenzen vermieden werden (Standard 12.1; AERA et al. 2014).

11.6.1 Fairness

Neben den eingangs erwähnten bedeutsamen Kriterien der Validität und Reliabilität ist gerade die Fairness (vgl. ► Kap. 2) ein bedeutsames Kriterium, das bei der Entwicklung von Tests im Allgemeinen und bei pädagogischen Tests im Besonderen berücksichtigt werden sollte. Die Standards for Educational and Psychological Testing widmen dem Thema „Fairness“ in Chapter 3 insgesamt 21 Standards, die für jede Phase der Testentwicklung und des Testeinsatzes von Bedeutung sein können. „Unfairness“ wird verschieden definiert, meist jedoch als Einfluss konstruktirrelevanter Varianz, wodurch Subgruppen systematisch benachteiligt oder bevorteilt werden (vgl. AERA et al. 2014; Crocker 2003).

Die Standards unterscheiden vier Aspekte, die im Rahmen der Fairnessanalyse beachtet werden sollten:

1. Testdesign, -entwicklung, -administration und Auswertungsverfahren, die möglichst valide Ergebnisinterpretationen bei vielen Individuen und relevanten Subgruppen gewährleisten (Test Design, Development, Administration, and Scoring Procedures that minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups) (Standards 3.1–3.5)
2. Validität der Testwertinterpretationen für die vorgesehenen Anwendungen bei der Zielpopulation (Validity of test score interpretations for intended uses for the intended examinee population) (Standards 3.6–3.8)
3. Möglichkeiten, um konstruktirrelevante Hindernisse zu entfernen und eine valide Testwertinterpretation für die vorgesehenen Anwendungen zu unterstützen (Accommodations to remove construct-irrelevant barriers and support valid interpretations of scores for their intended uses) (Standards 3.9–3.14)
4. Schutzmaßnahmen gegen unangemessene Testwertinterpretationen für die vorgesehenen Anwendungen (Safeguards against inappropriate score interpretations for intended uses) (Standards 3.15–3.20)

Gruppen von Testpersonen, bei denen kein Unterschied bei der Vorbereitung der Lernleistung bezüglich des zu messenden Konstrukts zu erwarten ist (z. B. männliche/weibliche Schüler im Hinblick auf ihr Fachwissen in Geschichte etc.), sollten auch in einem Test zur Erfassung der Lernleistung keine Unterschiede der Testleistungen aufweisen. Mit anderen Worten: Es wäre unfair, wenn die Testpersonen aufgrund solcher Merkmale (z. B. Geschlecht) beim Testen benachteiligt oder bevorteilt würden (vgl. ► Kap. 9). Empirische Überprüfungen, ob Testpersonen ein ungewöhnlich hohes oder geringes Aufgabenlösungsverhalten aufweisen, das nicht allein auf Veränderung ihrer Fähigkeit (z. B. Fachwissen), sondern auf die weiteren Merkmale, die keinen Zusammenhang mit dem Lösungsverhalten haben sollten (z. B. Muttersprache), zurückgeführt werden kann, können z. B. mit Analysen der Messinvarianz oder auf Basis differenzierlicher Itemfunktionsweisen (Differential Item Functioning, DIF), ermittelt werden (vgl. ► Kap. 16 und 18).

Weitere Merkmale, die ebenfalls keinen Zusammenhang mit dem Aufgabenlösungsverhalten beim pädagogischen Testen aufweisen sollten, sind z. B. Testangst oder aufgabenbearbeitungsstrategisches Verhalten wie Raten (Standard 12.7; Crooks et al. 1996).

Probleme der Fairness können nicht nur durch die Eigenschaften der Testpersonen in Erscheinung treten. Technische Fehler beim Testeinsatz oder der Bewertung bzw. des Scorings sowie interpretative Fehler beim Ziehen von Schlussfolgerungen aus den Testergebnissen können ebenso durch die Testleiter oder -anwender hervorgerufen werden (vgl. ► Kap. 2). Die nachweisliche und zertifizierte Eignung der

Testleiter und -anwender (insbesondere durch Erwerb entsprechender Lizenzen gemäß DIN 33430, ► Kap. 10) ist hierbei stets unabdingbare Voraussetzung, bevor sie mit entsprechenden Aufgaben zur pädagogischen Diagnostik betraut werden können (Standards 12.14–12.16).

11.6.2 Transparenz des Untersuchungs- und Interpretationsgegenstands (Konstrukte, Anforderungen und Inhalte)

Gerade beim pädagogischen Testen sind viele Bezugsgruppen an der Testung beteiligt oder an Informationen aus der Testung interessiert (Häcker et al. 1998). Um valide Testwertinterpretationen zu ermöglichen und fehlerhafte Testanwendungen zu vermeiden, ist eine eindeutige Kommunikation und Verbreitung von Anforderungen und Inhalten, die in den Tests erfasst werden, erforderlich. Mehrere Standards weisen auf den Informationsbedarf der Bezugsgruppen hin. Beispielsweise sind die Zwecke der Testung, die Relevanz und Repräsentativität der zu messenden und zu operationalisierenden Konstrukte in den Phasen der Testentwicklung (Standard 12.3 und 12.4) und möglicherweise verwendete Norm- und Schwellenwerte (Standard 12.5) allen Bezugsgruppen mitzuteilen. Ferner sind, wie im Rahmen des Datenmanagements deutlich wird, umfassende Dokumentationen zu den einzelnen Entwicklungs- und Interpretationsschritten, wie sie in den Designs zum pädagogischen Testen (► Tab. 11.2) zum Ausdruck kommen, anzufertigen (Standards 12.6 und 12.18; AERA et al. 2014). Auch vonseiten der Testentwickler sollten Instruktionen, Dokumentationen sowie Trainings zur Verfügung gestellt werden, die den Testanwendern den Testeinsatz erleichtern (Standard 12.16).

11.6.3 Variation der Prüfungsformen

Berücksichtigung der Lerngelegenheiten

Werden Tests eingesetzt, die eine summativ Bewertung des Lernerfolgs vornehmen und mit der erfolgreichen Absolvierung einer Bildungsinstitution verbunden sind (z. B. Abiturprüfungen, Abschlussprüfungen in der Berufsausbildung oder im Studium), muss der Testinhalt auch Gegenstand vorgelagerter Lerngelegenheiten sein (Standard 12.8; AERA et al. 2014). Dies erfordert Lerngelegenheiten, die für die Lernenden entwickelt und von ihnen absolviert werden (z. B. Unterrichtseinheiten, Vorlesungen, Übungen, Trainings) und in denen anhand von Lehrbüchern, Übungsaufgaben und anderen Lernmaterialien entsprechende Inhalte erlernt werden können. Dies stellt einen Unterschied zu verschiedenen psychologischen Tests (z. B. Persönlichkeitstests oder klinischen Verfahren) dar, da in diesen eine themenspezifische Vorbereitung eher unerwünscht ist (vgl. ► Kap. 4).

Gerade beim pädagogischen Testen zur Feststellung von Kompetenzen (zum Kompetenzbegriff s. Pant et al. 2013) ist es oftmals erforderlich, eine Bandbreite von verschiedenen Testformen einzusetzen, um die Komplexität der erworbenen Lerngegenstände angemessen und umfassend abzubilden. So gehen auch die Standards 12.9 und 12.10 (AERA et al. 2014) explizit auf eine Verwendung alternativer Testformen ein. Daher sollten Entscheidungen, ob eine Ausbildung oder eine Lerneinheit erfolgreich absolviert wurde, und Entscheidungen, die maßgeblich die berufliche und gesellschaftliche Teilhabe mitbestimmen (z. B. erfolgreiche Absolvierung eines Ausbildungsgangs), sowohl im Schul- als auch im Hochschulbereich stets auf Ergebnissen verschiedener Prüfungsformen basieren. Nicht nur im Schul-, sondern auch im Hochschulbereich werden Tests sowie auch mündliche Prüfungen, Präsentationen, Diskussionen und andere Prüfungsformate, in denen je nach fokussierter Kompetenz variierte Testformen eingesetzt werden, genutzt. Bei-

11.6 · Weitere Implikationen der Standards für pädagogisches Testen

spielsweise sind soziale Kompetenzen eher in kollaborierenden Testformen (z. B. Gruppendiskussion) prüfbar.

Ein zentrales Thema der generellen Standards für pädagogisches Testen stellt auch die Beachtung des in ► Abschn. 11.2.3 vorgestellten *Constructive Alignment* (Biggs 1999) dar. Die Übereinstimmung der Testergebnisse mit instruktionalen Zielen des Curriculums, die in vielfältigen Lerngelegenheiten in Schule oder Hochschule auch tatsächlich gelernt werden, muss sich in der Ausprägung der Testergebnisse widerspiegeln. Schüler und Studierende, die eine höhere Kompetenz in den lernbezogenen Kontexten entwickeln, sollten daher auch in den Testergebnissen zur Prüfung dieser Kompetenzen höhere Werte erzielen. Beispielsweise erreichen Studierende, die einen Lehrberuf an einer kaufmännischen berufsbildenden Schule anstreben, nach der Absolvierung fachdidaktischer Veranstaltungen im Studium auch in einem Test zum fachdidaktischen Wissen höhere Punktzahlen (Kuhn 2014, S. 234). Um verschiedene instruktionale Ziele in Tests abzubilden, ist daher ebenfalls eine Vielzahl von Prüfungsformen einzusetzen. Die Bedeutsamkeit der Übereinstimmung ist insbesondere in sogenannten High-Stakes-Testungen (► Abschn. 11.2.2), in denen eine Entscheidung über eine Zulassung, einen Abschluss der Lernenden oder die Steuerung von Bildungssystemen, Schulen und Schulbezirken getroffen werden soll, elementar (Standard 12.13, AERA et al. 2014).

Constructive Alignment

Übersichten, wie verschiedene Prüfungsformen mit Kompetenz- oder Lernzielen sowie den entsprechenden Lehr-Lern-Methoden in Übereinstimmung gebracht werden können, sind zahlreich und finden sich z. B. mit Bezug zum Hochschulbereich in Bachmann (2014).

11.6.4 Feedback

Rückmeldungen (Feedback) über Testergebnisse zu Zwecken der Optimierung von Unterrichtseinheiten oder individuellen Lernprozessen und zur Steuerung von Bildungsinstitutionen und -systemen sollen ebenfalls an alle an der Testung beteiligten Bezugsgruppen gerichtet werden. Um dieses Ziel zu erreichen, ist es notwendig, die Ergebnisse des Testens direkt an die Lernenden, die Lehrenden oder das familiäre und soziale Umfeld adressatengerecht zurückzumelden.

In mehreren Chapters der Standards for Educational and Psychological Testing (AERA et al. 2014) finden sich Hinweise, die beim Feedback zu beachten sind. In Chapter 6 sind die Standards 6.10–6.16 explizit auf die Berichterstattung sowie die Testwertinterpretationen gerichtet. Ebenso sind in Chapter 12 die Standards 12.17–12.19 mit explizitem Bezug zum pädagogischen Testen und auf das Rückmelden von Ergebnissen gerichtet.

Sowohl für den einzelnen Lernenden, aber auch für die Institutionen ist es wichtig, adressatengerechte Rückmeldungen der Testergebnisse zu erhalten, um die mit bestimmten formativen oder summativen Prüfungsformen verbundenen Ziele pädagogischen Testens zu erreichen (zur ethischen Perspektive s. Standard 11.8).

Definition

Unter **Feedback** kann die informative Rückmeldung durch eine Person oder einen Gegenstand (z. B. einen Lehrer, einen Professor, einen Schüler, Kommilitonen, Eltern oder auch ein Buch oder Computerprogramm) verstanden werden, die in ihrer Art und Struktur auf die Testleistung oder das Verständnis eines Lernenden abzielt, wobei verschiedene Aggregationsebenen (z. B. die Klasse oder eine Schulform) ebenfalls Feedback erhalten können. Die übliche Form des Feedbacks ist jedoch auf den einzelnen Lernenden bezogen (Hattie und Timperley 2007, S. 81). Generell zielt Feedback daher darauf ab, einem Lernenden eine Information zurückzumelden, die ihm dabei hilft, die Diskrepanz zwischen Ist- und Sollzustand in seinem Lernen

Aspekte des Feedbacks

zu überwinden (Müller und Ditton 2014). Gold (2015, S. 96) formuliert drei Fragen, die der Feedbackgebende bei jedem Feedback beantworten können sollte:

- Was sollen die Lernenden können? (Soll)
- Was können sie gegenwärtig? (Ist)
- Wie kann der Lehrende die Lernenden zur Erreichung der Ziele führen? (Plan)

Feedback kann einfach (im Sinne von richtig oder falsch) oder komplex (im Sinne einer Begründung der Bewertung sowie einer Instruktion zur Verbesserung des Lernverhaltens) sein.

Feedback kann sich auf vier Aspekte des pädagogischen Testens beziehen (Behnke 2016; Gold 2015; Hattie und Timperley 2007; Müller und Ditton 2014), und zwar auf

- die finalen Aufgabenlösungen,
- den Prozess der Aufgabenbearbeitung,
- die Selbstregulation des Bearbeitungsprozesses,
- die persönliche Ebene des Lernenden.

Eine optimale Rückmeldung sollte dabei die vier Aspekte immer in einem ausgewogenen Verhältnis einbeziehen, sodass sowohl die Richtigkeit und die Angemessenheit der Lösungsstrategie als auch die metakognitive Ebene der Handlungssteuerung und der Selbstregulation angesprochen werden (Gold 2015). Aus inhaltlicher Perspektive gilt zudem, dass Feedbacks immer in einem Kontext gegeben werden sollten, dabei der Grad der Diskrepanz von Ist- und Sollzustand einzubeziehen ist und der Weg zum korrekten Verständnis beschrieben sowie die kulturellen Hintergründe der Lernenden berücksichtigt werden sollten (Hattie und Timperley 2007). Beispielsweise ist es in den USA üblich, eher ein direktes personenbezogenes Feedback zu einer Testleistung zu geben, während in asiatischen Kulturen eine direkte Ansprache eher vermieden wird (Hattie und Timperley 2007).

11.7 Standards zum Management und zur Archivierung von Daten pädagogischen Testens

Datenmanagement und Datenmanagementplan

Mit dem zunehmenden Einsatz von Tests und der damit einhergehenden Zunahme qualitativen und quantitativen Datenmaterials werden Fragen der Organisation, Strukturierung, Verarbeitung, Speicherung und Sicherung von Testdaten immer wichtiger. Das *Datenmanagement* stellt dabei den Rahmen für Maßnahmen dar, die zu einer effizienten Handhabung der Daten von Bedeutung sind. Neben einfachen, operativen Fragestellungen, z.B. zum Hinzufügen, Ändern oder Sortieren neuer Variablen und Werte, stehen auch strategische Fragestellungen wie der Zugriff auf und die Organisation von Daten im Zentrum des Datenmanagements. Hierbei ist es wichtig, dass klar ist, wie umfangreich die Daten sind und wie sie technisch gespeichert und ggf. Dritten zur Verfügung gestellt werden sollen (z.B. ob eine einfache Speicherung in einer Excel-Tabelle ausreicht oder ob umfassende Datenbanken erzeugt werden müssen).

Definition

Wie Jensen (2012, S. 21) betont, bilden die Organisation der Abläufe sowie die Erhebung, Aufbereitung und Dokumentation der Daten den Kern des **Datenmanagements**.

Gerade bei großen Datensätzen, die nicht nur von einer Person, sondern von Projektverbünden generiert werden und auf die eine große Zahl von Akteuren Zugriff haben oder an dessen Informationen viele Akteure interessiert sind, ist es wichtig, die Verantwortlichkeiten in einem *Datenmanagementplan* zu regeln (Jensen

2012). In der Systematik eines Datenmanagementplans sollten alle Fragen zum Datenmanagement geregelt sein. Jensen (2012, S. 19 f.) nennt in einer Checkliste für Forschungsdaten folgende acht Kernbereiche, die im Rahmen des Datenmanagementplans zu berücksichtigen sind:

1. Richtlinien zum Datenmanagement durch den Geldgeber
2. Datenbeschreibung
3. Metadaten und Dokumentation
4. Datenschutz
5. Qualitätssicherung
6. Aufgaben- und Verantwortungsbereiche
7. Sach- und Personalkosten des Datenmanagements
8. Datensicherung, Langzeitarchivierung und Datenbereitstellung

In verschiedenen Modellen wird das Vorgehen des Datenmanagements detailliert beschrieben und reicht zum Teil deutlich über den Testentwicklungs- und Testanwendungsprozess hinaus – von der ersten Vorbereitung eines Forschungsprozesses bis hin zur Langzeitarchivierung von Daten. Ein Modell, das diesen Prozess beschreibt, ist z. B. das *Curation-Lifecycle-Modell* (Rümpel 2011).

Mehrere Institute in Deutschland bieten die Möglichkeit, Forschungsdaten zu archivieren und ggf. weiteren Wissenschaftlern für Sekundäranalysen zur Verfügung zu stellen. Neben der Möglichkeit Open-Source-Daten bereitzustellen – also ohne besondere Zugriffsregelungen –, ist es möglich, Daten lediglich zu archivieren, ohne die Vergabe von Zugriffsrechten oder einen nur eingeschränkten Datenzugriff zu erlauben.

Zur Sicherung von Forschungsdaten zum psychologischen Testen kann z. B. das Testarchiv des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID; ► <https://www.leibniz-psychology.org>) genutzt werden.

Datenarchive zur Speicherung von Forschungsdaten zum pädagogischen und psychologischen Testen bieten u. a. die folgenden Institute an:

- Leibniz-Institut für Sozialwissenschaften (GESIS; ► <https://www.gesis.org>)
- Deutsches Institut für Internationale Pädagogische Forschung (DIPF; ► <https://www.dipf.de>)
- Institut zur Qualitätsentwicklung im Bildungswesen (IQB; ► <https://www.iqb.hu-berlin.de>)

Bei Forschungsdaten wird der Schutz der personenbezogenen Daten immer wichtiger. Werden Daten digital gespeichert, durch verschiedene Personen ausgewertet oder in verschiedener Form publiziert, ist zu prüfen, inwiefern der Schutz der Person (hier der Testperson) gesichert ist, d. h. welche Daten (z. B. Geburtstag) wie detailliert (z. B. mit Jahresdatum) von wem (z. B. Regelung der Zugriffsrechte) analysiert werden dürfen. In Abgrenzung zur Datensicherheit ist der Datenschutz immer auf die einzelne Person bezogen. Zur Prüfung des Datenschutzes ist die derzeit gültige Form des Bundesdatenschutzgesetzes (BDSG) heranzuziehen, bei Analysen von internationalen Daten gelten zudem die Datenschutzgesetze der jeweiligen Länder. Seit 2018 ist für den europäischen Raum eine übergreifende Datenschutzgrundverordnung (EU-DSGVO) wirksam, die die Datenschutzregelungen in Europa vereinheitlicht.

Datenschutz und Datenmanagement

Auch wenn im Allgemeinen das Datenmanagement noch nicht umfassend in die Standards for Educational and Psychological Testing Eingang gefunden hat, wird u. a. in Chapter 9 explizit auf das Datenmanagement eingegangen und insbesondere die Verantwortlichkeiten von Testanwendern definiert (Standards 9.15–9.23).

11.8 Standards für Forschungsethik

Während ethische Fragen des Forschungsprozesses im Rahmen von Datenmanagementplänen umfassend u. a. beim Datenschutz behandelt werden, sind sie in den Standards for Educational and Psychological Testing sonst nicht explizit zu finden. Dies ist dem Umstand geschuldet, dass die Vielzahl an Ethikrichtlinien den Rahmen der Standards, die primär auf die Entwicklung und Anwendung von Tests ausgerichtet sind, deutlich übersteigen würde. Die AERA veröffentlicht daher zusätzlich zu den Standards den „Code of Ethics“ (AERA 2011).

- !** Die ethischen Standards umfassen sowohl Rechte der Untersuchungsteilnehmenden, die in ihrer Würde und ihrem Wohlergehen geschützt werden sollen, als auch Regeln über gute wissenschaftliche Praxis.

Forschungsethik und Wissenschaftsethik

Unterschieden werden hierbei die Forschungs- und Wissenschaftsethik. Die *Forschungsethik* befasst sich mit den Rechten der Untersuchungsteilnehmenden und der Vermeidung von Beeinträchtigungen der Involvierten durch den Forschungsprozess (Döring und Bortz 2016, S. 48). Die *Wissenschaftsethik* befasst sich mit Kriterien, nach denen jeder wissenschaftliche Erkenntnisgewinn einer guten Praxis folgen und überprüfbar sein muss. Demnach dürfen Forschungsergebnisse nicht unkritisch behauptet oder übernommen werden (Döring und Bortz 2016, S. 122).

Die ethischen Richtlinien der AERA (2011) bestehen aus fünf Prinzipien und daraus abgeleiteten spezifischen Standards. Die Prinzipien beschreiben dabei die Idealvorstellungen für professionelles Verhalten von Wissenschaftlern. Die darin enthaltenen Standards sind Regeln für ethisches Handeln, wobei die AERA keinen Anspruch auf Vollständigkeit erhebt, sodass eine Verhaltensweise, die nicht in den Standards aufgeführt ist, nicht automatisch unethisch sein muss (AERA 2011, S. 146).

Ethikprinzipien und Ethikstandards der AERA für das Verhalten von Wissenschaftlern (AERA 2011, S. 146 f.)

- A. *Professionelle Kompetenz*: Wissenschaftler sollten sich der Grenzen ihrer Expertise bewusst sein. Sie sollten nur solche Aufgaben ausführen, für die sie eine Qualifikation aufweisen und falls nötig in den Austausch mit anderen Experten treten.
- B. *Integrität*: Wissenschaftler sollten sich in ihrer professionellen Arbeit stets rechtsschaffend, ehrlich und respektvoll gegenüber anderen verhalten.
- C. *Professionelle und wissenschaftliche Verantwortung*: Wissenschaftler sollten die höchsten wissenschaftlichen und professionellen Standards an ihr Handeln anlegen und die Verantwortlichkeit für ihre Arbeit kennen sowie einen professionellen Austausch mit anderen Wissenschaftlern, auch im kritischen Diskurs, pflegen.
- D. *Respekt vor den Rechten, der Würde und der Vielfalt von Menschen*: Wissenschaftler respektieren die Rechte, die Würde und den Wert aller Menschen und gefährden diese nicht durch ihre Arbeit.
- E. *Soziale Verantwortlichkeit*: Wissenschaftler kennen ihre Verantwortung gegenüber der Gesellschaft. Sie sind darum bemüht, ihr Wissen und ihre Fortschritte der Öffentlichkeit zur Verfügung zu stellen.

Forschungsethische Standards

Weitere forschungsethische Standards liegen u. a. von folgenden Organisationen vor (s. auch Döring und Bortz 2016, S. 130):

- American Educational Research Association (AERA)
- Australian Association for Research in Education (AARE)

11.9 · Zusammenfassung

- Deutsche Forschungsgemeinschaft (DFG)
- Deutsche Gesellschaft für Erziehungswissenschaft (DGfE)
- Deutsche Gesellschaft für Psychologie (DGP)

In der aktuellen Praxis des pädagogischen Testens gewinnen die ethischen Standards zunehmend an Bedeutung und an Bildungsinstitutionen (wie Universitäten) oder dafür verantwortlichen Institutionen (wie Ministerien) etablieren sich z. B. Ethikkommissionen, die auf die Einhaltung der entsprechenden Standards achten.

11.9 Zusammenfassung

Die Standards for Educational and Psychological Testing bieten eine umfangreiche Darstellung von über 240 Standards, die zur Entwicklung, Durchführung und Evaluation pädagogischer und psychologischer Tests praktische Handlungsempfehlungen geben.

Zur Berücksichtigung der Besonderheiten der Zielsetzung und Entwicklung pädagogischer Tests können die Standards insbesondere zu Fragen der Validität einen Beitrag leisten. Weitere Implikationen aus den Standards für Anforderungen an pädagogisches Testen lassen sich u. a. zum Standardsetting, zur Fairness, zur Transparenz des Untersuchungsgegenstands und Interpretation, zu Formen der Diagnostik, zum Feedback sowie zum Datenmanagement finden. Unabhängig von den Standards gibt die AERA mit dem „Code of Ethics“ ethische Richtlinien zu Fragen der Forschungsethik heraus, denen aktuell eine immer größere Bedeutung zukommt.

11.10 Kontrollfragen

- ② Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Welche Standards sind im Hinblick auf die Ziele pädagogischen Testens zu beachten?
 2. Beschreiben Sie den wesentlichen Unterschied zwischen formativem und summativem Testen und geben Sie je ein Beispiel.
 3. Wie könnte ein Test zur Diagnostik der Rechtschreibkompetenz von Schülern unter Bezug auf die fünf Validitätsaspekte untersucht werden?
 4. Welche generellen Unterschiede bestehen zwischen den Verfahren zum Standardsetting? Nennen Sie die Chapters der Standards for Educational and Psychological Testing, die Ihnen Informationen hierzu liefern können.
 5. Wodurch kann die Fairness eines Tests beschrieben werden?
 6. Warum ist ein Datenmanagementplan notwendig und was beinhaltet er?

Literatur

- Achtenhagen, F. (2004). Prüfung von Leistungsindikatoren für die Berufsbildung sowie zur Ausdifferenzierung beruflicher Kompetenzprofile nach Wissensarten. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *Bildungsreform Band 8: Expertisen zu den konzeptionellen Grundlagen für einen Nationalen Bildungsbericht – Berufliche Bildung und Weiterbildung /Lebenslanges Lernen* (S. 11–32). Bonn: BMBF.
- American Educational Research Association (AERA). (2011). Code of ethics. *Educational Researcher*, 40, 145–156.
- 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.

- 11**
- Anderson, L. W. & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508–600). Washington, DC: American Council on Education.
- Bachmann, H. (Hrsg.). (2014). *Forum Hochschuldidaktik und Erwachsenenbildung: Bd. 1. Kompetenzorientierte Hochschullehre: Die Notwendigkeit von Kohärenz zwischen Lernzielen, Prüfungsformen und Lehr-Lern-Methoden* (2. Aufl.). Bern: hep.
- Behnke, K. (2016). *Umgang mit Feedback im Kontext Schule: Erkenntnisse aus Analysen der externen Evaluation und des Referendariats. Psychologie in Bildung und Erziehung: Vom Wissen zum Handeln*. Wiesbaden: Springer.
- Biggs, J. B. (1999). What the Student does: teaching for enhanced learning. *Higher Education Research & Development*, 18, 57–75.
- Biggs, J. B. & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain* (Vol. 19). New York, NY: David McKay.
- Biggs, J. & Tang, C. (2011). *Teaching for Quality Learning at University*. Maidenhead: McGraw-Hill and Open University Press.
- Böttcher, W., Bos, W., Döbert, H. & Holtappels, H. G. (Hrsg.). (2008). *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive: Dokumentation zur Herbsttagung der Kommission Bildungsorganisation, -planung, -recht (KBBB)*. Münster, New York, München, Berlin: Waxmann.
- Brand, W., Hofmeister, W. & Tramm, T. (2005). Auf dem Weg zu einem Kompetenzstufenmodell für die berufliche Bildung – Erfahrungen aus dem Projekt ULME. Verfügbar unter http://www.bwpat.de/ausgabe8/brand_etabwpat8.pdf [20.12.2019]
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education*, 17, 59–88.
- Brückner, S., Zlatkin-Troitschanskaia, O. & Förster, M. (2014). Relevance of adaptation and validation for international comparative research on competencies in higher education – A methodological overview and example from an international comparative project within the KoKoHs research program. In F. Musekamp & G. Spötl (Eds.), *Competence in higher education and the working environment. National and international approaches for assessing engineering competence. Vocational education and training: Research and practice* (Vol. 12, pp. 133–152). Frankfurt am Main: Lang.
- Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22, 5–11.
- Crooks, T. J., Kane, M. T. & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3, 265–286.
- Crowe, A., Dirks, C. & Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sciences Education*, 7, 368–381.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer.
- Dubs, R. (1978). *Aspekte des Lehrerverhaltens. Theorie, Praxis, Beobachtung. Ein Beitrag zum Unterrichtsgespräch*. Aarau: Sauerländer.
- Fink, L. D. (2003). *Creating significant learning experiences: An integrated approach to designing college courses*. New York, NY: Routledge.
- Freunberger, R. & Yanagida, T. (2012). Kompetenzdiagnostik in Österreich: Der Prozess des Standard-Settings. *Psychologie in Österreich*, 32, 396–403.
- Gold, A. (2015). *Guter Unterricht: Was wir wirklich darüber wissen*. Bristol: Vandenhoeck & Ruprecht.
- Grotjahn, R. (2010). *Der C-Test: Beiträge aus der aktuellen Forschung. Language testing and evaluation* (Vol. 18). Frankfurt am Main: Peter Lang.
- Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18, 5–9.
- Harlen, W. & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4, 365–379.
- Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of Competencies in Educational Contexts*. Göttingen: Hogrefe.
- Hattie, J., Jaeger, R. M. & Bond, L. (1999). Chapter 11: Persistent methodological questions in educational testing. *Review of Research in Education*, 24, 393–446.
- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112.

Literatur

- Heinisch, I., Romeike, R. & Eichler, K.-P. (2016). Outcome-orientierte Neuausrichtung der Hochschullehre für das Fach Mathematik. In A. Hoppenbrock, R. Biehler, R. Hochmuth & H.-G. Rück (Hrsg.), *Konzepte und Studien zur Hochschuldidaktik und Lehrerbildung Mathematik. Lehren und Lernen von Mathematik in der Studieneingangsphase* (S. 261–275). Wiesbaden: Springer Spektrum.
- Heubert, J. P. & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.
- Institut zur Qualitätsentwicklung im Bildungswesen (IQB). (2013). *Kompetenzstufenmodelle zu den Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss*. Verfügbar unter https://www.iqb.hu-berlin.de/bista/ksm/KSM_Chemie.pdf [20.12.2019]
- Jensen, U. (2012). *Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten*. Köln: GESIS.
- Kaftandjieva, F. (2010). *Methods for Setting Cut Scores in Criterion-referenced Achievement Tests*. Arnhem: Cito.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al (2007). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise (Bildungsforschung Band 1)*. Berlin: BMBF. Verfügbar unter: http://edudoc.ch/record/33468/files/develop_standards_nat_form_d.pdf [20.12.2019]
- Klieme, E., Bürgermeister, A., Harks, B., Blum, W., Leiß, D. & Rakoczy, K. (2010). Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA1. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik* (S. 64–74). Weinheim: Beltz.
- Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich: Modellbasierte Testentwicklung und Validierung*. Dissertation. Empirische Berufsbildungs- und Hochschulforschung: Vol. 2. Landau: Verlag Empirische Pädagogik.
- Marzano, R. J. & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2nd ed.). Thousand Oaks, CA: Corwin Press.
- Metzger, C., Waibel, R., Henning, C., Hodel, M. & Luzi, R. (1993). *Anspruchsniveau von Lernzielen und Prüfungen im kognitiven Bereich*. St. Gallen: Institut für Wirtschaftspädagogik.
- Mislevy, R. J. (2016). How Developments in Psychology and Technology Challenge Validity Argumentation. *Journal of Educational Measurement*, 53, 265–292.
- Mislevy, R. J. & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Moosbrugger, M. (1985). Das Niveau der Aufgaben in Lehrbüchern: Eine Analyse österreichischer Geschichtsbücher für die Hauptschule. *Unterrichtswissenschaft*, 13, 116–129.
- Müller, A. & Ditton, H. (2014). Feedback: Begriff, Formen und Funktionen. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 11–28). Münster: Waxmann.
- Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. PISA, Paris: OECD Publishing.
- Overwien, P. D. B. (2005). Stichwort: Informelles Lernen. *Zeitschrift für Erziehungswissenschaft*, 8, 339–355.
- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.
- Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. Projekt Standardsetting. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik* (S. 175–188). Weinheim: Beltz.
- Pellegrino, J. W. (2010). *The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement*. Princeton: Educational Testing Service.
- Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: a meta analysis*. Dissertation. University of Iowa. Retrieved from <http://ir.uiowa.edu/cgi/viewcontent.cgi?article=2639&context=etd> [20.12.2019]
- Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.). (2016). *PISA 2015: Eine Studie zwischen Kontinuität und Innovation*. Münster, New York, NY: Waxmann.
- Rümpel, S. (2011). Der Lebenszyklus von Forschungsdaten. In S. Büttner, H.-C. Hobohm & L. Müller (Hrsg.), *Handbuch Forschungsdatenmanagement* (S. 25–34). Bad Honnef: Bock + Herchen.
- Walstad, W. B., Rebeck, K. & Butters, R. B. (2013). *Test of economic literacy: Examiners' manual* (4th ed.). New York, NY: National Council on Economic Education.
- Wild, E. & Möller, J. (Hrsg.). (2009). *Pädagogische Psychologie*. Berlin, Heidelberg: Springer.

- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., Hansen, M. & Happ, R. (2013). Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In *Lehrerbildung auf dem Prüfstand* (S. 108–133). Landau: Empirische Pädagogik e.V.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S. & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher Education Learning Outcomes Assessment: International Perspectives* (pp. 175–197). Frankfurt am Main: Peter Lang.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M. & Brückner, S. (2017). *Modeling and measuring competencies in higher education: Approaches to challenges in higher education policy and practice*. Wiesbaden: Springer VS.
- Zumbo, B. D. (2016). Standard-setting methodology: Establishing performance standards and setting cut-scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism*, 41, 74–82.

Testtheorien

Inhaltsverzeichnis

- Kapitel 12 Testtheorien im Überblick – 251**
Helfried Moosbrugger, Karin Schermelleh-Engel, Jana C. Gäde und Augustin Kelava
- Kapitel 13 Klassische Testtheorie (KTT) – 275**
Helfried Moosbrugger, Jana C. Gäde, Karin Schermelleh-Engel und Wolfgang Rauch
- Kapitel 14 Klassische Methoden der Reliabilitätsschätzung – 305**
Jana C. Gäde, Karin Schermelleh-Engel und Christina S. Werner
- Kapitel 15 Modellbasierte Methoden der Reliabilitätsschätzung – 335**
Karin Schermelleh-Engel und Jana C. Gäde
- Kapitel 16 Einführung in die Item-Response-Theorie (IRT) – 369**
Augustin Kelava und Helfried Moosbrugger
- Kapitel 17 Interpretation von Testwerten in der Item-Response-Theorie (IRT) – 411**
Dominique Rauch und Johannes Hartig
- Kapitel 18 Überblick über Modelle der Item-Response-Theorie (IRT) – 425**
Augustin Kelava, Stefano Noventa und Alexander Robitzsch
- Kapitel 19 Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT) – 447**
Norman Rose
- Kapitel 20 Computerisiertes adaptives Testen – 501**
Andreas Frey



Testtheorien im Überblick

Helfried Moosbrugger, Karin Schermelleh-Engel, Jana C. Gäde und Augustin Kelava

Inhaltsverzeichnis

- 12.1 Einleitung – 252**
- 12.2 Klassische Testtheorie (KTT) – 254**
 - 12.2.1 Annahmen der KTT – 254
 - 12.2.2 Itemcharakteristik und Spezifische Objektivität – 255
 - 12.2.3 Testwertvariablen – 257
 - 12.2.4 Reliabilität – 257
 - 12.2.5 Modelltests – 258
 - 12.2.6 Eindimensionale Messmodelle – 259
 - 12.2.7 Mehrdimensionale Messmodelle – 260
- 12.3 Item-Response-Theorie (IRT) – 260**
 - 12.3.1 Annahmen der IRT – 261
 - 12.3.2 Itemcharakteristische Funktion und Spezifische Objektivität – 261
 - 12.3.3 Summenscores – 263
 - 12.3.4 Reliabilität – 264
 - 12.3.5 Modelltests – 264
 - 12.3.6 Eindimensionale IRT-Modelle – 265
 - 12.3.6.1 Modelle für separierbare Parameter – 265
 - 12.3.6.2 Modelle für nicht separierbare Modellparameter – 267
 - 12.3.7 Mehrdimensionale IRT-Modelle – 267
- 12.4 Klassische Testtheorie (KTT) vs. Item-Response-Theorie (IRT) – 268**
 - 12.4.1 Wesentliche Charakteristika der KTT und der IRT – 268
 - 12.4.2 Übergreifendes Konzept – 270
- 12.5 Zusammenfassung – 271**
- 12.6 Kontrollfragen – 271**
- Literatur – 271**

i Die am häufigsten verwendeten Testtheorien sind die Klassische Testtheorie (KTT) und die Item-Response-Theorie (IRT). Beide Theorien verfolgen sehr ähnliche Ziele bei der Konstruktion von Testverfahren zur Messung individueller Merkmalsausprägungen. Die KTT ist primär für Testitems mit kontinuierlichem (oder zumindest vielstufigem) Antwortformat konzipiert und konzentriert sich auf die Gewinnung von wahren Merkmalsausprägungen (True-Scores) sowie auf die Reliabilität und Validität der Testwerte. Die IRT hingegen wurde primär für Testitems mit dichotomen (und polytom geordneten) Antwortkategorien entwickelt und hat ihren Schwerpunkt auf der Schätzung latenter Personenparameter und Itemparameter, um Rückschlüsse auf interessierende Fähigkeitsmerkmale (seltener Einstellungs- und Persönlichkeitsmerkmale) und Itemcharakteristika zu ziehen. In den letzten Jahrzehnten näherten sich die KTT und die IRT jedoch zunehmend aneinander an, sodass sie inzwischen viele Gemeinsamkeiten aufweisen. Mit der KTT und der IRT liegen somit zwei Testtheorien vor, die sich in der Vergangenheit bestens bewährt haben. Beide Theorien ergänzen sich vorteilhaft, können aber auch als Spezialfälle eines gemeinsamen Modells aufgefasst werden. Beide Theorien weisen als Spezialfälle jedoch immer noch einige charakteristische Unterschiede auf, sodass sie ihre eigenständigen Berechtigungen haben.

12.1 Einleitung

Mit der Klassischen Testtheorie (KTT) und der jüngeren Item-Response-Theorie (IRT) liegen zwei bewährte Testtheorien vor, die sich bei der Konstruktion von Testverfahren und der Interpretation von Testwerten vorteilhaft ergänzen.

Sowohl die KTT (► Kap. 13) als auch die IRT (► Kap. 16) postulieren, dass die Beantwortung der Aufgaben (Items) eines Tests von latenten Einstellungs-, Persönlichkeits- oder Fähigkeitsmerkmalen abhängt, die das Testverhalten bestimmen, aber nicht direkt beobachtbar sind. Die Itemantworten werden als beobachtbares Verhalten in den manifesten Itemvariablen erfasst und stellen Indikatoren für die individuelle Ausprägung des zugrunde liegenden Merkmals (Konstrukt, latente Variable) dar oder – im mehrdimensionalen Fall – mehrerer zugrunde liegender Merkmale.

Die IRT wurde primär für Testitems mit dichotomen (und polytom geordneten) Antwortkategorien entwickelt und hat ihren Schwerpunkt auf der Schätzung latenter Personenparameter und Itemparameter, um Rückschlüsse auf interessierende Einstellungs-, Persönlichkeits- oder Fähigkeitsmerkmale zu ziehen.

Die KTT ist primär für Testitems mit kontinuierlichem Antwortformat konzipiert und konzentriert sich bei der Messung individueller Merkmalsausprägungen auf die Gewinnung von Testwerten zur Schätzung der wahren Werte (True-Scores) sowie deren Reliabilität und Validität. Unter Verwendung geeigneter Messmodelle können – in Analogie zur IRT – Personenparameter in Form von Faktorwerten (Factor-Scores) sowie Itemleichtigkeitsparameter geschätzt werden.

Im letzten Jahrhundert war die KTT in der psychologischen Forschung zunächst die bestimmende Testtheorie, auf deren Grundlage ein- und mehrdimensionale Messinstrumente zur Messung ein- oder mehrdimensionaler Merkmale entwickelt wurden. Die Popularität der KTT resultierte u. a. aus der sehr eingängigen Vorstellung, dass sich ein manifestes Itemwert aus einem wahren Wert und einem Fehlerwert zusammensetzt. Eine wesentliche Annahme dieser Theorie bestand darin, dass die Beziehungen zwischen kontinuierlichen beobachtbaren Itemvariablen und einer kontinuierlichen latenten Variablen (Fähigkeit, Einstellung, Persönlichkeitsmerkmal) linear sind, sodass für die Bestimmung und Interpretation der Kennwerte der KTT keine vertieften statistischen Kenntnisse nötig waren. Nach Aufsummierung der einzelnen Itemwerte zu Test(summen)werten war die Schätzung der Reliabilität und Validität der Messungen relativ einfach. Ein Nachteil war da-

12.1 · Einleitung

durch gegeben, dass die auf der KTT basierenden Kennwerte auf strengen Modellannahmen beruhen, deren Gültigkeit zunächst nicht überprüft werden konnte.

Weil die KTT primär für kontinuierliche Antwortformate und lineare Beziehungen zwischen den Itemvariablen und den latenten Konstrukten konzipiert war, wurde mit der IRT eine weitere Testtheorie entwickelt, die ihren Fokus vor allem auf dichotome oder kategorial gestufte Itemvariablen legte und für deren Zusammenhang mit den latenten Konstrukten probabilistische Modelle formulierte. Wege[n] der zugrunde gelegten kurvilinearen, zumeist logistischen Zusammenhänge erforderte die Schätzung der Modellparameter (Kennwerte für die Schwierigkeit und die Trennschärfe der Items sowie für die Merkmalsausprägungen der Personen) fundierte statistische Kenntnisse, sodass die IRT-Modelle zunächst seltener zur Anwendung kamen als die Modelle der KTT. Durch die Entwicklung spezieller Statistikprogramme für die IRT und durch die verbesserte universitäre Methodenausbildung nahm der Einsatz der IRT für die Testentwicklung in der späten zweiten Hälfte des 20. Jahrhunderts jedoch sprunghaft zu. Zudem hatte die IRT den Vorteil, dass die modellimplizierten Annahmen der IRT-Modelle hinsichtlich ihres Zutreffens mit speziellen Modelltests überprüft und somit – falls nicht zutreffend – auch verworfen werden konnten. Diese Tatsache trug wohl am meisten dazu bei, dass die IRT über Jahre als die elaboriertere der beiden Testtheorien angesehen wurde.

Erst relativ spät setzte sich in der breiten Anwendercommunity die Erkenntnis durch, dass auch die Modellannahmen der KTT (► Kap. 13; Steyer und Eid 2001) – analog zur IRT – in Form von Messmodellen formuliert und anhand der konfirmatorischen Faktorenanalyse (CFA, ► Kap. 24) hinsichtlich ihres Zutreffens überprüft werden können. Zuvor kamen vornehmlich die struktursuchende exploratorische Faktorenanalyse (EFA, ► Kap. 23) zum Einsatz, mit der eine Überprüfung der Modellannahmen nicht möglich war.

Eine Ursache für die zum Teil noch immer geäußerte Kritik an „klassisch“ konstruierten Tests kann darin gesehen werden, dass die im Rahmen der KTT entwickelten Methoden der Reliabilitätsschätzung oftmals unkritisch und folglich nicht korrekt angewandt wurden, weil die jeweiligen modelltheoretischen Implikationen verschiedener Reliabilitätskoeffizienten keine Berücksichtigung fanden. Inzwischen setzt sich jedoch zunehmend auch im Rahmen der KTT die Überprüfung der Messmodelle durch, um die Charakteristika der Items, insbesondere die Schwierigkeit (in der KTT eigentlich Leichtigkeit) und die Trennschärfe nicht nur deskriptivstatistisch (► Kap. 7), sondern auch als Parameter der Messmodelle zu bestimmen. Darauf aufbauend lassen sich dann adäquate Reliabilitätsmaße schätzen (vgl. ► Kap. 14 und 15).

Inzwischen ist es sowohl für die KTT als auch für die IRT möglich, die modellimplizierten Annahmen mit geeigneten Modelltests hinsichtlich ihres Zutreffens empirisch zu überprüfen und ggf. auch zu verwerfen. Der lange konstatierte Vorteil der IRT gegenüber der KTT erweist sich somit als obsolet und die immer noch vernehmbare Kritik an der KTT als ungerechtfertigt, da sie bei einer korrekten Anwendung der KTT keinen Bestand hat (Eid und Schmidt 2014; Raykov und Marcoulides 2011, 2016; Steyer und Eid 2001).

In den letzten 30 Jahren wurden die Beziehungen zwischen den Grundlagen der KTT und der IRT mehrfach untersucht. Hierbei wurden viele Übereinstimmungen gefunden (vgl. Kamata und Bauer 2008; McDonald 1999; Takane und de Leeuw 1987), in denen die engen Beziehungen zwischen der KTT und der IRT deutlich werden (Kohli et al. 2015; Raykov und Marcoulides 2016; Raykov et al. 2019).

Nachfolgend sollen die beiden Testtheorien kurz vorgestellt und der Schwerpunkt auf einige wesentliche Übereinstimmungen gelegt werden. Die eigenständigen Charakteristika werden in den nachfolgenden ► Kap. 13 bis 19 ausführlich behandelt; hierbei wird auch auf methodische Ansätze zugegriffen, die in den ► Kap. 22 bis 27 näher erläutert werden.

IRT: logistische Zusammenhänge zwischen kategorialen Itemvariablen und einer latenten Variablen

Kritik an der KTT nicht mehr gerechtfertigt

Übereinstimmungen zwischen KTT und IRT

12.2 Klassische Testtheorie (KTT)

Die KTT (► Kap. 13) stellt seit mehr als 70 Jahren die theoretische Basis zur Konstruktion psychodiagnostischer Tests und zur Interpretation der resultierenden Testergebnisse dar. Erste Grundlagen der KTT finden sich bereits bei Spearman (1904a, 1904b), die Prinzipien der KTT wurden aber erst von Gulliksen (1950), Lord und Novick (1968) sowie Zimmerman (1975, 1976) entwickelt und von Steyer (1989) sowie Steyer und Eid (2001) formalisiert und weiter ausgearbeitet.

Aufbauend auf der KTT verfügt die psychometrische Diagnostik über sehr bewährte Ansätze zur Beurteilung der Reliabilität und Validität von Tests und Messverfahren. Die relativ ökonomische und praktikable Umsetzung der testtheoretischen Anforderungen ist ein wesentlicher Grund dafür, dass sich die KTT in hohem Maße durchgesetzt hat.

Messfehlertheorie

Die KTT ist im Wesentlichen eine Messfehlertheorie (Eid und Schmidt 2014; Lord und Novick 1968; Steyer und Eid 2001; Zimmerman 1976). Mit dieser Bezeichnung kommt bereits zum Ausdruck, dass Messungen in der Regel mit einem Fehler behaftet sind und dass ein wesentliches Ziel der KTT darin besteht, den Messfehler vom wahren Wert (True-Score) zu trennen.

In der KTT ist der wahre Wert τ_{vi} einer Person v für Item i als Erwartungswert einer intraindividuellen Verteilung von Itemwerten y_{vi} definiert. Die Definition des wahren Wertes impliziert verschiedene Eigenschaften der Messfehlervariablen und der True-Score-Variablen, die in empirischen Anwendungen nicht falsch sein können und deshalb nicht überprüft werden müssen (Eid und Schmidt 2014; Steyer und Eid 2001; Zimmerman 1975, 1976; s. auch ► Kap. 13).

■■ Grundgleichung der KTT

Grundgleichung der KTT

Auf der Ebene der Items nimmt die Grundgleichung der KTT jeweils eine Aufteilung der beobachteten Itemvariablen in eine True-Score-Variable und eine Fehlervariable vor. Die Grundgleichung der KTT für ein Item y_i ($i = 1, \dots, p$) lautet somit (vgl. ► Kap. 13):

$$y_i = \tau_i + \varepsilon_i, \quad (12.1)$$

wobei y_i die Itemvariable (Variable mit den Antworten der untersuchten Personen für Item i), τ_i die True-Score-Variable (Variable der wahren Werte) und ε_i die Messfehlervariable (Variable der Fehlerwerte) bezeichnet.

Die Aufteilung eines beobachteten Wertes in einen wahren Anteil und einen Fehleranteil sowie die damit verbundenen Eigenschaften der Messfehler- und True-Score-Variablen wurden früher als „Axiome“ bezeichnet. Heute wird nicht mehr von „Axiomen“ gesprochen, da sich die Ansicht durchgesetzt hat, dass aus der Definition des wahren Wertes als Erwartungswert einer intraindividuellen Verteilung von Itemwerten alle weiteren Eigenschaften der True-Score- und der Fehlervariablen abgeleitet werden können (Eid und Schmidt 2014; Lord und Novick 1968; Steyer und Eid 2001).

12.2.1 Annahmen der KTT

Die Grundgleichung der KTT stellt insoweit kein Modell im engeren Sinn dar, da die Grundgleichung für sich alleine nicht überprüfbar ist. Erst die Verwendung mehrerer Items zur Messung desselben Konstrukt eröffnet die Definition eines Messmodells und die Bestimmung der Kennwerte der KTT.

Bei der Definition des Messmodells wird berücksichtigt, dass das Antwortverhalten der Testpersonen auf die Items nicht nur von der Ausprägung des latenten Konstrukt, sondern auch von den Messeigenschaften des jeweiligen Items abhängt. Diese Zusammenhänge können in einer Messmodellgleichung (Gl. 12.2)

bzw. 12.3) ausgedrückt werden, in der zusätzliche Modellparameter (z. B. Itemschwierigkeit bzw. -leichtigkeit, Itemtrennschärfe) die Messeigenschaften der Items abbilden; sie findet sich analog auch in der IRT (► Abschn. 12.3).

Die implizierten Modellannahmen können hinsichtlich ihres Zutreffens mit Modelltests (► Abschn. 12.2.5) überprüft und somit ggf. auch falsifiziert werden. Typischerweise wird angenommen, dass alle Itemvariablen dasselbe Konstrukt (dieselbe latente Variable) messen und die Messfehler unkorreliert sind (Annahme der Eindimensionalität). Des Weiteren wird in der Regel angenommen, dass die Beziehungen zwischen den beobachtbaren (manifesten) Itemvariablen und der latenten Variablen linear sind.

Unter der Annahme der Eindimensionalität lässt sich somit für alle Items eine gemeinsame latente Variable η definieren, die den Ausprägungen der True-Score-Variablen τ_i ($i = 1, \dots, p$) und somit auch den Ausprägungen der manifesten Itemvariablen y_i zugrunde liegt.

Die True-Score-Variablen τ_i eines Items i kann als Summe eines Leichtigkeitsparameters α_i (in umgekehrter Polung auch als „Schwierigkeitsparameter“ bezeichnet) und der mit dem Trennschärfe-/Diskriminationsparameter λ_i gewichteten latenten Variablen η dargestellt werden (Eid und Schmidt 2014; s. auch ► Kap. 13):

$$\tau_i = \alpha_i + \lambda_i \cdot \eta \quad (12.2)$$

Durch Einsetzen von Gl. (12.2) in Gl. (12.1) erweitert sich die Grundgleichung der KTT wie folgt:

$$y_i = \alpha_i + \lambda_i \cdot \eta + \varepsilon_i \quad (12.3)$$

12.2.2 Itemcharakteristik und Spezifische Objektivität

Die in Gl. (12.2) dargestellte lineare Beziehung zwischen der latenten Variablen η und der True-Score-Variablen τ_i kann – in Analogie zur IRT (► Abschn. 12.3) – für jedes Item in Form einer Itemcharakteristik ausgedrückt werden, wobei der Diskriminationsparameter λ_i die Steilheit und der Leichtigkeitsparameter α_i die Lage der Itemcharakteristiken entlang der Ordinate abbildet (► Abb. 12.1). Je größer der Wert von λ_i ist, desto steiler ist die Itemcharakteristik und desto besser („schärfer“) differenziert/trennt das Item zwischen Personen mit einer hohen und Personen mit einer niedrigen Ausprägung im latenten Merkmal η . Je größer bzw. kleiner α_i ist, desto leichter bzw. schwieriger ist die Antwort auf das Item bei gegebener Ausprägung in η im Sinne des Konstruktts (d. h. richtig/zustimmend/symptomatisch).

Sofern alle Items – wenn auch auf unterschiedlichen Anforderungs-/Leichtigkeitsniveaus – gleich gut zur Differenzierung zwischen den Merkmalsausprägungen geeignet sind, verlaufen die Itemcharakteristiken parallel (► Abb. 12.1). Die Itemcharakteristiken unterscheiden sich dann nicht in ihrer Steilheit λ_i (die Diskriminationsparameter weisen für alle Items denselben Wert auf), wohl aber in ihrer Leichtigkeit α_i , da die Anforderungen der Items an die Testpersonen unterschiedlich sind (Eid und Schmidt 2014). Die Parallelität der Itemcharakteristiken wird z. B. in den Modellen essentiell τ -äquivalenter oder essentiell τ -paralleler Itemvariablen vorausgesetzt (zu den verschiedenen Modellen s. ► Tab. 12.1, ► Abschn. 12.2.6 sowie ► Kap. 13).

In ► Abb. 12.1 wurde der Erwartungswert von η zur Normierung auf null fixiert, sodass die Leichtigkeitsparameter den Erwartungswerten der Itemvariablen entsprechen:

$$\begin{aligned} E(y_i) &= E(\tau_i + \varepsilon_i) \\ &= E(\tau_i) + E(\varepsilon_i) \\ &= E(\alpha_i + \lambda_i \cdot \eta) + E(\varepsilon_i) \\ &= E(\alpha_i) + \lambda_i \cdot E(\eta) + E(\varepsilon_i) \\ &= \alpha_i + \lambda_i \cdot 0 + 0 = \alpha_i \end{aligned} \quad (12.4)$$

Gemeinsame latente Variable η

Beziehung zwischen der True-Score-Variablen τ_i und der latenten Variablen η

Itemcharakteristik

Tabelle 12.1 Charakteristika ausgewählter eindimensionaler KTT-Modelle der Messäquivalenz mit Angabe der adäquaten Methoden der Reliabilitätsschätzung

Modell	Charakteristika (Messeigenschaften)
τ-kongenerisches Modell	<ul style="list-style-type: none"> – Voraussetzung: Eindimensionalität – Diskriminationsparameter (Faktorladungen): frei geschätzt – Leichtigkeitsparameter (Interzepte): frei geschätzt – Fehlervarianzen: frei geschätzt – Adäquates Reliabilitätsmaß: McDonalds Omega
Essentiell τ-äquivalentes Modell	<ul style="list-style-type: none"> – Voraussetzung: Eindimensionalität – Diskriminationsparameter: für alle Items identisch (parallele Itemcharakteristiken) – Leichtigkeitsparameter: frei geschätzt – Fehlervarianzen: frei geschätzt – Adäquate Reliabilitätsmaße: Cronbachs Alpha und McDonalds Omega
Essentiell τ-paralleles Modell	<ul style="list-style-type: none"> – Voraussetzung: Eindimensionalität – Diskriminationsparameter: für alle Items identisch (parallele Itemcharakteristiken) – Leichtigkeitsparameter: frei geschätzt – Fehlervarianzen: für alle Items identisch – Adäquate Reliabilitätsmaße: Spearman-Brown-Formel der Testverlängerung, Cronbachs Alpha und McDonalds Omega

Anmerkung: Zur Normierung wurde hier die latente Variable η standardisiert (vgl. ► Kap. 24).

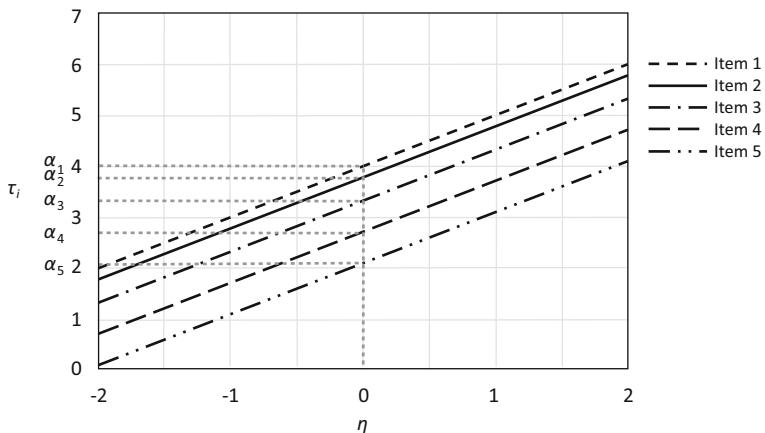


Abb. 12.1 Parallel Itemcharakteristiken im essentiell τ -äquivalenten Modell: Abhängigkeit der True-Score-Variablen τ_i von der Ausprägung der latenten Variablen η bei konstanten Diskriminationsparametern λ_i , und unterschiedlichen Leichtigkeitsparametern α_i der fünf Items. Die Leichtigkeitsparameter α_i entsprechen den Mittelwerten der Itemvariablen: $\alpha_1 = 4.0$, $\alpha_2 = 3.8$, $\alpha_3 = 3.3$, $\alpha_4 = 2.7$ und $\alpha_5 = 2.1$

Da der Erwartungswert einer Konstanten gleich der Konstanten ist ($E(\alpha_i) = \alpha_i$), der Erwartungswert der Fehlervariablen per Definition null sein muss ($E(\varepsilon_i) = 0$) und der Erwartungswert von η auf null fixiert wurde ($E(\eta) = 0$), folgt, dass der Erwartungswert einer Itemvariablen gleich dem Leichtigkeitsparameter ist.

Liegt Modellkonformität für essentielle τ -Äquivalenz oder essentielle τ -Parallellität vor, ist davon auszugehen, dass die Items in ► Abb. 12.1 die Eigenschaft für spezifisch objektive Vergleiche besitzen (zur Überprüfung der Modellkonformität anhand von Modelltests ► Abschn. 12.2.5 sowie ► Kap. 13 und 24). Für einen Pool modellkonformer Items bedeutet diese Eigenschaft, dass sowohl der Vergleich zweier Personen hinsichtlich ihrer latenten Merkmalsausprägungen nicht von den ausgewählten Items abhängt als auch der Vergleich zweier Items hinsichtlich ih-

Spezifische Objektivität der Vergleiche setzt mindestens eine essentielle τ -Äquivalenz voraus

rer Leichtigkeitsparameter nicht von den ausgewählten Personen abhängt (Eid und Schmidt 2014; s. auch ► Kap. 13).

12.2.3 Testwertvariablen

In der KTT liegt der Fokus in der Regel auf den Testwertvariablen. Die Werte der Testwertvariablen werden gewonnen, indem die individuellen Itemwerte y_{vi} von Person v in mehreren Items i , die dasselbe Merkmal messen, zu einem Summenscore aufsummiert werden, der als Testwert Y_v der Person v verwendet wird. Für den Erwartungswert von Y_v gilt $E(Y_v) = \hat{T}_v$, wobei \hat{T}_v der geschätzte True-Score der Testwertvariablen Y einer Person v ist (vgl. hierzu ► Kap. 2 und 13). Die Testwerte können dazu verwendet werden, Testpersonen untereinander sowie auch mit den Werten einer Normpopulation zu vergleichen (► Kap. 9).

Das ungewichtete Aufsummieren der Itemwerte zu Testwerten ist vor allem dann angemessen, wenn anhand von Modelltests (► Abschn. 12.2.5) festgestellt wurde, dass die strengen messtheoretischen Voraussetzungen der Eindimensionalität (essentielle τ -Äquivalenz oder essentielle τ -Parallelität) und der Unkorreliertheit der Messfehler erfüllt sind.

Im Zusammenhang mit der Reliabilitätsbestimmung (► Abschn. 12.2.4) wird dargelegt, unter welchen Bedingungen die ungewichtete Aufsummierung der Items suboptimal ist und wie in solchen Fällen vorgegangen werden sollte; sind nur weniger strenge Äquivalenzannahmen erfüllt, liefern die Summenscores dennoch wesentliche Informationen über die individuellen Merkmalsausprägungen.

Summenscores dienen als Testwerte

Für die Summenbildung müssen messtheoretische Voraussetzungen überprüft werden

12.2.4 Reliabilität

Da in den Testwerten auch die Messfehler der einzelnen Itemvariablen enthalten sind, ist die Schätzung des Messfehleranteils in den Testwerten anhand der Reliabilität notwendig.

Die Reliabilität als Maß der Messgenauigkeit einer Testwertvariablen Y ist definiert als Verhältnis der Varianz der True-Score-Variablen T zur Varianz der Testwertvariablen Y :

$$Rel(Y) = \frac{Var(T)}{Var(Y)} = \frac{Var(T)}{Var(T) + Var(E)} \quad (12.5)$$

Die True-Score-Varianz $Var(T)$ setzt sich zusammen aus den aufsummierten True-Score-Varianzen der Itemvariablen und den Kovarianzen zwischen den True-Score-Variablen der Itemvariablen (vgl. ► Kap. 13 und 14). Wird hier ein τ -kongenerisches Messmodell mit unterschiedlichen Diskriminationsparametern der Itemvariablen mit $\tau_i = \alpha_i + \lambda_i \cdot \eta$ verwendet (Gl. 12.2), so ergibt sich die folgende Varianzaufteilung, wobei der Leichtigkeitsparameter α_i nicht berücksichtigt werden muss, weil dessen Varianz gleich null ist:

$$\begin{aligned} Var(T) &= \sum_{i=1}^p Var(\tau_i) + 2 \cdot \sum_{i < i'} Cov(\tau_i, \tau_{i'}) \\ &= \sum_{i=1}^p Var(\lambda_i \cdot \eta) + 2 \cdot \sum_{i < i'} Cov(\lambda_i \cdot \eta, \lambda_{i'} \cdot \eta) \\ &= \sum_{i=1}^p \lambda_i^2 \cdot Var(\eta) + 2 \cdot \sum_{i < i'} \lambda_i \cdot \lambda_{i'} \cdot Cov(\eta, \eta) \\ &= \sum_{i=1}^p \lambda_i^2 \cdot Var(\eta) + 2 \cdot \sum_{i < i'} \lambda_i \cdot \lambda_{i'} \cdot Var(\eta) \end{aligned} \quad (12.6)$$

Klassische und modellbasierte Methoden der Reliabilitätsschätzung

Konfidenzintervalle um geschätzte wahre Werte \hat{T}_v

Nachteile des Summierungsverfahrens

Direkte Schätzung latenter Personenwerte (Factor-Scores)

Beurteilung der Modellgüte

Wird die latente Varianz aus Normierungsgründen auf eins fixiert, so verdeutlicht Gl. (12.6) auch, dass die True-Score-Varianz $Var(T)$ über Aufsummierung der quadrierten Faktorladungen sowie über Aufsummierung der Produkte der Faktorladungen aller Itemvariablen berechnet wird.

Zur Schätzung des Anteils der True-Score-Varianz an der Gesamtvarianz der Testwertvariablen Y werden klassische und modellbasierte Methoden der Reliabilitätsschätzung unterschieden (► Kap. 14 und 15). Die klassischen Reliabilitätsmaße (Cronbachs Alpha, Spearman-Brown-Formel der Testverlängerung) beruhen auf sehr restriktiven Modellannahmen hinsichtlich der Messäquivalenz der Items (► Abschn. 12.2.6), die zunächst anhand von Modelltests (► Abschn. 12.2.5) überprüft werden müssen. Da diese strengen Voraussetzungen empirisch oft nicht gegeben sind, werden als Alternative modellbasierte Reliabilitätsmaße wie die verschiedenen Omega-Koeffizienten (► Kap. 15) empfohlen, die auf weniger strengen Annahmen beruhen.

Mithilfe der Reliabilität kann der *Standardmessfehler* berechnet/geschätzt werden, mit dem die Bildung von *Konfidenzintervallen* um die geschätzten True-Scores \hat{T}_v jeder Person möglich ist (► Kap. 13). Das Konfidenzintervall umfasst den Wertebereich, in dem sich 95 bzw. 99 % aller möglichen wahren Werte befinden, die den geschätzten wahren Wert \hat{T}_v erzeugt haben könnten (vgl. Bortz und Döring 2006, S. 414 ff.).

Die beobachteten Testwerte Y_v stellen nur dann genaue Schätzungen der latenten Personenwerte η_v dar, wenn die Itemvariablen essentielle τ -Parallelität (vgl. Eid und Schmidt 2014, S. 294), also gleiche True-Score- und gleiche Fehleranteile aufweisen. In diesem Fall würden die beobachteten Testwerte Y_v als Punktschätzungen der wahren Werte T_v eine lineare Funktion der geschätzten Personenwerte der latenten Variablen η_v darstellen. Die Schätzungen anhand von Testwerten sind jedoch oftmals nicht optimal, weil die Testwerte einerseits Messfehleranteile beinhalten und andererseits auf einer ungewichteten Aufsummierung der Itemwerte beruhen.

Da Items aufgrund der unterschiedlichen Iteminhalte jedoch nur selten essentiell τ -parallel oder zumindest essentiell τ -äquivalent sind, bietet sich die Möglichkeit an, die latenten Personenwerte η_v direkt anhand der Factor-Scores der latenten Merkmalsvariablen η zu schätzen. Hier können verschiedene Methoden zur Factor-Score-Schätzung herangezogen werden, wobei eine häufig verwendete Methode die Regressionsmethode ist (vgl. Eid und Schmidt 2014, S. 293). In diesem Fall wird empfohlen, die Genauigkeit der Schätzungen – analog zur IRT – auch für die direkt geschätzten latenten Personenwerte η_v mithilfe ihrer Standardfehler durch Bildung von Konfidenzintervallen zu bestimmen.

Eid und Schmidt (2014, S. 292) betonen, dass im Rahmen der KTT „das Konzept der Reliabilität auf die Schätzungen der latenten Personenwerte übertragen werden kann. Diese Reliabilität spiegelt wider, inwieweit sich Unterschiede in den geschätzten Personenwerten auf wahre Personenunterschiede zurückführen lassen“.

12.2.5 Modelltests

Zur Beurteilung der Modellkonformität (Modellgüte, Modellfit) können als Gütemaße ein inferenzstatistischer Modelltest, der χ^2 -Test, sowie mehrere deskriptive Gütemaße herangezogen werden (► Kap. 24). Weisen die Ergebnisse auf einen guten Modellfit hin, passt das Modell zu den Daten und die einzelnen Modellparameter dürfen interpretiert und dazugehörige Einzelhypothesen geprüft werden. Nur bei zufriedenstellendem Modellfit dürfen Personenwerte, Reliabilitätskoeffizienten und Konfidenzintervalle bestimmt und interpretiert werden.

Des Weiteren kann die Gültigkeit des gewählten Messmodells in verschiedenen Subpopulationen oder zu verschiedenen Messzeitpunkten geprüft werden, indem die Itemparameter über die Gruppen oder die Messzeitpunkte hinweg als gleich (invariant) definiert werden (Messinvarianz, vgl. ▶ Kap. 24). Das Modell mit den Gleichheitsrestriktionen wird mit dem Modell ohne Gleichheitsrestriktionen (d.h. mit frei geschätzten Parametern) mittels χ^2 -Differenztest verglichen, wobei ein nicht signifikantes Ergebnis für die Messinvarianz sprechen würde.

Messinvarianz

12.2.6 Eindimensionale Messmodelle

Basierend auf der KTT lassen sich verschiedene Messmodelle formulieren, die unterschiedlich restriktive Anforderungen an die Messäquivalenz der Items stellen (Eid und Schmidt 2014; Steyer und Eid 2001; s. auch ▶ Kap. 13 und 14). Im Wesentlichen lassen sich die Modelle der τ -Kongenerität, der essentiellen τ -Äquivalenz und der essentiellen τ -Parallelität der Messungen unterscheiden (► Tab. 12.1). Diese Messmodelle sind Spezialfälle des allgemeinen faktorenanalytischen Modells (Eid und Schmidt 2014, S. 249) und können anhand der CFA (► Kap. 24) hinsichtlich des Zutreffens ihrer Modellannahmen beurteilt werden.

Anforderungen an die Messäquivalenz

Im eindimensionalen Fall beruhen die Messmodelle auf der Annahme, dass alle Items dieselbe latente Variable messen und die Fehlervariablen unkorreliert sind, wobei die Annahme der unkorrelierten Fehlervariablen in begründeten Fällen gelockert werden kann (s. Methodeneffekte, ► Kap. 24; s. Bollens Omega, ► Kap. 15). Des Weiteren wird meist angenommen, dass die manifesten Itemvariablen und die latenten Variablen kontinuierlich sind und lineare Beziehungen aufweisen.

In ► Tab. 12.1 sind die Spezialfälle des eindimensionalen Messmodells mit ihren unterschiedlich strengen Restriktionen dargestellt. Anhand von Modelltests kann entschieden werden, ob die jeweiligen Modellannahmen hinsichtlich der Gleichheit bzw. Ungleichheit der Diskriminationsparameter (Faktorladungen), der Leichtigkeitsparameter (Interzepte) und der Fehlervarianzen der Itemvariablen auf die empirischen Daten zutreffen.

Die in ► Tab. 12.1 aufgeführten Modelle setzen voraus, dass alle Itemvariablen y_i genau eine gemeinsame latente Variable η messen (Eindimensionalität der Items). Im Modell τ -kongenerischer Itemvariablen können sich die Items hinsichtlich ihrer Messeigenschaften (Diskriminations- und Leichtigkeitsparameter sowie Fehlervarianzen) unterscheiden. Wird dagegen angenommen, dass einzelne Messeigenschaften der Items gleich sind, resultieren die strengeren Modelle essentiell τ -äquivalenter oder essentiell τ -paralleler Variablen.

In diesen Modellen wird angenommen, dass die Diskriminationsparameter λ_i der Items invariant (identisch) sind und die Itemcharakteristiken damit parallel verlaufen (vgl. ► Abb. 12.1). Diese Annahme impliziert, dass die Items das zugrunde liegende latente Merkmal η im gleichen Ausmaß messen; hinsichtlich ihrer Leichtigkeit (Interzept α_i) können sich die Items jedoch unterscheiden. Werden auch die Fehlervarianzen der Itemvariablen als identisch angenommen, so bedeutet dies, dass sich die manifesten Variablen nicht in ihren Varianzen unterscheiden, und dass somit alle Itemvariablen identische Eigenschaften bei der Aufsummierung zur Testwertvariablen und Schätzung des wahren Wertes mitbringen (vgl. Eid und Schmidt 2014, S. 308).

Essentielle τ -Äquivalenz

Für ein essentiell τ -äquivalentes Modell, in dem die Diskriminationsparameter λ_i aller Items auf den Wert eins fixiert wurden, würde sich die Messmodellgleichung (Gl. 12.3) wie folgt vereinfachen:

$$y_i = \tau_i + \varepsilon_i = \alpha_i + 1 \cdot \eta + \varepsilon_i \quad (12.7)$$

Ein Vorteil des essentiell τ -äquivalenten Modells besteht darin, dass spezifisch objektive Vergleiche von Personen und Items möglich sind, da der Vergleich zweier

Vorteil der Spezifischen Objektivität

Adäquate Methoden der Reliabilitätsschätzung

Personen bezüglich ihrer geschätzten Personenwerte nicht von den ausgewählten Items abhängt und ebenso auch nicht der Vergleich zweier Items bezüglich ihrer Leichtigkeitsparameter nicht von den ausgewählten Personen (Eid und Schmidt 2014; s. auch ► Kap. 16).

Die Überprüfung des Zutreffens der unterschiedlich restriktiven Modellannahmen mittels Modelltests ist auch deshalb erforderlich, um eine adäquate Methode der Reliabilitätsschätzung zu wählen. Wurde z. B. essentielle τ -Äquivalenz nachgewiesen, so kann die Reliabilität anhand von Cronbachs Alpha bestimmt werden. Wird zusätzlich nachgewiesen, dass auch die Fehlervarianzen gleich sind, kann auch die Spearman-Brown-Formel der Testverlängerung (Brown 1910; Spearman 1910) zur Reliabilitätsschätzung verwendet werden (vgl. ► Kap. 14). Weist die Modellüberprüfung dagegen darauf hin, dass sich die Diskriminationsparameter und die Fehlervarianzen der Items unterscheiden, so darf die Reliabilität weder anhand der Spearman-Brown-Formel der Testverlängerung noch anhand von Cronbachs Alpha, sondern nur anhand von McDonalds Omega bestimmt werden (vgl. ► Kap. 15).

12.2.7 Mehrdimensionale Messmodelle

Neben den eindimensionalen Modellen wurden inzwischen auch verschiedene mehrdimensionale Modelle entwickelt, die auf der KTT aufbauen und explizit mehrere systematische Varianzquellen berücksichtigen, die einen Einfluss auf das manifeste Antwortverhalten haben (s. z. B. Rauch und Moosbrugger 2011).

Mehrdimensionale Messmodelle erlauben die Bestimmung der Reliabilität (und der Validität) mehrdimensionaler Tests sowie die Bildung der Testwerte für die einzelnen Dimensionen. Wird mit den manifesten Itemvariablen ein mehrdimensionales Konstrukt erfasst, z. B. Perfektionismus, Prüfungsangst oder Intelligenz, so können z. B. anhand des Bifaktormodells (► Kap. 24) verschiedene Omega-Koeffizienten als Reliabilitätsmaße geschätzt werden (► Kap. 15).

Mehrdimensionale Modelle auf Basis der KTT liegen formal aber auch dann vor, wenn ein Konstrukt zu mehreren Messgelegenheiten gemessen wird. In diesem Fall wird mit den manifesten Itemvariablen nicht nur das gemeinsame Konstrukt („Trait“), sondern auch ein situationsspezifischer Anteil („State-Residuum“) gemessen. Anhand der Latent-State-Trait-Theorie (LST-Theorie, ► Kap. 26; Steyer et al. 2015) kann die Reliabilität in einen traitspezifischen Anteil (Konsistenz) und einen situationspezifischen Anteil (Spezifität) aufgeteilt werden.

Werden mehrere Verfahren (z. B. Selbst-/Fremdbeurteilung) zur Messung eines Konstrukt eingesetzt, so kann die Methodenspezifität als ein vom Trait unabhängiger Effekt im Rahmen der Multitrait-Multimethod-Analyse (MTMM-Analyse) geschätzt werden (Eid et al. 2008; s. auch ► Kap. 25 und 27). Die Reliabilität kann hier in einen trait- und einen methodenspezifischen Anteil aufgeteilt werden.

Insgesamt stellt die KTT die Basis für eine Vielzahl von Messmodellen dar, anhand derer psychometrische Überprüfungen der Item- und Testwertvariablen vorgenommen werden können.

12.3 Item-Response-Theorie (IRT)

Vergleichbar zur KTT wird auch in der IRT (► Kap. 16) postuliert, dass die Beantwortung der Aufgaben (Items) eines Tests von einer latenten Fähigkeit oder Persönlichkeitseigenschaft abhängt, die das Testverhalten bestimmt und die nicht direkt beobachtbar ist. Das mit den Testitems erfasste beobachtbare Verhalten stellt somit lediglich einen Indikator für die individuelle Ausprägung des zugrunde liegenden, nicht direkt beobachtbaren Merkmals (latente Variable, latentes Konstrukt) dar.

Die IRT wurde entwickelt, um Rückschlüsse auf die individuelle Ausprägung dieser latenten Konstrukte (z. B. Fähigkeiten) zu ziehen, und zwar vor allem für den Fall, dass von den Testpersonen typischerweise dichotome (oder auch kategorial geordnete) Antworten („Responses“) auf verschiedene Testitems vorliegen (Fischer 1996; Hambleton und Swaminathan 1985; Lord 1980; Lord und Nowick 1968).

Die der IRT zugrunde liegende Idee wurde bereits von Thurstone (1925, 1928) entwickelt, wonach die Lösung einer Aufgabe von der latenten Merkmalsausprägung einer Person und der Aufgabenschwierigkeit abhängt (im Gegensatz zur KTT wird in der IRT die Schwierigkeit und nicht die Leichtigkeit der Aufgaben/Items modelliert). Die IRT gewann aber erst in der zweiten Hälfte des vergangenen Jahrhunderts zunehmend an Bedeutung, nachdem das „Rasch-Modell“ (Rasch 1960) und das „Birnbaum-Modell“ (Birnbaum 1968) entwickelt worden waren (► Abschn. 12.3.6) und benutzerfreundliche Computerprogramme zur Schätzung der Parameter zur Verfügung standen.

Im Rahmen der IRT wurde eine Vielzahl von Untermodellen entwickelt (vgl. ► Kap. 16 und 18). Die Modelle der IRT unterscheiden sich u. a. bezüglich der Kategorienanzahl der Antwortvariablen (z. B. dichotom, polytom mit geordneten Antwortkategorien) und der Anzahl der zu schätzenden Itemparameter (z. B. Diskriminations-, Schwierigkeits- oder Schwellenparameter).

Rückschlüsse auf latente Konstrukte

12.3.1 Annahmen der IRT

Wesentliche Annahmen der IRT sind u. a., dass die Itemvariablen (Antwortvariablen der Items) eindimensional (homogen) sind, also nur *eine* latente Variable erfassen, und dass der Zusammenhang zwischen der latenten Variablen und der Wahrscheinlichkeit, ein Item positiv zu beantworten, durch die *Itemcharakteristische Funktion* (IC-Funktion) in Form einer kurvilinearen, meist logistischen Beziehung adäquat beschrieben wird. Die Antworten auf unterschiedliche Items dürfen – außer von der Itemschwierigkeit – nur von der Ausprägung der latenten Variablen abhängen, was sich darin äußert, dass die Antworten auf jeder lokalen Stufe der latenten Variablen unabhängig voneinander sind; es muss also *lokale stochastische Unabhängigkeit* (auch als bedingte stochastische Unabhängigkeit bezeichnet) gegeben sein (vgl. Eid und Schmidt 2014; ► Kap. 16).

Liegt lokale stochastische Unabhängigkeit vor, so hängt die Lösungswahrscheinlichkeit einer Aufgabe (oder die Wahrscheinlichkeit, eine Aufgabe im Sinne des Konstrukt zu beantworten) ausschließlich von der Ausprägung des latenten Merkmals ab und wird nicht durch das Lösen oder Nichtlösen (bzw. die Beantwortung) einer vorangegangenen Aufgabe beeinflusst. Diese Annahme hat den Vorteil, dass die Berechnung der gemeinsamen Lösungswahrscheinlichkeit aller Items stark vereinfacht wird, da die Einzelwahrscheinlichkeiten miteinander multipliziert werden können.

In Latent-Trait-Modellen wird die latente Variable meist als kontinuierlich angenommen. Wenn es sich bei der zu messenden latenten Variablen jedoch nicht um eine kontinuierliche, sondern um eine kategoriale Variable handelt (z. B. Perfektionismustypen), kommen Latent-Class-Modelle zur Anwendung (► Kap. 22).

Lokale stochastische Unabhängigkeit

Latent-Trait- vs. Latent-Class-Modelle

12.3.2 Itemcharakteristische Funktion und Spezifische Objektivität

Die IRT-Modelle stellen einen Zusammenhang her zwischen den Itemvariablen (Itemantworten, Responses) und einer latenten Personenvariablen („Latent Trait“)

Antworten abhängig von Itemschwierigkeit und Personenfähigkeit

Logistische IC-Funktion

Gemeinsame Skala für Personenfähigkeit und Itemschwierigkeit

Rasch-Modell

Spezifisch objektive Vergleiche

im kontinuierlichen Fall (► Kap. 18) oder einer latenten Klassenvariablen („Latent Class“) im kategorialen Fall (► Kap. 22). Diese Zusammenhänge werden über kurvilineare Beziehungen abgebildet, die IC-Funktionen.

Die Wahrscheinlichkeit, mit der eine Person ein konkretes Antwortverhalten zeigt, ist dabei einerseits von der Schwierigkeit des Items und andererseits von der Fähigkeit oder Merkmalsausprägung der Person abhängig. Somit kann für jede Person und für jedes Item die Wahrscheinlichkeit bestimmt werden, mit der bei Leistungstests eine Aufgabe gelöst bzw. bei Persönlichkeitstests ein Item symptomatisch im Sinne des Konstrukt beantwortet wird.

Im einfachsten Fall mit dichotomen Antwortkategorien (nein/ja, nicht gelöst/gelöst, nicht zugestimmt/zugestimmt) und einer zweiseitigen, 0/1-kodierten Itemvariablen wird als Beziehung zwischen der latenten Merkmalsausprägung und der Wahrscheinlichkeit, das Item zu lösen, eine monotone, kurvilineare (meist logistische) IC-Funktion gewählt. Die IC-Funktion besagt, dass bei einer höheren Merkmalsausprägung (z. B. Fähigkeit oder Persönlichkeitseigenschaft) eine höhere Lösungswahrscheinlichkeit bzw. eine höhere symptomatische Beantwortungswahrscheinlichkeit resultiert, die asymptotisch gegen eins geht, wohingegen bei einer niedrigeren Merkmalsausprägung eine niedrigere symptomatische Beantwortungswahrscheinlichkeit resultiert, die asymptotisch gegen null geht. Die logistische Beziehung lässt sich anhand folgender Gleichung ausdrücken:

$$P(y_i = 1|\eta) = \frac{e^{\lambda_i \cdot (\eta - \beta_i)}}{1 + e^{\lambda_i \cdot (\eta - \beta_i)}} \quad (12.8)$$

Hierbei bezeichnet y_i die Itemvariable des Items i , e die Euler'sche Zahl ($= 2.718$), η die latente Variable (Personenvariable), λ_i den Diskriminationsparameter und β_i den Schwierigkeitsparameter des Items i .

Wie Gl. (12.8) verdeutlicht, stellt die Differenz zwischen der latenten Merkmalsausprägung η einer Person und der jeweiligen Itemschwierigkeit β_i die entscheidende Größe für die Lösungswahrscheinlichkeit dar. Die Personenfähigkeit und die Itemschwierigkeit sind somit differenzskaliert und lassen sich auf einer gemeinsamen Skala (Joint Scale) darstellen. Hierdurch können individuelle Personenwerte durch ihre Abstände zu den Itemschwierigkeiten interpretiert werden (► Kap. 16 und 17).

Weisen alle Diskriminationsparameter λ_i der Items eines Tests den konstanten Wert eins auf, so vereinfacht sich Gl. (12.8) auf die IC-Funktion des Rasch-Modells (Rasch 1960; ► Kap. 16; Strobl 2012). Konstante Diskriminationsparameter bedeuten, dass alle Items gleich gut zwischen Personen mit einer niedrigen und Personen mit einer hohen Ausprägung im latenten Konstrukt differenzieren. Die IC-Funktionen aller Items verlaufen dann parallel in dem Sinne, dass sie sich durch Parallelverschiebung entlang der η -Achse ineinander überführen lassen.

Für das Beispiel in □ Abb. 12.2 wurde der Erwartungswert von η zur Normierung auf null fixiert. Um den Annahmen des Rasch-Modells zu entsprechen, wurden alle Diskriminationsparameter λ_i auf eins fixiert. Somit verlaufen die IC-Funktionen der fünf Items parallel. Item 1 ist in diesem fiktiven Beispiel das leichteste Item, Item 5 dagegen das schwerste Item.

■ ■ Spezifische Objektivität

Eine vorteilhafte Eigenschaft z. B. des Rasch-Modells oder des Partial-Credit-Modells (PCM; Masters 1982) ist die Spezifische Objektivität. Diese zeigt sich daran, dass alle Diskriminationsparameter λ_i denselben Wert aufweisen und die logistischen Itemcharakteristiken parallel verlaufen. Wie □ Abb. 12.2 zeigt, unterscheiden sich die Itemcharakteristiken nicht in ihrer Steilheit, sondern nur hinsichtlich ihrer Schwierigkeitsparameter β_i . In diesem Fall sind spezifisch objektive Vergleiche möglich, und zwar sowohl hinsichtlich der Merkmalsausprägungen zweier Personen unabhängig von den verwendeten Items als auch hinsichtlich der

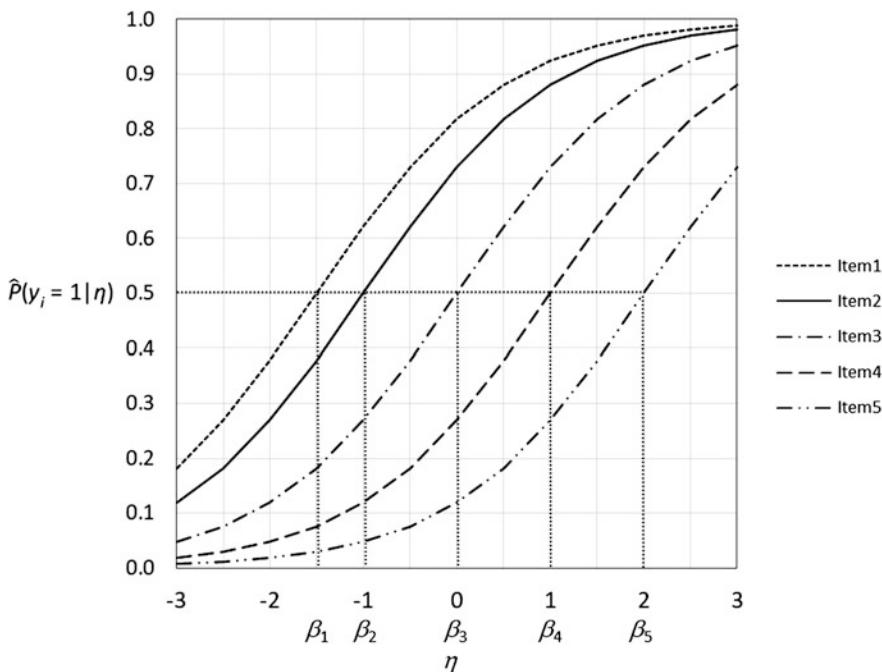


Abb. 12.2 Parallelle IC-Funktionen im Rasch-Modell: Geschätzte Lösungswahrscheinlichkeit $\hat{P}(y_i = 1 | \eta)$ von fünf Items y_i in Abhängigkeit von der Ausprägung der latenten Variablen η bei konstanten Diskriminationsparametern λ_i und unterschiedlichen Schwierigkeitsparametern β_i . Die fünf Items unterscheiden sich nur in ihren Schwierigkeitsparametern β_i : $\beta_1 = -1.5$, $\beta_2 = -1.0$, $\beta_3 = 0$, $\beta_4 = 1.0$, $\beta_5 = 2.0$.

Schwierigkeit zweier Items unabhängig von den getesteten Personen. Sprechen die Modelltests (► Abschn. 12.3.5; vgl. ► Kap. 16 und 19) für Modellkonformität, kann die Eigenschaft der Spezifischen Objektivität als gegeben angenommen werden.

12.3.3 Summenscores

Die Anwendung von IRT-Modellen ermöglicht eine Schätzung der individuellen Ausprägung der latenten Variablen für jede Person (Personenparameter, ► Kap. 16, s. auch ► Kap. 17).

Im Rasch-Modell und weiteren Modellen mit separierbaren Modellparametern (► Abschn. 12.3.6.1) wird die Anzahl der von einer Testperson gelösten Aufgaben bzw. der im Sinne des untersuchten Konstrukt beantworteten Items zu einem Summenwert (Summenscore) aufaddiert. Die aufsummierten Werte stellen – z. B. im Rasch-Modell mit 0/1-kodierten Itemvariablen – eine suffiziente („erschöpfende“) Schätzung der Personenparameter dar (► Kap. 16).

Aus der Höhe der Summenscores kann daher unmittelbar auf die latente Merkmalsausprägung geschlossen werden, da bei gleicher Anzahl gelöster Aufgaben trotz unterschiedlicher Antwortmuster identische Schätzungen der Personenparameter resultieren. Bei Gültigkeit des Rasch-Modells ist also nicht entscheidend, welche der Rasch-homogenen Aufgaben gelöst bzw. bejaht wurden, sondern lediglich deren Anzahl.

Individuelle Schätzung des Personenparameters

Summenscore als suffiziente Statistik

12.3.4 Reliabilität

Testinformation als Maß für die Messgenauigkeit

Konfidenzintervallbreite der Personenparameter variiert in Abhängigkeit von der Ausprägung der latenten Variablen

Marginale Reliabilitätskoeffizienten

Personenseparierbarkeit

Adaptives Testen

Überprüfung der Stichprobenunabhängigkeit

Vergleich der beobachteten und erwarteten Häufigkeiten von Antwortmustern

Im Unterschied zur KTT werden in der IRT die Konfidenzintervalle für die geschätzten Personenparameter nicht mithilfe der Reliabilitätskoeffizienten, sondern mithilfe der Testinformation gebildet. Die Testinformation beruht auf den einzelnen Iteminformationen, die jeweils angeben, welchen Beitrag ein Item zur Schätzgenauigkeit eines latenten Personenwertes leistet (► Kap. 16). Die Testinformation erlaubt die Bestimmung der Standardfehler der geschätzten Personenparameter und stellt somit ein Maß für die Messgenauigkeit (genauer: Präzision) dar, mit der die latenten Werte (Personenparameter) geschätzt werden. In gewisser Weise entspricht die Testinformation somit der Reliabilität in der KTT (vgl. z. B. Bandalos 2018, S. 429; ► Kap. 19).

Im Unterschied zur KTT ist die Messgenauigkeit in der IRT nicht für alle Personenwerte gleich; vielmehr variiert die Messgenauigkeit der Testwertvariablen in Abhängigkeit von der Ausprägung der latenten Variablen η . Dies bedeutet, dass die Testinformation in unterschiedlichen Wertebereichen der latenten Variablen höher bzw. niedriger ist und somit die Konfidenzintervallbreite der Personenparameter variieren (vgl. Jabrayilov et al. 2016; ► Kap. 19).

In Analogie zur Reliabilität der KTT sind aber auch in der IRT durchschnittliche („marginale“) Reliabilitätskoeffizienten als Kennwerte der Messgenauigkeit von Tests entwickelt worden (vgl. ► Kap. 19).

Ähnlich wie in der KTT wird die marginale Reliabilität bestimmt als

$$Rel(\hat{\eta}) = \frac{Var(\eta)}{Var(\hat{\eta})} = 1 - \frac{Var(\varepsilon)}{Var(\hat{\eta})} \quad (12.9)$$

Das Verhältnis der Varianz der wahren Personenwerte $Var(\eta)$ an der Varianz der geschätzten Personenwerte $Var(\hat{\eta})$ gibt an, inwieweit die geschätzten Personenwerte wahre Unterschiede zwischen den Personen wiedergeben. Dieses Verhältnis wird auch *Personenseparierbarkeit* genannt (Eid und Schmidt 2014, S. 181).

Die Testinformation kann durch Erhöhung der Itemanzahl oder durch gezielte Verwendung von Items mit einer hohen Iteminformation gesteigert werden. Letztere Überlegung findet beim adaptiven Testen Verwendung, indem bei einer Testung nur diejenigen Items sukzessive dargeboten werden, deren Iteminformationen für die untersuchte Person optimal zur Schätzung ihrer Fähigkeit oder ihres Persönlichkeitsmerkmals beitragen (vgl. ► Kap. 20).

12.3.5 Modelltests

Die Modellkonformität (Gültigkeit) von IRT-Modellen zur Beantwortung der Frage, ob die Beziehung zwischen den Itemvariablen (Responses) und dem latenten Konstrukt durch das jeweilige IRT-Modell zutreffend beschrieben wird, kann anhand verschiedener Modelltests überprüft werden. Beispielsweise kann zur Überprüfung der Stichprobenunabhängigkeit, d. h., ob die Itemparameter in verschiedenen Subpopulationen denselben Wert aufweisen, der bedingte Likelihood-Quotienten-Test nach Andersen (1973) verwendet werden, dessen Prüfgröße χ^2 -verteilt ist. Des Weiteren werden häufig auch der grafische Modelltest und der Wald-Test eingesetzt (► Kap. 16).

Im Rasch-Modell werden z. B. die Antwortmuster für den Modelltest verwendet. Werden die Wahrscheinlichkeiten der möglichen Antwortmuster mit der Stichprobengröße multipliziert, so resultieren die erwarteten Häufigkeiten der Antwortmuster. Anhand des Pearson- χ^2 -Tests kann entschieden werden, ob die Abweichungen zwischen den beobachteten und den erwarteten Häufigkeiten im Zufallsbereich liegen, was für Modellkonformität (d. h. Eindimensionalität) spricht, oder

ob es überzufällig starke Abweichungen gibt, die der Modellkonformität entgegenstehen (vgl. Eid und Schmidt 2014; ► Kap. 16).

Die geschätzten Parameter dürfen nur interpretiert werden, wenn die Modellkonformität hinsichtlich eines eindimensionalen Merkmals (► Abschn. 12.3.6) bzw. hinsichtlich eines mehrdimensionalen Merkmals (► Abschn. 12.3.7) anhand eines Modelltests nachgewiesen wurde.

12.3.6 Eindimensionale IRT-Modelle

Im Rahmen eindimensionaler IRT-Modelle können anhand der Skalierung der Itemvariablen z. B. Modelle mit dichotomem Antwortmodus von Modellen mit polytomem Antwortmodus mit geordneten Antwortkategorien unterscheiden werden. Die Modelle unterscheiden sich darüber hinaus durch unterschiedlich strenge Annahmen, die sich u. a. auf die Gleichheit oder Ungleichheit der Diskriminationsparameter sowie bei mehrstufigen Antwortkategorien auf die jeweiligen Schwellenparameter beziehen.

Ein weiteres Unterscheidungskriterium besteht in der Separierbarkeit der Modellparameter (vgl. Müller 1999), die nur in Modellen mit identischen Diskriminationsparametern gegeben ist. Diese Eigenschaft ermöglicht spezifisch objektive Vergleiche zwischen Personen und zwischen Items. Die Separierbarkeit der Modellparameter wird auch als *Stichprobenunabhängigkeit* bezeichnet, die als Basis für die Überprüfung der Modellkonformität eine wesentliche Rolle spielt (vgl. ► Kap. 16). Eine Übersicht über ausgewählte eindimensionale IRT-Modelle findet sich in □ Abb. 12.3.

Separierbarkeit der Modellparameter

12.3.6.1 Modelle für separierbare Parameter

Als separierbar werden Modellparameter dann bezeichnet, wenn sie *nicht* in Abhängigkeit von η variieren. Zur Gruppe der Modelle für separierbare Modellparameter zählen u. a. das Rasch-Modell (Rasch 1960), das Partial-Credit-Modell (PCM, Masters 1982) und das Rating-Scale-Modell (RSM, Andrich 1978).

■■ Rasch-Modell

Das Rasch-Modell (vgl. ► Kap. 16) basiert auf der Annahme, dass alle Items eines Tests eindimensional sind, d. h. eine gemeinsame latente Variable (z. B. Fähigkeit, Einstellung, Persönlichkeitseigenschaft) messen. Diese Annahme wird auch als *Rasch-Homogenität* bezeichnet.

Rasch-Homogenität

Dabei werden die Diskriminationsparameter aller Items auf den Wert eins fixiert, während sich die Itemschwierigkeiten unterscheiden dürfen. Aufgrund der Fixierung der Diskriminationsparameter auf denselben Wert ist die Separierbarkeit der Item- und Personenparameter gegeben. Im Rasch-Modell wird somit für die logistische IC-Funktion jedes Items jeweils nur *ein* (Item-)Parameter (die Itemschwierigkeit) geschätzt, weshalb das Rasch-Modell auch als *IPL-Modell* (Ein-Parameter-Logistisches-Modell) bezeichnet wird.

IPL-Modell

■■ Partial-Credit-Modell (PCM) und Rating-Scale-Modell (RSM)

Erweiterungen des Rasch-Modells für kategoriale Antwortvariablen mit geordneten Antwortkategorien sind u. a. das PCM (Masters 1982) und das RSM (Andrich 1978).

Erweiterungen des Rasch-Modells

Das PCM (Masters 1982) stellt eine Erweiterung des Rasch-Modells für kategoriale Itemvariablen mit geordneten Antwortkategorien dar, wobei die Annahmen des Rasch-Modells auf mehrere Schwellen zwischen den einzelnen Antwortkategorien übertragen werden (Schwellenwahrscheinlichkeiten; vgl. DeMars 2018; ► Kap. 18).

Ein Spezialfall des PCM ist das RSM (Andrich 1978), das ebenfalls als Erweiterung des Rasch-Modells aufgefasst werden kann. Das RSM stellt eine sparsamere

Antwortmodus	Separierbare Modellparameter	Nicht separierbare Modellparameter
Dichotom	<p>Rasch-Modell (Rasch, 1960)</p> <p><i>Diskriminationsparameter</i> für alle Items identisch (auf eins fixiert)</p> <p><i>Schwierigkeitsparameter</i> frei geschätzt</p>	<p>Birnbaum-Modell (Birnbaum, 1968)</p> <p>Erweiterung des Rasch-Modells</p> <p><i>Diskriminationsparameter</i> frei geschätzt</p> <p><i>Schwierigkeitsparameter</i> frei geschätzt</p>
Polytom, geordnete Antwortkategorien	<p>Partial-Credit-Modell (PCM; Masters, 1982)</p> <p>Erweiterung des Rasch-Modells</p> <p><i>Diskriminationsparameter</i> für alle Items identisch</p> <p><i>Schwellenparameter^a</i> frei geschätzt</p> <p><i>Anzahl der Kategorien</i> müssen für alle Items gleich sein</p>	<p>Generalized Partial-Credit-Modell (GPCM; Muraki, 1992)</p> <p>Erweiterung des PCM und des Birnbaum-Modells</p> <p><i>Diskriminationsparameter</i> frei geschätzt</p> <p><i>Schwellenparameter^a</i> frei geschätzt</p> <p><i>Anzahl der Kategorien</i> müssen für alle Items gleich sein</p>
	<p>Rating-Scale-Modell (RSM; Andrich, 1978)</p> <p>Erweiterung des Rasch-Modells und Spezialfall des PCM</p> <p><i>Diskriminationsparameter</i> für alle Items identisch</p> <p><i>Schwellenparameter^a</i> frei geschätzt</p> <p>Die Abstände zwischen den Schwellen innerhalb eines Items können unterschiedlich sein, ihre paarweisen Differenzen jedoch über die Items hinweg müssen jedoch identisch sein</p> <p><i>Anzahl der Kategorien</i> müssen für alle Items gleich sein</p>	<p>Graded-Response-Modell (GRM; Samejima, 1969)</p> <p>Erweiterung des RSM und des Birnbaum-Modells</p> <p><i>Diskriminationsparameter</i> frei geschätzt</p> <p><i>Schwellenparameter^b</i> frei geschätzt</p> <p><i>Anzahl der Kategorien</i> können unterschiedlich sein</p>

■ Abb. 12.3 Charakteristika ausgewählter eindimensionaler IRT-Modelle

Variante des PCM dar, da die Abstände zwischen den Schwellen innerhalb eines Items unterschiedlich sein können (z. B. Schwelle 2 vs. Schwelle 1), ihre paarweisen Differenzen jedoch über die Items hinweg identisch sein müssen. Das bedeutet, dass das RSM ein PCM ist, in dem zusätzlich zur Wahrscheinlichkeit der Kategorienewahl die Einschränkung besteht, dass identische Differenzen für jedes Paar von Schwierigkeitsparametern für sukzessiv aufeinanderfolgende Kategorien über alle Items hinweg bestehen (Raykov und Marcoulides 2018). Es handelt sich bei beiden Modellen um *direkte Modelle* (Embretson und Reise 2000, S. 105), da die Schwellenparameter direkt geschätzt werden. Von Müller (1987) wurde auch ein Rasch-Modell für kontinuierliche Ratingskalen entwickelt.

Direkte Modelle

12.3.6.2 Modelle für nicht separierbare Modellparameter

Als nicht separierbar werden Modellparameter bezeichnet, wenn sie in Abhängigkeit von η variieren. Zur Gruppe der Modelle für nicht separierbare Modellparameter zählen u. a. das Birnbaum-Modell (Birnbaum 1968), das Generalized Pratial-Credit-Modell (GPCM, Muraki 1992) und das Graded-Response-Modell (GRM; Samejima 1969).

■■ Birnbaum-Modell

Das Birnbaum-Modell (vgl. ▶ Kap. 16) stellt eine weniger strenge Variante des Rasch-Modells dar, da die Diskriminationsparameter nicht mehr als identisch angenommen werden, sondern frei geschätzt werden.

Im Birnbaum-Modell werden somit für die logistische IC-Funktion jedes Items jeweils *zwei* Parameter (der Schwierigkeitsparameter und der Diskriminationsparameter) geschätzt, weshalb das Birnbaum-Modell auch als 2PL-Modell (Zwei-Parameter-Logistisches-Modell) bezeichnet wird.

Das Birnbaum-Modell verwendet ebenfalls die logistische IC-Funktion, jedoch werden für alle Items unterschiedliche Steigungen der IC-Funktionen zugelassen. Die unterschiedlichen Steigungen bedeuten, dass die Items unterschiedlich gut zwischen Personen mit schwächerer oder stärkerer Merkmalsausprägung differenzieren können (▶ Kap. 16). Als Konsequenz sind keine spezifisch objektiven Vergleiche wie im Rasch-Modell möglich und die Separierbarkeit der Item- und Personenparameter ist nicht gegeben.

Erweiterung des Rasch-Modells mit weniger strengen Annahmen

2PL-Modell

Keine Spezifische Objektivität

■■ Generalized Partial-Credit-Modell (GPCM) und Graded-Response-Modell (GRM)

Ein bekanntes Modell für polytome Antwortvariablen ist das GPCM (Muraki 1992), das auf Grundlage des PCM entwickelt wurde (▶ Kap. 18). Das GPCM entspricht dem PCM unter Hinzunahme des Diskriminationsparameters, der zwischen den Items eines Tests variieren kann. Somit stellt das Modell eine polytome Erweiterung des Birnbaum-Modells dar, da die Abstände zwischen zwei aufeinanderfolgenden Schwellen über alle Items hinweg nicht gleich sein müssen und die Diskriminationsparameter – wie im Birnbaum-Modell – frei geschätzt werden können. Das GPCM ist wie das PCM und das RSM ein *direktes IRT-Modell*, das nur einen einstufigen Schätzprozess erfordert.

Erweiterung des Birnbaum-Modells

Das GRM (Samejima 1969) beschreibt eine ganze Modelfamilie für Items mit polytom geordneten Itemantworten. Im Gegensatz zum GPCM dürfen die Items eine unterschiedliche Anzahl an Antwortkategorien haben. Im einfachsten Fall postuliert das GRM einheitliche Steigungen der Antwortkategorienkurven innerhalb eines Items und nutzt zur Parameterschätzung eine über die Antwortkategorien kumulierende Schätzfunktion.

Modelfamilie für Items mit polytom geordneten Itemantworten

Das GRM wird als *indirektes IRT-Modell* oder Differenzmodell (Embretson und Reise 2000) bezeichnet, da die Schätzung der bedingten Wahrscheinlichkeit für eine Person, eine bestimmte Kategorie zu wählen, einen zweistufigen Prozess erfordert.

12.3.7 Mehrdimensionale IRT-Modelle

Neben den in □ Abb. 12.3 aufgeführten Modellen gibt es eine Vielzahl weiterer Modelle. Hierzu gehören die mehrdimensionalen IRT-Modelle (MIRT-Modelle), die auch als multidimensionale IRT-Modelle bezeichnet werden. Wie bei den mehrdimensionalen KTT-Modellen werden hier mehrere systematische Varianzquellen (z. B. mehrere Personenmerkmale) berücksichtigt, die einen Einfluss auf das manifeste Antwortverhalten haben. Die Wahrscheinlichkeit einer Antwort wird somit anhand von mindestens zwei latenten Merkmalen modelliert.

MIRT-Modelle erfassen mehrere Personenmerkmale

MIRT-Modelle werden seltener als eindimensionale Modelle verwendet, was eher pragmatische als theoretische Gründe hat. MIRT-Modelle sind nicht nur Erweiterungen der eindimensionalen IRT-Modelle, sondern auch Spezialfälle der Faktorenanalyse (Reckase 2009). Einen Überblick über mehrdimensionale Modelle geben z. B. Kelava, Robitzsch und Noventa in ▶ Kap. 18.

In der Kompetenzdiagnostik werden jedoch beispielsweise inzwischen zunehmend häufiger MIRT-Modelle eingesetzt (vgl. Hartig und Höhler 2008, 2009). Durch MIRT-Modelle sind eine differenzierte Diagnostik und ein Zugewinn an diagnostischer Information möglich. MIRT-Bifaktormodelle stellen eine weitere mehrdimensionale Alternative dar, wenn geprüft werden soll, ob ein Generalfaktor und zusätzliche spezifische Faktoren die Itemantworten beeinflussen (vgl. Reise et al. 2007; Reise und Waller 2009).

Mehrdimensionale Modelle auf Basis der IRT liegen formal auch dann vor, wenn ein Konstrukt im Längsschnitt, d. h. zu mehreren Messzeitpunkten, gemessen wird (vgl. Houts et al. 2018; Millsap 2010). Ein Längsschnittmodell wird z. B. benötigt, um die Messinvarianz eines Tests über die Zeit hinweg zu testen.

12.4 Klassische Testtheorie (KTT) vs. Item-Response-Theorie (IRT)

Enge Beziehungen zwischen KTT und IRT

Wie die bisherigen Ausführungen zur KTT und zur IRT zeigen, bestehen enge Beziehungen zwischen beiden Theorien. Entgegen der früher oft geäußerten Auffassung (Lord 1980; Hambleton und Swaminathan 1985; Fischer 1996) ist die IRT nicht als Alternative zur KTT zu sehen, sondern eher als Ergänzung. Voraussetzung dafür ist allerdings, dass die auf der KTT basierenden Modelle adäquat formuliert und überprüft werden (▶ Kap. 13).

12.4.1 Wesentliche Charakteristika der KTT und der IRT

In □ Tab. 12.2 werden wesentliche Charakteristika der beiden Theorien gegenübergestellt und die zahlreichen Übereinstimmungen zwischen der KTT und der IRT verdeutlicht, auf die bereits bei der bisherigen Darstellung der Theorien eingegangen wurde. Nachfolgend sollen jene drei Merkmale aus □ Tab. 12.2 kurz erläutert werden, in denen sich die Testtheorien etwas unterscheiden.

■■ Kategoriale und kontinuierliche Itemvariablen

Sowohl die KTT als auch die IRT eignen sich zur Analyse von geordneten kategorialen und kontinuierlichen Itemvariablen, auch wenn die KTT ihren Schwerpunkt bei kontinuierlichen Variablen und die IRT bei kategorialen Variablen hat. Das wurde nicht immer so gesehen.

Ein Kritikpunkt an der KTT bestand früher darin, dass eine manifeste Antwort-/Itemvariable zur Messung einer True-Score-Variablen mindestens Intervallskalenniveau aufweisen müsse und dass die KTT somit für kategoriale Antwortvariablen ungeeignet sei. Typischerweise wurden deshalb dichotome Antwortvariablen und Items mit geordneten Antwortkategorien anhand von IRT-Modellen analysiert, kontinuierliche Antwortvariablen dagegen anhand von Modellen auf Basis der KTT.

Diese Aufteilung wurde aber bereits in der Vergangenheit aufgeweicht, als sich innerhalb der IRT Ansätze für kontinuierliche Itemvariablen entwickelten: Bereits Rasch (1960), Samejima (1973) und Müller (1987) erweiterten vorhandene IRT-Modelle (Rasch-Modell, GRM und RSM) auf kontinuierliche Itemvariablen (vgl. auch Mellenbergh 2016). Die IRT ist somit nicht zwingend auf kategoriale Variablen beschränkt.

IRT-Modelle für kontinuierliche Variablen

Tabelle 12.2 Vergleich der KTT und der IRT bezüglich ausgewählter Charakteristika

	KTT	IRT
Kontinuierliche Itemvariablen	+	(+)
Kategoriale Itemvariablen	(+)	+
IC-Funktion	linear	logistisch
Lokale stochastische Unabhängigkeit	(+) schwächere Annahme unkorrelierter Messfehler	+
Spezifische Objektivität von Vergleichen	+	+ sofern Diskriminationsparameter identisch sind
Stichprobenunabhängigkeit der Parameterschätzungen	+	+
Reliabilität der Testwertvariablen	+	+ basierend auf Item- und Testinformation
Adaptives Testen	(+)	+
Modelltests	+	+
Eindimensionale Messmodelle	+	+
Mehrdimensionale Messmodelle	+	+
Einordnung in ein übergreifendes Konzept	+	+

Anmerkung: + = trifft zu, (+) = trifft bedingt zu

Ebenso ist die KTT nicht ausschließlich auf kontinuierliche Itemvariablen beschränkt. Raykov und Marcoulides (2011, S. 122) konstatieren:

- » There is neither such an assumption in CTT nor a requirement nor a need for the observed score X within the CTT framework to be a continuous measure.¹

Nach Muthén (2012) besteht zwischen der IRT und einem CFA-Modell für kategoriale Variablen kein grundsätzlicher Unterschied:

- » I don't think there is a difference between CFA of categorical variables and IRT. It is sometimes claimed but I don't agree.

Die KTT ist somit nicht auf intervallskalierte Itemvariablen beschränkt. Die manifesten Itemvariablen sowie die latenten Variablen können dabei wie in der IRT sowohl kontinuierlich als auch geordnet kategorial sein (Raykov et al. 2019; Raykov und Marcoulides 2016).

KTT-Modelle für kategoriale Variablen

■■ Lokale stochastische Unabhängigkeit

In der IRT stellt die lokale stochastische Unabhängigkeit eine wesentliche Voraussetzung der Modellkonformität dar (vgl. ► Kap. 16). Bei lokaler stochastischer Unabhängigkeit hängt die Lösungswahrscheinlichkeit einer Aufgabe (oder die Wahrscheinlichkeit, eine Aufgabe im Sinne des Konstrukt zu beantworten) ausschließlich von der Ausprägung des latenten Merkmals ab. In der IRT muss ein Modell diese Annahme erfüllen, damit sichergestellt ist, dass die Items nur ein einziges gemeinsames Konstrukt erfassen.

In der KTT wird die Eindimensionalität dagegen in der Regel auf Grundlage der weniger strengen Annahme unkorrelierter Fehlervariablen überprüft. Die An-

1 Hier bezeichnet CTT die klassische Testtheorie und X die beobachtbare Itemvariable.

nahme unkorrelierter Messfehler muss bei eindimensionalen Modellen erfüllt sein und deshalb überprüft werden.

Die Korrelation ist ein Maß für die lineare Abhängigkeit zweier Variablen. Korrelierende Messfehler sprechen daher für ein mehrdimensionales Modell, in dem weitere latente Variablen als systematische Varianzquellen zur Erklärung der linearen Beziehungen zwischen den Itemvariablen berücksichtigt werden. Läge jedoch eine kurvilineare (z. B. eine quadratische) Beziehung vor, so wäre die Korrelation null, obwohl tatsächlich eine Abhängigkeit zwischen den Itemvariablen bestünde, die im Modell aber nicht berücksichtigt wird. Aus der strengeren Annahme der stochastischen Unabhängigkeit folgt die korrelative Unabhängigkeit, nicht jedoch umgekehrt (Eid und Schmidt 2014, S. 262).

■ ■ Adaptives Testen

Adaptives Testen ist ein Verfahren zur Messung individueller Ausprägungen von Personenmerkmalen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am Antwortverhalten der untersuchten Person orientiert (► Kap. 20). Dabei werden nur solche Items sukzessive dargeboten, die optimal zur Schätzung der Fähigkeit oder des Persönlichkeitsmerkmals der untersuchten Person beitragen, indem sie nicht zu schwer und nicht zu leicht sind. Die Testung wird solange durchgeführt, bis die gewünschte Präzision der Schätzung des latenten Personenwertes oder ein anderes Kriterium erreicht ist. Der Hauptvorteil adaptiven Testens besteht in einer Messeffizienzsteigerung, die in den meisten Fällen beträchtlich ausfällt.

Adaptives Testen wird meist im Rahmen der IRT eingesetzt. Voraussetzung ist, dass alle Items Konformität mit den Annahmen des verwendeten IRT-Modells (z. B. des Rasch-Modells) aufweisen. In dem Fall können die resultierenden Personenparameter auch bei Vorgabe unterschiedlicher Items problemlos miteinander verglichen werden. Die Personenwerte werden nach jedem Item (oder einer Gruppe von Items) zusammen mit einem Standardfehler geschätzt. Bei IRT-Modellen kann die Messpräzision als Funktion der zu messenden latenten Variablen η variieren und ist durch die Testinformationsfunktion darstellbar.

Adaptives Testen ist auch im Rahmen der KTT möglich. Bei Verwendung der KTT ist dafür ebenfalls nötig, dass alle Items Konformität mit den Annahmen des verwendeten KTT-Modells (z. B. des Modells essentiell τ -paralleler Variablen) aufweisen. Auch hier können die resultierenden Personenparameter bei Vorgabe unterschiedlicher Items problemlos miteinander verglichen werden. Wie beim Rasch-Modell können die Iteminformationsfunktionen (Kehrwerte der Item-Fehlervarianzen) herangezogen werden, um den Standardfehler für die Schätzung eines latenten Personenwertes zu bestimmen. Im Gegensatz zum Rasch-Modell hängt die Testinformationsfunktion jedoch nicht von der Ausprägung der latenten Variablen η ab (vgl. Eid und Schmidt 2014, S. 291).

12.4.2 Übergreifendes Konzept

Die Annäherung zwischen KTT und IRT hat inzwischen eine recht lange Tradition und führte zur Entwicklung übergreifender, beide Theorien umfassender Konzepte. So entwickelte bereits Mellenbergh (1994) ein Konzept der „generalisierten linearen IRT“, das die KTT und die IRT als Spezialfälle eines gemeinsamen Modells auffasst. Holland und Hoskens (2003) stellten einen Ansatz vor, nach dem sich die KTT als Spezialfall der IRT darstellen lässt. Eid und Schmidt (2014) orientierten sich am Konzept von Mellenbergh (1994) und stellten die KTT und die IRT integrativ im Rahmen eines gemeinsamen Konzepts dar.

Durch diese integrative Darstellung wird deutlich, dass die KTT eine moderne Testtheorie ist, die der IRT in nichts nachsteht. Die KTT und die IRT sollten daher

12.5 Zusammenfassung

nicht als konkurrierende Ansätze angesehen werden, sondern als zwei Methoden, die sich gegenseitig vorteilhaft ergänzen und aufgrund einiger wesentlicher charakteristischer Unterschiede weiterhin ihre eigenständige Berechtigung haben.

12.5 Zusammenfassung

Die am häufigsten verwendeten Testtheorien in der Psychometrie sind die KTT und die IRT. Beide Theorien verfolgen sehr ähnliche Ziele bei der Konstruktion und Interpretation von eindimensionalen und mehrdimensionalen Testverfahren zur Messung individueller Merkmalsausprägungen.

Die KTT wird primär für Testitems mit kontinuierlichem (oder zumindest vielstufigem) Antwortformat angewendet und konzentriert sich bei der Messung individueller Merkmalsausprägungen auf die Gewinnung von Testwerten zur Schätzung der True-Scores sowie deren Reliabilität und Validität. Die IRT hingegen wird primär für Testitems mit dichotomen (oder auch polytom geordneten) Antwortkategorien angewendet und hat ihren Schwerpunkt auf der Schätzung latenter Personenparameter, um Rückschlüsse auf interessierende Einstellungs-, Persönlichkeits- oder Fähigkeitsmerkmale zu ziehen, sowie latenter Itemparameter.

In den letzten Jahrzehnten haben sich die KTT und die IRT aufgrund vieler Gemeinsamkeiten zunehmend angenähert. Aufgrund einiger charakteristischer Unterschiede haben beide Theorien ihre eigenständige Berechtigung und ergänzen einander vorteilhaft.

12.6 Kontrollfragen

?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Welche Art von Itemvariablen wird üblicherweise in der KTT, welche in der IRT verwendet?
2. Was wird unter dem Begriff „Spezifische Objektivität“ verstanden? Bei welcher Testtheorie spielt dieser Begriff eine Rolle?
3. Welche Definition der Reliabilität verwendet die KTT, welche die IRT?
4. Bei welchen IRT-Modellen ist die Separierbarkeit der Modellparameter gegeben? Welche Voraussetzung muss dafür gegeben sein?
5. Welche Testtheorie erlaubt die Durchführung von Modelltests?
6. Welche Gemeinsamkeit weisen die Itemcharakteristiken des Modells essentiell τ -äquivalenter Variablen und des Rasch-Modells auf?

Literatur

- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123–140.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York, NY: The Guilford Press.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical Theories of Mental Test Scores* (S. 395–479). Reading: Addison-Wesley.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- DeMars, C. E. (2018). Classical test theory and item response theory. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 49–74). Hoboken, NJ: Wiley.

- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M. & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, 13, 230–253.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fischer, G. H. (1996). IRT-Modelle als Forschungsinstrumente der Differentiellen Psychologie. In K. Pawlik (Hrsg.), *Grundlagen und Methoden der Differentiellen Psychologie* (S. 673–729). Göttingen: Hogrefe.
- Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer-Nijhoff Publishing.
- Hartig, J. & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of Psychology*, 216, 89–101.
- Hartig, J. & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, 35, 57–63.
- Holland, P. W. & Hoskens, M. (2003). Classical test theory as a first-order response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123–149.
- Houts, C. R., Morlock, R., Blum, S. I., Edwards, M. C. & Wirth, R. J. (2018). Scale development with small samples: a new application of longitudinal item response theory. *Quality of Life Research*, 27, 1721–1734.
- Jabrayilov, R., Emons, W. H. M. & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, 40, 559–572.
- Kamata, A. & Bauer, D. J. (2008). A note on the relationship between factor analytic and item response theory models. *Structural Equation Modeling*, 15, 136–153.
- Kohli, N., Koran, J. & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*, 75, 389–405.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum.
- Mellenbergh, G. J. (1994). Generalized item response theory. *Psychological Bulletin*, 115, 300–307.
- Mellenbergh, G. J. (2016). Models for continuous responses. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory. Volume I: Models*. Boca Raton, FL: CRC Press.
- Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, 4, 5–9.
- Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, 52, 165–181.
- Müller, H. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen*. Bern: Huber.
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muthén, B. (2012). *Binary CFA vs IRT* [Mplus discussion: Confirmatory Factor Analysis]. Retrieved from <http://www.statmodel.com/discussion/messages/9/10401.html?1347474605> [23.12.2019]
- Rasch, G. (1960). *Studies in Mathematical Psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.
- Rauch, W. & Moosbrugger, H. (2011). Klassische Testtheorie: Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik. Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden. Serie II, Psychologische Diagnostik* (Bd. 2, S. 1–86). Göttingen: Hogrefe.
- Raykov, T., Dimitrov, D. M., Marcoulides, G. A. & Harrison, M. (2019). On the connections between item response theory and classical test theory: A note on true score evaluation for polytomous items via item response modeling. *Educational and Psychological Measurement*, 79, 1198–1209.
- Raykov, T. & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.
- Raykov, T. & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, 76, 325–338.
- Raykov, T. & Marcoulides, G. A. (2018). *A Course in Item Response Theory and Modeling with Stata*. College Station, TX: Stata Press.
- Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

- Reise, S. P., Morizot, J. & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16* (Suppl. 1), 19–31.
- Reise, S. P. & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology, 5*, 27–48.
- Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika, 34*, Suppl. 1, 1–97.
- Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203–219.
- Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.
- Spearman, C. (1904b). General Intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201–292.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 171–195.
- Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3*, 25–60.
- Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin, Heidelberg: Springer.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A Theory of States and Traits – Revised. *Annual Review of Clinical Psychology, 11*, 71–98.
- Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis (Sozialwissenschaftliche Forschungsmethoden)*. Hampp-Verlag: Mering.
- Takane, Y. & de Leeuw, J. (1987). On the relation between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40*, 395–412.
- Zimmerman, D. W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement, 36*, 85–96.



Klassische Testtheorie (KTT)

Helfried Moosbrugger, Jana C. Gäde, Karin Schermelleh-Engel und Wolfgang Rauch

Inhaltsverzeichnis

- 13.1 Einleitung – 277**
- 13.2 Grundannahmen der KTT – 277**
 - 13.2.1 Arten von Messfehlern – 278
 - 13.2.2 Definition des wahren Wertes (True-Score) – 278
- 13.3 Zerlegung einer Itemvariablen in True-Score- und Messfehlervariable – 279**
 - 13.3.1 Grundgleichung der KTT – 279
 - 13.3.2 Eigenschaften der Messfehler- und True-Score-Variablen – 279
- 13.4 Testwertvariable Y und Testwerte Y_v – 280**
- 13.5 Das Gütekriterium der Reliabilität – 281**
 - 13.5.1 Reliabilität einer Itemvariablen – 281
 - 13.5.2 Reliabilität einer Testwertvariablen – 281
- 13.6 Messmodelle zur Schätzung der Reliabilität – 282**
 - 13.6.1 Unterschiedliche Stufen der Messäquivalenz – 283
 - 13.6.1.1 Modell τ -kongenerischer Variablen – 284
 - 13.6.1.2 Modell essentiell τ -äquivalenter Variablen – 284
 - 13.6.1.3 Modell essentiell τ -paralleler Variablen – 285
 - 13.6.2 Herleitung der wahren Item- und Testwertvarianzen in Abhängigkeit von den Modellparametern – 285
 - 13.6.2.1 Modellbasierte Zerlegung der Itemvarianz und -kovarianz – 285
 - 13.6.2.2 Modellbasierte Zerlegung der Testwertvarianz – 287
 - 13.6.3 Überprüfung der Messäquivalenz – 288
 - 13.6.4 Messäquivalenz-Implikationen ausgewählter Reliabilitätskoeffizienten – 289
 - 13.6.5 Konfidenzintervalle für Reliabilitätskoeffizienten – 291
- 13.7 Empirisches Beispiel – 291**
 - 13.7.1 Datenanalyse – 292
 - 13.7.2 Ergebnisse – 292

- 13.7.2.1 Reliabilität der Testwertvariablen – 293
- 13.7.2.2 Diskriminationsparameter/Faktorladungen – 293
- 13.7.2.3 Itemreliabilität – 293
- 13.7.2.4 Leichtigkeitsparameter/Interzept – 294

13.8 Schätzung individueller Merkmalsausprägungen – 294

- 13.8.1 Geschätzte wahre Testwerte \hat{T}_v – 294
- 13.8.1.1 Konfidenzintervall für T_v – 295
- 13.8.1.2 Kritische Differenz von Testwerten – 297
- 13.8.2 Geschätzte latente Personenwerte $\hat{\eta}_v$ – 298

13.9 Erweiterung der KTT – 298

- 13.9.1 Mehrdimensionale Ansätze – 298
- 13.9.2 Generalisierbarkeitstheorie – 299
- 13.9.2.1 Zusätzliche Varianzquellen – 299
- 13.9.2.2 Varianzzerlegung – 300
- 13.9.2.3 Generalisierbarkeitskoeffizienten – 301

13.10 Zusammenfassung – 302

13.11 EDV-Hinweise – 302

13.12 Kontrollfragen – 302

Literatur – 303

13.1 · Einleitung

i Die Klassische Testtheorie (KTT) stellt die theoretischen Grundlagen zur Konstruktion von Testverfahren und zur Interpretation von Testwerten zur Verfügung, die als theoretische Basis vieler psychodiagnostischer Tests verwendet werden. Die Bezeichnung „klassisch“ soll zum Ausdruck bringen, dass diese Testtheorie bereits vor ca. 70 Jahren entwickelt wurde. Sie umfasst im Wesentlichen Messmodelle für kontinuierliche manifeste und latente Variablen (vgl. Steyer und Eid 2001), die eine Zerlegung der Messwerte in wahre Werte und Fehlerwerte und darauf aufbauend die Bestimmung der Messgenauigkeit (Reliabilität) erlauben. Demgegenüber liefert die erst später entstandene Item-Response-Theorie (IRT, ▶ Kap. 16) insbesondere Messmodelle zur Beziehung zwischen dichotomen oder ordinalskalierten manifesten und kontinuierlichen latenten Variablen. Beide Theorien ergänzen sich vorteilhaft und können als Spezialfälle eines gemeinsamen Modells aufgefasst werden (▶ Kap. 12; Eid und Schmidt 2014; Rauch und Moosbrugger 2011; Raykov und Marcoulides 2016).

13.1 Einleitung

Die Zuverlässigkeit von Messungen, d. h. deren Reliabilität, ist ein zentrales Merkmal der Klassischen Testtheorie (KTT). Individuelle Merkmalsausprägungen können nur anhand reliabler Messungen zuverlässig erfasst werden. Mit der KTT wurden innerhalb der wissenschaftlichen Psychologie seit Beginn des 20. Jahrhunderts die messtheoretischen und statistischen Grundlagen entwickelt, um Merkmalsunterschiede zwischen Personen möglichst exakt und ökonomisch erfassen zu können. Die Bezeichnung „klassisch“ soll zum Ausdruck bringen, dass diese Testtheorie zeitlich früher entwickelt wurde als die Item-Response-Theorie (IRT, ▶ Kap. 12 und 16).

Erste Grundlagen der KTT finden sich bereits bei Spearman (1904a, 1904b), der den Messfehler aus der Korrelation zweier Variablen herausrechnete und einen für diese Korrektur benötigten Reliabilitätsindex entwickelte. Die Prinzipien der KTT wurden von Gulliksen (1950), Lord und Novick (1968) sowie Zimmerman (1975, 1976) entwickelt und von Steyer (1989) sowie Steyer und Eid (2001) formalisiert und ausgearbeitet.

13.2 Grundannahmen der KTT

Die KTT ist im Wesentlichen eine Messfehlertheorie (Steyer und Eid 2001). Mit dieser Feststellung kommt bereits zum Ausdruck, dass Messungen in der Regel mit einem Fehler behaftet sind. Damit Messungen in der Psychologie sinnvoll genutzt werden können (z. B. für diagnostische Zwecke), sollten sie möglichst genau und zuverlässig sein.

Messungen in der Psychologie sind mit Messungen in anderen Bereichen wie z. B. der Medizin vergleichbar. Wird beispielsweise der Blutdruck gemessen, sollte ein möglichst zuverlässiger Wert ermittelt werden, d. h., der Messfehler sollte möglichst klein sein. Eine fehlerhafte Messung könnte unter Umständen dazu führen, dass eine Person ein Medikament zur Senkung des Blutdrucks bekommt, obwohl der Blutdruck tatsächlich nicht erhöht ist, oder dass eine Person kein Medikament erhält, obwohl eine Medikation notwendig wäre.

KTT und Reliabilität

Messfehlertheorie

Unsystematische Messfehler**13.2.1 Arten von Messfehlern****■■ Unsystematische Messfehler**

Um eine zuverlässigere Messung zu erhalten, könnte die Blutdruckmessung mehrmals durchgeführt werden. Die Schwankungen der resultierenden Messungen können auf zufällige Messfehler zurückgeführt werden, wenn keine systematischen Einflüsse auf sie einwirken. Werden wiederholte Messungen durchgeführt und die Messergebnisse gemittelt, ergibt sich eine genauere Schätzung des tatsächlichen Wertes. Rein zufällige positive und negative Abweichungen von diesem Wert mitteln sich als *unsystematische Messfehler* bei wiederholten Messungen aus. Der mittlere Wert wiederholter Messungen liefert damit eine zuverlässigere (reliable) Schätzung des wahren Wertes (hier des Blutdrucks) als das Ergebnis nur einer Messung.

Systematische Messfehler**■■ Systematische Messfehler**

Messfehler können jedoch auch in systematischer Weise auftreten. Das ist z. B. dann der Fall, wenn das Messgerät nicht korrekt geeicht wurde und deshalb immer etwas erhöhte Werte angezeigt werden. Die Ursache systematischer Messfehler kann auch in der Person selbst liegen. Nimmt eine Person z. B. an, eine schwerwiegende Krankheit zu haben, und ist deshalb bei allen Blutdruckmessungen sehr aufgereggt, könnte die Blutdruckmessung dieser Person durch die blutdrucksteigernde Wirkung der Aufregung auch im Mittel systematisch erhöht sein. Der Einfluss solcher *systematischen Messfehler* würde sich durch Bildung des Mittelwertes über mehrfache Messungen *nicht* reduzieren.

Der Einfluss des nicht korrekt geeichten Messgeräts kann erst durch Hinzunahme eines weiteren Messinstruments (z. B. eines zweiten Blutdruckmessgeräts) identifiziert werden; der Einfluss der Aufregung könnte ggf. durch eine Langzeitblutdruckmessung identifiziert werden, wenn die Aufregung nach einiger Zeit abflacht.

Wahrer Wert und Fehlerwert**Definition des True-Scores τ_{vi}** **13.2.2 Definition des wahren Wertes (True-Score)**

Der KTT liegt die Annahme zugrunde, dass sich jeder beobachtbare Messwert y_{vi} einer Person v ($v = 1, \dots, n$) auf einer Variablen i (beispielsweise Item i , $i = 1, \dots, p$) zusammensetzt aus einem wahren Wert τ_{vi} (True-Score, kleines griechisches Tau) und einem Messfehler ε_{vi} (Fehlerwert, kleines griechisches Epsilon):

$$y_{vi} = \tau_{vi} + \varepsilon_{vi} \quad (13.1)$$

Der wahre Wert τ_{vi} einer Person v in einem Merkmal, gemessen mit einem Messinstrument/Item i , ist unbekannt und kann erst durch mehrfache Messung geschätzt werden. Der wahre Wert ist definiert als der personenbedingte Erwartungswert der Variablen y_{vi} (Guttman 1945; Lord und Novick 1968; Steyer und Eid 2001; Zimmerman 1975). Zur Verdeutlichung dieser Definition hilft das folgende Gedankenexperiment: Wird der Messwert y_{vi} einer Person v theoretisch unendlich oft mit demselben Item i erfasst, ergibt sich eine intraindividuelle Verteilung der Messwerte y_{vi} (gemessen mit Item i für Person v), deren Erwartungswert dem True-Score τ_{vi} , d. h. der wahren Ausprägung dieser Person auf Item i , entspricht:

$$\tau_{vi} := E(y_{vi}) \quad (13.2)$$

Während der wiederholten Messungen darf sich die Merkmalsausprägung der untersuchten Person nicht verändern und alle äußeren Einflüsse auf den Messwert müssen gleich bleiben, sodass Schwankungen in den Messungen ausschließlich

13.3 · Zerlegung einer Itemvariablen in True-Score- und Messfehlervariable

auf unsystematischen, d. h. zufälligen Messfehlern beruhen und nicht auf Veränderungen der Merkmalsausprägung oder Änderungen der Messbedingungen.

Der Messfehler einer Person ergibt sich entsprechend als Differenz aus dem beobachteten Wert und dem wahren Wert:

$$\varepsilon_{vi} = y_{vi} - \tau_{vi} \quad (13.3)$$

Anmerkung: In der Vergangenheit wurden die Grundannahmen der KTT oftmals als „Axiome“ bezeichnet. Diese Ansicht wird heute jedoch nur noch vereinzelt vertreten (z. B. von DeMars 2018), während sich allgemein die Einsicht durchgesetzt hat, dass aus der Definition der True-Score-Variablen als personenbedingtem Erwartungswert alle weiteren Eigenschaften der True-Score- und der Fehlervariablen abgeleitet werden können, sodass die frühere Bezeichnung der Grundannahmen als Axiome als obsolet betrachtet werden kann (Steyer und Eid 2001).

Messfehler einer Person

Axiome obsolet

13.3 Zerlegung einer Itemvariablen in True-Score- und Messfehlervariable

13.3.1 Grundgleichung der KTT

In einer Stichprobe von n Personen hat jede Person für jedes Item einen wahren Wert und einen Fehlerwert. Unter dieser Annahme kann die Itemvariable y_i , die die quantifizierten Antworten aller Personen auf ein Item i enthält, in die Variable der wahren Werte τ_i und in die Variable der Fehlerwerte ε_i zerlegt werden. Auf Ebene der Itemvariablen y_i lautet die Grundgleichung der KTT somit wie folgt:

$$y_i = \tau_i + \varepsilon_i \quad (13.4)$$

Hierbei bezeichnet τ_i die True-Score-Variable und ε_i die Messfehlervariable des Items i .

Grundgleichung der KTT auf Ebene der Itemvariablen

13.3.2 Eigenschaften der Messfehler- und True-Score-Variablen

Aus der Definition des wahren Wertes als personenbedingtem Erwartungswert (Gl. 13.2) lassen sich ohne weitere Annahmen verschiedene Eigenschaften der Messfehler- und True-Score-Variablen ableiten (Eid et al. 2017, S. 848 f.; Eid und Schmidt 2014; Raykov und Marcoulides 2011; Steyer und Eid 2001; Yousfi und Steyer 2006; Zimmerman 1975):

1. *Bedingter Erwartungswert der Messfehlervariablen: $E(\varepsilon_i | \tau_i) = 0$*

Der bedingte Erwartungswert der Messfehlervariablen ε_i , gegeben eine beliebige True-Score-Variable τ_i , ist null. Wird die Merkmalsausprägung einer Person mehrmals gemessen, so mitteln sich die Messfehler aus. Dasselbe Prinzip gilt auch für mehrere Personen, die jeweils einmal gemessen werden. Haben diese Personen denselben wahren Wert, so ist zu erwarten, dass die beobachteten Werte der Personen teilweise über und teilweise unter dem wahren Wert liegen werden und sich diese Messfehler über die Personen hinweg ausmitteln. Dieses Prinzip gilt auch bei Messung unterschiedlicher Merkmale: Der Erwartungswert des Messfehlers ist immer null und hängt weder von der True-Score-Variablen τ_i , noch von der True-Score-Variablen τ_j einer anderen Itemvariablen ab.

Eigenschaften der Messfehler- und True-Score-Variablen

2. *Unbedingter Erwartungswert der Messfehlervariablen: $E(\varepsilon_i) = 0$*
Aus Eigenschaft 1 folgt, dass auch der unbedingte Erwartungswert der Messfehlervariablen ε_i gleich null ist. Das bedeutet, dass der Erwartungswert der Messfehlervariablen unabhängig vom wahren Wert des gemessenen Merkmals null ist.
3. *Unkorreliertheit der Messfehler- und True-Score-Variablen: $Cov(\varepsilon_i, \tau_i) = 0$*
Da eine Messfehlervariable ε_i nur unsystematische Anteile enthält, ist sie mit jeder True-Score-Variablen τ_i oder τ_j unkorreliert. Damit ist auch die Kovarianz zwischen Messfehler- und True-Score-Variablen gleich null.
4. *Dekomposition der Varianz: $Var(y_i) = Var(\tau_i) + Var(\varepsilon_i)$*
Aus Eigenschaft 3 folgt, dass sich die Varianz einer Itemvariablen $Var(y_i)$ additiv zerlegen lässt in die Varianz der True-Score-Variablen $Var(\tau_i)$ und die Varianz der Messfehlervariablen $Var(\varepsilon_i)$.
5. *Dekomposition der Kovarianz: $Cov(y_i, y_{i'}) = Cov(\tau_i, \tau_{i'}) + Cov(\varepsilon_i, \varepsilon_{i'})$*
Auch die Kovarianz zwischen zwei Itemvariablen y_i und $y_{i'}$ lässt sich additiv zerlegen in die Kovarianz der True-Score-Variablen und die Kovarianz der Messfehlervariablen.

Bei Eindimensionalität muss die Zusatzannahme unkorrelierter Messfehler gelten

Zusätzlich wird häufig die Annahme getroffen, dass auch die Fehlervariablen untereinander unkorreliert sind ($Cov(\varepsilon_i, \varepsilon_{i'}) = 0$), was in empirischen Anwendungen aber nicht immer gegeben ist. Die Annahme unkorrelierter Fehler muss bei eindimensionalen Modellen erfüllt sein.

13.4 Testwertvariable Y und Testwerte Y_v

Werden latente Merkmale (Konstrukte) anhand mehrerer eindimensionaler Items gemessen, so entspricht dies einer Mehrfachmessung desselben Merkmals. Wird aus p Itemvariablen y_i ($i = 1, \dots, p$) durch Aufsummierung der individuellen Antworten eine *Testwertvariable Y* gebildet, so stellt diese Testwertvariable eine Linearkombination (ungewichtete Summe) der Itemvariablen dar.

$$Y = y_1 + y_2 + \dots + y_p \quad (13.5)$$

Entsprechend der Grundgleichung der KTT (Gl. 13.4) setzt sich jede der p Itemvariablen y_i aus einer True-Score-Variablen τ_i und einer Fehlervariablen ε_i zusammen. Somit setzt sich analog auch die Testwertvariable Y aus einer True-Score- und einer Fehlervariablen zusammen, wobei die True-Score-Variable T (großes griechisches Tau) der Testwertvariablen durch Aufsummierung der einzelnen Item-True-Scores τ_i und die Fehlervariable E (großes griechisches Epsilon) der Testwertvariablen durch Aufsummierung der einzelnen Item-Fehlervariablen ε_i gebildet wird:

$$Y = T + E = (\tau_1 + \dots + \tau_p) + (\varepsilon_1 + \dots + \varepsilon_p) \quad (13.6)$$

■ ■ **Beobachteter Testwert Y_v als Punktschätzung für den wahren Wert T_v**
Um das Testergebnis einer Person v zu bestimmen, addiert man die einzelnen individuellen Itemwerte y_{vi} zu einem Summenwert Y_v auf (► Kap. 7):

$$Y_v = \sum_{i=1}^p y_{vi} \quad (13.7)$$

Rohwert vs. Normwert

Dieser Summenwert wird üblicherweise als *Testwert* oder – im Unterschied zu Normwerten (► Kap. 9) – auch als *Rohwert* der Person v in dem jeweiligen Test bezeichnet.

Um zu zeigen, dass es sich beim Testwert Y_v gemäß Gl. (13.7) um eine Punktschätzung des gesuchten wahren Wertes T_v handelt, wird der Erwartungswert des

13.5 · Das Gütekriterium der Reliabilität

Testwertes untersucht (zu den verwendeten Rechenregeln s. z. B. Moosbrugger 1983, S. 47 f.).

$$E(Y_v) = E\left(\sum_{i=1}^p y_{vi}\right) = \sum_{i=1}^p E(y_{vi}) \quad (13.8)$$

Nach Gl. (13.2) lässt sich für $E(y_{vi})$ der wahre Wert τ_{vi} einsetzen:

$$E(Y_v) = \sum_{i=1}^p \tau_{vi} = T_v \quad (13.9)$$

Der Erwartungswert von Y_v entspricht damit dem wahren Wert T_v . Somit kann der Summenwert Y_v als Punktschätzung \hat{T}_v des wahren Wertes T_v einer bestimmten Person v verwendet werden:

$$Y_v = \hat{T}_v \quad (13.10)$$

Punktschätzungen sollten nur bei hoher Reliabilität interpretiert werden; eine Intervallschätzung für den wahren Wert wird z. B. unter Zuhilfenahme der Reliabilität in ► Abschn. 13.8.1.1 behandelt.

Punktschätzung des wahren Wertes T_v

13.5 Das Gütekriterium der Reliabilität

Ein wesentliches Ziel der KTT besteht darin, die wahre Varianz und die Fehlervarianz einer beobachteten Variablen (Itemvariable, Testwertvariable) zu schätzen, um eine Aussage über die Messgenauigkeit (Reliabilität) treffen zu können. Die Reliabilität ist ein äußerst wichtiges Gütekriterium eines Tests und wird als Anteil der wahren Varianz an der Gesamtvarianz geschätzt. Die Varianzzerlegung einer Testwertvariablen in die Varianz der True-Score-Variablen und in die Varianz der Fehlervariablen (vgl. Eigenschaft 4 der Messfehler- und True-Score-Variablen in ► Abschn. 13.3.2) ist für die Bestimmung der Reliabilität zentral.

Reliabilitätsbestimmung als wesentliches Ziel der KTT

Um die Reliabilität schätzen zu können, müssen mehrere Messungen eines Merkmals vorliegen. Dies können mehrere Items innerhalb eines Tests, mehrere Messungen anhand desselben Tests zu unterschiedlichen Messzeitpunkten oder mehrere Messungen anhand verschiedener Tests sein (vgl. z. B. ► Kap. 14; Bandalos 2018; Eid und Schmidt 2014).

Mehrfachmessungen nötig

13.5.1 Reliabilität einer Itemvariablen

Itemreliabilität

Da sowohl die Werte der True-Score-Variablen τ_i als auch die Werte der Fehlervariablen ε_i bei einer Messung unbekannt sind, können die True-Score- und die Messfehleranteile erst bei Mehrfachmessungen bestimmt werden. Anhand von Mehrfachmessungen ist die Bestimmung der Reliabilität sowohl für jede Itemvariable als auch für die Testwertvariable möglich. Die Reliabilität einer Itemvariablen wird bestimmt über den Anteil der True-Score-Varianz an der Gesamtvarianz dieser Itemvariablen: $Var(\tau_i) / Var(y_i)$.

13.5.2 Reliabilität einer Testwertvariablen

Das Gütekriterium der Reliabilität bezieht sich in der Regel auf Testwertvariablen, d. h. auf die zu Testwerten aufsummierten Itemwerte. Die Reliabilität ist eines der

Reliabilität als differentialpsychologisches Maß

wichtigsten Kennwerte der KTT und stellt ein Maß für die Zuverlässigkeit einer Messung und somit auch ein Maß der Messfehlerfreiheit dar. Die Reliabilität gibt an, in welchem Ausmaß beobachtbare interindividuelle Unterschiede der Messungen eines Merkmals durch wahre Merkmalsunterschiede bedingt sind. Somit ist die Reliabilität ein differentialpsychologisches Maß, das darüber Auskunft gibt, wie genau interindividuelle Unterschiede zwischen Personen erfasst werden können (Eid et al. 2017).

Definition der Reliabilität

Definition

Die **Reliabilität** ist ein Maß für die Genauigkeit einer Messung. Sie ist definiert als das Verhältnis der Varianz der True-Score-Variablen $Var(T)$ zur Gesamtvarianz der Testwertvariablen $Var(Y)$:

$$Rel(Y) = \frac{Var(T)}{Var(Y)} = \frac{Var(T)}{Var(T) + Var(E)} \quad (13.11)$$

Der Wertebereich der Reliabilität liegt zwischen 0 (keine Reliabilität) und 1 (höchste Reliabilität).

Wertebereich des Reliabilitätskoeffizienten

Das Verhältnis der Varianz der True-Score-Variablen zur Varianz der Testwertvariablen (Gl. 13.11) wird als *Reliabilitätskoeffizient* bezeichnet. Der Wertebereich des Reliabilitätskoeffizienten liegt zwischen 0 und 1. Der Maximalwert 1 bedeutet, dass die Testwertvarianz nur aus wahrer Varianz besteht, sodass gilt: $Var(Y) = Var(T)$. Der Minimalwert von 0 gibt dagegen an, dass die Testwertvarianz keine wahre Varianz enthält, sondern nur aus Fehlervarianz besteht, sodass gilt: $Var(Y) = Var(E)$. Ein Testergebnis, d. h. eine Messung anhand eines Tests, ist also umso reliabler, je größer der wahre Varianzanteil $Var(T)$ an der Gesamtvarianz $Var(Y)$ ist. Umgekehrt gilt auch, dass die Reliabilität bei zunehmender Fehlervarianz abnimmt (wegen $Var(T) = Var(Y) - Var(E)$; s. auch Gl. 13.4).

Die additive Varianzzerlegung stellt die Grundlage der Reliabilitätsbestimmung dar (► Abschn. 13.6.2): Sowohl die klassischen Methoden der Reliabilitätsschätzung (► Kap. 14) als auch die modellbasierten Methoden der Reliabilitätsschätzung (► Kap. 15) beruhen auf dieser Varianzzerlegung.

Gemäß Gl. (13.11) ist die Reliabilität ein Varianzverhältnis und kann als Determinationskoeffizient interpretiert werden, der angibt, zu welchem Anteil die Testwertvarianz durch die True-Score-Variable determiniert ist (Yousfi und Steyer 2006). Dieser Koeffizient stellt ein normiertes Effektgrößenmaß dar (Eid und Schmidt 2014); er ermöglicht es, Messungen anhand von unterschiedlichen Messinstrumenten miteinander zu vergleichen.

Schätzung der wahren Varianz und der Fehlervarianz

Wie können nun die unbekannten Varianzen $Var(T)$ und $Var(E)$ (Gl. 13.11) geschätzt werden? Unter Verwendung der konfirmatorischen Faktorenanalyse (CFA, ► Kap. 24) liefern die Messmodelle für τ -kongenerische Variablen, essentiell τ -äquivalente Variablen und essentiell τ -parallele Variablen (► Abschn. 13.6.1) diese Kennwerte.

13.6 Messmodelle zur Schätzung der Reliabilität

Liegen zur Messung eines latenten Konstrukt mehrere Messungen (z. B. in Form von Items) vor, kann anhand eines Messmodells geprüft werden, ob die Items tatsächlich dasselbe Merkmal messen. Diese und weitere Voraussetzungen müssen

13.6 · Messmodelle zur Schätzung der Reliabilität

erfüllt sein, damit die Reliabilität der Testwertvariablen (► Abschn. 13.8.1) und darüber hinaus auch die latenten Personenwerte (► Abschn. 13.8.2) geschätzt werden können.

Liegt den Messungen *ein* gemeinsames latentes Konstrukt („Faktor“) zugrunde, entspricht dies der Grundannahme eines *eindimensionalen Messmodells*. Das bedeutet, dass alle Items *ein* gemeinsames Merkmal η messen und die korrelativen Zusammenhänge zwischen den Itemvariablen somit nur von einer einzigen latenten Variablen η erklärt werden. Diese Grundannahme impliziert, dass die *Messfehler unkorreliert* sind.

Eindimensionalität

Messmodellgleichung

Die True-Score-Variable τ_i eines Items i kann als Funktion der latenten Variablen η dargestellt werden. Sie ergibt sich aus der Summe des Leichtigkeitsparameters (Interzept) α_i und der mit dem Diskriminationsparameter (Faktorladung λ_i) gewichteten latenten Variablen η :

$$\tau_i = \alpha_i + \lambda_i \cdot \eta \quad (13.12)$$

Setzt man Gl. (13.12) in Gl. (13.4) ein, lässt sich entsprechend auch die beobachtete Itemvariable y_i in Abhängigkeit von der latenten Variablen η und dem Messfehler ε_i in Form einer Messmodellgleichung darstellen:

$$y_i = \tau_i + \varepsilon_i = \alpha_i + \lambda_i \cdot \eta + \varepsilon_i \quad (13.13)$$

Die Messmodellgleichung verdeutlicht, dass die Beziehung zwischen der latenten Variablen η und der Itemvariablen y_i abhängig ist von der Ausprägung des Leichtigkeitsparameters α_i und der Ausprägung des Diskriminationsparameters λ_i .

13.6.1 Unterschiedliche Stufen der Messäquivalenz

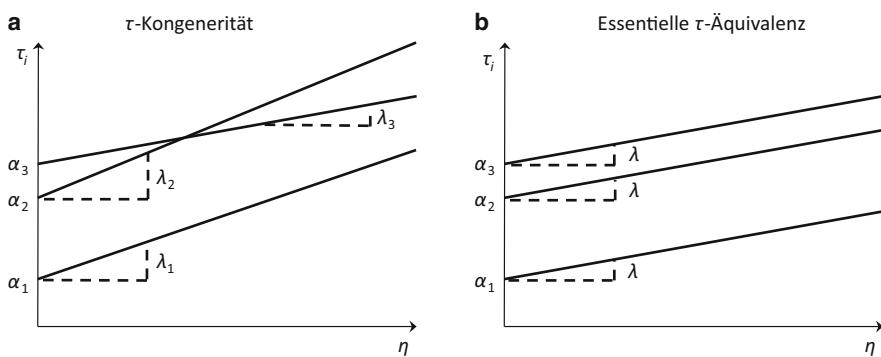
Die für die Reliabilität relevanten Messmodelle unterscheiden sich hinsichtlich der den Messungen zugrunde liegenden Stufen der Messäquivalenz. Je nach Stufe der Messäquivalenz weisen die Items andere Messeigenschaften auf, die im Messmodell definiert werden. Die Messmodelle unterscheiden sich bezüglich der Gleichheit oder Ungleichheit der Faktorladungen (*Diskriminationsparameter*), der Interzepte (*Leichtigkeitsparameter*) und der *Fehlervarianzen* der Itemvariablen. Die jeweils angenommenen Modelleigenschaften sind empirisch überprüfbar („testbar“). Je mehr Modellrestriktionen in Form von Gleichheitsannahmen für die Itemparameter getroffen werden, desto strenger bzw. restriktiver ist das zu prüfende Messmodell und desto weniger Parameter müssen geschätzt werden.

Messmodelle mit unterschiedlichen Modellrestriktionen

Erstrebenswert ist ein möglichst strenges Modell mit wenigen zu schätzenden Parametern, da es „sparsamer“ („parsimonious“) ist als Modelle mit vielen zu schätzenden Parametern. Nach dem wissenschaftlichen Parsimonitätsprinzip würde ein möglichst sparsames Modell bevorzugt werden, sofern es die Datenstruktur ebenso gut beschreibt wie ein weniger restriktives Modell (s. Eid et al. 2017, S. 787; Raykov und Marcoulides 2011).

Prinzip der Parsimonität

Drei Modelle mit zunehmend restriktiveren Annahmen bezüglich der Gleichheit der Parameter sollen nachfolgend etwas ausführlicher behandelt werden: das Modell τ -kongenerischer Variablen, das keine Gleichheitsrestriktionen aufweist (► Abschn. 13.6.1.1), das Modell essentiell τ -äquivalenter Variablen (► Abschn. 13.6.1.2) und das Modell essentiell τ -paralleler Variablen, das die meisten Gleichheitsrestriktionen aufweist (► Abschn. 13.6.1.3). Die Modelle sind hierarchisch



■ Abb. 13.1 Beziehung zwischen den True-Score-Variablen τ_i und der latenten Merkmalsvariablen η für drei Itemvariablen mit unterschiedlichen Interzepten α_1 bis α_3 und **a** unterschiedlichen Diskriminationsparametern λ_1 bis λ_3 (τ -Kongenerität) sowie **b** identischen Diskriminationsparametern $\lambda_i = \lambda$ (essentielle τ -Äquivalenz)

geschachtelt (vgl. Eid et al. 2017, S. 870). Weitere Modelle werden in Eid und Schmidt (2014) behandelt.

13.6.1.1 Modell τ -kongenerischer Variablen

τ-Kongenerität

Das Modell τ -kongenerischer Variablen (Jöreskog 1971) geht von eindimensionalen Messungen aus, bei denen sich die latente Variable η auf jede Itemvariable unterschiedlich stark auswirken kann. Es ist das am wenigsten restriktive Modell, da weder für die Diskriminationsparameter noch für die Leichtigkeitsparameter Gleichheitsrestriktionen vorliegen. Die Parameter dürfen im τ -kongenerischen Modell über die Items hinweg frei variieren und somit unterschiedliche Werte annehmen. Auch die Fehlervarianzen dürfen sich unterscheiden.

In faktorenanalytischer Terminologie zeigen sich die unterschiedlichen Diskriminationsparameter der Itemvariablen y_i in unterschiedlichen *Faktorladungen* (λ_i) (Näheres zu faktorenanalytischen Verfahren ► Kap. 23 sowie ► Kap. 24).

Des Weiteren können Items eines Leistungstests leichter oder schwerer zu lösen sein bzw. Items eines Persönlichkeitstests leichter oder schwerer symptomatisch im Sinne des Konstrukt zu beantworten sein. Für die Itemvariablen y_i werden deshalb unterschiedliche Leichtigkeitsparameter – oftmals etwas missverständlich auch als Itemschwierigkeiten bezeichnet – angenommen; in faktorenanalytischer Terminologie zeigen sich die Leichtigkeitsparameter in unterschiedlichen *Interzepten* α_i .

! In der einschlägigen Literatur wird das Interzept häufig mit α bezeichnet, daher wird diese Notation auch hier verwendet. Das Interzept α ist dabei weder zu verwechseln mit dem α -Fehlerrisiko in der Inferenzstatistik noch mit dem Reliabilitätskoeffizienten α von Cronbach (► Abschn. 13.6.2)!

In ■ Abb. 13.1 ist der Zusammenhang zwischen der latenten Variablen η und den True-Score-Variablen τ_i der i Items für τ -kongenerische Variablen (■ Abb. 13.1a) und für essentiell τ -äquivalente Variablen (■ Abb. 13.1b) dargestellt. Im Modell τ -kongenerischer Variablen (■ Abb. 13.1a) dürfen sich sowohl die Leichtigkeitsparameter (Interzepte α_1 bis α_3) als auch die Diskriminationsparameter (Faktorladungen λ_1 bis λ_3) der Items unterscheiden. Dies wird deutlich durch die unterschiedlichen Achsenabschnitte und Steigungen der Geraden. Dieses Modell setzt lediglich voraus, dass eindimensionale Messungen eines gemeinsamen latenten Konstrukt vorliegen.

13.6.1.2 Modell essentiell τ -äquivalenter Variablen

Essentielle τ -Äquivalenz

Im Modell essentiell τ -äquivalenter Variablen ist die Messäquivalenz strenger gefasst als im Modell τ -kongenerischer Variablen. Bei essentieller τ -Äquivalenz

(Abb. 13.1b) wird angenommen, dass die Diskriminationsparameter (Faktorladungen λ_i) aller Itemvariablen identisch sind und einen konstanten Wert λ aufweisen. Wie im τ -kongenerischen Modell dürfen sich die Leichtigkeitsparameter α_i unterscheiden. Dies wird deutlich durch unterschiedliche Achsenabschnitte (Interzepte α_1 bis α_3) bei identischen Steigungen und somit Parallelität der Geraden.

13.6.1.3 Modell essentiell τ -paralleler Variablen

Eine noch strengere Stufe der Messäquivalenz wird im Modell essentiell τ -paralleler Variablen angenommen. In diesem Fall dürfen sich die Itemvariablen wie in den beiden anderen Modellen in Bezug auf ihre Leichtigkeitsparameter (Interzepte) unterscheiden; die Diskriminationsparameter (Faktorladungen) hingegen müssen wie im Modell essentiell τ -äquivalenter Variablen gleich sein. Zusätzlich müssen jedoch auch die Fehlervarianzen aller Itemvariablen denselben Wert aufweisen: $Var(\varepsilon_1) = Var(\varepsilon_2) = \dots = Var(\varepsilon_p)$.

Essentielle τ -Parallelität

Anmerkung: Der Begriff „essentiell“ bedeutet, dass sich die Itemvariablen hinsichtlich der additiven Konstanten (Interzepte) unterscheiden können. Werden dagegen zusätzlich gleiche Interzepte für alle Messungen eines latenten Konstruktus angenommen, resultieren die noch restriktiveren Modelle der τ -Äquivalenz und der τ -Parallelität. Diese äußerst restriktiven Modelle spielen jedoch für die Varianz – und damit für die Reliabilitätsbestimmung – keine Rolle und werden daher hier nicht weiter behandelt.

13.6.2 Herleitung der wahren Item- und Testwertvarianzen in Abhängigkeit von den Modellparametern

Die beschriebenen Stufen der Messäquivalenz spielen bei der Bestimmung der modellbasierten Varianzen und Kovarianzen und damit bei der Reliabilitätsschätzung eine wichtige Rolle, da die verschiedenen Reliabilitätsmaße in Abhängigkeit von den Modellparametern jeweils unterschiedliche Messmodelle implizieren. Die Relevanz der je nach Messäquivalenzmodell unterschiedlichen modellbasierten Varianzen und Kovarianzen wird deutlich, wenn man sich die Definition der Reliabilität als Varianzverhältnis (Gl. 13.11) in Erinnerung ruft. Basiert ein Reliabilitätsmaß beispielsweise auf dem Messmodell essentiell τ -äquivalenter Variablen (d.h., es liegen identische Faktorladungen der Itemvariablen mit der impliziten Annahme identischer Itemkovarianzen vor), wäre die Schätzung der wahren Varianz und damit der Reliabilität verzerrt, wenn tatsächlich unterschiedliche Itemkovarianzen vorlägen. Deshalb soll im Folgenden gezeigt werden, wie die wahren Varianzen und Kovarianzen in Abhängigkeit von den Modellparametern ausgedrückt werden können.

Verschiedene Reliabilitätsmaße implizieren unterschiedliche Messmodelle

13.6.2.1 Modellbasierte Zerlegung der Itemvarianz und -kovarianz

Die *Itemvarianz* (Varianz der Itemvariablen y_i) setzt sich aus True-Score-Varianz und Fehlervarianz zusammen. Die True-Score-Varianz kann mit Gl. (13.13) modellbasiert anhand der mit dem Diskriminationsparameter (Faktorladung) λ_i gewichteten Varianz der latenten Variablen η geschätzt werden:

$$\begin{aligned} Var(y_i) &= Var(\tau_i) + Var(\varepsilon_i) \\ &= Var(\alpha_i + \lambda_i \cdot \eta) + Var(\varepsilon_i) \end{aligned} \quad (13.14)$$

Die Varianz von τ_i , ausgedrückt als Linearkombination ($\alpha_i + \lambda_i \cdot \eta$) in Gl. (13.14), setzt sich nach einer allgemeinen Regel aus vier Kovarianztermen zusammen:

$$\begin{aligned} \text{Var}(\tau_i) &= \text{Var}(\alpha_i + \lambda_i \cdot \eta) \\ &= \text{Cov}(\alpha_i, \alpha_i) + \text{Cov}(\alpha_i, \lambda_i \cdot \eta) + \text{Cov}(\lambda_i \cdot \eta, \alpha_i) \\ &\quad + \text{Cov}(\lambda_i \cdot \eta, \lambda_i \cdot \eta) \end{aligned}$$

Da alle Kovarianzen mit der Konstanten α_i gleich null sind, verbleibt von den vier Termen lediglich $\text{Cov}(\lambda_i \cdot \eta, \lambda_i \cdot \eta)$. Die Kovarianz der Variablen $\lambda_i \cdot \eta$ mit sich selbst ist gleich ihrer Varianz: $\text{Cov}(\lambda_i \cdot \eta, \lambda_i \cdot \eta) = \text{Var}(\lambda_i \cdot \eta)$. Für die Varianz des Produkts einer Konstanten (hier λ_i) mit einer Variablen (hier η) gilt weiter, dass die Konstante quadriert, vorgezogen und mit der Varianz der Variablen multipliziert wird. Somit gilt $\text{Var}(\tau_i) = \lambda_i^2 \cdot \text{Var}(\eta)$, woran man sieht, dass die Unterschiede in den wahren Varianzen der verschiedenen Items nur auf die Faktorladungen λ_i zurückgehen. Die modellbasierte Varianzzerlegung der Itemvariablen lautet somit:

$$\text{Var}(y_i) = \lambda_i^2 \cdot \text{Var}(\eta) + \text{Var}(\varepsilon_i) \quad (13.15)$$

Die *Kovarianz* zwischen zwei Itemvariablen y_i und $y_{i'}$, die dasselbe Konstrukt η messen, kann als Kovarianz zwischen den True-Score-Variablen der beiden Variablen, τ_i und $\tau_{i'}$, ausgedrückt werden (Gl. 13.16), da alle Kovarianzen der True-Score-Variablen τ_i und $\tau_{i'}$ mit den Fehlervariablen ε_i und $\varepsilon_{i'}$ null sind (vgl. Eigenschaft 1 der True-Score- und Fehlervariablen in ► Abschn. 13.3.2) und da im eindimensionalen Modell außerdem auch die Messfehler als unkorreliert angenommen werden. Somit reduziert sich die Gleichung der Kovarianz der bei beiden Itemvariablen auf die Kovarianz der True-Score-Variablen:

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \text{Cov}(\tau_i + \varepsilon_i, \tau_{i'} + \varepsilon_{i'}) \\ &= \text{Cov}(\tau_i, \tau_{i'}) \end{aligned} \quad (13.16)$$

Zerlegung der Itemkovarianz

Zerlegung der True-Score-Kovarianz

Die Kovarianz der True-Score-Variablen τ_i und $\tau_{i'}$ lässt sich nun als Produkt der Faktorladungen der beiden Items i und i' und der Varianz von η darstellen.

Nach Gl. (13.12) können die True-Score-Variablen τ_i und $\tau_{i'}$, die beide dieselbe latente Variable η messen, wie folgt ausgedrückt werden:

$$\tau_i = \alpha_i + \lambda_i \cdot \eta \quad \text{und} \quad \tau_{i'} = \alpha_{i'} + \lambda_{i'} \cdot \eta \quad (13.17)$$

Die Kovarianz zwischen den True-Score-Variablen ist somit:

$$\text{Cov}(\tau_i, \tau_{i'}) = \text{Cov}(\alpha_i + \lambda_i \cdot \eta, \alpha_{i'} + \lambda_{i'} \cdot \eta) \quad (13.18)$$

Da die Interzepte α_i und $\alpha_{i'}$ als Konstanten weder miteinander noch mit der Variablen η kovariieren können, entfallen diese Koeffizienten. Die Gleichung reduziert sich damit zu:

$$\text{Cov}(\tau_i, \tau_{i'}) = \text{Cov}(\lambda_i \cdot \eta, \lambda_{i'} \cdot \eta) \quad (13.19)$$

Die Kovarianz der latenten Variablen η mit sich selbst ist gleich ihrer Varianz: $\text{Cov}(\eta, \eta) = \text{Var}(\eta)$. Zur Berechnung der Kovarianz der True-Score-Variablen werden nun die Faktorladungen λ_i und $\lambda_{i'}$ vorgezogen und mit der Varianz von η multipliziert:

$$\text{Cov}(\tau_i, \tau_{i'}) = \lambda_i \cdot \lambda_{i'} \cdot \text{Var}(\eta) \quad (13.20)$$

Wie man sieht, lässt sich die Kovarianz der True-Score-Variablen τ_i und $\tau_{i'}$ als Produkt der Faktorladungen der beiden Items i und i' und der Varianz von η darstellen. Zusammenfassend konnte gezeigt werden, dass sowohl die Varianzen der Itemvariablen (Gl. 13.15) als auch die Kovarianzen der True-Score-Variablen (Gl. 13.20) in Parametern des Messmodells ausgedrückt werden können.

■■ Gleichheitsrestriktionen der Modellparameter

Aus den Gleichheitsrestriktionen der Modelle essentieller τ -Äquivalenz und essentieller τ -Parallelität ergeben sich sowohl für die empirischen Varianzen und Kovarianzen der Items als auch für die Varianz der latenten Variablen wesentliche und je nach Modell unterschiedliche Konsequenzen.

Gleichheitsrestriktion der Diskriminationsparameter λ_i : Sind die Diskriminationsparameter λ_i aller Items eines Tests identisch, so weisen die Itemvariablen denselben Anteil an wahrer Varianz ($\lambda_i^2 \cdot \text{Var}(\eta) = \text{Var}(\tau)$ Gl. 13.15) auf. Damit sind auch die Kovarianzen aller Itempaare identisch (Gl. 13.16). Sind die Diskriminationsparameter gleich 1, entspricht die Varianz der latenten Variablen $\text{Var}(\eta)$ der True-Score-Varianz der Items $\text{Var}(\tau)$ (Gl. 13.15) und gleichzeitig auch genau der Kovarianz der Items (Gl. 13.16).

Gleichheitsrestriktion der Fehlervarianz $\text{Var}(\varepsilon_i)$: Sind zusätzlich zu den Diskriminationsparametern auch die Fehlervarianzen aller Items identisch, so weisen alle Itemvariablen dieselbe Varianz auf (Gl. 13.15).

! Modellabhängige Implikationen für Varianzen und Kovarianzen

Aus Gl. (13.15) und (13.20) wird deutlich, dass die verschiedenen Messmodelle sehr unterschiedliche Annahmen hinsichtlich der Varianzen bzw. Kovarianzen der Items implizieren (vgl. ▶ Kap. 14):

- Im Modell der τ -Kongenerität sind Faktorladungen und Fehlervarianzen der Items variabel, sodass die Items unterschiedliche Varianzen und Kovarianzen aufweisen können.
- Im Modell essentieller τ -Äquivalenz sind alle Faktorladungen identisch; dies impliziert, dass alle Itemvariablen dieselbe Kovarianz untereinander aufweisen.
- Im Modell essentieller τ -Parallelität sind zusätzlich zu den Faktorladungen auch alle Fehlervarianzen identisch; dies impliziert, dass alle Itemvariablen dieselbe Kovarianz untereinander und außerdem auch dieselbe Varianz aufweisen.

Die Annahmen der essentiellen τ -Äquivalenz und essentiellen τ -Parallelität sind sehr streng, sodass empirische Daten diesen Modellannahmen oftmals nicht entsprechen.

13.6.2.2 Modellbasierte Zerlegung der Testwertvarianz

Die Testwertvarianz $\text{Var}(Y)$ setzt sich aus der True-Score-Varianz und der Fehlervarianz zusammen, da die True-Score-Variablen T mit den Messfehlervariablen E unkorreliert ist (Eigenschaft 3 der Messfehler- und True-Score-Variablen in ▶ Abschn. 13.3.2):

$$\text{Var}(Y) = \text{Var}(T + E) = \text{Var}(T) + \text{Var}(E) \quad (13.21)$$

True-Score- und Messfehlervarianz

Die Varianz der Fehlervariablen E ergibt sich somit als Differenz aus der Varianz der Testwertvariablen Y und der Varianz der True-Score-Variablen T :

$$\text{Var}(E) = \text{Var}(Y) - \text{Var}(T) \quad (13.22)$$

Da sich die Testwertvariable Y nach Gl. (13.6) aus den aufsummierten True-Score-Variablen und den aufsummierten Fehlervariablen der Items zusammensetzt, gilt für die Varianz der Testwertvariablen Y :

$$\text{Var}(Y) = \text{Var}\left(\sum_{i=1}^p \tau_i + \sum_{i=1}^p \varepsilon_i\right) \quad (13.23)$$

Aus der allgemeinen Regel, dass sich die Varianz einer Linearkombination aus den Varianzen der einzelnen Summanden plus der zweifachen Kovarianz der Summanden zusammensetzt, folgt:

$$\text{Var}(Y) = \sum_{i=1}^p \text{Var}(\tau_i) + \sum_{i=1}^p \text{Var}(\varepsilon_i) + 2 \cdot \sum_{i < i'} \text{Cov}(\tau_i + \varepsilon_i, \tau_{i'} + \varepsilon_{i'}) \quad (13.24)$$

Da alle Kovarianzen mit den Fehlervariablen gleich null sind, reduziert sich Gl. (13.24) zu

$$\text{Var}(Y) = \sum_{i=1}^p \text{Var}(\tau_i) + 2 \cdot \sum_{i < i'} \text{Cov}(\tau_i, \tau_{i'}) + \sum_{i=1}^p \text{Var}(\varepsilon_i) \quad (13.25)$$

Dabei entspricht die Summe der ersten beiden Summanden der True-Score-Varianz der Testwertvariablen $\text{Var}(T)$:

$$\text{Var}(T) = \sum_{i=1}^p \text{Var}(\tau_i) + 2 \cdot \sum_{i < i'} \text{Cov}(\tau_i, \tau_{i'}) \quad (13.26)$$

Die True-Score-Varianz der Testwertvariablen lässt sich unter Verwendung von Gl. (13.12) ($\tau_i = \alpha_i + \lambda_i \cdot \eta$) in Abhängigkeit von den Modellparametern darstellen, wobei das Interzept α_i als additive Konstante bei der Varianzberechnung entfällt, da die Varianz einer Konstanten null ist ($\text{Var}(\alpha_i) = 0$):

$$\begin{aligned} \text{Var}(T) &= \sum_{i=1}^p \text{Var}(\lambda_i \cdot \eta) + 2 \cdot \sum_{i < i'} \lambda_i \cdot \lambda_{i'} \cdot \text{Var}(\eta) \\ &= \sum_{i=1}^p \lambda_i^2 \cdot \text{Var}(\eta) + 2 \cdot \sum_{i < i'} \lambda_i \cdot \lambda_{i'} \cdot \text{Var}(\eta) \end{aligned} \quad (13.27)$$

! Mit dieser Herleitung der wahren Varianz der Testwertvariablen in Abhängigkeit von den Modellparametern ist der gesuchte Ausdruck $\text{Var}(T)$ im Zähler der Reliabilitätsdefinition (Gl. 13.11) gefunden und lässt sich im Rahmen der CFA schätzen.

13.6.3 Überprüfung der Messäquivalenz

Die Messmodelle mit ihren spezifischen Modellannahmen können explizit formuliert und anhand der CFA überprüft werden (► Kap. 24). Anhand eines Modelltests kann geprüft werden, welches Messmodell zu den Daten passt und welches Reliabilitätsmaß (► Abschn. 13.6.4) entsprechend verwendet werden darf.

Die vorgestellten Messmodelle zur Überprüfung der Messäquivalenz (► Tab. 13.1) können als Varianten eines einfaktoriellen Modells aufgefasst werden. Die Überprüfung der Modellgültigkeit und die Schätzung der Modellparameter können anhand von Schätzmethoden erfolgen, die von Computerprogrammen für lineare Strukturgleichungsmodelle, zu denen die CFA gehört, zur Verfügung gestellt werden, z. B. *Mplus* (Muthén und Muthén 2017) oder R-Paket *lavaan* (Rosseel 2012). Am häufigsten wird die Maximum-Likelihood-Methode verwendet. Diese beruht auf der Annahme multivariat-normalverteilter manifesten (beobachteten) Variablen (z. B. Itemvariablen) und bestimmt die Passung des Modells zu den Daten anhand des χ^2 -Tests.

Mit dem χ^2 -Test wird die Nullhypothese überprüft, dass sich die vom Modell implizierte Kovarianzmatrix (und falls von Interesse der Mittelwertevektor) nicht

Tabelle 13.1 Charakteristika der Messmodelle bei τ -Kongenerität, bei essentieller τ -Äquivalenz und bei essentieller τ -Parallelität sowie adäquate Reliabilitätsmaße

	τ -Kongenerität	Essentielle τ -Äquivalenz	Essentielle τ -Parallelität
Messmodell-gleichung für y_i	$y_i = \alpha_i + \lambda_i \cdot \eta + \varepsilon_i$	$y_i = \alpha_i + \lambda \cdot \eta + \varepsilon_i$	$y_i = \alpha_i + \lambda \cdot \eta + \varepsilon_i$
True-Score-Variablen τ_i	$\tau_i = \alpha_i + \lambda_i \cdot \eta$	$\tau_i = \alpha_i + \lambda \cdot \eta$	$\tau_i = \alpha_i + \lambda \cdot \eta$
Annahmen	1. Eindimensionalität	1. Eindimensionalität 2. Identische Diskriminationsparameter (Faktorladungen): $\lambda_i = \lambda_{i'} = \lambda$	1. Eindimensionalität 2. Identische Diskriminationsparameter (Faktorladungen): $\lambda_i = \lambda_{i'} = \lambda$ 3. Identische Fehlervarianzen: $Var(\varepsilon_i) = Var(\varepsilon_{i'})$
Reliabilitätsmaß	<i>McDonalds Omega</i>	<i>Cronbachs Alpha</i>	<i>Spearman-Brown-Formel der Testverlängerung</i>

Anmerkung: Der Index i bezeichnet das Item ($i = 1, \dots, p; i \neq i'$), α den Leichtigkeitsparameter (Interzept) und λ die Faktorladung (Diskriminationsparameter).

von der beobachteten Kovarianzmatrix (und dem beobachteten Mittelwertvektor) unterscheidet. Die resultierende Prüfgröße ist χ^2 -verteilt mit $df = s - t$ Freiheitsgraden (df), wobei s die Anzahl der empirischen Informationen (Varianzen, Kovarianzen und ggf. Mittelwerte) und t die Anzahl zu schätzender Modellparameter (Diskriminationsparameter, Fehlervarianzen, Varianz der latenten Variablen und ggf. Leichtigkeitsparameter) bezeichnet. Ein *nicht* signifikanter χ^2 -Wert ($p > .01$) zeigt an, dass das Modell zu den Daten passt, während ein signifikanter p -Wert ($p \leq .01$) zur Ablehnung des Modells führen würde.

Ist die Normalverteilungsannahme verletzt, wie es bei grobstufigen Itemvariablen häufig der Fall ist, so kann die robuste Maximum-Likelihood-Methode verwendet werden, mit der eine Korrektur des χ^2 -Tests und der geschätzten Standardfehler der Parameter vorgenommen wird. Alternativ kann auf verteilungsfreie Methoden zur Schätzung der Modellparameter zurückgegriffen werden (für einen kurzen Überblick ► Kap. 24).

Der χ^2 -Test ist der einzige inferenzstatistische Test zur Beurteilung der Modellgüte (vgl. Bollen 1989; Schermelleh-Engel et al. 2003) und sollte deshalb immer berichtet werden.

Weiter stehen deskriptive Gütekriterien zur Modellbeurteilung zur Verfügung, z. B. der Root Mean Square Error of Approximation (RMSEA) oder der Comparative Fit Index (CFI, ► Kap. 24), die anhand von Daumenregeln zur Beurteilung der approximativen Modellpassung interpretiert werden.

Robuste Maximum-Likelihood-Methode

Deskriptive Gütemaße

13.6.4 Messäquivalenz-Implikationen ausgewählter Reliabilitätskoeffizienten

Wie beschrieben implizieren die verschiedenen Stufen der Messäquivalenz entweder die Gleichheit oder Ungleichheit der Itemvarianzen und -kovarianzen. Da die Reliabilität als Varianzverhältnis definiert ist (Gl. 13.11) und den verschiedenen Reliabilitätsmaßen jeweils unterschiedliche Annahmen der Messäquivalenz zugrunde liegen, sind auch diese Unterschiede bezüglich der modellbasierten Varianzen und Kovarianzen für die Reliabilitätsschätzung zentral.

Verzerrte Reliabilitätsschätzung bei Verletzung der Voraussetzungen

Reliabilitätsschätzung abhängig von der Stufe der Messäquivalenz

McDonalds Omega

Cronbachs Alpha

Spearman-Brown-Formel der Testverlängerung

Basiert ein Reliabilitätsmaß beispielsweise auf dem Modell essentiell τ -äquivalenter Variablen (d. h. auf einem Messmodell mit identischen Faktorladungen der Itemvariablen und folglich mit der impliziten Annahme identischer Itemkovarianzen, ► Abschn. 13.6.2.1), wäre die Schätzung der wahren Varianz und damit der Reliabilität verzerrt, wenn tatsächlich unterschiedliche Itemkovarianzen vorlägen.

Daher muss zunächst anhand der CFA (► Kap. 24) überprüft werden, welche Stufe der Messäquivalenz vorliegt (► Abschn. 13.6.3). So kann entschieden werden, wie die True-Score-Varianz der Testwertvariablen $Var(T)$ zu berechnen ist und welcher der im Folgenden aufgeführten Reliabilitätskoeffizienten zur Bestimmung der Messgenauigkeit verwendet werden soll. Hierbei handelt es sich um Punktschätzungen der Reliabilität (zur Bestimmung der zugehörigen Konfidenzintervalle ► Abschn. 13.6.5).

McDonalds Omega (► Kap. 15) basiert auf dem Modell der τ -Kongeneritität und führt zu einer angemessenen Schätzung der Reliabilität der Testwertvariablen, wenn alle Itemvariablen eindimensionale Messungen desselben Konstrukts darstellen. Dieses Maß setzt keine Gleichheitsrestriktionen bezüglich der Itemparameter (Faktorladungen und Fehlervarianzen) voraus.

Cronbachs Alpha (Cronbach 1951; Guttman 1945) basiert auf dem strengeren Modell essentieller τ -Äquivalenz und führt daher nur dann zu einer angemessenen Schätzung der Reliabilität, wenn alle λ_i identisch sind und folglich alle Itemvariablen dieselbe Kovarianz untereinander aufweisen.

Die *Spearman-Brown-Formel der Testverlängerung* (Brown 1910; Spearman 1910; Steyer und Eid 2001) basiert auf dem noch strengeren Modell essentieller τ -Parallelität und führt daher nur dann zu einer angemessenen Schätzung der Reliabilität, wenn die Itemvariablen dieselbe Kovarianz untereinander sowie zusätzlich dieselbe Varianz aufweisen.

Die Zusammenhänge zwischen den Messmodellen und den verschiedenen Reliabilitätsmaßen werden ausführlicher dargestellt in den Kapiteln zur Reliabilität (► Kap. 14 und 15) und im Kapitel zur Einführung in die CFA (► Kap. 24).

In □ Tab. 13.1 sind Charakteristika der drei Messäquivalenzmodelle zusammenfassend dargestellt: die Messmodellgleichungen und Modellannahmen, die modellbasierten True-Score-Variablen sowie angemessene Reliabilitätsmaße. Wird der Parameterindex i in den Gleichungen weggelassen, so bedeutet dies, dass der Parameter über die Items hinweg als konstant angenommen wird.

Geschachtelte Modelle

Die drei Modelle in □ Tab. 13.1 sind aufgrund der zunehmend restiktiveren Annahmen ineinander geschachtelt (vgl. ► Kap. 24). Das Modell essentiell τ -paralleler Variablen ist mit zwei Gleichheitsannahmen (gleiche Faktorladungen und gleiche Fehlervarianzen) das restiktivste Modell. Wird die Restriktion der gleichen Fehlervarianzen gelockert, so resultiert das Modell essentiell τ -äquivalenter Variablen; werden zusätzlich noch die Faktorladungen frei geschätzt, so resultiert das Modell τ -kongenerischer Variablen. Das Modell essentieller τ -Parallelität ist als restiktivstes Modell im Modell essentieller τ -Äquivalenz geschachtelt. Die beiden Modelle essentieller τ -Parallelität und essentieller τ -Äquivalenz sind wiederum im am wenigsten restiktiven Modell der τ -Kongenerität geschachtelt.

Für jedes der drei Modelle gibt es einen adäquaten Reliabilitätskoeffizienten, der genau zu den entsprechenden Annahmen passt. Die Reliabilitätskoeffizienten der jeweils weniger restiktiven Modelle können jedoch auch verwendet werden, wenn ein restiktiveres Modell einen guten Modellfit aufweist: Wenn also die (essentielle) τ -Parallelität der Messungen anhand der CFA nachgewiesen wurde, so kann die Reliabilität auch über McDonalds Omega oder Cronbachs Alpha angemessen geschätzt werden, da diese Maße auf schwächeren Modellannahmen basieren.

13.6.5 Konfidenzintervalle für Reliabilitätskoeffizienten

Wie eingangs am Beispiel der Blutdruckmessung verdeutlicht wurde, können Messfehler bei der Diagnostik mitunter zu fehlerhaften Entscheidungen führen und sollten daher bei diagnostischen Fragestellungen stets beachtet werden. Punktschätzungen von Reliabilitätskoeffizienten bilden den Populationswert nur selten exakt ab, da sie mit einer Unsicherheit behaftet sind. Deshalb können die Punktschätzungen der Reliabilitätskoeffizienten durch eine Intervallschätzung in Form eines zweiseitigen 95 %-Konfidenzintervalls ergänzt werden, um eine Aussage über die Präzision der Schätzung zu erhalten (Kelley und Pornprasertmanit 2016; Raykov 2002; Raykov und Marcoulides 2011).

Ein Konfidenzintervall gibt eine untere und eine obere Grenze des Wertebereichs an, in dem der Populationswert mit einer bestimmten Sicherheit, z. B. 95 %, zu liegen kommt. Da Reliabilitätskoeffizienten auf den Wertebereich zwischen null und eins begrenzt sind, sollte ein *asymmetrisches Konfidenzintervall* verwendet werden (vgl. auch Eid und Schmidt 2014; ► Kap. 15).

Unter Verwendung entsprechender Statistik-Software (► Abschn. 13.11) kann für den jeweiligen Reliabilitätskoeffizienten auch ein geeigneter Standardfehler $SE(Re)$ geschätzt werden. Für die Schätzung des Standardfehlers sollte entweder die Delta-Methode oder eine von verschiedenen Bootstrap-Methoden verwendet werden, wodurch eine im Vergleich zu herkömmlichen Methoden genauere Schätzung des Standardfehlers möglich ist (vgl. Kelley und Pornprasertmanit 2016; Raykov 2002; Raykov und Marcoulides 2004).

Intervallschätzung

Asymmetrisches Konfidenzintervall

13.7 Empirisches Beispiel

Die drei auf der KTT aufbauenden Modelle der Messäquivalenz (τ -Kongenerität, essentielle τ -Äquivalenz und essentielle τ -Parallelität) sollen nachfolgend anhand eines empirischen Beispiels miteinander verglichen werden. Je nach Passung der Modelle zu den Daten kann entschieden werden, welches Reliabilitätsmaß für die vorliegenden Daten adäquat ist.

Datensatz

Der hier verwendete empirische Datensatz von $N = 250$ Personen stammt aus einer Untersuchung von Amend (2015), in der u. a. die Skala Neurotizismus der Kurzversion des Big Five Inventory (BFI-K; Rammstedt und John 2005) mit 4 Items eingesetzt wurde. Die Testpersonen sollten auf einer 5-stufigen Ratingskala von „1 = sehr unzutreffend“ bis „5 = sehr zutreffend“ ankreuzen, in welchem Ausmaß die Aussagen der Items auf sie zutreffen; die Itemvariablen haben also einen Wertebereich von 1–5.

Die Items lauten:

Ich ...

- BFI-K-4. werde leicht deprimiert, niedergeschlagen.
- BFI-K-9. bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.
(–)
- BFI-K-14. mache mir viele Sorgen.
- BFI-K-19. werde leicht nervös und unsicher.

(–) Invers formuliertes Item

Skala Neurotizismus des BFI-K

13.7.1 Datenanalyse

Nacheinander wurden die drei ineinander geschachtelten Modelle der Messäquivalenz hinsichtlich ihrer Passung zu den Daten überprüft. Die Parameter wurden mit dem Programm *Mplus* (Muthén und Muthén 2017) geschätzt. Da die Antwortvariablen von der Normalverteilung abweichen, wurde die robuste Maximum-Likelihood-Methode verwendet.

Sukzessive Modelltestung

Zur sukzessiven Modelltestung wurde zunächst das am wenigsten restriktive Modell τ -kongenerischer Variablen getestet und bei gutem Modellfit sukzessive die strengeren Modellannahmen des Modells essentiell τ -äquivalenter und essentiell τ -paralleler Variablen hinzugenommen, um zu prüfen, ob auch bei strengeren Modellannahmen weiterhin ein guter Modellfit besteht. Dem Sparsamkeitsprinzip entsprechend wurde das restaktivste Modell beibehalten, das noch gut zu den Daten passt.

13.7.2 Ergebnisse

Wie die Ergebnisse zeigen (Abb. 13.2), passt das τ -kongenerische Modell gut zu den Daten, da der χ^2 -Wert von 1.88 bei 2 Freiheitsgraden nicht signifikant ist ($p = .39$). Das bedeutet, dass die Struktur der empirischen Daten durch ein eindimensionales Messmodell ohne Gleichheitsrestriktionen – weder hinsichtlich der Faktorladungen noch hinsichtlich der Fehlervarianzen – gut beschrieben werden kann.

Die Modelltests der beiden strengeren Modelle der essentiellen τ -Äquivalenz und essentiellen τ -Parallelität weisen dagegen auf signifikante Abweichungen zwischen dem Modell und den Daten hin, wie die signifikanten χ^2 -Werte ($p < .01$) anzeigen. Diese Modelle müssen deshalb als nicht passend verworfen werden. Nur das τ -kongenerische Modell kann beibehalten werden.

13
Nur das τ -kongenerische Modell
passt hier zu den Daten

	Messäquivalenz		
	τ -Kongenerität	Essentielle τ -Äquivalenz	Essentielle τ -Parallelität
Modellfit	$\chi^2(2) = 1.88, p = .39$	$\chi^2(6) = 23.73, p = .00$	$\chi^2(8) = 31.35, p = .00$
Faktorladungen (standardisiert)	$\lambda_1 = 1.012 (.807)$ $\lambda_2 = .615 (.545)$ $\lambda_3 = .940 (.796)$ $\lambda_4 = .780 (.666)$	$\lambda_1 = .848 (.717)$ $\lambda_2 = .848 (.679)$ $\lambda_3 = .848 (.743)$ $\lambda_4 = .848 (.710)$	$\lambda_1 = .834 (.704)$ $\lambda_2 = .834 (.704)$ $\lambda_3 = .834 (.704)$ $\lambda_4 = .834 (.704)$
Fehlervarianzen (standardisiert)	$Var(\varepsilon_1) = .550 (.349)$ $Var(\varepsilon_2) = .894 (.703)$ $Var(\varepsilon_3) = .511 (.366)$ $Var(\varepsilon_4) = .764 (.557)$	$Var(\varepsilon_1) = .680 (.486)$ $Var(\varepsilon_2) = .841 (.539)$ $Var(\varepsilon_3) = .582 (.447)$ $Var(\varepsilon_4) = .707 (.496)$	$Var(\varepsilon_1) = .708 (.505)$ $Var(\varepsilon_2) = .708 (.505)$ $Var(\varepsilon_3) = .708 (.505)$ $Var(\varepsilon_4) = .708 (.505)$
Reliabilität	McDonalds Omega = .805 95%-KI: [.763; .841]	Cronbachs Alpha = .804 95%-KI: [.762; .840]	Spearman-Brown-Reliabilität = .797 95%-KI: [.753; .835]
<i>Anmerkung:</i> Die Parameter in den grau unterlegten Zellen dürfen nicht interpretiert werden, da die Modelltests einen nicht zufriedenstellenden Modellfit anzeigen. 95%-KI = 95%-Konfidenzintervall. Zur Modellidentifikation wurde die Varianz von η auf eins fixiert.			

Abb. 13.2 Unstandardisierte und standardisierte Parameterschätzungen (in Klammern) der Modelle τ -kongenerischer, essentiell τ -äquivalenter und essentiell τ -paralleler Variablen am Beispiel der Skala Neurotizismus (4 Items) der Kurzform des Big Five Inventory (BFI-K; Rammstedt und John 2005) an einer Stichprobe von $N = 250$

13.7.2.1 Reliabilität der Testwertvariablen

Die Reliabilität der Testwertvariablen der Skala Neurotizismus darf ausschließlich für das τ -kongenerische Messmodell berechnet werden, das einen guten Modellfit zeigt. Als geeignetes Reliabilitätsmaß ist in diesem Fall nur McDonalds Omega zu verwenden, da die anderen Reliabilitätskoeffizienten auf strengeren Annahmen beruhen, die aufgrund der Modelltests verworfen wurden. Der Wert von $\omega = .805$ mit dem 95 %-Konfidenzintervall [.763; .841] zeigt eine zufriedenstellende Höhe für eine Skala, die aus nur vier Items besteht.

Die Spearman-Brown-Formel und Cronbachs Alpha sind für das Beispiel nur zum Vergleich aufgeführt. Auch wenn sich in diesem Beispiel die Reliabilitäts schätzungen anhand der verschiedenen Koeffizienten nur wenig unterscheiden, sollte nur McDonalds Omega als das passende Reliabilitätsmaß interpretiert werden. Die relativ geringen numerischen Unterschiede zwischen den Reliabilitätskoeffizienten können vor allem darauf zurückgeführt werden, dass sich die geschätzten Faktorladungen und Fehlervarianzen der Modelle nicht so stark von einander unterscheiden, dass sie bei nur vier Items zu großen Unterschieden führen.

Bestimmung der Reliabilität über McDonalds Omega

13.7.2.2 Diskriminationsparameter/Faktorladungen

Im Modell τ -kongenerischer Variablen werden alle Faktorladungen frei geschätzt. Die geringste Faktorladung hat das einzige invers formulierte Item (Item 2) mit einer standardisierten Faktorladung von $\lambda_2 = .545$. Bei invers formulierten Items wird häufig eine geringere Faktorladung beobachtet als bei Items, die im Sinne des Konstrukt sformuliert sind, da invertierte Items oftmals neben der vom zu messenden Konstrukt erzeugten Varianz noch Methodenvarianz aufweisen, die im Messfehler enthalten ist (vgl. dazu auch ► Kap. 25).

13.7.2.3 Itemreliabilität

Die quadrierten standardisierten Faktorladungen (unstandardisierte vs. standardisierte Parameterschätzungen, s. u.) stellen jeweils die *Reliabilität der einzelnen Itemvariablen* dar (► Abschn. 13.5.1). Die Itemreliabilität gibt den Anteil der True-Score-Varianz einer Itemvariablen an der Gesamtvarianz dieser Itemvariablen an, wobei die Gesamtvarianz bei standardisierten Variablen 1.0 beträgt. Die Itemreliabilität kann auch über die standardisierte Fehlervarianz berechnet werden, indem der Wert der standardisierten Fehlervarianz von eins subtrahiert wird. Wie die Ergebnisse des τ -kongenerischen Modells zeigen, weist Item 2 die geringste standardisierte Faktorladung ($\lambda_2 = .545$) und damit auch die geringste Itemreliabilität ($\lambda_2^2 = .297$) auf, während Item 1 mit der höchsten Faktorladung ($\lambda_1 = .807$) auch die höchste Itemreliabilität ($\lambda_1^2 = .651$) aufweist.

Itemreliabilität als quadrierte standardisierte Faktorladung

Unstandardisierte und standardisierte Parameterschätzungen

Aufgrund des schlechten Modellfits dürfen weder die unstandardisierten noch die standardisierten Parameterschätzungen der Modelle essentiell τ -äquivalenter und essentiell τ -paralleler Variablen interpretiert werden. Die nachfolgenden Erläuterungen sollen jedoch kurz die Zusammenhänge zwischen unstandardisierten und standardisierten Parameterschätzungen verdeutlichen.

Modell essentiell τ -äquivalenter Variablen

Im Modell essentiell τ -äquivalenter Variablen sind die unstandardisierten Faktorladungen/Diskriminationsparameter durch die Gleichheitsrestriktion für alle Items identisch (.848), während sich die standardisierten Faktorladungen unterscheiden (.679–.743). Die standardisierten Faktorladungen entsprechen den Korrelationen zwischen den Itemvariablen y_1 bis y_4 und der latenten Variablen η und werden

quadriert als Itemreliabilität interpretiert. Die standardisierten Faktorladungen (und damit auch die Itemreliabilitäten) unterscheiden sich aufgrund der unterschiedlichen geschätzten Itemvarianzen (zur Standardisierung einer Faktorladung ► Kap. 24).

Modell essentiell τ -paralleler Variablen

Im Modell essentiell τ -paralleler Variablen sind sowohl die unstandardisierten als auch die standardisierten Parameterschätzungen über die Itemvariablen hinweg identisch, da zusätzlich zu den Faktorladungen auch die Fehlervarianzen gleichgesetzt wurden, sodass sich die geschätzten Itemvarianzen, die bei der Standardisierung relevant sind, nicht mehr unterscheiden. Bei essentieller τ -Parallelität wird somit die Annahme getestet, dass die Itemvariablen identische Varianzen und eine identische Reliabilität aufweisen, was aber – wie der schlechte Modellfit im Beispiel zeigt – hier nicht der Fall ist.

13.7.2.4 Leichtigkeitsparameter/Interzept

Leichtigkeitsparameter entsprechen hier den Mittelwerten der Itemvariablen

Im τ -kongenerischen Modell (wie auch in den anderen Modellen) dürfen sich die Leichtigkeitsparameter bzw. die Interzepte unterscheiden. Die Parameter wurden in allen Modellen auf folgende Werte geschätzt: $\alpha_1 = 2.924$, $\alpha_2 = 3.164$, $\alpha_3 = 3.524$, $\alpha_4 = 3.012$. Das leichteste Item ist also Item 3, das schwierigste ist Item 1. Die Interzepte sind mit den Mittelwerten der Itemvariablen identisch, da in allen Modellen der Erwartungswert der latenten Variablen η zur Normierung auf null fixiert wurde (vgl. ► Kap. 24).

13.8 Schätzung individueller Merkmalsausprägungen

Zentrales Anliegen der Psychodiagnostik

Manifeste Testwerte vs. latente Personenwerte

Testwerte als Summe der Itemwerte

Ein zentrales Anliegen der Psychodiagnostik besteht darin, individuelle Ausprägungen latenter Merkmale zu messen, um interindividuelle oder ggf. auch intraindividuelle Vergleiche hinsichtlich eines oder mehrerer latenter Merkmale vornehmen zu können. Da die latenten Merkmale nicht direkt gemessen werden können, muss deren Ausprägung über empirisch verfügbare Indikatoren, d. h. über die Werte der Itemvariablen geschätzt werden.

Als Schätzwerte (Punktschätzungen) der individuellen wahren Merkmalsausprägungen werden zumeist die individuellen Werte der Testwertvariablen (manifeste Testwerte) verwendet. Testwerte lassen sich als Summe der Itemwerte einfach berechnen (► Abschn. 13.4) und stellen ein wichtiges Maß in der psychodiagnostischen Praxis dar. Alternativ können unter Verwendung der CFA latente Personenwerte (Faktorwerte bzw. Factor-Scores η_v) als Schätzwerte für die individuelle Ausprägung des untersuchten latenten Merkmals bestimmt werden (► Abschn. 13.8.2). Unter bestimmten Bedingungen (d. h. insbesondere im Fall τ -kongenerischer Itemvariablen) führt dieses Verfahren zu präziseren Schätzungen der individuellen Merkmalsausprägungen.

13.8.1 Geschätzte wahre Testwerte \hat{T}_v

Zur Gewinnung von individuellen Testwerten für einzelne Personen wurde die Aufsummierung der einzelnen Itemwerte zu einem Test(summen)wert beschrieben, um darüber eine Punktschätzung des wahren Testwertes T_v zu erhalten (► Abschn. 13.4).

Die in der Praxis am häufigsten angewendete Methode zur Schätzung des wahren Testwertes T_v ist die Verwendung des individuellen Testwertes Y_v (s. Gl. 13.10).

13.8 · Schätzung individueller Merkmalsausprägungen

Erfassen alle Items eines Tests relevante Aspekte eines Merkmals auf inhaltlich plausible und nach Aspekten der Testkonstruktion sinnvolle Art und Weise, so sollte durch jede Itemvariable mehr True-Score-Varianz als Fehlervarianz in den Testwert eingehen. Das heißt, der Anteil der True-Score-Varianz erhöht sich durch jedes Item stärker als die Fehlervarianz, sodass auch die Reliabilität der Testwertvariablen zunimmt. Die Schätzung der Reliabilität sollte daher routinemäßig vorgenommen werden. Je reliabler die Testwertvariable ist, desto genauer ist die Schätzung der wahren Werte der Personen anhand dieser Testwerte. Nur reliable Messungen führen zu aussagekräftigen Punktschätzungen, die besonders für diagnostische Fragestellungen eine zentrale Voraussetzung darstellen, z. B. beim Vergleich der individuellen Testwerte mit den Normwerten der Eichstichprobe (T -Werte, Z -Werte; ► Kap. 9).

13.8.1.1 Konfidenzintervall für T_v

Da der Testwert Y_v aber nur eine Punktschätzung des wahren Wertes T_v ist, besteht eine gewisse Unsicherheit, ob der anhand von Y_v geschätzte Wert \hat{T}_v tatsächlich mit dem wahren Wert T_v übereinstimmt.

Das Ausmaß der Unsicherheit kann durch ein Konfidenzintervall um den geschätzten wahren Wert \hat{T}_v quantifiziert werden. Es umfasst denjenigen Bereich der Merkmalsausprägungen, in dem sich mit hoher statistischer Sicherheit alle möglichen wahren Werte T_v befinden, die den Stichprobenschätzwert \hat{T}_v erzeugt haben können. Die Konfidenzintervallbreite ist umso kleiner, je höher die Reliabilität des Tests ist.

Zur vereinfachten Berechnung des Konfidenzintervalls muss der *Standardmessfehler* $SD(E)$ geschätzt werden, der als Wurzel aus der Varianz des Messfehlers definiert ist (Fehlervarianz, s. Varianzzerlegung ► Abschn. 13.6.2.2, Gl. 13.22). Bei bekannter Höhe des Reliabilitätskoeffizienten $Rel(Y)$ kann die Zerlegung der Testwertvarianz $Var(Y)$ in die True-Score- und Fehlervarianz sehr einfach erfolgen, wobei die wahre Varianz $Var(T)$ als Produkt $Rel(Y) \cdot Var(Y)$ ausgedrückt werden kann:

$$\begin{aligned} Var(Y) &= Var(T) + Var(E) \\ &= Rel(Y) \cdot Var(Y) + Var(E) \end{aligned} \quad (13.28)$$

Löst man Gl. (13.28) nach der Fehlervarianz $Var(E)$ auf, so wird deutlich, dass $Var(E)$ den unerklärten Fehlervarianzanteil an der Testwertvarianz $Var(Y)$ darstellt:

$$\begin{aligned} Var(E) &= Var(Y) - Rel(Y) \cdot Var(Y) \\ &= Var(Y) \cdot [1 - Rel(Y)] \end{aligned} \quad (13.29)$$

Zieht man aus der Fehlervarianz (Gl. 13.29) die Wurzel, so erhält man den gesuchten Standardmessfehler $SD(E)$, der entsprechend aus der Standardabweichung der Testwerte multipliziert mit der Wurzel aus der Unreliabilität $(1 - Rel(Y))$ gebildet werden kann¹:

$$SD(E) = SD(Y) \cdot \sqrt{1 - Rel(Y)} \quad (13.30)$$

Wie zu erkennen ist, wird der Standardmessfehler mit ansteigender Reliabilität kleiner und mit abnehmender Reliabilität größer.

Gl. (13.31) zeigt, wie die Grenzen und somit auch die Breite des Konfidenzintervalls bestimmt werden können. Das Konfidenzintervall umfasst denjenigen Bereich eines Merkmals, in dem sich $(1 - \alpha) \cdot 100\%$ aller möglichen wahren Werte T_v befinden, die den Stichprobenschätzwert \hat{T}_v erzeugt haben können. Unter der Annahme normalverteilter Fehler findet man das Konfidenzintervall für

Hohe Reliabilität als Voraussetzung für diagnostische Aussagen

Unsicherheit der Punktschätzung

Konfidenzintervallbreite

Standardmessfehler als Wurzel aus der Fehlervarianz

Bestimmung der Breite des Konfidenzintervalls

¹ Wurde die Reliabilität anhand der CFA geschätzt, so ist ein adäquater Standardfehler vorhanden, der nicht nach diesem vereinfachten Verfahren berechnet werden muss.

große Stichproben mithilfe des Wertes $z_{1-\alpha/2}$ aus der Standardnormalverteilung (z -Verteilung, Anhang):

$$\hat{T}_v - z_{1-\alpha/2} \cdot SD(E) \leq T_v \leq \hat{T}_v + z_{1-\alpha/2} \cdot SD(E) \quad (13.31)$$

Der wahre Wert T_v kommt mit einer Wahrscheinlichkeit von $(1 - \alpha)$ in diesem Intervall zu liegen, wobei die Irrtumswahrscheinlichkeit α in der Regel mit .05 bzw. mit .01 angenommen wird, was einer statistischen Sicherheit von 95 bzw. 99 % entspricht (► Beispiel 13.1).

Beispiel 13.1: Intelligenzmessung zur Diagnostik von Hochbegabung

In einem Intelligenztest mit einem arithmetischen Mittelwert $\bar{Y} = 100$ und einer Standardabweichung $SD(Y) = 15$ erzielte eine Person einen Testwert (Intelligenzquotient, IQ) von $Y_v = \hat{T}_v = 130$.

$Rel(Y) = 1.00$

Legt man diesem Test eine Reliabilität von 1.0 und einen Standardmessfehler von 0 zugrunde, würde man die Person zur Gruppe der Hochbegabten zählen, denn laut Definition gelten Personen mit einem $IQ \geq 130$ als hochbegabt.

$Rel(Y) = .89$

Nehmen wir nun an, der Test habe realistischerweise lediglich eine Reliabilität von $Rel(Y) = .89$. Dann ist der Standardmessfehler nicht null, sondern berechnet sich gemäß Gl. (13.30):

$$\begin{aligned} SD(E) &= 15 \cdot \sqrt{1 - .89} \\ &= 15 \cdot .33 \\ &\approx 5 \end{aligned}$$

Mithilfe des Standardmessfehlers bildet man nach Gl. (13.31) ein Konfidenzintervall und wählt eine statistische Sicherheit $(1 - \alpha)$, mit der der wahre Wert in diesem Intervall liegen soll. Wird die Irrtumswahrscheinlichkeit $\alpha = .05$ und damit $z_{\alpha/2} \cong 2$ gewählt, so erhält man folgendes Konfidenzintervall für T_v :

$$\begin{aligned} 130 - 2 \cdot 5 &\leq T_v \leq 130 + 2 \cdot 5 \\ 120 &\leq T_v \leq 140 \end{aligned}$$

Man erkennt, dass bei diesem Test der wahre IQ der Person mit 95 %iger Wahrscheinlichkeit (bzw. mit einem α -Risiko von 5 %) im Bereich zwischen 120 und 140 diagnostiziert wird. Trotz des erzielten Testwertes von $Y_v = 130$ kann der wahre Wert T_v sowohl deutlich oberhalb, aber auch deutlich unterhalb von 130 liegen. In diesem Fall kann daher nicht mit statistischer Sicherheit gesagt werden, dass diese Person zu den Hochbegabten zählt.

Welchen Testwert müsste die Testperson bei $Rel(Y) = .89$ erreichen, um mit statistischer Sicherheit zur Gruppe der Hochbegabten gezählt werden zu können? Ihr Testwert müsste zumindest bei 140 oder darüber liegen, denn dann würde die untere Grenze des Konfidenzintervalls $\hat{T}_v - 2 \cdot SD(E)$ keinen niedrigeren IQ umfassen als den für Hochbegabung festgelegten Schwellenwert von 130.

$Rel(Y) = .96$

In einem IQ-Test mit einer höheren Reliabilität von $Rel(Y) = .96$ beträgt der Standardmessfehler nur $SD(E) \approx 3$. In diesem Test müsste die Person nur einen IQ von 136 erzielen, um mit 95 %iger Sicherheit einen wahren Wert T_v von 130 oder höher aufzuweisen.

Vorsicht bei Tests mit $Rel(Y) < .80$

Da mit sinkender Reliabilität die Konfidenzintervalle nicht nur sehr breit, sondern auch die Punktschätzungen ungenau werden (Näheres s. Fischer 1974, S. 40 f.), sollten Tests mit einer Reliabilität $Rel(Y) < .80$ für die Individualdiagnostik möglichst nicht verwendet werden.

13.8.1.2 Kritische Differenz von Testwerten

Um verschiedene Testwerte von zwei Personen, Y_1 und Y_2 , dahingehend beurteilen zu können, ob sich auch ihre wahren Werte T_1 und T_2 mit statistischer Sicherheit voneinander unterscheiden, müssen *kritische Differenzen* D_{krit} gebildet werden. Hierbei handelt es sich um Differenzen der Testwerte, die erreicht oder überschritten werden müssen, um die Unterschiede auch für T behaupten zu können. Die kritischen Differenzen geben an, um wie viele Einheiten sich die Testwerte unter Berücksichtigung der Messfehler unterscheiden müssen, damit von unterschiedlichen wahren Werten ausgegangen werden kann.

Bei der Berechnung der kritischen Differenzen ist folgende Fallunterscheidung zweckmäßig:

■ ■ Fall 1: Zwei Testwerte stammen aus demselben Test, d. h., den Messungen liegt dieselbe Reliabilität zugrunde

Hier bemisst sich die kritische Differenz wie folgt (s. Eid und Schmidt 2014, S. 382):

Vergleiche innerhalb eines Tests

$$D_{\text{krit}} = z_{1-\alpha/2} \cdot SD(Y) \cdot \sqrt{2 \cdot [1 - Rel(Y)]} \quad (13.32)$$

Ist die empirische Differenz der beiden Testwerte gleich oder größer als diese kritische Differenz, dann sind die wahren Werte mit einer statistischen Sicherheit von $(1 - \alpha)$ verschieden (► Beispiel 13.2).

Beispiel 13.2: Vergleich der Ergebnisse in einem Konzentrationstest

In einem Konzentrationstest mit einer Reliabilität $Rel(Y) = .96$, der auf den Mittelwert $\bar{Y} = 100$ und die Standardabweichung $SD(Y) = 15$ normiert ist, erzielte eine Person A einen Testwert $Y_A = 115$ und eine Person B einen Testwert $Y_B = 106$, die Differenz ist folglich 9. Kann man mit einer statistischen Sicherheit von 95 % davon ausgehen, dass sich auch die wahren Werte T_A und T_B voneinander unterscheiden?

Nach Gl. (13.32) erhalten wir für den Konzentrationstest mit $z_{1-\alpha/2} \cong 2$ folgende kritische Differenz:

$$D_{\text{krit}} = 2 \cdot 15 \cdot \sqrt{2 \cdot [1 - .96]} = 30 \cdot .28 = 8.40$$

Da die empirische Differenz zwischen den beiden Testwerten (hier 9) größer ist als die kritische Differenz (hier 8.4), kann von einem signifikanten Unterschied der wahren Werte ausgegangen werden.

■ ■ Fall 2: Die beiden Testwerte stammen aus verschiedenen Tests, d. h., den Messungen liegen verschiedene Reliabilitäten zugrunde

Hier bemisst sich die kritische Differenz wie folgt (s. Schmidt-Atzert und Amelang 2012, S. 53), wobei die beiden Testwertvariablen gleiche Standardabweichungen aufweisen müssen, was im Bedarfsfall durch geeignete Testwertnormierungen erzielt werden kann (► Kap. 9):

Vergleiche zwischen verschiedenen Tests

$$D_{\text{krit}} = z_{1-\alpha/2} \cdot SD(Y) \cdot \sqrt{2 - [Rel(Y_1) + Rel(Y_2)]} \quad (13.33)$$

Ist die Differenz der beiden Testwerte gleich oder größer als diese kritische Differenz, dann sind die wahren Werte mit einer statistischen Sicherheit von $(1 - \alpha)$ verschieden.

13.8.2 Geschätzte latente Personenwerte $\hat{\eta}_v$

Faktorwerte, Factor-Scores $\hat{\eta}_v$

Auf Basis der beschriebenen Messmodelle (► Abschn. 13.6) können zur Bestimmung der Merkmalsausprägung nicht nur Test(summen)werte, sondern auch geschätzte latente Personenwerte (Faktorwerte, Factor-Scores) $\hat{\eta}_v$ als Maß der individuellen Ausprägung des latenten Merkmals η_v geschätzt und Personen anhand dieser Werte verglichen werden.

Dazu werden die Personenwerte modellbasiert, d. h. auf der Basis des zugrunde gelegten Messmodells, für jede Person geschätzt. Ein Vorteil dieses Vorgehens besteht darin, nicht auf die Verwendung messfehlerbehafteter Testwerte angewiesen zu sein und das Niveau der Messäquivalenz modellbasiert überprüfen und berücksichtigen zu können.

Zur Schätzung der Personenwerte stehen verschiedene Schätzmethoden zur Verfügung (vgl. Eid und Schmidt 2014, S. 290 ff.). In Mplus ist beispielsweise eine Schätzmethode implementiert (Muthén und Muthén 2017, S. 47 f.; s. auch Skrondal und Rabe-Hesketh 2014), die auf dem Bayes-Ansatz beruht und der Regressionsmethode nach Thomson (1938) und Thurstone (1935) entspricht.

Die latenten Personenwerte werden für jede Person anhand ihrer beobachteten Itemwerte geschätzt. Die beobachteten Itemwerte gehen als unabhängige Variablen zusammen mit den geschätzten Modellparametern (Leichtigkeitsparameter α , Diskriminationsparameter λ) und dem modellimplizierten geschätzten Erwartungswert des Faktors η in die Regressionsgleichung mit den latenten Personenwerten als Kriterium ein (Eid et al. 2017, S. 872 f.). Zusätzlich kann ein Konfidenzintervall für jeden latenten Personenwert geschätzt werden.

Bezüglich der Beziehung zwischen den manifesten Test(summen)werten \hat{T}_v und den latenten Personenwerten $\hat{\eta}_v$ lässt sich feststellen, dass nur bei essentiell τ -parallelen Itemvariablen die manifesten Testwerte eine unverzerrte Schätzung der latenten Personenwerte darstellen und direkt proportional zu diesen sind (s. Eid et al. 2017, S. 874; Eid und Schmidt 2014, S. 294). Vor allem im Fall der weniger strengen τ -Kongeneritität sind die Testwerte und die latenten Personenwerte nicht direkt vergleichbar, da sich die manifeste Testwertvariable aufgrund unterschiedlicher Fehlervarianzen und/oder Diskriminationsparameter der Itemvariablen (► Abschn. 13.6.1) aus weiteren Varianzquellen zusammensetzt. Die in ► Abschn. 13.8.1 behandelte ungewichtete Aufsummierung der Items zur Testwertvariablen könnte in diesem Fall allerdings durch eine Gewichtung mit den Diskriminationsparametern präzisiert werden.

Aufgrund der leichten Verfügbarkeit stellt die manifeste Testwertvariable aber auch bei τ -kongenerischen Itemvariablen ein praktisches Maß zur Schätzung der individuellen Merkmalsausprägungen dar.

Gewichtung der Itemvariablen anhand der Diskriminationsparameter möglich

Eindimensionale vs. mehrdimensionale Modelle

13.9 Erweiterung der KTT

13.9.1 Mehrdimensionale Ansätze

Im Rahmen der bisher vorgestellten eindimensionalen Modelle wird angenommen, dass das Antwortverhalten der untersuchten Personen von einer einzigen latenten Variable abhängt. Neben diesen eindimensionalen Ansätzen gibt es inzwischen auch mehrdimensionale Ansätze, die explizit mehrere latente Variablen simultan berücksichtigen.

■■ Unsystematische vs. systematische Anteile des Messfehlers

In der KTT wird allgemein angenommen, dass sich ein beobachteter Wert aus einem wahren Anteil und einem zufälligen Fehleranteil zusammensetzt. Alle Effekte,

die nicht auf das zu messende latente Merkmal zurückgehen, sind im eindimensionalen Modell zunächst im Messfehler enthalten. Da außer den unsystematischen Einflüssen auch systematische Einflüsse auf die Messungen vorliegen können, die unabhängig vom untersuchten Merkmal sind, enthält der Messfehler somit sowohl unsystematische (zufällige) als auch systematische Anteile (Raykov und Marcoulides 2011).

Diese unsystematischen und systematischen Anteile können nur dann voneinander getrennt werden, wenn zusätzlich geeignete Mehrfachmessungen vorgenommen werden. Aufbauend auf der KTT können dann mehrdimensionale Modelle definiert werden, in denen nicht nur eine einzige latente Variable enthalten ist, sondern mehrere latente Variablen, die sich auf die beobachtbaren Variablen auswirken und den Messfehler reduzieren. Hierzu liegen bereits viele Modellentwicklungen im Rahmen der mehrdimensionalen CFA vor (für die Grundlagen mehrdimensionaler CFA-Modelle ► Kap. 24).

Besteht z. B. die Annahme, dass sich die Situation, in der die Messung stattfindet, systematisch auf die Messung auswirkt, so müssten mehrere Messungen in unterschiedlichen Situationen durchgeführt werden, um einen Situationseffekt aufdecken zu können. Werden Messungen zu mehreren Messzeitpunkten erhoben, kann die Situationsspezifität als ein vom Trait unabhängiger Effekt geschätzt werden (s. Latent-State-Trait-Theorie, LST-Theorie, ► Kap. 26; Steyer et al. 2015).

Weiter kann die Annahme bestehen, dass das verwendete Messinstrument einen eigenständigen Effekt auf die Messung hat. Um einen solchen Methodeneffekt aufzudecken zu können, müssten mehrere verschiedene Messinstrumente für die Messung des latenten Merkmals verwendet werden. Werden mehrere Messmethoden zur Messung eines Konstrukt eingesetzt, kann die Methodenspezifität als ein vom Trait unabhängiger Effekt geschätzt werden (s. Multitrait-Multimethod-Analyse, MTMM-Analyse, ► Kap. 25; Eid et al. 2008).

Eine Kombination aus Längsschnitt- und Querschnittmodellen ermöglicht die Schätzung sowohl der Situationsspezifität als auch der Methodenspezifität in einem gemeinsamen Modell (vgl. u. a. ► Kap. 26 und 27; Geiser und Lockhart 2012).

13.9.2 Generalisierbarkeitstheorie

Ein bahnbrechender Ansatz, der als einer der ersten die Mehrdimensionalität von Messungen berücksichtigte, war die *Generalisierbarkeitstheorie* (Cronbach et al. 1963, 1972; Gleser et al. 1965). Dieser Theorie zufolge geht man davon aus, dass multiple systematische Fehlervarianzquellen („Facetten“) existieren, die sich gemeinsam auf die Messung auswirken und deren Varianzkomponenten anhand der Varianzanalyse voneinander separiert werden können. In den letzten Jahren hat die Generalisierbarkeitstheorie an Bedeutung gewonnen (vgl. Brennan 2001, 2011; Marcoulides 1996; Raykov und Marcoulides 2006, 2011; Vispoel et al. 2018, 2019), weshalb sie nachfolgend kurz dargestellt werden soll.

Das Hauptziel der Generalisierbarkeitstheorie (G-Theorie) besteht darin, im Rahmen von *Generalisierbarkeitsstudien* den Einfluss unterschiedlicher Varianzquellen/Facetten auf die Messungen zu schätzen und somit die Gesamtvarianz in verschiedene Varianzkomponenten zu zerlegen. Im anschließenden Schritt kann diese Information in Entscheidungsstudien genutzt werden, um Messprozeduren für einen spezifischen Anwendungszweck optimal zu gestalten, z. B. um die Genauigkeit einer Messung unter Kostenrestriktionen zu maximieren.

13.9.2.1 Zusätzliche Varianzquellen

Die Annahme, dass sich die Varianz beobachteter Item- oder Testwertvariablen nur aus der merkmalsspezifischen True-Score-Varianz und der zufälligen Fehlervarianz zusammensetzt, erweist sich in der psychodiagnostischen Praxis häufig als

Mehrfachmessungen

Situationsspezifität

Methodenspezifität

Multiple systematische Fehlervarianzquellen

Generalisierbarkeits- und Entscheidungsstudien

Eindimensionalität oftmals nicht gegeben

Facetten als weitere Varianzquellen

Ein-Facetten-Design

Varianzanalytisches Design

unrealistisch, da Messprozeduren genutzt werden müssen, von denen bekannt ist, dass sie zusätzliche Varianzquellen beinhalten.

So werden beispielsweise in der empirischen Bildungsforschung bei Testaufgaben zum Leseverständnis typischerweise mehrere Fragen (Items) zu derselben Textpassage („Testlets“) gestellt. In diesem Fall ist jedoch die Unkorreliertheit der Messfehler als Voraussetzung der Eindimensionalität nicht mehr gegeben (► Abschn. 13.3.2), beispielsweise wenn eine Person alle Testaufgaben zu einer Textpassage nicht beantworten kann, weil sie die ganze Passage nicht angemessen verstanden hat. Dennoch ist ein solches Vorgehen bei Leseverständnisaufgaben aus inhaltlichen und ökonomischen Gründen meist erforderlich, um hinreichend lange und inhaltlich komplexe Texte vorgeben zu können.

Auch andere praktische Anforderungen führen zu systematischen Anteilen im Messfehler, etwa wenn im Rahmen eines eignungsdiagnostischen Assessment-Centers die Leistung eines Bewerbers oder einer Bewerberin in mehreren Aufgaben (Items) durch unterschiedliche Beobachter (Rater) beurteilt wird. In beiden Fällen (Leseverständnis, Assessment-Center) muss zusätzlich zur wahren Varianz und zur Fehlervarianz eine weitere Varianzquelle/Facette angenommen werden, um Effekte der Textpassagen bzw. der unterschiedlichen Rater abzubilden.

13.9.2.2 Varianzzerlegung

Der G-Theorie zufolge können mehrere systematische Varianzquellen existieren, die sich gemeinsam auf die Messung auswirken. Hierzu wird die Grundgleichung der KTT (Gl. 13.4) erweitert, indem angenommen wird, dass nicht nur die True-Score-Variablen T und die zufällige Messfehlervariable E , sondern auch Facetten der Messprozedur, z. B. Rater oder Situationen, die Testwertvariable beeinflussen.

Anhand des Beispiels eines Assessment-Centers könnte z. B. davon ausgegangen werden, dass der Testwert Y_{vr} einer Person v anhand der Einschätzung durch Rater r gemessen wird. Wären mehrere Rater zufällig aus einem größeren Pool von Ratern („Universum“ von Ratern) gezogen worden, könnte die Facette „Rater“ mit in das Modell aufgenommen werden. Ein solches Ein-Facetten-Design wird als „One-Facet-Person-crossed-with-Items-Design“ ($v \times r$) bezeichnet (vgl. Raykov und Marcoulides 2011, S. 226), da die Facette „Rater“ die einzige Quelle systematischer Fehlervarianz darstellt.

Das Ein-Facetten-Design kann als ein varianzanalytisches Design aufgefasst werden. Der Testwert der Person v , eingeschätzt durch den Beurteiler/Rater r , kann wie folgt zerlegt werden:

$$Y_{vr} = \mu + (\mu_v - \mu) + (\mu_r - \mu) + (Y_{vr} - \mu_v - \mu_r + \mu) \quad (13.34)$$

Hierbei setzt sich der Testwert additiv zusammen aus einem Gesamtmittelwert μ , einem Personeneffekt ($\mu_v - \mu$), einem Ratereffekt ($\mu_r - \mu$) und einem Residuum ($Y_{vr} - \mu_v - \mu_r + \mu$), das alle weiteren potentiellen Varianzquellen repräsentiert.

Bei der varianzanalytischen Zerlegung der Testwertvarianz (vgl. z. B. Brennan 2001, 2011) ist zu beachten, dass in der G-Theorie die Annahme der Unabhängigkeit der Messfehler zwischen den Personen gelten muss. Die Gesamtvarianz der Testwerte kann in folgende Varianzkomponenten zerlegt werden (vgl. Cronbach et al. 1972):

- Eine Komponente $Var(T)$ der wahren Varianz, z. B. ein gemessener Trait
- Eine Komponente $Var(R)$ der wahren Varianz mehrerer Rater (Beurteiler, allgemein: Erfassungsmethoden)
- Die Fehlervarianz $Var(E)$, die im Rahmen der G-Theorie dem höchstmöglichen Interaktionseffekt entspricht, hier der Interaktion zwischen Personen und Ratern, konfundiert mit dem Messfehler

Somit kann die Varianz einer Testwertvariablen wie folgt zerlegt werden:

Varianzanalytische Zerlegung

$$\text{Var}(Y) = \text{Var}(T) + \text{Var}(R) + \text{Var}(E) \quad (13.35)$$

Neben diesem sehr einfachen Beispiel sind natürlich weitere Facetten denkbar, die sich auf die Messungen auswirken können und die folglich mitberücksichtigt werden können.

13.9.2.3 Generalisierbarkeitskoeffizienten

Mithilfe der verschiedenen Varianzkomponenten können schließlich Generalisierbarkeitskoeffizienten bestimmt werden, um die Messgenauigkeit einer Messprozedur zu beschreiben. Diese Koeffizienten nehmen Werte zwischen null und eins an. Unterschieden werden zwei Arten der Fehlervarianz, die sich auf relative und absolute Entscheidungen beziehen (vgl. Brennan 2001, 2011).

Relative Entscheidungen beziehen sich auf interindividuelle Unterschiede zwischen Personen, d. h. die Rangordnung der Personen. In die relative Fehlervarianz gehen mit Ausnahme von $\text{Var}(T)$ alle Varianzanteile ein, die sich auf Informationen über die Rangordnung der Personen beziehen. Die relative Fehlervarianz, die auch als „ δ -type error“ bezeichnet wird (vgl. Raykov und Marcoulides 2011), ist in einem Ein-Facetten-Design wie folgt definiert:

$$\text{Var}(E_\delta) = \frac{\text{Var}(E)}{n_r} \quad (13.36)$$

Dabei bezeichnen $\text{Var}(E_\delta)$ die relative Fehlervarianz, $\text{Var}(E)$ die Fehlervarianz, die mögliche Interaktionen und Messfehler enthält, und n_r die Rateranzahl.

Zwei Arten der Fehlervarianz

Relative Entscheidungen mit „ δ -type error“

Absolute Entscheidungen beziehen sich dagegen auf das Fähigkeitsniveau der Personen. Wenn getestet werden soll, ob eine Testperson ein bestimmtes Fähigkeitsniveau erreicht, das z. B. für einen Job benötigt wird, so wird die absolute Fehlervarianz verwendet. Diese beinhaltet sowohl Informationen über die Rangordnung der Personen als auch Unterschiede in den durchschnittlichen Werten. Die absolute Fehlervarianz, die auch als „ Δ -type error“ bezeichnet wird (vgl. Raykov und Marcoulides 2011), ist in einem Ein-Facetten-Design wie folgt definiert:

$$\text{Var}(E_\Delta) = \frac{\text{Var}(R)}{n_r} + \frac{\text{Var}(E)}{n_r} \quad (13.37)$$

Absolute Entscheidungen mit „ Δ -type error“

Dabei sind $\text{Var}(E_\Delta)$ die absolute Fehlervarianz, $\text{Var}(R)$ die Ratervarianz, $\text{Var}(E)$ die Fehlervarianz, die mögliche Interaktionen und Messfehler enthält, und n_r die Rateranzahl.

Der Generalisierbarkeitskoeffizient für relative Entscheidungen g_{rel} wird wie folgt berechnet:

$$g_{\text{rel}} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E_\delta)} \quad (13.38)$$

Generalisierbarkeitskoeffizient für relative Entscheidungen

Dieser Koeffizient entspricht Cronbachs Alpha, wenn die Voraussetzung der essentiellen τ -Äquivalenz für Alpha auch hier erfüllt ist. Auch andere klassische Reliabilitätsmaße können anhand der G-Theorie berechnet werden, wenn die jeweiligen Voraussetzungen erfüllt sind (Vispoel et al. 2018).

Der Generalisierbarkeitskoeffizient für absolute Entscheidungen g_{abs} wird analog berechnet:

$$g_{\text{abs}} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E_\Delta)} \quad (13.39)$$

Generalisierbarkeitskoeffizient für absolute Entscheidungen

Standardfehler und Konfidenzintervalle für diese Koeffizienten können ebenfalls angegeben werden (Raykov und Marcoulides 2011; Vispoel et al. 2018). Sowohl

klassische Reliabilitätskoeffizienten als auch Intraklassen-Korrelationskoeffizienten (ICC) stellen Spezialfälle der Koeffizienten der G-Theorie dar (Revelle und Condon 2018).

Die G-Theorie erweitert den Rahmen der klassischen Reliabilitätsmethoden, indem konzeptuelles und mathematisch-statistisches Handwerkszeug für die Analyse unterschiedlichster Einflüsse auf die Variabilität von Testwerten bereitgestellt wird (Rauch und Moosbrugger 2011). Anders als die Koeffizienten der MTMM-Analyse (► Kap. 25) und der LST-Theorie (► Kap. 26), die in der Regel anhand der CFA geschätzt werden, basiert die G-Theorie allerdings auf manifesten und nicht auf latenten Variablen.

13.10 Zusammenfassung

Die KTT liefert die theoretischen Grundlagen zur Konstruktion psychologischer Testverfahren mit in der Regel kontinuierlichen Itemvariablen sowie zur Interpretation der Testwerte, die durch Aufsummierung der Itemwerte gewonnen werden. Dabei stellt die Aufteilung der beobachteten Messwerte in einen wahren Wert und einen Fehlerwert den zentralen theoretischen Ausgangspunkt dar. Liegen mehrere Messungen desselben Merkmals vor, lassen sich aufbauend auf der KTT verschiedene eindimensionale Messmodelle formulieren, die auf unterschiedlich restriktiven, testbaren Annahmen basieren.

Anhand dieser Messmodelle kann mithilfe der CFA überprüft werden, welche Stufe der Messäquivalenz den Itemvariablen (allgemein: den Messungen) zugrunde liegt. Zur Beurteilung der Messgenauigkeit einer Testwertvariablen können abhängig von der gegebenen Stufe der Messäquivalenz verschiedene Reliabilitätskoeffizienten geschätzt werden, die zusätzlich durch ein Konfidenzintervall ergänzt werden sollten.

Zur Schätzung der individuellen Merkmalsausprägungen werden zumeist manifeste Testwerte verwendet. Zur individualdiagnostischen Beurteilung eines Testwertes bzw. zur Beurteilung kritischer Differenzen zwischen mehreren Testwerten sollte auf eine hohe Reliabilität und schmale Konfidenzintervalle geachtet werden. Alternativ können latente Personenwerte Verwendung finden, die mittels CFA als Factor-Scores geschätzt werden können.

Neben eindimensionalen Modellen gibt es inzwischen auch mehrdimensionale Ansätze wie die Generalisierbarkeitstheorie, die auf der KTT aufbauen und explizit mehrere latente Variablen als systematische Varianzquellen berücksichtigen.

13.11 EDV-Hinweise

Alle in diesem Kapitel behandelten Reliabilitätsmaße können mit gängigen EDV-Programmen, z. B. SPSS oder R, analysiert werden. Die Voraussetzungen lassen sich beispielsweise mit den Programmen *Mplus* (Muthén und Muthén 2017) oder dem R-Paket *lavaan* (Rosseel 2012) testen.

13.12 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was ist das „Klassische“ an der KTT?
2. Erläutern Sie kurz die Eigenschaften der Messfehler- und True-Score-Variablen.
3. Wie ist der Reliabilitätskoeffizient in der KTT definiert?

4. Warum soll zur Bestimmung des wahren Wertes auch ein Konfidenzintervall gebildet werden?
5. Auf welchen Annahmen basiert das Modell essentiell τ -äquivalenter Variablen?
6. Worin besteht das Hauptziel der Generalisierbarkeitstheorie?

Literatur

- Amend, N. (2015). *Who's perfect? Pilotstudie zur Untersuchung potenzieller Korrelate des Merkmals Perfektionismus*. Unveröffentlichte Bachelorarbeit, Institut für Psychologie, Goethe Universität, Frankfurt am Main.
- Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York, NY: The Guilford Press.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, 1–21.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York, NY: Wiley.
- Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137–163.
- DeMars, C. E. (2018). Classical test theory and item response theory. In P. Irving, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 49–74). Hoboken, NJ: Wiley.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
- Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M. & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, 13, 230–253.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Geiser, C. & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, 17, 255–283.
- Gleser, G. C., Cronbach, L. J. & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30, 395–418.
- Gulliksen, H. (1950). *Theory of Mental Tests*. New York, NY: Wiley.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Kelley, K. & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for homogeneous composite measures. *Psychological Methods*, 21, 69–92.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Marcoulides, G. A. (1996). Estimating variance components in generalizability theory. *Structural Equation Modeling*, 3, 290–299.
- Moosbrugger, H. (1983). Modelle zur Beschreibung statistischer Zusammenhänge in der psychologischen Forschung. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten. Enzyklopädie der Psychologie, Serie I: Forschungsmethoden der Psychologie* (Bd. 4, S. 1–58). Göttingen: Hogrefe.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Rammstedt, B. & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. *Diagnostica*, 51, 195–206.
- Rauch, W. & Moosbrugger, H. (2011). Klassische Testtheorie: Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik. Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden. Serie II, Psychologische Diagnostik* (Bd. 2, S. 1–86). Göttingen: Hogrefe.
- Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, 37, 89–103.

- Raykov, T. & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, 11, 621–637.
- Raykov, T. & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, 6, 81–95.
- Raykov, T. & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.
- Raykov, T. & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, 76, 325–338.
- Revelle, W. & Condon, D. M. (2018). Reliability. In P. Irving, T. Booth & D. Hughes (Eds.), *The Wiley-Blackwell Handbook of Psychometric Testing*. West Sussex, UK: Blackwell Publishing Ltd.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48, 1–36.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8, 23–74.
- Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin, Heidelberg: Springer.
- Skrondal, A. & Rabe-Hesketh, S. (2014). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models* (2nd ed.). Boca Raton: Chapman & Hall.
- Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1904b). “General Intelligence”, objectively determined and measured. *American Journal of Psychology*, 15, 201–292.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 171–195.
- Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, 3, 25–60.
- Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin, Heidelberg: Springer.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A Theory of States and Traits – Revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- Thomson, G. H. (1938). *The factorial analysis of human ability*. London: University of London Press.
- Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago, IL: University of Chicago Press.
- Vispoel, W. P., Morris, C. A. & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, 23, 1–26.
- Vispoel, W. P., Morris, C. A. & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, 24, 153–178.
- Yousfi, S. & Steyer, R. (2006). Klassische Testtheorie. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 288–303). Göttingen: Hogrefe.
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, 40, 395–412.
- Zimmerman, D. W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement*, 36, 85–96.



Klassische Methoden der Reliabilitätsschätzung

Jana C. Gäde, Karin Schermelleh-Engel und Christina S. Werner

Inhaltsverzeichnis

- 14.1 Was ist Reliabilität? – 307**
 - 14.1.1 Definition der Reliabilität – 307
 - 14.1.2 Klassische vs. modellbasierte Methoden der Reliabilitätsschätzung – 308
 - 14.1.3 Beziehung zu Objektivität und Validität – 309
- 14.2 Grundlagen – 309**
 - 14.2.1 Klassische Testtheorie (KTT) – 309
 - 14.2.2 Eindimensionalität – 311
 - 14.2.3 Messeigenschaften der Itemvariablen – 312
 - 14.2.4 Messäquivalenz – 312
 - 14.2.5 Schätzung der True-Score-Varianz einer Itemvariablen – 313
- 14.3 Cronbachs Alpha – 314**
 - 14.3.1 Berechnung – 314
 - 14.3.2 Voraussetzungen von Cronbachs Alpha – 317
 - 14.3.3 Überprüfung der Voraussetzungen von Cronbachs Alpha – 318
 - 14.3.4 Probleme von Cronbachs Alpha – 318
 - 14.3.4.1 Fehlerkovarianzen durch Methodeneffekte – 319
 - 14.3.4.2 Unklare Bedeutung des Begriffs „interne/innere Konsistenz“ – 319
 - 14.3.4.3 Untere Schranke der Reliabilität – 322
- 14.4 Test-Test-Korrelation – 322**
 - 14.4.1 Berechnung – 322
 - 14.4.2 Voraussetzungen der Test-Test-Korrelation – 324
 - 14.4.3 Überprüfung der Voraussetzungen der Test-Test-Korrelation – 326
 - 14.4.4 Probleme der Test-Test-Korrelation – 327
 - 14.4.4.1 Probleme der Retest-Reliabilität – 327
 - 14.4.4.2 Probleme der Paralleltest- und Split-Half-Reliabilität – 328
- 14.5 Vergleichbarkeit der Reliabilitätsmaße – 329**
- 14.6 Einflüsse auf die Reliabilität – 330**

- 14.7 Anzustrebende Höhe der Reliabilität – 330**
 - 14.8 Auswahl eines geeigneten Reliabilitätsmaßes – 331**
 - 14.9 Zusammenfassung – 332**
 - 14.10 EDV-Hinweise – 333**
 - 14.11 Kontrollfragen – 333**
- Literatur – 333**

14.1 · Was ist Reliabilität?

i Eine zentrale Aufgabe der psychologischen Diagnostik besteht darin, latente Konstrukte (Merkmale) zuverlässig zu messen. Soll z. B. anhand eines Tests untersucht werden, ob eine Frau nach der Geburt eines Kindes eine behandlungsbedürftige postpartale Depression aufweist, so muss einerseits sichergestellt sein, dass mit dem Test tatsächlich Depressivität und nicht z. B. Angst gemessen wird (Validität); andererseits muss die Messung aber auch zuverlässig und messgenau sein (Reliabilität). Werden z. B. mehrere Messungen in kurzen Zeitabständen hintereinander durchgeführt, so sollten etwaige geringfügige Unterschiede zwischen den Messungen ausschließlich aus unsystematischen Messfehlern resultieren, wenn das zu messende Merkmal tatsächlich unverändert geblieben ist. Die Zuverlässigkeit von Messungen ist ein ausgesprochen wichtiges Gütekriterium der Klassischen Testtheorie (KTT), das als Reliabilität bezeichnet wird (vgl. ▶ Kap. 2 und 13).

14.1 Was ist Reliabilität?

Wann immer die Ausprägung eines Merkmals gemessen wird, sollte das dazu verwendete Messinstrument (Test, Fragebogen, Skala, Item) eine möglichst hohe Messgenauigkeit aufweisen. Ein Messinstrument hat dann eine hohe Messgenauigkeit, wenn die resultierenden Messergebnisse nur mit einem geringen Messfehler behaftet sind. Die Reliabilität ist ein Maß für die Messgenauigkeit der mit einem Messinstrument erzielten Messergebnisse. Sie ist außerdem ein normiertes Effektgrößenmaß (Eid und Schmidt 2014, S. 267), das den Vergleich verschiedener Messinstrumente hinsichtlich der Zuverlässigkeit ihrer Messungen ermöglicht. So kann man beispielsweise prüfen, ob die Messung des Gewichts einer Person mit einer neuen, geeichten Digitalwaage mit einer höheren Messgenauigkeit möglich ist als mit einer alten Analogwaage.

Ein Messinstrument ist perfekt reliabel, wenn die resultierenden Messwerte frei von zufälligen Messfehlern sind. In diesem Fall würde bei jeder Wiederholungsmessung (z. B. des Gewichts derselben Person) genau derselbe Messwert resultieren und dem wahren Wert (True-Score) exakt entsprechen. Ein Messinstrument ist dagegen umso weniger reliabel, je größer der Anteil zufälliger Messfehler ist. In diesem Fall würden bei wiederholten Messungen unterschiedliche Messwerte resultieren, die sich aus dem wahren Wert und unsystematischen Messfehlern zusammensetzen.

Im Rahmen der Testkonstruktion interessiert in der Regel die Reliabilität der Testwertvariablen Y , die durch Aufsummierung der einzelnen Itemvariablen y_i ($i = 1, \dots, p$) gebildet wird (d. h., pro Testperson werden die Antworten auf die einzelnen Items i aufsummiert).

Wird beispielsweise das latente (d. h. nicht direkt beobachtbare) Merkmal Neurotizismus über zehn Items eines Tests gemessen, so werden die zehn Werte der Itemantworten pro Person jeweils zu einem Testwert aufsummiert, der die Ausprägung einer Person im Merkmal Neurotizismus messen soll. In der Regel interessiert die Reliabilität der resultierenden Testwertvariablen Y ; sie gibt an, wie zuverlässig die Neurotizismusausprägung anhand der zehn aufsummierten Items gemessen werden kann.

14.1.1 Definition der Reliabilität

Zur Bestimmung der Reliabilität wird die Gesamtvarianz der Testwerte ($Var(Y)$) in den Anteil der wahren Varianz (True-Score-Varianz, $Var(T)$) und den Anteil der zufälligen Messfehlervarianz (Fehler-/Error-Varianz, $Var(E)$) zerlegt. Die Reliabilität einer Testwertvariablen Y wird als Anteil der True-Score-Varianz an der Gesamtvarianz der Testwertvariablen geschätzt (Gl. 14.1). In analoger Weise lässt

Reliabilität als Maß der Messgenauigkeit

Reliabilität, wahrer Wert und Messfehler

Die Testwertvariable ergibt sich als Summe der Itemvariablen

sich auch die Reliabilität eines einzelnen Items bestimmen (Gl. 14.2). Im Folgenden wird es vorrangig um die Reliabilität von Testwertvariablen gehen.

Reliabilität als Varianzverhältnis

Definition

Die **Reliabilität einer Testwertvariablen** ist definiert als das Verhältnis der Varianz der True-Score-Variablen T (griechischer Großbuchstabe Tau) zur Gesamtvarianz der Testwertvariablen Y , wobei sich die Gesamtvarianz der Testwertvariablen aus der Varianz der True-Score-Variablen T und der Varianz der Fehlervariablen E (griechischer Großbuchstabe Epsilon) zusammensetzt:

$$Rel(Y) = \frac{Var(T)}{Var(Y)} = \frac{Var(T)}{Var(T) + Var(E)} \quad (14.1)$$

Definition

Die **Reliabilität einer Itemvariablen** y_i lässt sich analog bestimmen als das Verhältnis der Varianz der Item-True-Score-Variablen τ_i (griechischer Kleinbuchstabe tau) zur Gesamtvarianz der Itemvariablen y_i , wobei sich die Gesamtvarianz der Itemvariablen wiederum aus der Varianz der True-Score-Variablen τ_i und der Varianz der Fehlervariablen ε_i des Items (griechischer Kleinbuchstabe epsilon) zusammensetzt:

$$Rel(y_i) = \frac{Var(\tau_i)}{Var(y_i)} = \frac{Var(\tau_i)}{Var(\tau_i) + Var(\varepsilon_i)} \quad (14.2)$$

Definition

Der **Wertebereich der Reliabilität** liegt zwischen 0 (fehlende Messgenauigkeit) und 1 (höchste Messgenauigkeit).

14.1.2 Klassische vs. modellbasierte Methoden der Reliabilitätsschätzung

Zur Schätzung der Reliabilität der Testwertvariablen werden im Wesentlichen folgende Methoden unterschieden:

- Klassische Methoden* der Reliabilitätsschätzung, die auf Stichprobenkennwerten beruhen und bestimmt werden über die
 - Kovarianzen der Itemvariablen eines Tests/einer Skala (► Abschn. 14.3) oder
 - Korrelation der Testwertvariablen zweier paralleler Tests (► Abschn. 14.4).
- Modellbasierte Methoden* der Reliabilitätsschätzung (► Kap. 15), die auf Modellparametern beruhen, die im Rahmen der konfirmatorischen Faktorenanalyse (CFA, ► Kap. 24) anhand von expliziten Messmodellen geschätzt werden.

Beide Ansätze beruhen auf der Klassischen Testtheorie (KTT, ► Kap. 12 und 13; vgl. auch Eid et al. 2017; Eid und Schmidt 2014; Raykov und Marcoulides 2011; Steyer und Eid 2001).

In diesem Kapitel werden die klassischen Methoden der Reliabilitätsschätzung behandelt, die auf Stichprobenkennwerten (Kovarianzen oder Korrelationen) beruhen, die nicht modellbasiert gewonnen wurden. Allerdings setzen auch die klassischen Methoden implizite Messmodelle voraus, die statistisch überprüft werden müssen.

Überprüfung von Messmodellen

■■ Messmodelle

Ein Messmodell stellt die formale Zuordnung mehrerer Messungen (Items) zu einem zu messenden Merkmal, dem latenten Konstrukt, her. Die klassischen Reliabilitätsmaße setzen implizit voraus, dass die einzelnen Messungen jeweils sehr ähnliche Messeigenschaften aufweisen (Modelle der Messäquivalenz, ▶ Kap. 13). Das Zutreffen dieser strengen Voraussetzungen muss gewährleistet sein, damit die Reliabilität anhand der Stichprobenkennwerte zuverlässig geschätzt werden kann.

Für die modellbasierten Reliabilitätsmaße werden hingegen die Stichprobenkennwerte (Kovarianzen oder Korrelationen) zur Gewinnung der relevanten Modellparameter (Faktorladungen und Fehlervarianzen der Itemvariablen y_i) im Rahmen einer CFA verwendet. Das zugrunde liegende explizite Messmodell wird statistisch formalisiert und empirisch überprüft. Bei diesem Ansatz können die Meseigenschaften der einzelnen Messungen modellbasiert berücksichtigt werden und dürfen sich daher unterscheiden (Modelle der Messäquivalenz, ▶ Kap. 13 und 24).

Auf Basis der CFA können modellbasierte Reliabilitätskoeffizienten für ein- oder mehrdimensionale Modelle geschätzt werden. Die modellbasierten Methoden werden in ▶ Kap. 15 ausführlich dargestellt. Im Gegensatz zu den modellbasierten Methoden beziehen sich die klassischen Methoden nur auf eindimensionale Modelle.

Implizite Messmodelle der klassischen Methoden

Explizite Messmodelle der modellbasierten Methoden

14.1.3 Beziehung zu Objektivität und Validität

Im Kontext der Testkonstruktion kann eine hohe Reliabilität einer Testwertvariablen als Eigenschaft eines Tests (also eines Messinstruments) nur dann erzielt werden, wenn die Messbedingungen (Testsituation, Testdurchführung und Testauswertung) standardisiert werden. Die Kontrolle der Messbedingungen ist Gegenstand des Gütekriteriums der Objektivität (▶ Kap. 2). Die Objektivität stellt eine wesentliche Voraussetzung für die Reliabilität dar.

Objektivität

Bei der Bestimmung der Reliabilität eines Tests kommt es nicht in erster Linie darauf an, ob der Test inhaltlich tatsächlich genau das Merkmal misst, das gemessen werden soll – dies ist eine Frage der Validität (▶ Kap. 21, vgl. auch ▶ Kap. 2). Der Fokus der Reliabilität liegt ausschließlich auf der Messgenauigkeit, die für die Validität eines Tests eine wesentliche Voraussetzung darstellt.

Validität

14.2 Grundlagen

14.2.1 Klassische Testtheorie (KTT)

Die Bestimmung der Anteile der True-Score- und Messfehlervarianz an der Gesamtvarianz einer Itemvariablen y_i basiert auf der KTT, die im Wesentlichen eine Messfehlertheorie ist. Der KTT liegt die Annahme zugrunde, dass sich jeder beobachtete Messwert y_{vi} einer Person v auf einer Variablen i (z. B. Item i) additiv zusammensetzt aus einem wahren Wert (True-Score) τ_{vi} und einem Messfehler ε_{vi} (Grundgleichung der KTT, s. ▶ Kap. 13, ▶ Abschn. 13.3.1).

Grundgleichung der KTT

Definition

Der wahre Wert τ_{vi} einer Person ist definiert als der personenbedingte Erwartungswert der Variablen y_{vi} (Guttman 1945; Lord und Novick 1968; Zimmerman 1975, 1976). Aus der Definition des wahren Wertes als personenbedingtem Erwartungswert und der Grundgleichung der KTT folgen für die True-Score- und Fehlervariablen einige wichtige Eigenschaften, die in ▶ Kap. 13 erläutert werden (vgl. auch Eid und Schmidt 2014; Eid et al. 2017; Lord und Novick 1968; Steyer und Eid 2001; Zimmerman 1976).

Definition des wahren Wertes

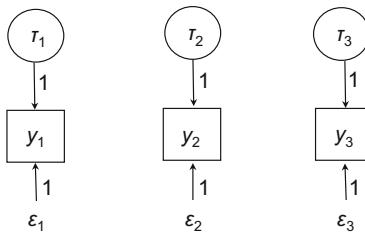


Abb. 14.1 Schematische Darstellung der Grundgleichung der KTT für drei Itemvariablen y_i , die sich jeweils aus einer True-Score-Variablen τ_i und einer Fehlervariablen ε_i zusammensetzen ($i = 1, 2, 3$)

Aus der Grundgleichung der KTT folgt, dass sich jede Itemvariable y_i ($i = 1, \dots, p$) additiv zusammensetzt aus der True-Score-Variablen τ_i und der Fehlervariablen ε_i :

$$y_i = \tau_i + \varepsilon_i \quad (14.3)$$

Damit lässt sich auch die Varianz einer Itemvariablen in die True-Score- und die Fehlervarianz zerlegen:

$$\text{Var}(y_i) = \text{Var}(\tau_i) + \text{Var}(\varepsilon_i) \quad (14.4)$$

In **Abb. 14.1** ist die Grundgleichung der KTT für drei Itemvariablen y_1, y_2 und y_3 schematisch dargestellt: Jede Itemvariable setzt sich aus einer mit dem Wert eins gewichteten True-Score-Variablen τ_i sowie einer mit dem Wert eins gewichteten Fehlervariablen ε_i zusammen.

■■ Mehrfachmessungen zur Bestimmung der Reliabilität

Theoretisch ist die Reliabilität als Varianzverhältnis eindeutig definiert. In der Praxis kann sie jedoch nicht exakt berechnet werden. Zwar kann ein einzelner Messwert als Schätzwert für den wahren Wert dienen, jedoch sind Rückschlüsse auf die Präzision der Messung nicht möglich, da sich die wahren Werte und Messfehler bei nur einer einzigen Messung einer einzelnen Person nicht bestimmen lassen.

Aber auch ohne die wahren Werte einzelner Personen zu kennen, kann das gesuchte Varianzverhältnis als Maß für die Messgenauigkeit geschätzt werden, wenn man die Ebene der einzelnen Personen und einzelnen Items verlässt und stattdessen alle Items, aus denen sich ein Test zusammensetzt, sowie die Messungen mehrerer Personen betrachtet: Wird ein latentes Merkmal anhand mehrerer Items gemessen, so liegen Mehrfachmessungen desselben Merkmals mit unterschiedlichen, aber ähnlichen Messinstrumenten/Items vor, die zu einer Testwertvariablen aufsummiert werden können, sofern sie zumindest die Bedingung der Eindimensionalität erfüllen. Wird diese Testwertvariable an einer Stichprobe erhoben, lassen sich die True-Score- und die Messfehlervarianz der Testwertvariablen bestimmen.

Alle Methoden der Reliabilitätsschätzung basieren darauf, dass mehr als eine Messung des untersuchten Merkmals vorliegt. Hierbei kann es sich um

- verschiedene Items innerhalb eines Tests (► Abschn. 14.3) oder
- wiederholte Messungen anhand desselben oder anhand verschiedener Tests handeln (► Abschn. 14.4; vgl. z. B. Bandalos 2018).

14.2.2 Eindimensionalität

Eindimensionalität bedeutet, dass alle Items eines Tests nur ein einziges, gemeinsames latentes Merkmal erfassen bzw. dass der wahre Varianzanteil der Items auf genau ein latentes Merkmal zurückzuführen ist. Sind die Items eines Tests oder Fragebogens eindimensional, so lassen sich die systematischen Zusammenhänge (Kovarianzen/Korrelationen) der Items y_i durch ein einziges latentes Merkmal erklären.

Die Eindimensionalität eines Tests bzw. der Testitems lässt sich anhand der CFA überprüfen (► Kap. 24). Im Rahmen der CFA wird das latente Merkmal als gemeinsamer erklärender Faktor η modelliert und dessen Anteil aus den Beziehungen der Itemvariablen auspartialisiert (Abb. 14.2).

Die Partialkorrelationen der Itemvariablen (d. h. die verbleibenden Korrelationen nach Auspartialisierung des Einflusses der latenten Variablen η) sollten bei Eindimensionalität der Items null sein. Bleiben jedoch Restkorrelationen bestehen (sichtbar in Kovarianzen der Messfehler), liegen weitere systematische Zusammenhänge zwischen den Itemvariablen vor, die nicht durch das latente Merkmal erklärt werden. Die Unkorriertheit der Messfehlervariablen ist somit eine zentrale Voraussetzung für die Eindimensionalität eines Tests. Somit lautet die Annahme, dass die Fehlerkovarianzen aller Itemvariablen i und i' gleich null sind:

$$\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0 \quad (14.5)$$

! Da die Annahme der Unkorriertheit der Fehler in empirischen Daten verletzt sein kann, muss sie mittels CFA überprüft werden.

■■ Kovarianzen der True-Score-Variablen werden durch das latente Merkmal erklärt

Die beobachteten Werte der Itemvariablen lassen sich jeweils zerlegen in einen wahren Anteil und einen Fehleranteil (vgl. Gl. 14.3). Die wahren Anteile kovariieren miteinander, da sie dasselbe Merkmal erfassen. Diese Kovarianzen zwischen den True-Score-Variablen werden komplett durch die latente Variable η erklärt, sodass die Partialkorrelationen der True-Score-Variablen null sind; d. h., die verbleibenden Korrelationen zwischen den True-Score-Variablen nach Auspartialisierung des Einflusses der latenten Variablen η sind null.

Items messen ein gemeinsames latentes Merkmal

Unkorrierte Messfehler als Voraussetzung für Eindimensionalität

Partialkorrelationen der True-Score-Variablen sind null

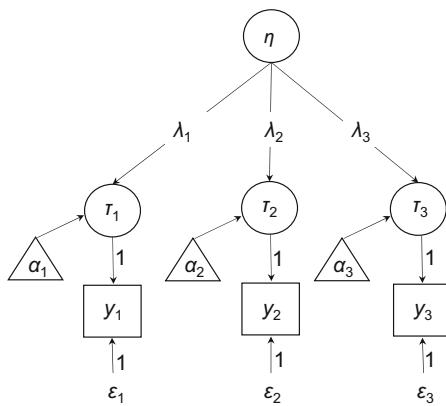


Abb. 14.2 Schematische Darstellung eines eindimensionalen Messmodells mit einer latenten Variablen η , gemessen anhand von drei Items y_i , die sich jeweils aus einem wahren Anteil τ_i und einem Fehleranteil ε_i zusammensetzen ($i = 1, 2, 3$). Die messspezifischen Itemparameter α_i (Leichtigkeitsparameter, Interzept) und λ_i (Diskriminationsparameter, Faktorladung) geben an, wie die latente Variable η mit den True-Score-Variablen τ_i und den beobachteten Variablen y_i verknüpft ist

Somit lässt sich die True-Score-Variable τ_i jeder Itemvariablen y_i ($i = 1, \dots, p$) in Abhängigkeit von der latenten Variablen η und den messspezifischen Itemparametern α_i (Leichtigkeitsparameter, Interzept) und λ_i (Diskriminationsparameter, Faktorladung) darstellen:

$$\tau_i = \alpha_i + \lambda_i \cdot \eta \quad (14.6)$$

Setzt man Gl. (14.6) in Gl. (14.3) ein, lässt sich auch die beobachtete Itemvariable y_i in Abhängigkeit von der latenten Variablen η , den messspezifischen Itemparametern α_i und λ_i sowie dem Fehleranteil ε_i darstellen:

$$y_i = \tau_i + \varepsilon_i = \alpha_i + \lambda_i \cdot \eta + \varepsilon_i \quad (14.7)$$

Die Items eines Tests können unterschiedliche Messeigenschaften aufweisen (► Abschn. 14.2.3) und sich dahingehend voneinander unterscheiden, wie sich Veränderungen der latenten Variablen η auf den wahren Wert τ_i der jeweiligen Itemvariablen auswirken.

14.2.3 Messeigenschaften der Itemvariablen

Die Itemvariablen y_i können sich in den messspezifischen Parametern α_i und λ_i und in der Fehlervarianz $Var(\varepsilon_i)$ unterscheiden:

- Die Itemvariablen y_i können sich in einem additiven Term (Leichtigkeitsparameter; Interzept α_i) unterscheiden. Dies ist der Fall, wenn die Itemvariablen y_i unterschiedliche Mittelwerte (Leichtigkeiten) aufweisen.
- Die Itemvariablen können sich in einem multiplikativen Term (Diskriminationsparameter; Faktorladung λ_i) unterscheiden. Dies ist der Fall, wenn sich die Höhe des Zusammenhangs zwischen den Itemvariablen y_i und der latenten Variablen η unterscheidet. Die quadrierte standardisierte Faktorladung entspricht der *Itemreliabilität*.
- Die Itemvariablen können sich in der Höhe der Fehlervarianz $Var(\varepsilon_i)$ unterscheiden. Dies ist der Fall, wenn sich der Einfluss von Messfehlern zwischen den Items unterscheidet.

Skalenniveau

Die klassischen Reliabilitätsmethoden setzen streng genommen voraus, dass die Itemvariablen metrisch, d. h. mindestens intervallskaliert, sind. In der Praxis sind Itemvariablen jedoch häufig kategorial mit geordneten Antwortkategorien, d. h. ordinalskaliert. Für die Reliabilitätsschätzung ordinalskalierter Variablen mit zwei oder mehr Antwortkategorien stehen verschiedene Methoden zur Verfügung, die von Schermelleh-Engel und Gäde kurz in ► Kap. 15 behandelt werden. Methoden der Reliabilitätsschätzung im Rahmen der Item-Response-Theorie (IRT) finden sich z. B. in ► Kap. 19 sowie bei Eid und Schmidt (2014).

14.2.4 Messäquivalenz

Die Messäquivalenz beschreibt, ob die messspezifischen Parameter mehrerer Items zur Messung einer latenten Variablen identische Werte annehmen und ob gleiche Fehlervarianzanteile vorliegen. Sind sowohl die Faktorladungen als auch die Fehlervarianzen der Items identisch, bedeutet dies, dass diese Items dieselbe Reliabilität haben.

Leichtigkeitsparameter α_i (Interzept)

Diskriminationsparameter λ_i (Faktorladung) und Itemreliabilität

Fehlervarianz $Var(\varepsilon_i)$

Metrische Itemvariablen

Für die Reliabilitätsbestimmung einer Testwertvariablen muss geklärt werden, welche Art (Stufe) der Messäquivalenz (s. ▶ Kap. 13, ▶ Abschn. 13.6.1) vorliegt, da die verschiedenen Reliabilitätsmaße auf unterschiedlich strengen Annahmen hinsichtlich der Gleichheit der messspezifischen Parameter und der Fehlervarianzen der Items beruhen. Diese Annahmen müssen überprüft und erfüllt sein, um eine zuverlässige Schätzung der Reliabilität zu gewährleisten.

- ! Die klassischen Reliabilitätsmaße setzen im Wesentlichen die strengen Annahmen der essentiellen τ -Äquivalenz und der essentiellen τ -Parallelität voraus, die mittels CFA getestet werden können. Der Begriff „essentiell“ bedeutet, dass sich die Itemvariablen im additiven konstanten Term α_i (Interzept) unterscheiden dürfen.

Da die Varianz einer Konstanten null ist und die Kovarianz jeder anderen Variablen mit einer Konstanten ebenfalls null ist, wirkt sich die additive Konstante α_i bei der Bestimmung der Reliabilität als Verhältnis der erklärten Varianz zur Gesamtvarianz nicht weiter aus. Die empirischen Mittelwerte der Itemvariablen dürfen sich daher unterscheiden.

**Interzepte für
Reliabilitätsbestimmung irrelevant**

14.2.5 Schätzung der True-Score-Varianz einer Itemvariablen

Die True-Score-Varianz einer Itemvariablen y_i lässt sich anhand ihrer Kovarianz mit einer anderen Itemvariablen $y_{i'}$ bestimmen. Die Kovarianz zwischen zwei Itemvariablen $Cov(y_i, y_{i'})$ ist in einem eindimensionalen Messmodell allein auf die Kovarianz ihrer wahren Werte $Cov(\tau_i, \tau_{i'})$ zurückzuführen, da True-Score- und Fehlervariablen unkorreliert sind (s. Eigenschaft 3 der True-Score- und Messfehlervariablen in ▶ Kap. 13, ▶ Abschn. 13.3.2) und da bei Eindimensionalität auch keine Fehlerkovarianzen vorliegen (Gl. 14.5). Somit kovariieren nur die wahren Werte:

$$Cov(y_i, y_{i'}) = Cov(\tau_i + \varepsilon_i, \tau_{i'} + \varepsilon_{i'}) = Cov(\tau_i, \tau_{i'}) \quad (14.8)$$

In einem eindimensionalen Modell entspricht die Kovarianz zwischen zwei Itemvariablen der True-Score-Varianz $Var(\tau_i)$ der beiden Items (Gl. 14.9).

**Itemkovarianz entspricht
der wahren Varianz**

$$Cov(y_i, y_{i'}) = Cov(\tau_i, \tau_{i'}) = Var(\tau_i) = Var(\tau_{i'}) \quad (14.9)$$

Berechnet man die Kovarianzen eines Items mit allen anderen Items eines Tests und stellt fest, dass jeweils unterschiedliche Werte resultieren, wird deutlich, dass die True-Score-Varianz des einzelnen Items über die empirischen Kovarianzen nicht eindeutig bestimmt werden kann. Dies wäre nur dann der Fall, wenn die Kovarianzen aller Itempaare identisch wären. Nur dann ließe sich die True-Score-Varianz des einzelnen Items eindeutig anhand der für alle Itempaare identischen Kovarianz bestimmen.

- ! Dies macht deutlich, dass strenge Gleichheitsannahmen erfüllt sein müssen, wenn die Reliabilität anhand der empirischen Kovarianzen bestimmt werden soll.

Die Kovarianz zwischen zwei Items berechnet sich in einem eindimensionalen Modell aus dem Produkt der Faktorladungen und der Varianz der latenten Variablen η :

$$\begin{aligned} Cov(y_i, y_{i'}) &= Cov(\lambda_i \cdot \eta, \lambda_{i'} \cdot \eta) \\ &= \lambda_i \cdot \lambda_{i'} \cdot Var(\eta) \end{aligned} \quad (14.10)$$

Da die Varianz der latenten Variablen η für alle Itemvariablen identisch ist, werden Unterschiede in den Kovarianzen zwischen den Items allein durch deren Faktorladungen erklärt. Die Zerlegung der Itemkovarianz in die Modellparameter wurde etwas ausführlicher in ▶ Kap. 13, ▶ Abschn. 13.6.2.1 dargestellt.

14.3 Cronbachs Alpha

Das wohl bekannteste und am häufigsten verwendete Reliabilitätsmaß ist Cronbachs Alpha (Cronbach 1951). Dieser Reliabilitätskoeffizient wurde nach Lee J. Cronbach (1916–2001) benannt, wurde tatsächlich aber bereits früher als „Koeffizient λ_3 “ entwickelt (Guttman 1945), der eine Verallgemeinerung der für dichotome Variablen entwickelten „Kuder-Richardson-Formel 20 (KR-20)“ darstellt (Kuder und Richardson 1937). Auch Hoyt (1941) entwickelte eine mit Cronbachs Alpha identische Formel, die auf der Varianzanalyse beruht.

14.3.1 Berechnung

Cronbachs Alpha (α) wird meist in Form von Gl. (14.11) berechnet, wobei p die Anzahl der Items, y_i die Itemvariablen ($i = 1, \dots, p$) und Y die Testwertvariable, d. h. die Variable der aufsummierten Itemantworten, bezeichnet:

$$Rel(Y) = \alpha = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^p Var(y_i)}{Var(Y)} \right) \quad (14.11)$$

In die Berechnung von Cronbachs Alpha gehen die Anzahl p der Items, die Summe der Itemvarianzen $Var(y_i)$ sowie die Gesamtvarianz der Testwertvariablen $Var(Y)$ ein (Gl. 14.11). Aus der Gleichung geht nicht unmittelbar hervor, dass Cronbachs Alpha neben den Itemvarianzen auch die *Itemkovarianzen* nutzt, um die True-Score-Varianz der Testwertvariablen zu ermitteln. Dies wird deutlich, wenn man sich erinnert, dass die Testwertvariable eine Linearkombination (Summe) der Itemvariablen ist, deren Varianz nach allgemeinen Rechenregeln bestimmt werden kann. Die Varianz einer solchen Linearkombination ergibt sich durch Addition der Varianzen der einzelnen Summanden (hier also der Itemvariablen) und der zweifachen Summe ihrer nicht redundanten Kovarianzen. Die Varianz der Testwertvariablen entspricht damit der Summe der Itemvarianzen plus der zweifachen Summe der Itemkovarianzen¹:

$$\begin{aligned} Var(Y) &= Var(y_1 + y_2 + \dots + y_p) \\ &= \sum_{i=1}^p Var(y_i) + 2 \cdot \sum_{i < i'} Cov(y_i, y_{i'}) \end{aligned} \quad (14.12)$$

Varianz der Testwertvariable

Setzt man in Gl. (14.12) $y_i = \tau_i + \varepsilon_i$ ein (vgl. Gl. 14.3), lässt sich die Varianz der Testwertvariablen entsprechend als Summe der True-Score- und Fehlervarianzen der Items ausdrücken. Da alle Kovarianzen mit den Fehlervariablen null sind, gilt für $Var(Y)$:

$$Var(Y) = \underbrace{\sum_{i=1}^p Var(\tau_i)}_{Var(T)} + 2 \cdot \underbrace{\sum_{i < i'} Cov(\tau_i, \tau_{i'})}_{0} + \underbrace{\sum_{i=1}^p Var(\varepsilon_i)}_{Var(E)} \quad (14.13)$$

¹ Allgemeine Rechenregeln für Varianzen, s. z. B. Steyer und Eid (2001, S. 343) oder Eid et al. (2017, S. 196).

- ! Die True-Score-Varianzen und -kovarianzen der *Itemvariablen* liefern somit relevante Informationen zur Bestimmung der True-Score- und Fehlervarianz der *Testwertvariablen*.

Die *Itemkovarianzen* liefern dabei die notwendigen Informationen zur Bestimmung der True-Score-Varianz der Testwertvariablen, da in einem eindimensionalen Modell nur die wahren Anteile der Itemvariablen systematisch kovariieren. Diese Kovarianzinformation wird neben der Gesamtvarianz bei der Reliabilitätsschätzung anhand von Cronbachs Alpha genutzt. Cronbachs Alpha setzt essentielle τ -Äquivalenz der Itemvariablen voraus, damit die Reliabilität der Testwertvariablen anhand der Itemvarianzen und -kovarianzen korrekt geschätzt werden kann.

Cronbachs Alpha setzt essentielle τ -Äquivalenz der Itemvariablen voraus

Wo steckt die True-Score-Varianz der Testwertvariablen in der Formel von Cronbachs Alpha?

Die True-Score-Varianz der Testwertvariablen $Var(T_i)$ ergibt sich aus den Kovarianzen der Itemvariablen, da diese der Item-True-Score-Varianz $Var(\tau_i)$ entspricht (Gl. 14.9). Die Itemkovarianzen sind wiederum in der Gesamtvarianz der Testwertvariablen enthalten, wie sich leicht zeigen lässt (vgl. Gl. 14.12).

Ausgehend von der Formel von Cronbachs Alpha (Gl. 14.11) lässt sich in Gl. (14.14) für die Testwertvariable im Zähler der Gleichung der Ausdruck aus Gl. (14.12) einsetzen (Zeile 2), sodass im Zähler der Gleichung die zweifache Summe der Itemkovarianzen resultiert (Zeile 4):

$$\begin{aligned}
 \alpha &= \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^p Var(y_i)}{Var(Y)} \right) \\
 &= \frac{p}{p-1} \cdot \left(\frac{Var(Y)}{Var(Y)} - \frac{\sum_{i=1}^p Var(y_i)}{Var(Y)} \right) \\
 &= \frac{p}{p-1} \cdot \frac{\left(\sum_{i=1}^p Var(y_i) + 2 \cdot \sum_{i < i'} Cov(y_i, y_{i'}) \right) - \sum_{i=1}^p Var(y_i)}{Var(Y)} \\
 &= \frac{p}{p-1} \cdot \frac{2 \cdot \sum_{i < i'} Cov(y_i, y_{i'})}{Var(Y)} \tag{14.14}
 \end{aligned}$$

Da im Modell essentiell τ -äquivalenter Messungen die Kovarianzen aller Itempaare identisch sind, kann die zweifache Summe der nicht redundanten Kovarianzen durch $p \cdot (p-1) \cdot Cov(y_i, y_{i'})$ ersetzt werden. Die alternative Schreibweise von $p \cdot (p-1)$ statt der zweifachen Summe der Kovarianzen wird verständlich, wenn man die Kovarianzmatrix in ► Beispiel 14.1 betrachtet: Die drei Itemkovarianzen in der unteren Dreiecksmatrix sind identisch mit den drei Itemkovarianzen in der oberen Dreiecksmatrix, sodass die zweifache Summe der Kovarianzen einer Dreiecksmatrix der Gesamtsumme aller Kovarianzen der Matrix entspricht. Insgesamt liegen im Beispiel sechs Kovarianzen vor, wobei sich die Anzahl aus $p \cdot (p-1) = 3 \cdot (3-1)$ ergibt. Somit erhält man die Gesamtsumme aller Kovarianzen durch $p \cdot (p-1) \cdot Cov(y_i, y_{i'})$.

Da die Kovarianzen zwischen den Itemvariablen den Item-True-Score-Varianzen entsprechen (vgl. Gl. 14.9), erhält man durch Kürzen und mit Gl. (14.9) die mit p^2 multiplizierte True-Score-Varianz der Itemvariablen als wahre Varianz der

Identische True-Score-Varianzen aller Itemvariablen

Testwertvariablen $Var(T)$ im Zähler:

$$\begin{aligned}\alpha &= \frac{p}{p-1} \cdot \frac{p(p-1) \cdot Cov(y_i, y_{i'})}{Var(Y)} \\ &= \frac{p^2 \cdot Var(\tau)}{Var(Y)} \\ &= \frac{Var(T)}{Var(Y)}\end{aligned}\quad (14.15)$$

Wie man sieht, ist der Term $p/(p-1)$ in der Formel von Cronbachs Alpha kein Korrekturfaktor, wie häufig angenommen wird, sondern notwendig, um den korrekten Anteil der wahren Varianz an der Gesamtvarianz der Testwertvariablen bestimmen zu können. Nur bei essentieller τ -Äquivalenz der Itemvariablen sind die True-Score-Varianzen aller Itemvariablen identisch und ihre Kovarianzen jeweils gleich der True-Score-Varianz. Nur in diesem Fall geht durch diesen Term die True-Score-Varianz p^2 -mal in die Berechnung von Alpha ein.

Beispiel 14.1: Berechnung von Cronbachs Alpha

Folgendes Zahlenbeispiel zeigt die Itemvarianzen und -kovarianzen für drei essentiell τ -äquivalente Itemvariablen y_1 , y_2 und y_3 (links) und die Aufteilung der empirischen Itemvarianzen in True-Score- und Fehlervarianz (rechts).

Itemvarianzen und -kovarianzen			Aufteilung der Itemvarianzen und -kovarianzen in True-Score- und Fehlervarianzen		
2.10	1.50	1.50	1.50 + 0.60	1.50	1.50
1.50	2.00	1.50	1.50	1.50 + 0.50	1.50
1.50	1.50	1.90	1.50	1.50	1.50 + 0.40
<i>Rel = .90</i>					

Bei p Items gibt es $p \cdot (p-1)$ Kovarianzen, in diesem Beispiel also $3 \cdot (3-1) = 6$ Kovarianzen, die für die essentiell τ -äquivalenten Variablen alle denselben Wert aufweisen und der Item-True-Score-Varianz entsprechen (hier 1.50).

Zur Berechnung von Alpha nach Gl. (14.11) werden die Anzahl der Items ($p = 3$), die Kovarianz der Items ($Cov(y_i, y_{i'}) = 1.50$) sowie die Varianzen der Items ($Var(y_1) = 2.10$; $Var(y_2) = 2.00$; $Var(y_3) = 1.90$) benötigt. Die Varianz der Testwertvariablen lässt sich dafür zunächst nach Gl. (14.12) anhand der Itemvarianzen und -kovarianzen bestimmen:

$$\begin{aligned}Var(Y) &= \sum_{i=1}^p Var(y_i) + 2 \cdot \sum_{i < i'} Cov(y_i, y_{i'}) \\ &= (2.10 + 2.00 + 1.90) + 2 \cdot (1.50 + 1.50 + 1.50) = 6 + 9 = 15\end{aligned}$$

Werden die Werte in Gl. (14.11) eingesetzt, resultiert für Alpha folgendes Ergebnis:

$$\alpha = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^p Var(y_i)}{Var(Y)} \right) = \frac{3}{3-1} \cdot \left(1 - \frac{6}{15} \right) = \frac{3}{2} \cdot \frac{9}{15} = .90$$

14.3 · Cronbachs Alpha

Bei Berechnung von Alpha anhand von Gl. (14.15) resultiert entsprechend dasselbe Ergebnis:

$$\alpha = \frac{p^2 \cdot \text{Cov}(y_i, y_{i'})}{\text{Var}(Y)} = \frac{p^2 \cdot \text{Var}(\tau)}{\text{Var}(Y)} = \frac{3^2 \cdot 1.50}{15} = \frac{13.50}{15} = .90$$

Die anhand von Cronbachs Alpha geschätzte Reliabilität für das empirische Zahlenbeispiel ist somit $\alpha = .90$.

14.3.2 Voraussetzungen von Cronbachs Alpha

Cronbachs Alpha beruht auf der Grundidee, dass für einen aus mehreren Items bestehenden Test die Testwertvariable als Summe der Itemvariablen gebildet und die Reliabilität für diese Summenvariable (Testwertvariable) anhand der Varianzen und Kovarianzen der Items bestimmt werden kann.

! Damit Cronbachs Alpha tatsächlich die Reliabilität als Verhältnis der True-Score-Varianz zur Gesamtvarianz schätzt, müssen als Voraussetzung die Eindimensionalität und die essentielle τ -Äquivalenz der Itemvariablen gegeben sein.

Eindimensionalität und essentielle τ -Äquivalenz

Um dem Modell essentieller τ -Äquivalenz zu entsprechen, müssen die Kovarianzen aller Itempaare identisch sein. Die Kovarianz der Items entspricht der True-Score-Varianz $\text{Var}(\tau)$ eines Items (vgl. Gl. 14.9); sie ist im Modell essentieller τ -Äquivalenz für alle Items identisch:

$$\text{Cov}(y_i, y_{i'}) = \text{Cov}(\tau_i, \tau_{i'}) = \text{Var}(\tau) \quad (14.16)$$

Identische Itemkovarianzen

Die Items weisen somit genau dann denselben Anteil an wahrer Varianz auf, wenn für alle Items essentielle τ -Äquivalenz gegeben ist. Empirisch zeigt sich dies in näherungsweise identischen Kovarianzen für alle Itempaare. In diesem Fall müssen für alle Items identische Faktorladungen $\lambda_i = \lambda_{i'} = \lambda$ vorliegen. Die Gleichheit der Faktorladungen lässt sich gezielt überprüfen (► Abschn. 14.3.3). Somit gilt im Modell essentieller τ -Äquivalenz für die True-Score-Varianz der Itemvariablen:

$$\text{Cov}(y_i, y_{i'}) = \lambda \cdot \lambda \cdot \text{Var}(\eta) = \text{Var}(\tau) \quad (14.17)$$

Die wahre Varianz ist in dem Fall für alle Items identisch, während sich der Anteil der Fehlervarianz über die Items hinweg unterscheiden kann:

$$\text{Var}(\varepsilon_i) \neq \text{Var}(\varepsilon_{i'}) \quad (14.18)$$

Die Varianz jedes Items y_i lässt sich somit zerlegen in die für alle Items identische True-Score-Varianz und eine jeweils itemspezifische Fehlervarianz.

■■ Kovarianzmatrix bei essentieller τ -Äquivalenz

Korrespondierend zu dem ► Beispiel 14.1 werden in den folgenden drei linken Spalten die Kovarianzmatrix von drei Items y_1 , y_2 und y_3 mit ihren drei Varianzen und sechs Kovarianzen und in den drei rechten Spalten die dazugehörige Varianzaufteilung bei essentieller τ -Äquivalenz dargestellt.

$\text{Var}(y_1)$	$\text{Cov}(y_1, y_2)$	$\text{Cov}(y_1, y_3)$	$\text{Var}(\tau) + \text{Var}(\varepsilon_1)$	$\text{Var}(\tau)$	$\text{Var}(\tau)$
$\text{Cov}(y_2, y_1)$	$\text{Var}(y_2)$	$\text{Cov}(y_2, y_3)$	$\text{Var}(\tau)$	$\text{Var}(\tau) + \text{Var}(\varepsilon_2)$	$\text{Var}(\tau)$
$\text{Cov}(y_3, y_1)$	$\text{Cov}(y_3, y_2)$	$\text{Var}(y_3)$	$\text{Var}(\tau)$	$\text{Var}(\tau)$	$\text{Var}(\tau) + \text{Var}(\varepsilon_3)$
Für alle $\text{Var}(\tau)$ gilt: $\text{Var}(\tau) = \lambda^2 \cdot \text{Var}(\eta)$ (vgl. ► Gl. 14.17)					

Die drei Varianzen setzen sich jeweils aus der True-Score- und der Fehlervarianz zusammen, wobei die True-Score-Varianzen der drei Items identisch sind, während sich die Fehlervarianzen unterscheiden. Durch die unterschiedlichen Fehlervarianzen unterscheiden sich auch die empirischen Varianzen der Items. Bei essentieller τ -Äquivalenz der Items sind die Kovarianzen aller Itempaare identisch und entsprechen der True-Score-Varianz der Items.

Die essentielle τ -Äquivalenz kann anhand von ► Beispiel 14.1 verdeutlicht werden. Wie die empirischen Werte zeigen, sind die Varianzen der drei essentiell τ -äquivalenten Itemvariablen unterschiedlich groß (2.10, 2.00 und 1.90) und setzen sich jeweils zusammen aus derselben True-Score-Varianz (1.50) und unterschiedlichen Fehlervarianzen (0.60, 0.50 und 0.40). Alle sechs Kovarianzen weisen bei essentiell τ -äquivalenten Variablen denselben Wert auf und entsprechen der True-Score-Varianz der Items (hier 1.50).

! Cronbachs Alpha entspricht der Reliabilität der Testwertvariablen nur bei gegebener essentieller τ -Äquivalenz, d. h. nur im Fall eindimensionaler Messungen und identischer True-Score-Varianzen aller Itemvariablen (vgl. auch Novick und Lewis 1967; Raykov und Marcoulides 2011; Steyer und Eid 2001). Andernfalls ist Cronbachs Alpha kein geeignetes Reliabilitätsmaß.

14.3.3 Überprüfung der Voraussetzungen von Cronbachs Alpha

Vor Anwendung von Cronbachs Alpha zur Schätzung der Reliabilität einer Skala sollten die notwendigen Voraussetzungen mittels CFA überprüft werden (vgl. ► Kap. 24). Die erforderliche Eindimensionalität und die essentielle τ -Äquivalenz der Itemvariablen lassen sich überprüfen, indem ein eindimensionales Messmodell spezifiziert wird, in dem alle Itemvariablen auf einer gemeinsamen latenten Variablen η laden und keine Fehlerkovanzen vorliegen (Eindimensionalität).

Entsprechend der Annahme der essentiellen τ -Äquivalenz müssen für alle Itemvariablen die multiplikativen Terme λ_i identisch sein, d. h., die Faktorladungen müssen im Modell für alle Itemvariablen gleichgesetzt werden. Diese Gleichsetzung impliziert, dass die empirischen Kovarianzen aller Itemvariablen identisch sind. Anhand des χ^2 -Tests wird überprüft, ob das Modell mit diesen Restriktionen einen guten Modellfit aufweist, d. h., ob diese Modellannahmen gut zu den empirischen Daten passen. Zusätzlich kann mittels χ^2 -Differenztest geprüft werden, ob sich der Modellfit dieses restriktiven Modells signifikant vom Modellfit eines Modells ohne die Restriktion gleicher Faktorladungen unterscheidet (vgl. ► Kap. 24).

14.3.4 Probleme von Cronbachs Alpha

Die Aussagekraft der Reliabilitätsschätzung anhand von Cronbachs Alpha hängt davon ab, ob die notwendigen Voraussetzungen erfüllt sind. Nur bei Eindimensionalität (d. h., alle Items messen nur ein gemeinsames Merkmal) und essentieller τ -Äquivalenz (d. h. identische Diskriminationsparameter λ) wird die Reliabilität durch Cronbachs Alpha korrekt geschätzt. Sind diese Voraussetzungen nicht erfüllt, kann die Schätzung der Reliabilität unter Umständen deutlich verzerrt sein. Ganz allgemein sollte Cronbachs Alpha nicht verwendet werden, wenn die Voraussetzungen verletzt sind (► Abschn. 14.8).

! Die Eindimensionalität der Itemvariablen ist eine wichtige Voraussetzung für die Verwendung von Cronbachs Alpha; umgekehrt kann von Cronbachs Alpha aber nicht auf die Dimensionalität der Messungen geschlossen werden (► Exkurs 14.1).

14.3.4.1 Fehlerkovarianzen durch Methodeneffekte

Eine artifizielle Unter- oder Überschätzung der Reliabilität anhand von Cronbachs Alpha kann aufgrund von *Methodeneffekten* resultieren (► Kap. 15). Methodeneffekte können beispielsweise durch invers formulierte Items entstehen. Invers formulierte Items werden oftmals verwendet, um Antworttendenzen (z. B. Akquieszenz, ► Kap. 4) zu eliminieren. Solche Items können jedoch problematisch sein: Faktorenanalytische Untersuchungen haben gezeigt, dass invers formulierte Items unabhängig vom jeweiligen Iteminhalt einen eigenen Faktor bilden können (vgl. u. a. Podsakoff et al. 2003). In einem Fragebogen zur Schmerzregulation (Schermelleh-Engel 1995) finden sich beispielsweise in der Subskala *Ablenkung* zwei invers formulierte Items: „Wenn ich Schmerzen habe, kann ich an nichts anderes mehr denken“ und „Wenn ich Schmerzen habe, kann ich mich durch nichts davon ablenken“. Diese beiden Items beinhalten neben der Merkmalsvarianz zusätzliche systematische Methodenvarianz, die in einer positiven Fehlerkovarianz resultiert. Eine solche Methodenvarianz würde der essentiellen τ -Äquivalenz der Items widersprechen und Cronbachs Alpha würde in diesen Fällen die tatsächliche Reliabilität der Skalen überschätzen (► Exkurs 14.1, Beispiel C).

Auch ähnliche Itemformulierungen können dazu führen, dass neben der wahren Merkmalsvarianz zusätzlich eine systematische Varianz entsteht, die als Methodeneffekt interpretiert werden kann. In der Subskala *Concern over Mistakes* der Mehrdimensionalen Perfektionismus-Skala (MPS-F) von Frost et al. (1990) in der deutschen Fassung von Stöber (1995) finden sich z. B. die Items „Wenn ich bei der Arbeit/beim Studium versage, dann bin ich auch als Mensch ein Versager“ und „Wenn ich nicht so gut bin wie andere, dann heißt das, dass ich als Mensch weniger wert bin“. Wie eine CFA zeigen konnte (► Kap. 15), besteht eine zusätzliche positive Fehlerkovarianz zwischen den Itemvariablen, die aufgrund der ähnlichen Formulierungen dieser Items („ich als Mensch“) als Methodenvarianz interpretiert werden kann.

Invers formulierte Items

Ähnliche Itemformulierungen

14.3.4.2 Unklare Bedeutung des Begriffs „interne/innere Konsistenz“

Cronbachs Alpha wird häufig als ein Maß der internen bzw. inneren Konsistenz bezeichnet. Dieser Begriff ist jedoch unklar und nicht genau definiert. Die Begriffe interne (oder innere) Konsistenz, Homogenität und Eindimensionalität werden häufig gleichbedeutend verwendet (Cronbach 1951; McDonald 1999; Peters 2014; Schmitt 1996; Sijtsma 2009), ebenso wie Homogenität, Eindimensionalität und Faktorsättigung (Cronbach 1951; McDonald 1999; Schmitt 1996; Sijtsma 2009). Nach Raykov und Marcoulides (2011, S. 155) bedeutet interne Konsistenz in Bezug auf Cronbachs Alpha:

- » Alpha is an index of internal consistency, i.e., the degree to which a set of components is interrelated, in the sense of inter-item covariance per unit of overall sum variance.

Ein hoher Wert von Cronbachs Alpha wird somit oftmals gleichgesetzt mit einer hohen internen Konsistenz und damit mit hohen Korrelationen zwischen den Items, obwohl auch andere Erklärungen für einen hohen Alpha-Wert möglich sind (vgl. z. B. Cho und Kim 2015; Cortina 1993; Cronbach 1951; Grayson 2004; McDonald 1999). Wie die Beispiele im ► Exkurs 14.1 zeigen, kann ein hoher Alpha-Wert nicht nur aus hoch miteinander korrelierenden Items resultieren, sondern auch aufgrund von Mehrdimensionalität (Beispiel B) oder zusätzlichen Methodeneffekten (Beispiel C).

Mehrdimensionalität und Methodeneffekte

Exkurs 14.1**Bedeutet ein hoher Alpha-Wert, dass die zugrunde liegenden Skala eindimensional ist?**

Im Folgenden sind vier verschiedene Beispiele für sechs Items eines Tests mit den jeweiligen Kovarianz- und Korrelationsmatrizen der Items aufgeführt. Die Hauptdiagonalen aller vier Kovarianzmatrizen, d. h. die Varianzen der Itemvariablen y_1 – y_6 , sind über alle Beispiele hinweg identisch und variieren über die Items hinweg zwischen 1.5 und 2.5:

- Bei gegebener essentieller τ -Äquivalenz (Beispiel A) weist jede Itemvariable dieselbe True-Score-Varianz auf, während die Fehlervarianzen unterschiedlich sein kön-

nen, was sich in identischen Itemkovarianzen, aber unterschiedlichen Itemkorrelationen äußert.

- In den Beispielen B, C und D sind dagegen die Voraussetzungen der Eindimensionalität und essentiellen τ -Äquivalenz der Itemvariablen verletzt.

Für jede der vier Kovarianzmatrizen wurde Cronbachs Alpha beispielhaft berechnet – unabhängig davon, ob dessen Voraussetzungen erfüllt sind oder nicht. Die Beispiele verdeutlichen, dass von einem hohen Alpha-Wert nicht auf die Eindimensionalität der Itemvariablen geschlossen werden kann.

Beispiel A: Essentielle τ -Äquivalenz gegeben – Die Berechnung von Cronbachs Alpha ist angemessen

Itemvarianzen und -kovarianzen						Itemkorrelationen					
1.5	0.8	0.8	0.8	0.8	0.8	1.00	.46	.41	.53	.46	.41
0.8	2.0	0.8	0.8	0.8	0.8	.46	1.00	.36	.46	.40	.36
0.8	0.8	2.5	0.8	0.8	0.8	.41	.36	1.00	.41	.36	.32
0.8	0.8	0.8	1.5	0.8	0.8	.53	.46	.41	1.00	.46	.41
0.8	0.8	0.8	0.8	2.0	0.8	.46	.40	.36	.46	1.00	.36
0.8	0.8	0.8	0.8	0.8	2.5	.41	.36	.32	.41	.36	1.00
$\alpha = .80$											

In diesem Beispiel ist die Voraussetzung der essentiellen τ -Äquivalenz der sechs Itemvariablen erfüllt, was an den identischen Kovarianzen zu erkennen ist. Die Reliabilität des Tests kann daher über Cronbachs Alpha angemessen geschätzt werden. Mit einem Wert von $\alpha = .80$ ist die Reliabilität zufriedenstellend hoch. Die

zugehörige Korrelationsmatrix zeigt, dass bei essentieller τ -Äquivalenz keinesfalls auch die Korrelationen identisch sein müssen. Identische Korrelationen ergänzen sich nur für essentiell τ -parallele Messungen, d. h., wenn zusätzlich alle Fehlervarianzen und damit alle Itemvarianzen identisch wären (► Abschn. 14.4.2).

Beispiel B: Fehlende Eindimensionalität – Die Berechnung von Cronbachs Alpha ist unangemessen

Itemvarianzen und -kovarianzen						Itemkorrelationen					
1.5	1.3	1.3	0.4	0.4	0.4	1.00	.75	.67	.27	.23	.21
1.3	2.0	1.3	0.4	0.4	0.4	.75	1.00	.58	.23	.20	.18
1.3	1.3	2.5	0.4	0.4	0.4	.67	.58	1.00	.21	.18	.16
0.4	0.4	0.4	1.5	1.5	1.5	.27	.23	.21	1.00	.87	.77
0.4	0.4	0.4	1.5	2.0	1.5	.23	.20	.18	.87	1.00	.67
0.4	0.4	0.4	1.5	1.5	2.5	.21	.18	.16	.77	.67	1.00
$\alpha = .80$											

In diesem Beispiel ist keine Eindimensionalität gegeben, da die Itemvariablen y_1 bis y_3 ein Konstrukt und die Itemvariablen y_4 bis y_6 ein weiteres Konstrukt messen; die Konstrukte korrelieren gering miteinander ($r = .29$). Die Voraussetzung der essentiellen τ -Äquivalenz ist somit nicht erfüllt und Cronbachs Alpha sollte nicht

berechnet werden. Wird Cronbachs Alpha dennoch berechnet, ergibt sich zwar der identische Wert für Cronbachs Alpha ($\alpha = .80$) wie für in Beispiel A, jedoch wird hier ersichtlich, dass von einem hohen Alpha-Wert nicht auf die Eindimensionalität der Messungen geschlossen werden kann.

Beispiel C: Positive Fehlerkovarianz – Die Berechnung von Cronbachs Alpha ist unangemessen

Itemvarianzen und -kovarianzen						Itemkorrelationen					
1.5	0.8	0.8	0.8	0.8	0.8	1.00	.46	.41	.53	.46	.41
0.8	2.0	2.0 ^a	0.8	0.8	0.8	.46	1.00	.89	.46	.40	.36
0.8	2.0 ^a	2.5	0.8	0.8	0.8	.41	.89	1.00	.41	.36	.32
0.8	0.8	0.8	1.5	0.8	0.8	.53	.46	.41	1.00	.46	.41
0.8	0.8	0.8	0.8	2.0	0.8	.46	.40	.36	.46	1.00	.36
0.8	0.8	0.8	0.8	0.8	2.5	.41	.36	.32	.41	.36	1.00
$\alpha = .825$											
^a Die Kovarianz zwischen Itemvariablen y_2 und y_3 in Höhe von 2.0 setzt sich zusammen aus der wahren Varianz (0.8) und einer positiven Fehlerkovarianz in Höhe von 1.2.											

In diesem Beispiel ist die Voraussetzung der essentiellen τ -Äquivalenz nicht erfüllt, da zwei Itemvariablen eine zusätzliche positive Fehlerkovarianz aufweisen: Die Kovarianz zwischen y_2 und y_3 in Höhe von 2.0 setzt sich zusammen aus der wahren Varianz (0.8) und einer positiven Fehlerkovarianz in Höhe von 1.2. Die

Kovarianz der beiden Items entspricht damit nicht mehr der wahren Varianz der übrigen Items. Wird Cronbachs Alpha in unangemessener Weise dennoch berechnet, so wird Alpha gegenüber der eindimensionalen Messung in Beispiel A überschätzt ($\alpha = .825$).

Beispiel D: Negative Fehlerkovarianz – Die Berechnung von Cronbachs Alpha ist unangemessen

Itemvarianzen und -kovarianzen						Itemkorrelationen					
1.5	0.8	0.8	0.8	0.8	0.8	1.00	.46	.41	.53	.46	.41
0.8	2.0	-0.4 ^a	0.8	0.8	0.8	.46	1.00	-.18	.46	.40	.36
0.8	-0.4 ^a	2.5	0.8	0.8	0.8	.41	-.18	1.00	.41	.36	.32
0.8	0.8	0.8	1.5	0.8	0.8	.53	.46	.41	1.00	.46	.41
0.8	0.8	0.8	0.8	2.0	0.8	.46	.40	.36	.46	1.00	.36
0.8	0.8	0.8	0.8	0.8	2.5	.41	.36	.32	.41	.36	1.00
$\alpha = .771$											
^a Die Kovarianz zwischen Itemvariablen y_2 und y_3 in Höhe von -0.4 setzt sich zusammen aus der wahren Varianz (0.8) und einer negativen Fehlerkovarianz in Höhe von -1.2.											

Hier ist die Voraussetzung der essentiellen τ -Äquivalenz nicht erfüllt, da zwei Itemvariablen eine zusätzliche negative Fehlerkovarianz aufweisen: Die Kovarianz zwischen y_2 und y_3 in Höhe von -0.4 setzt sich zusammen aus der wahren Varianz (0.8) und einer negativen Fehlerkovarianz in Höhe von -1.2. Die Kovarianz der beiden Items entspricht damit nicht mehr der wahren Varianz der übrigen Items. Wird Cronbachs Alpha in unangemessener Weise dennoch berechnet, so wird Cronbachs Alpha gegenüber der eindimensionalen Messung in Beispiel A unterschätzt ($\alpha = .771$).

Wie die Beispiele zeigen, sagt ein hoher Alpha-Wert nichts über die Dimensionalität der Messungen aus. Vielmehr sollten die Eindimensionalität sowie die essentielle τ -Äquivalenz der Messungen zunächst immer als Voraussetzung zur Anwendung von Cronbachs Alpha mittels CFA überprüft werden (► Kap. 24).

Die Dimensionalität muss neben der essentiellen τ -Äquivalenz der Itemvariablen immer getestet werden, bevor Cronbachs Alpha als Reliabilitätskoeffizient interpretiert werden darf.

Itemanzahl

Weiter wirkt sich auch die Anzahl der Items auf die Höhe des Alpha-Wertes aus. Wie bereits Cortina (1993) zeigen konnte, erreicht Cronbachs Alpha selbst bei nur gering korrelierten Items (im Durchschnitt $r = .30$) bei 6 Items bereits einen Wert von $\alpha = .72$, bei 12 Items einen Wert von $\alpha = .84$ und bei 18 Items einen Wert von $\alpha = .88$.

Die interne Konsistenz und Alpha gleichzusetzen, erscheint somit nicht sehr sinnvoll (vgl. Raykov und Marcoulides 2011; Streiner 2003). Der Begriff der internen Konsistenz wird daher hier nicht für Cronbachs Alpha verwendet. Bei Bedarf kann kontextspezifisch auf die Eindimensionalität oder die hohen Korrelationen zwischen Items Bezug genommen werden.

14.3.4.3 Untere Schranke der Reliabilität

Oftmals wird Cronbachs Alpha als untere Schranke der Reliabilität bezeichnet. Bereits Guttman (1945) bezeichnete seine sechs Lambda-Koeffizienten, von denen λ_3 dem Koeffizienten Alpha entspricht, als Maße der unteren Schranke der Reliabilität. Das Beispiel C in ► Exkurs 14.1 zeigt allerdings, dass diese Aussage problematisch ist, wenn die Voraussetzungen für Cronbachs Alpha nicht erfüllt sind. Der Begriff der „unteren Schranke der Reliabilität“ erscheint daher nicht sehr hilfreich und nicht mehr zeitgemäß. Nach Novick und Lewis (1967) stellt Cronbachs Alpha eine untere Schranke der Reliabilität dar, wenn zwar die Voraussetzung der essentiellen τ -Äquivalenz verletzt ist, aber zumindest keine Methodeneffekte vorliegen (s. auch Steyer und Eid 2001). Bei Verletzung der Annahme der essentiellen τ -Äquivalenz sollten die Itemvariablen jedoch hohe Faktorladungen λ_i aufweisen (vgl. Raykov 1997; Raykov und Marcoulides 2015, 2019).

14.4 Test-Test-Korrelation

Die Test-Test-Korrelation stellt eine weitere Möglichkeit der Reliabilitätsbestimmung dar (Brown 1910; Guttman 1945; Spearman 1910). Zu dieser Art der klassischen Reliabilitätsbestimmung zählen die

- Retest-Reliabilität,
- Paralleltest-Reliabilität und
- Split-Half-Reliabilität.

Bei Bestimmung der Reliabilität anhand der Test-Test-Korrelation wird ein Merkmal in derselben Stichprobe zweimal gemessen und die Korrelation der resultierenden Testwertvariablen Y_1 und Y_2 bestimmt.

Hierbei kann das Merkmal entweder zu zwei Messzeitpunkten anhand desselben Tests bzw. anhand von zwei parallelen Testversionen oder auch zu einem Messzeitpunkt mit zwei Testhälften (Halbtests) gemessen werden. Die Messgenauigkeit kann entsprechend als Retest-, Paralleltest- oder Split-Half-Reliabilität anhand der Korrelation der aus den zwei Messungen gewonnenen Testwertvariablen (bzw. Halbtestwertvariablen) bestimmt werden.

Merkmal wird zweimal gemessen**14.4.1 Berechnung**

Zur Berechnung der Reliabilität werden die zwei Testwertvariablen korreliert

$$\text{Corr}(Y_1, Y_2) = \text{Rel}(Y) \quad (14.19)$$

! Um eine Korrelation zwischen zwei (Halb-)Testwertvariablen als Reliabilität interpretieren zu dürfen, müssen strenge Voraussetzungen erfüllt sein (► Abschn. 14.4.2), da nur bei Zutreffen der Voraussetzungen im Zähler des Korrelationskoeffizienten

14.4 · Test-Test-Korrelation

die True-Score-Varianz und im Nenner die Gesamtvarianz der Testwertvariablen stehen.

Dies wird deutlich, wenn man die Gleichung in einzelne Schritte zerlegt betrachtet. Die Korrelation zweier Variablen ist allgemein definiert als die Kovarianz beider Variablen geteilt durch das Produkt ihrer Standardabweichungen, die sich aus der positiven Quadratwurzel ihrer Varianzen ergeben:

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(Y_1)} \sqrt{\text{Var}(Y_2)}} \quad (14.20)$$

Die Testwertvariablen setzen sich aus ihren True-Score- und Fehlervariablen zusammen, sodass gilt:

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Cov}(T_1 + E_1, T_2 + E_2)}{\sqrt{\text{Var}(Y_1)} \sqrt{\text{Var}(Y_2)}} \quad (14.21)$$

Da nur die True-Score-Variablen korrelieren und alle Kovarianzen mit den Fehlervariablen entfallen, reduziert sich die Gleichung folgendermaßen:

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Cov}(T_1, T_2)}{\sqrt{\text{Var}(Y_1)} \sqrt{\text{Var}(Y_2)}} \quad (14.22)$$

Ist die True-Score-Varianz beider Testwertvariablen identisch ($\text{Var}(T_1) = \text{Var}(T_2)$), so kann die Kovarianz der True-Score-Variablen $\text{Cov}(T_1, T_2)$ im Zähler der Gleichung durch die Varianz der True-Score-Variablen $\text{Var}(T)$ ersetzt werden.

Sind auch die Varianzen beider Testwertvariablen identisch, resultiert aus dem Produkt im Nenner der Gleichung die Varianz der Testwertvariablen. Dies ist jedoch nur dann der Fall, wenn die beiden Testwertvariablen identische True-Score- und Fehlervarianzen aufweisen (essentielle τ -Parallelität).

- ! Die True-Score-Varianz der Testwertvariablen im Zähler und ihre Gesamtvarianz im Nenner der Gleichung ergeben sich somit nur, wenn die Voraussetzungen der essentiellen τ -Parallelität erfüllt sind.

Somit ist gezeigt, dass die Korrelation zwischen zwei Testwertvariablen nur unter der Bedingung der essentiellen τ -Parallelität tatsächlich die Reliabilität als Varianzverhältnis der True-Score-Varianz zur Gesamtvarianz schätzt.

Bei essentieller τ -Parallelität kann die Reliabilität anhand der Test-Test-Korrelation wie folgt geschätzt werden:

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Var}(T)}{\text{Var}(Y)} = \text{Rel}(Y) \quad (14.23)$$

Die True-Score-Varianz und die Gesamtvarianz der Testwertvariablen ergeben sich wiederum aus den Itemvarianzen und Itemkovarianzen. Daher gilt auf Itemebene, dass die Voraussetzung der essentiellen τ -Parallelität der Testwertvariablen nur erfüllt ist, wenn die korrespondierenden Itemvariablen essentielle τ -Parallelität aufweisen (► Abschn. 14.4.2).

- ! Die Korrelation der Testwertvariablen stellt nur bei essentieller τ -Parallelität der Testwertvariablen bzw. bei essentieller τ -Parallelität korrespondierender Itemvariablen eine adäquate Schätzung der Reliabilität dar.

Eindimensionalität und essentielle τ -Parallelität der Testwertvariablen über die Messungen hinweg

Spearman-Brown-Formel der Testverlängerung

Die Split-Half- oder Testhalbierungs-Reliabilität kann geschätzt werden, wenn ein Test aus einer größeren Anzahl von Items besteht. Hierzu werden die Items des Tests in zwei parallele Testhälften aufgeteilt und derselben Stichprobe von Testpersonen vorgegeben. Die Itemantworten pro Testhälfte werden zu je einem *Halbtestwert* aufsummiert und die Korrelation der Halbtestwertvariablen Y_a und $Y_{a'}$ zur Schätzung der Split-Half-Reliabilität der Testhälften verwendet.

$$\text{Corr}(Y_a, Y_{a'}) = \text{Split-Half-Rel}(Y_a) \quad (14.24)$$

Halbtest-Korrelation

Spearman-Brown-Formel der Testverlängerung

Da die Halbtest-Korrelation (Split-Half-Reliabilität) nur der Reliabilität eines Tests halber Länge entspricht, muss die Halbtest-Korrelation zur Schätzung der Reliabilität des Gesamttests $\text{Rel}(Y)$ rechnerisch auf die volle Testlänge aufgewertet werden. Dies geschieht mithilfe der Spearman-Brown-Formel der Testverlängerung (Gl. 14.25). Diese Formel kann allerdings nur verwendet werden, wenn die Variablen der Halbtestwerte essentiell τ -parallel sind (vgl. Raykov und Marcoulides 2011, S. 141) bzw. wenn die korrespondierenden Itemvariablen der Testhälften essentiell τ -parallel sind.

$$\text{Rel}(Y) = \frac{2 \cdot \text{Corr}(Y_a, Y_{a'})}{1 + \text{Corr}(Y_a, Y_{a'})} = \frac{2 \cdot \text{Rel}(Y_a)}{1 + \text{Rel}(Y_a)} \quad (14.25)$$

14.4.2 Voraussetzungen der Test-Test-Korrelation

Damit die Korrelation zweier (Halb-)Testwertvariablen als Reliabilität des Tests interpretiert werden darf, müssen einige strenge Voraussetzungen erfüllt sein. Mit der Test-Test-Korrelation ist nur dann eine adäquate Schätzung der Reliabilität des Tests möglich, wenn die Itemvariablen beider Messungen eine einzige latente Variable η messen, d. h., wenn sie eindimensional sind und keine Fehlerkovarianzen zwischen den Itemvariablen vorliegen (Raykov et al. 2015). Außerdem müssen die Itemvariablen *in beiden Messungen* dieselbe Reliabilität haben, d. h., sie müssen über die Messungen hinweg essentiell τ -parallel sein (vgl. auch die Erläuterungen zu Gln. (14.19) bis (14.23), die sich von der Ebene der Testwertvariablen analog auf die Ebene der Itemvariablen übertragen lassen).

Bei essentieller τ -Parallelität sind sowohl die True-Score- als auch die Fehlervarianzen der Items über die Messungen hinweg identisch. Nur bei identischen True-Score- und Fehlervarianzen der *Itemvariablen* folgt gleichzeitig, dass auch die True-Score- und Fehlervarianzen der *Testwertvariablen* (und somit deren Reliabilität) bei beiden Messungen identisch sind. Unterscheidet sich dieses Varianzverhältnis zwischen den Messungen, darf die Test-Test-Korrelation nicht als Reliabilität interpretiert werden.

Die essentielle τ -Parallelität der Testwertvariablen kann anhand der Messeigenschaften der einzelnen Items geprüft werden. Das Messmodell der Testwertvariablen selbst kann τ -kongenerisch sein (► Kap. 13). Das heißt, dass sich die Faktorladungen und Fehlervarianzen der Items innerhalb eines Tests unterscheiden dürfen, jedoch muss über die Messungen hinweg *strikte Messinvarianz* gegeben sein, d. h., dass bei beiden Messungen dasselbe τ -kongenerische Messmodell gültig sein muss (vgl. ► Kap. 24).

Strikte Messinvarianz bedeutet, dass die Faktorladung und somit auch die True-Score-Varianz jeder Itemvariablen bei beiden Messungen identisch, d. h. für korrespondierende Items invariant sind. Nur dann ist sichergestellt, dass anhand der Items dasselbe Merkmal gemessen wird. Zusätzlich muss die Fehlervarianz jedes

14.4 · Test-Test-Korrelation

Items bei beiden Messungen identisch (invariant) sein; auch die Varianz der latenten Variablen muss gleich sein. Damit ist sichergestellt, dass die Varianzen der Items über die Messungen hinweg identisch sind. Nur bei strikter Messinvarianz sind die Items *über die Messungen hinweg* essentiell τ -parallel und weisen bei jeder Messgelegenheit dieselbe Reliabilität auf. Bei strikter Messinvarianz der Items über die Messungen hinweg erfüllt auch die Testwertvariable die Anforderungen der essentiellen τ -Parallelität. Auch hier gilt, dass die Messfehler über die Messungen hinweg nicht miteinander korrelieren dürfen.

Das Modell der essentiellen τ -Parallelität basiert auf strengerer Annahmen als das Modell essentieller τ -Äquivalenz. Das Modell setzt neben Eindimensionalität der Messungen voraus, dass jede Itemvariable über die Messungen hinweg denselben Anteil an wahrer Varianz *und* denselben Anteil an Fehlervarianz aufweist:

$$\begin{aligned} \text{Var}(\tau_{i_1}) &= \text{Var}(\tau_{i_2}) = \text{Var}(\tau_i) \\ \text{Var}(\varepsilon_{i_1}) &= \text{Var}(\varepsilon_{i_2}) = \text{Var}(\varepsilon_i) \end{aligned} \quad (14.26)$$

Die Laufindizes i_1 und i_2 bezeichnen Item i bei Messung 1 und 2. Bei Wiederholungsmessung handelt es sich dabei um dasselbe Item, das wiederholt vorgegeben wird, bei Parallel- oder Halbtets handelt es sich um jeweils korrespondierende Items der zwei verwendeten Testversionen (► Beispiel 14.2). Die empirischen Varianzen der Itemvariablen sind somit bei essentieller τ -Parallelität ebenfalls bei beiden Messungen identisch.

Ein Item hat in beiden Messungen genau dann dieselbe True-Score-Varianz, wenn die Faktorladungen des Items bei beiden Messungen identisch sind ($\lambda_{i_1} = \lambda_{i_2} = \lambda_i$), was sich empirisch überprüfen lässt (► Abschn. 14.4.3):

$$\text{Cov}(y_{i_1}, y_{i_2}) = \text{Cov}(\lambda_i \cdot \eta, \lambda_i \cdot \eta) = \lambda_i^2 \cdot \text{Var}(\eta) = \text{Var}(\tau_i) \quad (14.27)$$

Beispiel 14.2: Kovarianzmatrix bei essentieller τ -Parallelität der Testwertvariablen

Das folgende Beispiel zeigt, wie die Itemkovarianzmatrix bei essentieller τ -Parallelität der Testwertvariablen y_1 und y_2 über die Messungen hinweg aussehen würde. Die Test-Test-Korrelation dürfte in diesem Beispiel als Reliabilität interpretiert werden.

Retest- bzw. Paralleltest-Reliabilität

Als Beispiel liegt ein Test mit sechs Items vor. Im Folgenden zu sehen sind die Kovarianzmatrizen der sechs essentiell τ -äquivalenten Itemvariablen y_1 bis y_6 zur Messung eines eindimensionalen Konstruktts bei strikter Messinvarianz der Items (bzw. essentieller τ -Parallelität der Testwertvariablen) über zwei Messungen hinweg (Wiederholungsmessung mit demselben Test oder Testung mittels Paralleltest).

	Messung 1							Messung 2					
	y_{11}	y_{21}	y_{31}	y_{41}	y_{51}	y_{61}		y_{12}	y_{22}	y_{32}	y_{42}	y_{52}	y_{62}
y_{11}	1.5	0.8	0.8	0.8	0.8	0.8	y_{12}	1.5	0.8	0.8	0.8	0.8	0.8
y_{21}	0.8	2.0	0.8	0.8	0.8	0.8	y_{22}	0.8	2.0	0.8	0.8	0.8	0.8
y_{31}	0.8	0.8	2.5	0.8	0.8	0.8	y_{32}	0.8	0.8	2.5	0.8	0.8	0.8
y_{41}	0.8	0.8	0.8	1.5	0.8	0.8	y_{42}	0.8	0.8	0.8	1.5	0.8	0.8
y_{51}	0.8	0.8	0.8	0.8	2.0	0.8	y_{52}	0.8	0.8	0.8	0.8	2.0	0.8
y_{61}	0.8	0.8	0.8	0.8	0.8	2.5	y_{62}	0.8	0.8	0.8	0.8	0.8	2.5
$Rel = .80$													

Die Kovarianzmatrix der sechs Items der ersten Messung deutet darauf hin, dass die Items essentiell τ -äquivalent sind, was sich an identischen Kovarianzen, aber unglei-

Strikte Messinvarianz der Itemvariablen über die Messungen hinweg

chen Varianzen zeigt. Wird dieser Test an derselben Stichprobe nun ein zweites Mal durchgeführt oder derselben Stichprobe eine parallele Testversion vorgelegt, lässt sich auch für diese zweite Messung die Kovarianzmatrix der sechs Items ermitteln.

Wie man sieht, sind die Kovarianzmatrizen der Itemvariablen y_1 bis y_6 bei beiden Messungen identisch. Dies ist dann der Fall, wenn die Items über die Messungen, d. h. über die Messzeitpunkte bzw. über die Paralleltests hinweg, strikte Messinvarianz aufweisen. Die Voraussetzung der essentiellen τ -Parallelität der Testwertvariablen ist in dem Fall gegeben und die Test-Test-Korrelation kann zur Reliabilitätsschätzung genutzt werden. Die Test-Test-Reliabilität beträgt in diesem Beispiel $Rel = .80$.

Split-Half-Reliabilität

Anhand des Beispiels lässt sich auch zeigen, welche Itemkovarianzmatrix der Split-Half-Reliabilität zugrunde liegt, wenn die Voraussetzung der essentiellen τ -Parallelität der Halbtestwertvariablen bzw. die strikte Messinvarianz korrespondierender Items über die Halbtests hinweg gegeben ist. Aus dem Test mit sechs Items der ersten Messung werden zwei Halbtests a und a' mit jeweils drei Items gebildet. Die Items innerhalb des jeweiligen Halbtests sind wiederum essentiell τ -äquivalent, was sich an identischen Kovarianzen, aber ungleichen Varianzen innerhalb der Halbtests zeigt. Dargestellt sind im Folgenden die Kovarianzmatrizen der zwei essentiell τ -parallelen Halbtests a und a' mit jeweils drei essentiell τ -äquivalenten Itemvariablen y_1 bis y_3 bzw. y_4 bis y_6 zur Messung eines eindimensionalen Konstruktts.

	Halbtest a				Halbtest a'		
	y_1	y_2	y_3		y_4	y_5	y_6
y_1	1.5	0.8	0.8		1.5	0.8	0.8
y_2	0.8	2.0	0.8		0.8	2.0	0.8
y_3	0.8	0.8	2.5		0.8	0.8	2.5

Halbtest-Reliabilität: $Split\text{-}Half\text{-}Rel = .667$;

Gesamttest-Reliabilität: $Rel = (2 \cdot .667) / (1 + .667) = 1.334 / 1.667 = .80$

Wie man sieht, sind die Kovarianzmatrizen der Itemvariablen y_1 bis y_3 und y_4 bis y_6 beider Halbtests identisch. Dies ist dann der Fall, wenn korrespondierende Items über die Halbtests hinweg strikte Messinvarianz aufweisen. Die Voraussetzung der essentiellen τ -Parallelität der Halbtestwertvariablen ist in dem Fall gegeben und die Korrelation der Halbtests kann als Grundlage zur Reliabilitätsschätzung genutzt werden. Die Reliabilität des Halbtests beträgt in diesem Beispiel $Split\text{-}Half\text{-}Rel = .667$ und kann über die Spearman-Brown-Formel (Gl. 14.25) zur Reliabilität des Gesamttests aufgewertet werden. Die Reliabilität des Gesamttests beträgt in diesem Beispiel $Rel = .80$, wie auch anhand der Test-Test-Korrelation des Gesamttests ermittelt werden konnte.

14.4.3 Überprüfung der Voraussetzungen der Test-Test-Korrelation

Die Voraussetzungen zur Schätzung der Reliabilität einer Skala anhand der Test-Test-Korrelation können mit der CFA empirisch überprüft werden (vgl. ► Kap. 24). Dazu wird ein Modell mit zwei korrelierten Faktoren spezifiziert, in dem die Itemvariablen einer Messgelegenheit (Messzeitpunkt, Testversion bzw. Testhälfte) je-

14.4 · Test-Test-Korrelation

weils auf einem eigenen Faktor laden. Weiter wird die Korrelation der Faktoren auf eins fixiert, um zu überprüfen, ob die Faktoren über die Messgelegenheiten hinweg einen gemeinsamen Faktor bilden, der inhaltlich dem zu messenden latenten Merkmal entspricht. Als weitere Voraussetzung der Eindimensionalität dürfen keine Fehlerkovarianzen zwischen den Itemvariablen vorliegen.

Zur Überprüfung der strikten Messinvarianz werden die Faktorladungen und die Fehlervarianzen der jeweils korrespondierenden Items über beide Faktoren hinweg gleichgesetzt. Zur Testung der essentiellen τ -Parallelität der Messungen werden zusätzlich noch die Varianzen der latenten Variablen gleichgesetzt, indem diese z. B. auf eins fixiert werden.

Passt das Modell zu den Daten, ist also der Modellfit hinreichend gut, was anhand des χ^2 -Tests beurteilt werden kann, darf die Test-Test-Korrelation als Reliabilität interpretiert werden. Aufgrund der strengen Voraussetzungen wird ein guter Modellfit für empirische Daten jedoch häufig nicht erreicht (► Abschn. 14.8).

14.4.4 Probleme der Test-Test-Korrelation

Die Schätzung der Reliabilität anhand der Korrelation der Testwertvariablen ist aufgrund der sehr restriktiven Voraussetzungen (► Abschn. 14.4.2) in vielen Fällen nicht anwendbar. Zusätzlich bestehen weitere praktische Probleme, die die Verwendung der Test-Test-Korrelation als Reliabilität erschweren.

14.4.4.1 Probleme der Retest-Reliabilität

■■ Stabilität des latenten Merkmals

Ein wesentliches Problem der Retest-Reliabilität besteht darin, dass sich das zu messende latente Merkmal über die Zeit nicht verändern darf, da nur dann die Kovarianz zwischen den Messungen als wahre Varianz interpretiert werden kann. Stabilität ist bei Traits (z. B. Extraversion) eher zu erwarten als bei situationsabhängigen Merkmalen (z. B. Heiterkeit, Stimmung). Hinsichtlich der Merkmalsveränderung im Zeitverlauf können zwei Fälle unterschieden werden:

Systematische Veränderungen der gemessenen Werte, die bei allen getesteten Personen gleich ausfallen, sind grundsätzlich kein Problem bei der Bestimmung der Retest-Reliabilität (Beispiel: Alle Testpersonen lernen zwischen dem ersten und zweiten Zeitpunkt genau gleich viel dazu). Durch das Addieren oder Subtrahieren eines konstanten Betrags bei allen Messwerten würden sich die wahren Varianzen und damit auch die Korrelation zwischen dem ersten und zweiten Messzeitpunkt nicht ändern.

Eine *unsystematische* Veränderung zwischen den Messgelegenheiten hat dagegen einen verfälschenden Einfluss auf die Test-Test-Korrelation. Unsystematische Veränderungen können beispielsweise bei instabilen Merkmalen durch interindividuell unterschiedliche Entwicklungsverläufe oder auch durch situationsspezifische Effekte, auf die die Testpersonen unterschiedlich reagieren, zustande kommen (vgl. Steyer 1987; Steyer et al. 1999; s. auch ► Kap. 26).

Höhere Stabilität bei Traits

Systematische Veränderungen

Unsystematische Veränderungen

■■ Erinnerungseinflüsse

Bedingt durch Erinnerungseinflüsse kann allerdings auch nur eine Scheinstabilität vorliegen: Bei einer Testwiederholung können sich Personen unter Umständen daran erinnern, was sie bei der ersten Testung geantwortet haben. Geben sie *nur deswegen* wieder genau die gleichen Antworten auf die Items, wäre die Retest-Reliabilität künstlich überhöht, da diese Tendenz, stabil erscheinende Antworten zu geben, empirisch nicht von dem eigentlich interessierenden Merkmal zu unterscheiden wäre.

Erinnerungseffekte

■■ Wahl des optimalen Retest-Intervalls

Die Länge des Zeitintervalls zwischen den Testungen spielt sowohl für Erinnerungseffekte als auch für unsystematische Veränderungen des Merkmals eine entscheidende Rolle. Eine allgemeingültige Regel für optimale Retest-Intervalle kann es nicht geben, da das relative Risiko für Merkmalsveränderungen und Erinnerungseffekte vom jeweiligen Testinhalt bzw. Merkmal abhängt. Das Intervall zwischen zwei Messungen sollte einerseits lang genug sein, um Erinnerungseffekte zu reduzieren, andererseits sollte es aber auch kurz genug sein, um deutliche Merkmalsveränderungen zu verhindern (vgl. Raykov und Marcoulides 2011). Um Retest-Reliabilitätskoeffizienten genauer beurteilen zu können, sollten die Retest-Intervalle (z. B. in Testmanualen) stets angegeben werden.

14.4.4.2 Probleme der Paralleltest- und Split-Half-Reliabilität

Parallele Formen eines Tests lassen sich erstellen, indem die Items den Testversionen aufgrund ähnlicher Inhalte und vergleichbarer statistischer Kennwerte (Diskriminationsparameter bzw. Faktorladung, Itemvarianz und Leichtigkeitsparameter bzw. Interzept, vgl. ► Kap. 13) zugeordnet werden. Anhand einer CFA kann überprüft werden, ob die beiden Testversionen tatsächlich als parallele Testversionen betrachtet werden können.

■■ Erstellung von Parallelformen eines Tests

In der Praxis existieren nur für relativ wenige Testverfahren geprüfte Parallelformen. Die Konstruktion paralleler Tests stellt ein wesentliches Problem dar. Zur Konstruktion von Parallelformen eines Persönlichkeitsfragebogens wäre ein großer Itempool nötig, der so viele gleichermaßen gut geeignete Items enthalten müsste, dass sich daraus gleich zwei zueinander parallele Formen konstruieren ließen. Bereits kleine Unterschiede der Itemformulierungen können jedoch dazu führen, dass die Faktorladungen der korrespondierenden Items in den beiden Testversionen unterschiedlich sind und sich somit die Bedeutung der jeweils gemessenen latenten Variablen zwischen den Testversionen unterscheidet. Parallelformen für Leistungstests sind etwas leichter zu realisieren, wenn diese aus vielen gleichartigen Items konstruiert sind. Hier können oft schon geringfügige Abwandlungen der verwendeten Aufgaben nutzbare Parallelformen erzeugen (z. B. strukturell vergleichbare Rechenaufgaben mit unterschiedlichen Zahlen).

Das gleiche Problem besteht, wenn parallele Testhälften aus den Items eines Tests gebildet werden sollen. Eine Aufteilung der Items per Zufall oder anhand von statistischen Kriterien führt meist nicht zu essentiell τ -parallelen Halbtests. In der Praxis gelingt die Konstruktion auch nur annähernd essentiell τ -paralleler Testhälften oder Testversionen daher häufig nicht. Da die strengen Annahmen der Paralleltest-Reliabilität überwiegend nicht erfüllt sind, sollte die Korrelation zwischen den Testwertvariablen von zwei Testversionen nicht ohne Überprüfung der Voraussetzungen als Reliabilität interpretiert werden (► Abschn. 14.8).

■■ Stabilität des latenten Merkmals und Übertragungseffekte

Bei längeren Intervallen zwischen der Bearbeitung von Parallelformen eines Tests können wie bei der Retest-Reliabilität Veränderungen der Merkmalsausprägungen auftreten, sodass die Korrelation zwischen den Testwertvariablen der beiden Testversionen nicht mehr als Reliabilität interpretiert werden kann. Bei einer unmittelbar aufeinanderfolgenden Bearbeitung der Tests können durch situative Effekte systematische Übertragungseffekte von einer auf die andere Parallelform auftreten. Um dies auszuschließen, können die Formen ausbalanciert präsentiert werden, indem jeweils einer Hälfte der Testpersonen zuerst Testform A gefolgt von Testform B und der anderen Hälfte zuerst Testform B gefolgt von Testform A gegeben wird.

Parallelformen für Leistungstests leichter realisierbar als für Persönlichkeitstests

Überprüfung der Voraussetzungen nötig

14.5 Vergleichbarkeit der Reliabilitätsmaße

Sind sowohl die essentielle τ -Äquivalenz der Itemvariablen als auch gleichzeitig die essentielle τ -Parallelität der Testwertvariablen zweier Messgelegenheiten gegeben, so sind die verschiedenen Reliabilitätsmethoden ineinander überführbar und für alle Methoden resultiert derselbe Reliabilitätswert der Testwertvariablen Y :

$$\begin{aligned} Rel(Y) &= \alpha = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^p Var(y_i)}{Var(Y)} \right) = \frac{p^2 \cdot Var(\tau)}{Var(Y)} \\ &= Corr(Y_1, Y_2) = Corr(Y_A, Y_B) = \frac{2 \cdot Rel(Y_a)}{1 + Rel(Y_a)} = \frac{Var(T)}{Var(Y)} \quad (14.28) \end{aligned}$$

Hierbei bezeichnet p die Anzahl der Items, aus denen der Test besteht; $p^2 \cdot Var(\tau)$ ist die gesamte True-Score-Varianz der Testwertvariablen $Var(T)$, die sich aus der für alle Items identischen True-Score-Varianz der Items $Var(\tau)$ ergibt. Y_1 und Y_2 bezeichnen die Testwertvariablen der wiederholten Messungen zu zwei Messzeitpunkten, Y_A und Y_B die Testwertvariablen der parallelen Tests A und B, und Y_a die Testwertvariable des Halbtests. Mithilfe der Spearman-Brown-Formel ($2 \cdot Rel / (1 + Rel)$) erfolgt die Aufwertung der Split-Half-Reliabilität zur Reliabilität des Gesamttests (vgl. Gl. 14.25).

Als Zahlenbeispiel soll hier die Kovarianzmatrix der sechs essentiell τ -äquivalenten Variablen zur Messung eines eindimensionalen Konstrukt aus Beispiel A in ► Exkurs 14.1 verwendet werden, die auch in ► Beispiel 14.2 genutzt wird. In diesem Beispiel beträgt die durchschnittliche Varianz der Itemvariablen 2.0, die True-Score-Varianz der Itemvariablen und damit auch alle Kovarianzen jeweils 0.8. Somit beträgt die gesamte Varianz der Testwertvariablen $Var(Y) = 36$, die sich aus der True-Score-Varianz $Var(T) = 28.8$ und der Fehlervarianz $Var(E) = 7.2$ zusammensetzt. Wie gezeigt werden konnte, resultiert daraus $\alpha = .80$.

Wird die Testung dagegen zweimal durchgeführt, z. B. mittels *Retestung* oder anhand eines *Paralleltests*, und weist die zweite Messung dieselbe Kovarianzmatrix auf wie die erste Messung (essentielle τ -Parallelität), wie das ► Beispiel 14.2 zeigt, kann die Korrelation zwischen den Testwertvariablen aus beiden Messungen als Reliabilität interpretiert werden; sie beträgt ebenfalls .80.

Die *Split-Half-Korrelation* zwischen den Halbtestwerten der Items 1 bis 3 und den Halbtestwerten der Items 4 bis 6 bezieht sich nur auf die Testhälften des Tests (► Beispiel 14.2). Die Korrelation der Testhälften beträgt .667. Die Aufwertung über die Spearman-Brown-Formel ergibt $2 \cdot .667 / (1 + .667) = .80$ und somit ebenfalls dasselbe Ergebnis.

Wie dieses Beispiel zeigt, gibt es nur eine einzige Reliabilität der Testwertvariablen (bezogen auf eine bestimmte Population) und nicht etwa verschiedene, von der verwendeten Methode abhängige Reliabilitäten. Die verschiedenen klassischen Reliabilitätsmaße stellen also trotz der unterschiedlichen Bezeichnungen immer dasselbe Gütemaß dar. Die Methoden der Reliabilitätsschätzung unterscheiden sich nur hinsichtlich des zugrunde liegenden Messmodells (z. B. essentielle τ -Äquivalenz der Itemvariablen oder essentielle τ -Parallelität der Testwertvariablen).

Verschiedene Bezeichnungen für dasselbe Gütemaß

- ! Sind die geforderten Voraussetzungen für jedes klassische Reliabilitätsmaß erfüllt, führen alle Methoden zu demselben Ergebnis. In der Praxis sollte immer eine Methode zur Reliabilitätsschätzung gewählt werden, deren Voraussetzungen erfüllt sind (► Abschn. 14.8).

14.6 Einflüsse auf die Reliabilität

Merkmalsvariabilität in verschiedenen Populationen

Reliabilitätskoeffizienten sind abhängig von der Variabilität eines Merkmals innerhalb einer Population. Wird z. B. die Reliabilität eines Tests zur Messung der Depressivität in einer Normalstichprobe bestimmt, so kann sich ein anderer Wert ergeben als bei Bestimmung der Reliabilität in einer klinischen Stichprobe. Einer der Gründe dafür ist die höhere bzw. niedrigere Variabilität (Heterogenität bzw. Homogenität) des untersuchten Merkmals in den verschiedenen Subpopulationen. Da die Merkmalsvarianz und somit auch die Varianz der True-Score-Variablen in homogenen Stichproben stärker eingeschränkt sind als in heterogenen Stichproben, wird die geschätzte Reliabilität in der klinischen Stichprobe geringer ausfallen als in der Normalstichprobe. Für eine homogene Stichprobe von Personen mit stark ausgeprägter Depressivität könnte die geschätzte Reliabilität daher sogar gegen null tendieren, obwohl die Messung sehr zuverlässig ist und sich der Messfehler nicht von den Messungen in anderen Stichproben unterscheidet.

! Da ein Test oftmals in verschiedenen Subpopulationen eingesetzt werden soll, muss die Reliabilität für jede Subpopulation bestimmt werden (vgl. Diagnostik- und Testkuratorium 2018).

Homogene vs. heterogene Items

Ist ein Konstrukt inhaltlich eher klar umschrieben und eng definiert, z. B. das Konstrukt Besorgnis, lässt es sich durch homogene Items messen. Im Gegensatz dazu wird ein weiter gefasstes Konstrukt, z. B. das Konstrukt Angst, durch heterogene Items gemessen. Bei gleicher Itemanzahl wäre zu erwarten, dass die Messung eines Konstruktcs anhand homogener Items im Vergleich zur Messung anhand heterogener Items reliabler ist. Das zeigt natürlich ein Problem bei der Itemgenerierung auf: Je ähnlicher die Items formuliert werden, desto höher ist die Reliabilität des Tests. Im Extremfall kann dies aber dazu führen, dass der Inhalt nur eines Items durch geringfügige Abwandlungen praktisch vervielfältigt wird, sodass zwar die Reliabilität gegen eins geht, die Validität jedoch sehr gering ist, weil nur noch ein sehr eng umschriebener Teilaspekt des Konstruktcs gemessen wird.

Auch die Anzahl der Items im Test wirkt sich auf die Höhe der Reliabilität aus. Je mehr Items ein Test umfasst, desto höher ist dessen Reliabilität. Diese Beziehung wird auch durch die Spearman-Brown-Formel der Testverlängerung (Gl. 14.25) für essentiell τ -parallele Tests deutlich (Brown 1910; Spearman 1910). Allerdings hat die Testverlängerung Grenzen, da es sich um eine Sättigungsfunktion handelt (vgl. Lord und Novick 1968, S. 138; Steyer und Eid 2001, S. 130). Der Anstieg der Reliabilität ist bei einer geringen Anzahl an Items hoch, während der Anstieg bei einer großen Anzahl von Items vernachlässigbar ist.

14.7 Anzustrebende Höhe der Reliabilität

Leistungs- vs. Persönlichkeitstests

Die anzustrebende Höhe der Reliabilität kann nicht auf einen einzigen Wert reduziert werden, da die Reliabilität von der Homogenität bzw. Heterogenität des Konstruktcs und von der untersuchten Population abhängt (► Abschn. 14.6). Eine allgemeingültige Empfehlung kann daher nicht gegeben werden. In der Praxis hängt die Höhe von verschiedenen Bedingungen ab, die hier kurz angesprochen werden sollen.

In der Literatur werden anzustrebende Werte genannt, die jedoch nicht in jedem Anwendungsbereich gleichermaßen erzielt werden können. Die Reliabilität etablierter Leistungstests liegt für globale Intelligenzmaße meist in einem hohen Bereich (um .90), während sie bei Persönlichkeitstests deutlich niedriger liegt (für einzelne Skalen oftmals nur um .70).

14.8 · Auswahl eines geeigneten Reliabilitätsmaßes

Zur Bestimmung der Reliabilität wurden in der Vergangenheit zumeist die klassischen Methoden Cronbachs Alpha, Retest-Reliabilität oder Split-Half-Reliabilität verwendet, deren Voraussetzungen allerdings nur in seltenen Fällen überprüft werden sein dürften. Gerade bei umfangreicheren Tests mit vielen Items ist jedoch bereits die Voraussetzung der Eindimensionalität selten gegeben. Daher wäre hier die Überprüfung der Reliabilität anhand adäquater Reliabilitätsmaße interessant (z. B. modellbasierte Reliabilitätsmaße, ▶ Kap. 15), um entsprechende Empfehlungen hinsichtlich der anzustrebenden Höhe der Reliabilität geben zu können.

Screening-Tests, die zur groben Einschätzung eines Merkmals mit möglichst geringem Aufwand dienen, sind in der Regel so kurz wie möglich gehalten und weisen daher oft nicht dieselbe Reliabilität wie umfangreichere Tests auf. Längere und reliablere Tests wären wissenschaftlich wünschenswert, können in der Praxis jedoch nicht immer eingesetzt werden, da beispielsweise bei betrieblichen Untersuchungen ein längerer Arbeitsausfall zu teuer oder bei klinischen Untersuchungen die Bearbeitung langer Fragebogen für Patienten überfordernd wäre. Durch den Einsatz computerisierter adaptiver Tests (▶ Kap. 20) lässt sich die Interdependenz von Reliabilität und Testlänge verringern.

Die Daumenregeln zur Beurteilung der Reliabilität gehen auf Nunnally (1967, 1978) zurück, der zunächst Werte von .50 oder .60 als hinreichend für die explorative Forschung ansah (Nunnally 1967). Diese Werte wurden später auf .70 erhöht (Nunnally 1978). Allerdings scheint für die Wahl dieser Werte keine rationale Begründung vorzuliegen (vgl. Cho und Kim 2015).

Trotz dieser Probleme haben sich Konventionen durchgesetzt, nach denen

- für Screening-Tests und bei heterogenen Konstrukten eine Reliabilität von mindestens .70 als angemessen betrachtet wird,
- bei homogenen Konstrukten eine Reliabilität zwischen .80 und .90 in den meisten Fällen als hinreichend hoch eingeschätzt wird und
- bei Leistungstests Werte > .90 als sehr gut eingestuft werden.

Diese Daumenregeln sollten jedoch nicht überbewertet werden, da je nach diagnostischer Fragestellung im Einzelfall auch andere Werte sinnvoll sein können.

Um eine Aussage über die Präzision der Reliabilitätsschätzung treffen zu können, ist die *Schätzung entsprechender Konfidenzintervalle* notwendig. Da jedoch die Standardfehler nur anhand der CFA adäquat geschätzt werden können, stehen ohne diese keine präzisen Konfidenzintervalle zur Verfügung. Mit größerem Stichprobenumfang wird die Schätzung zwar genauer, bei Verletzung der Voraussetzungen konvergiert die Punktschätzung jedoch nicht auf den wahren Reliabilitätswert (Raykov und Marcoulides 2011). Die Bestimmung des Konfidenzintervalls unter Verwendung des Standardmessfehlers ermöglicht aber zumindest eine grobe Schätzung der Präzision der Reliabilitätsschätzung (▶ Kap. 13).

Screening-Tests

Daumenregeln

Notwendigkeit von Konfidenzintervallen

Verwendung des Standardmessfehlers

14.8 Auswahl eines geeigneten Reliabilitätsmaßes

Die klassischen Methoden der Reliabilitätsschätzung unterscheiden sich hinsichtlich des jeweils zugrunde liegenden Messmodells (z. B. essentielle τ -Äquivalenz der Itemvariablen oder essentielle τ -Parallelität der Testwertvariablen). Das Zutreffen der vorausgesetzten Modelleigenschaften wird in der Praxis oftmals nicht überprüft, obwohl diese Prüfung zwingend erforderlich wäre, da die Koeffizienten nur dann als Maße der Reliabilität interpretiert werden dürfen, wenn die jeweiligen Voraussetzungen erfüllt sind. Andernfalls liefern die klassischen Reliabilitätsmaße nur unpräzise Punktschätzungen der Reliabilität. Daher sollten die Voraussetzungen immer zunächst geprüft werden, um ein für die gegebenen Daten geeignetes Reliabilitätsmaß zu verwenden.

Prüfung der Voraussetzungen notwendig

Tabelle 14.1 Stufen der Messäquivalenz sowie geeignete klassische und modellbasierte Reliabilitätsmaße für eindimensionale Modelle.
Voraussetzung: Das Ergebnis der CFA spricht für Eindimensionalität (keine Fehlervarianzen)

	τ -Kongenerität der Itemvariablen	Essentielle τ -Äquivalenz der Itemvariablen	Essentielle τ -Parallelität der Testwertvariablen
Modellrestriktionen	–	Diskriminationsparameter (Faktorladungen) identisch: $\lambda_i = \lambda_{i'} = \lambda$	– über Messzeitpunkte/Testhälften korrespondierende Diskriminationsparameter (Faktorladungen) identisch: $\lambda_{i_1} = \lambda_{i_2} = \lambda_i$ – und über Messzeitpunkte/Testhälften korrespondierende Fehlervarianzen identisch: $Var(\varepsilon_{i_1}) = Var(\varepsilon_{i_2}) = Var(\varepsilon_i)$
Reliabilitätskoeffizient	<i>McDonalds Omega</i>	<i>Cronbachs Alpha</i>	<i>Test-Test-Korrelation/Spearman-Brown-Formel</i>

Anmerkung: Der Index i bezeichnet die Items ($i = 1, \dots, p; i \neq i'$).

Modellbasierte Reliabilitätskoeffizienten

Da die Voraussetzungen der klassischen Reliabilitätsmaße in der Praxis oftmals verletzt sind, stehen alternative Methoden zur Verfügung, die auf weniger strengen Voraussetzungen beruhen. Hierzu gehören die modellbasierten Reliabilitätskoeffizienten (vgl. ► Kap. 15), die im Gegensatz zu den klassischen Methoden auf den geschätzten Parametern eines faktorenanalytischen Modells beruhen.

Soll ein Konstrukt z. B. anhand von Items gemessen werden, die *nicht* essentiell τ -äquivalent sind, so sollte Cronbachs Alpha nicht zur Reliabilitätsschätzung verwendet werden. Eine Alternative wäre in diesem Fall McDonalds Omega, da dieses Maß auf weniger strengen Annahmen basiert als Cronbachs Alpha. Revelle und Zinbarg (2009) fanden in einem Vergleich verschiedener Reliabilitätsmaße (u. a. Cronbachs Alpha, Guttmans Lambda-Koeffizienten und McDonalds Omega) eine klare Präferenz für McDonalds Omega. Ein Überblick über die verschiedenen Reliabilitätsmaße sowie deren Voraussetzungen findet sich in ► Tab. 14.1 (vgl. auch ► Kap. 12 und 13).

Anmerkung: Ist die Voraussetzungsprüfung aus praktischen Gründen nicht möglich, da beispielsweise die Stichprobe für die Durchführung einer CFA zu klein ist, wird zur Reliabilitätsschätzung oftmals ungeprüft auf klassische Maße zurückgegriffen. In diesem Fall ist in aller Regel Cronbachs Alpha gegenüber Maßen, die auf Test-Test-Korrelationen basieren, zu bevorzugen, da es die vergleichsweise weniger strengen Annahmen trifft. Weiter konnte gezeigt werden, dass bei mäßiger Verletzung der Annahme der essentiellen τ -Äquivalenz (d. h. bei Eindimensionalität und ähnlichen, hohen Faktorladungen auf dem gemeinsamen Faktor) die Reliabilität anhand von Cronbachs Alpha noch sinnvoll geschätzt werden kann (vgl. Raykov 1997; Raykov und Marcoulides 2015, 2019).

14.9 Zusammenfassung

Die klassischen Reliabilitätsmaße haben die Testkonstruktion und die Psychometrie für lange Zeit stark bestimmt. Sie sind historisch wertvoll, da sie zur Wissenschaftlichkeit der Testkonstruktion maßgeblich beigetragen haben, indem sie eine Bestimmung der Messgenauigkeit psychologischer Tests ermöglicht haben, die für Messinstrumente in anderen Forschungsbereichen, z. B. der Physik, Standard ist. Die Schätzung der Reliabilität basiert bei den klassischen Reliabilitätsmethoden auf den Varianzen und Kovarianzen der Itemvariablen innerhalb eines Tests (Cronbachs Alpha) oder der Korrelation der Testwertvariablen mehrerer Messungen (Retest-, Paralleltest-, Split-Half-Reliabilität).

Allerdings schätzen die klassischen Maße die Reliabilität nur dann adäquat, wenn strenge Voraussetzungen der Eindimensionalität und Messäquivalenz erfüllt

14.10 · EDV-Hinweise

sind. Da diese strengen Voraussetzungen in der Praxis oftmals nicht gegeben sind, sollten vorzugsweise modellbasierte Reliabilitätskoeffizienten verwendet werden (► Kap. 15). Diese beruhen auf weniger strengen Voraussetzungen und stellen daher oftmals eine sinnvolle Alternative zu den klassischen Reliabilitätsmaßen dar. Die heutige Verbreitung und Anwendbarkeit entsprechender Software zur modellbasierten Reliabilitätsschätzung machen dies inzwischen möglich.

14.10 EDV-Hinweise

Alle in diesem Kapitel behandelten Reliabilitätsmaße können mit gängigen EDV-Programmen, z. B. SPSS oder R, analysiert werden. Die Voraussetzungen lassen sich am einfachsten mit den Programmen *Mplus* (Muthén und Muthén 2017) oder dem R-Paket lavaan (Rosseel 2012) testen. Die SPSS-, *Mplus*- und R-Syntax für Anwendungsbeispiele finden sich in den Zusatzmaterialien unter ► <http://www.lehrbuch-psychologie.springer.com>.

14.11 Kontrollfragen

- ?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Auf welchen Voraussetzungen beruht Cronbachs Alpha?
 2. Warum muss essentielle τ -Parallelität der Testwertvariablen gegeben sein, um die Korrelation zwischen den Testwertvariablen als deren Reliabilität zu interpretieren?
 3. Wie lässt sich die Reliabilität eines Tests anhand der Split-Half-Reliabilität bestimmen?
 4. Warum kann in der Regel nicht von „der Reliabilität“ eines Tests gesprochen werden, wenn diese für eine Stichprobe ermittelt wurde?
 5. Welche Aussage lässt sich über die Reliabilität einer Testwertvariablen treffen, wenn die Voraussetzungen der klassischen Reliabilitätsmaße nicht erfüllt sind?

Literatur

- Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York, NY: The Guilford Press.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Cho, E. & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18, 207–230.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- 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.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Frost, R. O., Marten, P., Lahart, C. & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research*, 14, 449–468.
- Grayson, D. A. (2004). Some myths and legends of quantitative psychology. *Understanding Statistics*, 3, 101–134.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153–160.

- Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Novick, M. R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1–13.
- Nunnally, J. C. (1967). Psychometric theory. New York, NY: McGraw-Hill.
- Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, NY: McGraw-Hill.
- Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity. Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist*, 16, 56–69.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88, 879–903.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184.
- Raykov, T. & Marcoulides, G. A. (2011). *Psychometric Theory*. New York: Routledge.
- Raykov, T. & Marcoulides, G. A. (2015). A direct latent variable modeling based for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement*, 75, 146–156.
- Raykov, T., Marcoulides G. A. & Partelis, T. (2015). The importance of the assumption of uncorrelated errors in psychometric theory. *Educational and Psychological Measurement*, 75, 634–647.
- Raykov, T. & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, 79, 200–210.
- Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 145–154.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48, 1–36.
- Schermelleh-Engel, K. (1995). *Fragebogen zur Schmerzregulation (FSR)*. Frankfurt am Main: Swets Test Services.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350–353.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 171–195.
- Steyer, R. (1987). Konsistenz und Spezifität: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 8, 245–258.
- Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin, Heidelberg: Springer Verlag.
- Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389–408.
- Stöber, J. (1995). Frost Multidimensional Perfectionism Scale-Deutsch (FMPS-D). Unveröff. Manuskript. Freie Universität Berlin, Institut für Psychologie.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80, 99–103.
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, 40, 395–412.
- Zimmerman, D. W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement*, 36, 85–96.



Modellbasierte Methoden der Reliabilitätsschätzung

Karin Schermelleh-Engel und Jana C. Gäde

Inhaltsverzeichnis

- 15.1 Klassische vs. modellbasierte Reliabilitätsschätzung – 337**
 - 15.1.1 Probleme der klassischen Reliabilitätsschätzung – 337
 - 15.1.2 Modellbasierte Reliabilitätsschätzung als Alternative – 338
- 15.2 Eindimensionale Modelle – 339**
 - 15.2.1 Koeffizienten zur Schätzung der Reliabilität eindimensionaler Tests – 340
 - 15.2.1.1 Cronbachs Alpha – 342
 - 15.2.1.2 McDonalds Omega – 342
 - 15.2.1.3 Bollens Omega – 343
 - 15.2.2 Empirisches Beispiel – 344
 - 15.2.2.1 Reliabilitätsschätzung mit Cronbachs Alpha – 345
 - 15.2.2.2 Reliabilitätsschätzung mit McDonalds Omega – 346
 - 15.2.2.3 Reliabilitätsschätzung mit Bollens Omega – 347
 - 15.2.3 Asymmetrisches Konfidenzintervall – 348
- 15.3 Mehrdimensionale Modelle – 350**
 - 15.3.1 Bifaktormodell – 350
 - 15.3.2 Koeffizienten zur Schätzung der Reliabilität mehrdimensionaler Tests – 351
 - 15.3.2.1 Omega-Koeffizienten des Gesamttests – 351
 - 15.3.2.2 Omega-Koeffizienten der Subskalen – 353
 - 15.3.3 Empirisches Beispiel – 355
 - 15.3.3.1 Omega-Koeffizienten des Gesamttests – 355
 - 15.3.3.2 Omega-Koeffizienten der Subskalen – 357
- 15.4 Omega-Koeffizienten im Rahmen weiterer Faktormodelle – 360**
- 15.5 Bewertung der modellbasierten Reliabilitätsschätzung – 361**
 - 15.5.1 Vorteile gegenüber der klassischen Reliabilitätsschätzung – 361
 - 15.5.2 Probleme der modellbasierten Reliabilitätsschätzung – 362

15.6	Reliabilitätsschätzung ordinalskalierter Variablen – 363
15.6.1	Variablen mit fünf oder mehr Antwortkategorien – 363
15.6.2	Variablen mit zwei bis vier Antwortkategorien – 363
15.6.3	Variablen mit zwei Antwortkategorien – 363
15.6.4	Item-Parcels – 364
15.6.5	IRT-Modelle – 364
15.7	Erste Empfehlungen zur Beurteilung der Omega-Koeffizienten – 364
15.8	Zusammenfassung – 365
15.9	EDV-Hinweise – 366
15.10	Kontrollfragen – 366
	Literatur – 366

i Modellbasierte Methoden der Reliabilitätsschätzung verwenden die konfirmatorische Faktorenanalyse (CFA) zur Schätzung der Reliabilitätskoeffizienten. Sie beruhen im Vergleich zu den klassischen Methoden der Reliabilitätsschätzung auf weniger strengen Annahmen. „Modellbasiert“ bedeutet zum einen, dass die Modelle und Annahmen der Reliabilitätskoeffizienten explizit anhand von Modellttests überprüft werden, und zum anderen, dass die Reliabilitätskoeffizienten im Rahmen der CFA anhand der Modellparameter geschätzt werden. Nachfolgend sollen zur Reliabilitätsschätzung verschiedene modellbasierte Reliabilitätskoeffizienten vorgestellt werden. Für eindimensionale Tests werden Cronbachs Alpha, McDonalds Omega und Bollens Omega erläutert und für mehrdimensionale Tests verschiedene Omega-Koeffizienten, die sich entweder auf den gesamten mehrdimensionalen Test oder auf einzelne Subskalen des Tests beziehen.

15.1 Klassische vs. modellbasierte Reliabilitätsschätzung

Sowohl die klassischen als auch die modellbasierten Methoden der Reliabilitätsschätzung beruhen bei eindimensionalen Modellen auf der additiven Zerlegung der gesamten Testwertvarianz in zwei Varianzanteile: die wahre Varianz und die Fehlervarianz. Die Reliabilität wird als das Verhältnis der wahren Varianz zur Gesamtvarianz bestimmt. Damit ein empirisches Varianzverhältnis als Reliabilität interpretiert werden darf, muss die Gültigkeit bestimmter Modellrestriktionen gegeben sein. Die klassischen und modellbasierten Methoden unterscheiden sich u. a. hinsichtlich der Art und Anzahl der geforderten Modellrestriktionen.

Additive Varianzzerlegung

15.1.1 Probleme der klassischen Reliabilitätsschätzung

Die klassischen Methoden der Reliabilitätsschätzung (vgl. ► Kap. 13 und 14) entstanden, als psychometrische Modelle auf Basis latenter Variablen noch nicht entwickelt worden waren. Die klassischen Methoden setzen implizit die Gültigkeit bestimmter Modellrestriktionen voraus, die sich zum einen ganz grundlegend auf die Dimensionalität der Messungen (Items) sowie zum anderen zusätzlich auf die Stufe der Messäquivalenz der Messungen beziehen (vgl. ► Kap. 13; Eid und Schmidt 2014; Steyer und Eid 2001). Das Zutreffen der implizit vorausgesetzten Modellrestriktionen wie Eindimensionalität, essentielle τ -Äquivalenz oder essentielle τ -Parallelität der Messungen werden in der Praxis meist nicht explizit überprüft, obwohl diese Prüfung zwingend erforderlich wäre (vgl. ► Kap. 14).

Voraussetzungen oft nicht überprüft

Die Reliabilitätsschätzung basiert bei klassischen Methoden im Vergleich zu den modellbasierten Methoden nicht auf den geschätzten Parametern eines faktorenanalytischen Modells, sondern entweder auf den empirischen Varianzen und Kovarianzen der Itemvariablen (Cronbachs Alpha) oder auf den empirischen Korrelationen zwischen den Test(summen)werten paralleler Tests (z. B. Retest-Reliabilität und Paralleltest-Reliabilität).

Klassische Methoden

Die klassischen Reliabilitätsmaße sind mit dem Problem verbunden, dass ihre strengen Voraussetzungen in der Praxis nur selten erfüllt werden können. Dies führt zur Verwendung von Reliabilitätsmaßen, die für viele Fragestellungen eigentlich nicht angemessen sind. Als Folge resultieren bei Voraussetzungsverletzungen unpräzise und verzerrte Punktschätzungen der Reliabilität. Da bei Voraussetzungsverletzungen zudem die Standardfehler nicht adäquat geschätzt werden, sind auch die daraus gebildeten Konfidenzintervalle nicht korrekt.

Voraussetzungen nur selten erfüllt

Ein Vorteil der klassischen Maße liegt darin, dass sie recht einfach zu berechnen sind und bei Erfüllung der Voraussetzungen präzise Schätzungen liefern. Sind die Voraussetzungen nicht zu stark verletzt, können die klassischen Reliabilitätsmaße auch weiterhin als grobe Schätzungen verwendet werden (► Kap. 14).

15.1.2 Modellbasierte Reliabilitätsschätzung als Alternative

Modelle der CFA

Als Alternative bieten sich die Methoden der modellbasierten Reliabilitätsschätzung an, die meist auf weniger strengen Voraussetzungen basieren. Diese Reliabilitätsmaße beruhen auf Modellen der konfirmatorischen Faktorenanalyse (CFA), deren Passung zu den Daten anhand von Modelltests als passend oder auch als nicht passend beurteilt werden kann, was zur Beibehaltung oder zur Ablehnung des Modells führen würde. „Modellbasiert“ bedeutet, dass einerseits verschiedene Messmodelle mit unterschiedlich strengen Annahmen, die den Reliabilitätsmaßen ein- oder mehrdimensionaler Tests zugrunde liegen, auf ihre Passung überprüft werden können und dass andererseits die Schätzung der Reliabilitätskoeffizienten auf Grundlage der geschätzten Modellparameter erfolgt.

Modellbasiert können bei eindimensionalen Tests Cronbachs Alpha, McDonalds Omega und Bollens Omega geschätzt werden. Beispielsweise ist *McDonalds Omega* wie Cronbachs Alpha ein Koeffizient zur Bestimmung der Reliabilität der Testwertvariablen eines eindimensionalen Tests, der allerdings als Stufe der Messäquivalenz lediglich τ -Kongeneritität der Itemvariablen voraussetzt. Diese Voraussetzung impliziert, dass alle Items eines Tests dasselbe Konstrukt messen, sich die Anteile der wahren Varianz der einzelnen Items aber ebenso unterscheiden dürfen wie die Fehlervarianzen. *Cronbachs Alpha* dagegen basiert auf der strengeren Annahme der essentiellen τ -Äquivalenz, bei der die Faktorladungen alle denselben Wert aufweisen, mit der Konsequenz, dass alle wahren Varianzen und damit auch Kovarianzen der Itemvariablen gleich groß sein müssen.

Reliabilitätskoeffizienten eindimensionaler Tests

Reliabilitätskoeffizienten mehrdimensionaler Tests

Die modellbasierten Reliabilitätskoeffizienten sind nicht wie die klassischen Maße auf eindimensionale Konstrukte beschränkt. Wird ein mehrdimensionales Konstrukt anhand eines Tests mit mehreren Subskalen erfasst, so kann z. B. anhand eines Bifaktormodells (► Kap. 24) die wahre Varianz der Testwertvariablen aufgeteilt werden in einen allgemeinen Varianzanteil, der sich auf das übergeordnete Konstrukt (den Generalfaktor) bezieht, und in einen subskalenspezifischen Anteil, der diejenigen Anteile der wahren Varianz aller Subskalen enthält, die unabhängig vom Generalfaktor sind (spezifische Faktoren). Werden diese wahren Varianzanteile durch die Gesamtvarianz der Testwerte geteilt, resultieren die Omega-Koeffizienten Omega-total (ω_T), Omega-hierarchisch (ω_H) und Omega-spezifisch (ω_S). Für jede einzelne Subskala kann ebenfalls bestimmt werden, welcher Anteil der wahren Varianz auf den Generalfaktor zurückgeführt werden kann und welcher Anteil spezifisch für die jeweilige Subskala ist. Werden diese wahren Varianzanteile durch die Gesamtvarianz der jeweiligen Subskala geteilt, so resultieren die skalenspezifischen Koeffizienten Omega-Subskala-total ($\omega_{\text{Skala-T}}$), Omega-Subskala-hierarchisch ($\omega_{\text{Skala-H}}$) und Omega-Subskala-spezifisch ($\omega_{\text{Skala-S}}$). Im Gegensatz zu den klassischen Maßen kann mit modellbasierten Maßen auch der korrelierte Messfehler bei der Reliabilitätsschätzung berücksichtigt werden (*Bollens Omega*). Die verschiedenen modellbasierten Reliabilitätskoeffizienten sind in □ Tab. 15.1 aufgelistet.

! Punktschätzungen der Reliabilitätskoeffizienten sind mit Unsicherheit behaftet und treffen den Populationswert nicht exakt, sondern nur erwartungstreu. Die Punktschätzung sollte deshalb immer durch ein Konfidenzintervall ergänzt werden. Ein Konfidenzintervall gibt eine untere und eine obere Grenze für einen Wertebereich an, in dem der Populationswert mit einer bestimmten Sicherheit, z. B. 95 %, zu liegen kommt. Da Reliabilitätskoeffizienten auf den Wertebereich zwischen null und eins begrenzt sind, muss ein asymmetrisches Konfidenzintervall verwendet werden.

Tabelle 15.1 Systematik der modellbasierten Reliabilitätskoeffizienten für ein- und mehrdimensionale Tests

Erläuterung	Reliabilitätskoeffizient
Eindimensionaler Test (► Abschn. 15.2.1)	
Cronbachs Alpha: Anteil der wahren Varianz an der Gesamtvarianz der Testwertvariablen (Cronbach 1951; Guttman 1945)	α
McDonalds Omega: Anteil der wahren Varianz an der Gesamtvarianz der Testwertvariablen (McDonald 1970, 1999)	ω
Bollens Omega: Anteil der wahren Varianz an der Gesamtvarianz der Testwertvariablen (die mindestens eine Fehlerkovarianz enthält; Bollen 1980)	ω^*
Mehrdimensionaler Test (Gesamttest, ► Abschn. 15.3.2.1)	
Omega-total: Anteil der gesamten wahren Varianz, d. h. der durch den Generalfaktor und durch alle spezifischen Faktoren erklärten Varianz, an der Gesamtvarianz der Testwertvariablen (Revelle und Zinbarg 2009)	ω_T
Omega-hierarchisch: Anteil der durch den Generalfaktor erklärten wahren Varianz an der Gesamtvarianz der Testwertvariablen (Zinbarg et al. 2005)	ω_H
Omega-spezifisch: Anteil der durch alle spezifischen Faktoren erklärten wahren Varianz an der Gesamtvarianz der Testwertvariablen (Rodriguez et al. 2016)	ω_S
	$\omega_T = \omega_H + \omega_S$
Mehrdimensionaler Test (Subskala, ► Abschn. 15.3.2.2)	
Omega-Subskala-total: Anteil der gesamten wahren Varianz einer Subskala, d. h. der durch den Generalfaktor und einen spezifischen Faktor erklärten Varianz, an der Gesamtvarianz der Subskala im Rahmen des mehrdimensionalen Modells (vgl. Reise et al. 2013a; Rodriguez et al. 2016)	$\omega_{Skala-T}$
Omega-Subskala-hierarchisch (oftmals auch als ω_{HS} bezeichnet): Anteil der durch den Generalfaktor erklärten wahren Varianz an der Gesamtvarianz einer Subskala im Rahmen des mehrdimensionalen Modells (vgl. Reise et al. 2013a)	$\omega_{Skala-H}$
Omega-Subskala-spezifisch: Anteil der durch den spezifischen Faktor erklärten wahren Varianz an der Gesamtvarianz einer Subskala im Rahmen des mehrdimensionalen Modells (vgl. Rodriguez et al. 2016)	$\omega_{Skala-S}$
	$\omega_{Skala-T} = \omega_{Skala-H} + \omega_{Skala-S}$
<i>Anmerkung:</i> Wird aus theoretischen Gründen im mehrdimensionalen Modell eine Fehlerkovarianz aufgenommen, so werden die Reliabilitätskoeffizienten des mehrdimensionalen Tests mit einem * versehen.	

15.2 Eindimensionale Modelle

Eine zentrale Voraussetzung zur Aufsummierung der Itemwerte zu einem Test-(summen)wert ist die Eindimensionalität der Items. Eindimensionalität bedeutet, dass alle Items ein- und dasselbe Konstrukt erfassen und dass somit die True-Score-Variablen der Items lineare Funktionen voneinander und damit auch des latenten Konstrukts sind (vgl. ► Kap. 13; Eid et al. 2017b). Die in diesem Abschnitt besprochenen Methoden zur Schätzung der Reliabilität basieren im Wesentlichen auf der Annahme der Eindimensionalität.

Zur Überprüfung der Eindimensionalität wird in der Regel die CFA (vgl. ► Kap. 24) verwendet. Die CFA ermöglicht neben der Parameterschätzung für die Berechnung der Reliabilitätskoeffizienten auch die explizite Überprüfung der messtheoretischen Voraussetzungen der jeweiligen Koeffizienten. Cronbachs Alpha, McDonalds Omega und Bollens Omega sind modellbasierte Methoden der Reliabilitätsschätzung, die auf eindimensionalen oder zumindest im Wesentlichen eindimensionalen Messungen beruhen (► Exkurs 15.1).

Eindimensionalität als Voraussetzung für die Bildung der Test(summen)werte

CFA

Exkurs 15.1**Eindimensionalität, Mehrdimensionalität und Methodeneffekte**

Wenn mit einem Test oder Fragebogen nur *ein* Merkmal, d. h. eine einzige latente Variable, gemessen werden soll, müssen die einzelnen Messungen (Items) eindimensional sein. Diese Voraussetzung ist zentral für die klassischen Reliabilitätsmaße und für die meisten modellbasierten Reliabilitätsmaße (zum Zusammenhang zwischen wahren Werten und latenten Variablen ► Kap. 24 und 13). *Eindimensionalität* liegt vor, wenn die latente Variable die korrelativen Zusammenhänge aller beobachteten Variablen (Itemvariablen) vollständig erklärt, sodass keine auf *Mehrdimensionalität* hindeutenden Restkorrelationen verbleiben. Die Partialkorrelationen zwischen den Itemvariablen bei Auspartialisierung des Einflusses der latenten Variable betragen somit null. In diesem Fall stellt die latente Variable die einzige systematische Einflussgröße auf die gemessenen Variablen dar.

Bestehen zwischen einzelnen Messungen jedoch zusätzlich zum Einfluss der gemeinsamen latenten Variablen weitere substantielle korrelative Zusammenhänge, so kann dies ein Hinweis auf die *Mehrdimensionalität* der Messungen sein. Die Mehrdimensionalität kann z. B. daraus resultieren, dass entweder zusätzlich zum intendierten Konstrukt ein weiteres inhaltlich bedeutsames Konstrukt (z. B. Erfassung der Lesefähigkeit zusätzlich zum logischen Denken) oder ein inhaltlich nicht bedeutsames Konstrukt gemessen wird, das als „*Methodeneffekt*“ bezeichnet wird.

Unter dem Begriff „*Methode*“ werden unterschiedliche systematische Varianzquellen subsumiert, die sich über die

latente Variable hinaus auf die Messungen auswirken können (vgl. ► Kap. 25 und 26; s. auch Eid et al. 2016). Im Gegensatz zu den Methoden, die z. B. in der Multitrait-Multimethod-Analyse (MTMM-Analyse) verwendet werden und Charakteristika der Messinstrumente, der Beurteiler oder der Situationen beinhalten, handelt es sich hier um Charakteristika der Items (vgl. Podsakoff et al. 2003; Podsakoff et al. 2012). Hierzu gehören u. a. ähnliche Itemformulierungen und invers oder sozial erwünscht formulierte Items, also Gemeinsamkeiten weniger Items, die sie untereinander, aber nicht mit den anderen Items des Tests teilen. Diese Zusammenhänge zwischen Items, die über den Einfluss der gemeinsamen latenten Variablen hinausgehen, können über Methodenfaktoren oder über Fehlerkovarianzen modelliert werden.

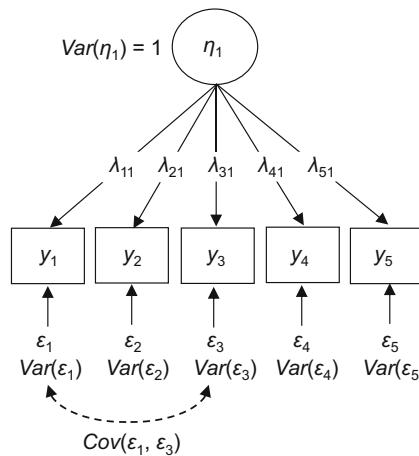
Aufgrund inhaltlicher Überlegungen sollte entschieden werden, ob es sich bei den Fehlerkovarianzen um inhaltlich irrelevante Methodeneffekte handelt, die der Eindimensionalität der Items nicht grundsätzlich entgegenstehen, oder ob eine Modifikation des eindimensionalen Konstrukts zu einem mehrdimensionalen Konstrukt vorgenommen werden sollte. Liegen inhaltlich irrelevante Methodeneffekte vor, so kann die eindimensionale Struktur beibehalten werden, jedoch muss zusätzlich mindestens eine *Fehlerkovarianz* in das Modell aufgenommen werden. In diesem Fall würden die Items „im Wesentlichen“ ein eindimensionales Konstrukt messen. Bei der Reliabilitätsschätzung sollten diese Fehlerkovarianzen dem Messfehler zugerechnet und Bollens Omega anstatt McDonalds Omega als Reliabilitätsmaß verwendet werden.

15.2.1 Koeffizienten zur Schätzung der Reliabilität eindimensionaler Tests

In der Vergangenheit wurde zur Schätzung der Reliabilität der Messungen einer eindimensionalen Skala fast ausschließlich Cronbachs Alpha (Cronbach 1951; von Guttman 1945 als „Koeffizient λ_3 “ bezeichnet) verwendet, obwohl dieses Maß auf der sehr strengen Annahme der essentiellen τ -Äquivalenz (► Kap. 13) beruht, die in empirischen Studien zumeist nicht erfüllt ist. Jedoch wurden schon vor mehreren Jahren sowohl auf Basis der exploratorischen Faktorenanalyse (EFA; McDonald 1970) als auch auf Basis der CFA (Jöreskog 1971; McDonald 1999; Werts et al. 1974) Alternativen zu Cronbachs Alpha entwickelt, die auf der weniger strengen Annahme der τ -Kongenerität beruhen (s. auch Bollen 1980, 1989). McDonald (1999) war der Erste, der auf Basis eines τ -kongenerischen Messmodells für den Anteil der wahren Varianz einer Testwertvariablen an der Gesamtvarianz die Bezeichnung Omega (ω) verwendete. Daher wird dieser ω -Koeffizient inzwischen häufig als „McDonalds Omega“ bezeichnet. Andere Autoren bezeichnen Omega auch als „composite reliability“ (z. B. Bentler 2009; Raykov 1997).

Ist die Voraussetzung der Eindimensionalität jedoch nur im Wesentlichen erfüllt, da alle Itemvariablen zwar auf einem gemeinsamen Faktor laden, zusätzlich aber eine (oder mehr als eine) Kovarianz zwischen Messfehlern aus theoretischen

McDonalds Omega als Alternative zu Cronbachs Alpha



■ Abb. 15.1 Eindimensionales Modell mit einem Faktor (η_1), fünf Itemvariablen (y_1 bis y_5) und allen Parametern zur Schätzung von Cronbachs Alpha, McDonalds Omega sowie Bollens Omega

Gründen ins Modell mit aufgenommen werden muss, sollte die Reliabilität über Bollens Omega (ω^*) anstatt über McDonalds Omega (ω) geschätzt werden (vgl. auch Raykov 2004; Raykov und Marcoulides 2016).

Die modellbasierten Reliabilitätskoeffizienten werden anhand der CFA geschätzt. In ■ Abb. 15.1 ist das Pfaddiagramm eines eindimensionalen Modells dargestellt. In diesem Modell laden fünf Itemvariablen (y_1 bis y_5) auf dem gemeinsamen Faktor η_1 . Alle potentiell zu schätzenden Parameter sind im Modell eingetragen.

Im eindimensionalen Modell mit fünf Indikatorvariablen (Itemvariablen y_1 bis y_5) (■ Abb. 15.1) sind fünf Faktorladungen (λ_{11} bis λ_{51}) und fünf Fehlervarianzen ($Var(\varepsilon_1)$ bis $Var(\varepsilon_5)$) zu schätzen, insgesamt also zehn Parameter. Wird aus theoretischen Gründen eine Fehlerkovarianz ins Modell aufgenommen, hier die Kovarianz zwischen den Fehlertern der Items y_1 und y_3 , kommt noch $Cov(\varepsilon_1, \varepsilon_3)$ als weiterer zu schätzender 11. Parameter hinzu. Die Varianz der latenten Variablen η_1 wurde hier aus Gründen der Normierung auf eins fixiert (vgl. ► Kap. 24) und wird nicht geschätzt.

Anhand der Parameter des eindimensionalen Faktormodells in ■ Abb. 15.1 können die Reliabilitätskoeffizienten Cronbachs Alpha, McDonalds Omega und Bollens Omega geschätzt werden, sofern der Modellfit für die jeweils zugrunde liegende Form der Messäquivalenz hinreichend gut ist.

Bollens Omega ω^* bei korrelierten Messfehlern

Guter Modellfit ist Voraussetzung

! Der Modellfit ist wesentlich für alle Reliabilitäts schätzungen. Die Reliabilitätskoeffizienten setzen unterschiedliche Modellannahmen und Restriktionen voraus, die erfüllt sein müssen. Nur wenn das Modell mit den entsprechenden Modellannahmen und Restriktionen gut zu den Daten passt, dürfen Reliabilitätskoeffizienten überhaupt berechnet und interpretiert werden.

In ■ Tab. 15.2 sind die Formeln (15.1)–(15.3) zur modellbasierten Schätzung von Cronbachs Alpha, McDonalds Omega und Bollens Omega aufgeführt.

Tabelle 15.2 Formeln der modellbasierten Reliabilitätskoeffizienten für eindimensionale Modelle

Reliabilitätskoeffizient	Formel	
Cronbachs Alpha: α Annahmen: Eindimensionalität, essentielle τ -Äquivalenz, unkorrelierte Messfehler	$\alpha = \frac{(p \cdot \lambda_1)^2 \cdot \text{Var}(\eta_1)}{(p \cdot \lambda_1)^2 \cdot \text{Var}(\eta_1) + \sum_{i=1}^p \text{Var}(\varepsilon_i)}$	(15.1)
McDonalds Omega: ω Annahmen: Eindimensionalität, τ -Kongeneritität, unkorrelierte Messfehler	$\omega = \frac{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot \text{Var}(\eta_1)}{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot \text{Var}(\eta_1) + \sum_{i=1}^p \text{Var}(\varepsilon_i)}$	(15.2)
Bollens Omega: ω^* Annahme: Eindimensionalität, τ -Kongeneritität, einzelne korrelierte Messfehler sind erlaubt	$\omega^* = \frac{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot \text{Var}(\eta_1)}{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot \text{Var}(\eta_1) + \sum_{i=1}^p \text{Var}(\varepsilon_i) + 2 \cdot \sum_{i < i'} \text{Cov}(\varepsilon_i, \varepsilon_{i'})}$	(15.3)

η_1 = latente Variable (Faktor); λ_1 = Faktorladung auf der latenten Variablen η_1 , die für alle τ -äquivalenten Items konstant ist; λ_{i1} = Faktorladung des Items i auf der latenten Variablen η_1 ; ε_i = Messfehlervariable des Items i ; Var = Varianz; Cov = Kovarianz; i = Laufindex der Items; p = Anzahl der Items

15.2.1.1 Cronbachs Alpha

Voraussetzungen für Cronbachs Alpha

Cronbachs Alpha (α) (Cronbach 1951; Guttman 1945) setzt voraus, dass alle Items dasselbe latente Merkmal erfassen (Eindimensionalität) und dass die Fehlervariablen unkorreliert sind. Zusätzlich setzt Cronbachs Alpha essentielle τ -Äquivalenz der Itemvariablen voraus, d. h., es wird angenommen, dass alle Items das latente Merkmal im gleichen Ausmaß erfassen und jeweils denselben Anteil an wahrer Varianz in der Population aufweisen. Ist die Annahme der essentiellen τ -Äquivalenz erfüllt, so sollten alle Itemvariablen untereinander identische Kovarianzen aufweisen (► Kap. 14). Im Rahmen der CFA impliziert diese Annahme, dass alle Itemvariablen identische Faktorladungen aufweisen und sich somit in Bezug auf ihre Diskriminationsfähigkeit nicht unterscheiden (vgl. Eid und Schmidt 2014, S. 313). In □ Abb. 15.1 müssten dementsprechend die Koeffizienten λ_{11} bis λ_{51} gleichgesetzt werden und alle Fehlerkovarianzen müssten null sein (auch die in □ Abb. 15.1 eingezeichnete Kovarianz zwischen den Messfehlern der Items y_1 und y_3). Die modellbasierte Formel für Cronbachs Alpha (Gl. 15.1 in □ Tab. 15.2) ist abgesehen von der Annahme gleicher Faktorladungen mit der Formel für McDonalds Omega identisch (► Abschn. 15.2.1.2). Ist der Modellfit eines CFA-Modells mit den entsprechenden Restriktionen (Eindimensionalität, gleiche Faktorladungen und unkorrelierte Messfehler) zufriedenstellend, kann davon ausgegangen werden, dass den Daten ein eindimensionales Modell zugrunde liegt und die Items jeweils denselben Anteil an wahrer Varianz in der Population aufweisen. Nur wenn alle oben genannten Annahmen erfüllt sind, ist Cronbachs Alpha als Reliabilitätsmaß angemessen.

Da diese Annahmen in Bezug auf empirische Daten aber nur selten erfüllt werden können, wird von der unkritischen Verwendung von Cronbachs Alpha inzwischen abgeraten (vgl. u. a. ► Kap. 14; Rauch und Moosbrugger 2011; Raykov und Marcoulides 2011; Revelle und Condon 2018; Sijtsma 2009).

15.2.1.2 McDonalds Omega

Modell τ -kongenerischer Variablen

Seit einigen Jahren hat sich das weniger strenge Modell τ -kongenerischer Variablen (Jöreskog 1971) durchgesetzt, das nicht auf der Annahme gleicher Faktorladungen beruht, sondern unterschiedlich hohe Faktorladungen zulässt (vgl. dazu ► Kap. 24 und 13). McDonalds Omega (ω) (Gl. 15.2 in □ Tab. 15.2) als modellba-

Exkurs 15.2**Herleitung der Formel für McDonalds Omega (ω)**

Zur Herleitung der Formel für McDonalds Omega (Gl. 15.2 in □ Tab. 15.2) müssen zunächst die Messmodellgleichungen für die Antwortvariablen der Items, die Itemvariablen y_1 bis y_5 in □ Abb. 15.1 aufgestellt werden. Angenommen wird, dass alle beobachteten Itemvariablen – wenn auch mit unterschiedlicher Messgenauigkeit – die latente Variable η_1 messen, sie also eindimensional sind:

$$\begin{aligned}y_1 &= \lambda_{11} \cdot \eta_1 + \varepsilon_1 \\y_2 &= \lambda_{21} \cdot \eta_1 + \varepsilon_2 \\y_3 &= \lambda_{31} \cdot \eta_1 + \varepsilon_3 \\y_4 &= \lambda_{41} \cdot \eta_1 + \varepsilon_4 \\y_5 &= \lambda_{51} \cdot \eta_1 + \varepsilon_5\end{aligned}\quad (15.4)$$

In Gl. (15.4) steht λ_{ij} ($i = 1, \dots, 5; j = 1$) für die Faktorladungen ($i = 1, \dots, p$ ist der Laufindex der Items, der zweite Index [hier 1] bezeichnet die latente Variable); die Fehlervariablen sind mit ε_i und die latente Variable mit η_1 bezeichnet.

Die Testwertvariable Y ergibt sich, indem die Werte der fünf Itemvariablen, die η_1 mit unterschiedlicher Messgenauigkeit messen, jeweils ungewichtet aufsummiert werden. Die Itemvariablen y_1 bis y_5 können in dieser Summe durch ihre Messmodellgleichungen aus Gl. (15.4) ersetzt werden:

$$\begin{aligned}Y &= y_1 + y_2 + y_3 + y_4 + y_5 \\&= (\lambda_{11} \cdot \eta_1 + \varepsilon_1) + (\lambda_{21} \cdot \eta_1 + \varepsilon_2) + (\lambda_{31} \cdot \eta_1 + \varepsilon_3) \\&\quad + (\lambda_{41} \cdot \eta_1 + \varepsilon_4) + (\lambda_{51} \cdot \eta_1 + \varepsilon_5)\end{aligned}\quad (15.5)$$

Umsortieren und Ausklammern von η_1 führen zu folgender Gleichung, wobei der erste Summand den wahren Anteil der Testwertvariablen Y und der zweite Summand den Fehleranteil darstellt:

$$\begin{aligned}Y &= (\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41} + \lambda_{51}) \cdot \eta_1 \\&\quad + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5)\end{aligned}\quad (15.6)$$

Der erste Summand in Gl. (15.6) stellt den wahren Anteil der Testwertvariablen Y dar, der zweite Summand den Fehleranteil. Die Varianz von Y kann nun als Summe aus wahrer Varianz und Fehlervarianz ausgedrückt werden, wobei angenommen wird, dass die Messfehlervariablen ε_1 bis ε_5 und

die latente Variable (Faktor) η_1 voneinander unabhängig sind (zu den Eigenschaften der True-Score- und Fehlervariablen s. ▶ Kap. 13):

$$\begin{aligned}Var(Y) &= Var[(\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41} + \lambda_{51}) \cdot \eta_1 \\&\quad + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5)] \\&= [(\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41} + \lambda_{51})^2 \cdot Var(\eta_1)] \\&\quad + [Var(\varepsilon_1) + Var(\varepsilon_2) + Var(\varepsilon_3) \\&\quad + Var(\varepsilon_4) + Var(\varepsilon_5)]\end{aligned}\quad (15.7)$$

Die Varianz des Produkts einer Konstanten (hier die Summe der Faktorladungen) mit der latenten Variablen wird gebildet, indem die Konstante quadriert und mit der Varianz der latenten Variablen multipliziert wird. Die Varianz der Summe der Fehlervariablen ist gleich der Summe der einzelnen Fehlervarianzen, da die Fehlervariablen voneinander unabhängig sind und somit keine Kovarianzen berücksichtigt werden müssen. Somit stellt der erste Summand in Gl. (15.7) die wahre Varianz der Testwertvariablen Y dar, der zweite Summand die Fehlervarianz.

Der Reliabilitätskoeffizient ω der Testwertvariablen Y wird bestimmt als Quotient aus der wahren Varianz und der gesamten Varianz, wobei sich die gesamte Varianz zusammensetzt aus der Summe der wahren Varianz und der Fehlervarianz:

$$\begin{aligned}\omega &= \frac{(\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41} + \lambda_{51})^2 \cdot Var(\eta_1)}{\left[(\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41} + \lambda_{51})^2 \cdot Var(\eta_1) \right]} \\&\quad + [Var(\varepsilon_1) + Var(\varepsilon_2) + Var(\varepsilon_3) + Var(\varepsilon_4) \\&\quad + Var(\varepsilon_5)]\end{aligned}\quad (15.8)$$

Verallgemeinert auf p Items ergibt sich in Gl. (15.9) die in □ Tab. 15.2 aufgeführte Gl. (15.2).

$$\omega = \frac{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot Var(\eta_1)}{\left(\sum_{i=1}^p \lambda_{i1} \right)^2 \cdot Var(\eta_1) + \sum_{i=1}^p Var(\varepsilon_i)}\quad (15.9)$$

wobei $i = 1, \dots, p$ der Laufindex der Items ist, und der zweite Index (hier 1) die latente Variable bezeichnet.

siertes Reliabilitätsmaß beruht lediglich auf der Annahme der Eindimensionalität und der Annahme unkorrelierter Messfehler. Im ▶ Exkurs 15.2 wird die Herleitung der Formel für McDonalds Omega vorgestellt.

15.2.1.3 Bollens Omega

Bollen (1980) erweiterte die Formel für McDonalds Omega, indem er von der strengen Annahme unkorrelierter Messfehler abwich und korrelierte Messfehler

Voraussetzung der Eindimensionalität gelockert

Bollens Omega vs. McDonalds Omega

als Teil der Fehlervarianz behandelte (► Abschn. 15.2.2.3). Somit muss die Voraussetzung der Eindimensionalität nur noch im Wesentlichen erfüllt sein, indem alle Itemvariablen zwar auf dem gemeinsamen Faktor laden, zusätzlich aber mindestens eine Kovarianz (oder Korrelation) zwischen Messfehlervariablen ins Modell aufgenommen wird. In □ Abb. 15.1 ist beispielsweise eine Kovarianz zwischen den Fehlervariablen ε_1 und ε_3 eingezeichnet. Anstatt über McDonalds Omega (ω) wird in diesem Fall die Reliabilität über Bollens Omega (ω^*) (Gl. 15.3 in □ Tab. 15.2) geschätzt (vgl. auch Raykov 2004; Raykov und Marcoulides 2016). Der Stern¹ als Zusatz von Omega (ω^*) soll anzeigen, dass Fehlerkovarianzen berücksichtigt werden, die aber inhaltlich als nicht relevant eingestuft und deshalb als Fehleranteile behandelt werden.

- ! Liegt eine gute theoretische Begründung für die Aufnahme einer Fehlerkovarianz (oder auch mehrerer Fehlerkovarianzen) in das Modell vor, so sollte anstatt des Koeffizienten ω der Koeffizient ω^* verwendet werden. Dies betrifft sowohl die Reliabilitätskoeffizienten ein- als auch mehrdimensionaler Modelle.

Wird aus theoretischen Gründen eine Kovarianz zwischen zwei Messfehlervariablen angenommen, so wird im Nenner der Formel von ω^* (Gl. 15.3 in □ Tab. 15.2) die Fehlerkovarianz aufgenommen (multipliziert mit 2, da alle Kovarianzen in einer Kovarianzmatrix doppelt vorkommen). Bei einer positiven Fehlerkovarianz nimmt der Koeffizient ω^* einen kleineren Wert als ω an, da sich der Nenner vergrößert, während ω^* bei einer negativen Kovarianz einen größeren Wert als ω annimmt, da sich der Nenner in der Gleichung verringert.

15.2.2 Empirisches Beispiel

Anwendungsbeispiel: Perfektionismus

Subskalen der MPS-F

Als empirisches Anwendungsbeispiel soll hier das mehrdimensionale Persönlichkeitsmerkmal Perfektionismus betrachtet werden, gemessen anhand der Mehrdimensionalen Perfektionismus-Skala (MPS-F) von Frost et al. (1990) in der deutschen Fassung von Stöber (1995). Nachfolgend werden drei der sechs Perfektionismus-Subskalen verwendet: *Personal Standards* (Hohe Standards), *Doubts about Actions* (Leistungsbezogene Zweifel) und *Concern over Mistakes* (Fehlersensibilität). *Personal Standards* wird der übergeordneten Dimension *Perfectionistic Strivings* zugeordnet, während *Doubts about Actions* sowie *Concern over Mistakes* der übergeordneten Dimension *Perfectionistic Concerns* zugeordnet werden. Diese Skalen erfassen zentrale Aspekte des Perfektionismus (vgl. u. a. Altstötter-Gleich und Bergemann 2006; Smith und Saklofske 2017; Stoeber 2014).

Der verwendete Datensatz von $N = 250$ Personen stammt aus einer Untersuchung von Amend (2015). Die Anzahl der Items wurde für das Beispiel reduziert, um die Analysen und Grafiken überschaubar zu halten: *Personal Standards* (PS) wird über drei Items gemessen (PS12, PS19, PS24), *Doubts about Actions* (DA) ebenfalls über drei Items (DA17, DA28, DA32) und *Concern over Mistakes* (CM) über fünf Items (CM09, CM21, CM23, CM25, CM34), wobei sich die Nummerierung auf die entsprechenden Items des Originalfragebogens bezieht. Zu beachten ist, dass die Verkürzung der Skalen nur aus didaktischen Gründen vorgenommen wird; es handelt sich also nicht um die Entwicklung einer Kurzskala.

Zunächst soll die Schätzung der Reliabilitätskoeffizienten für eindimensionale Konstrukte, Cronbachs Alpha, McDonalds Omega und Bollens Omega, anhand der Skala CM demonstriert werden. Hierfür müssen die messtheoretischen Voraussetzungen der Reliabilitätsmaße getestet werden, d. h. die Eindimensionalität

¹ Die Bezeichnung ω^* (Omega-Stern) wurde von Kenneth Bollen auf der 13. Tagung der Fachgruppe „Methoden und Evaluation“ der Deutschen Gesellschaft für Psychologie in Tübingen im Jahr 2017 persönlich autorisiert.

der Messungen und damit die Unkorreliertheit der Messfehler (Voraussetzungen für Cronbachs Alpha und McDonalds Omega) sowie die Gleichheit der Faktorladungen (zusätzliche Voraussetzung für Cronbachs Alpha). Später wird die Schätzung der Reliabilitätskoeffizienten des mehrdimensionalen Modells erläutert (► Abschn. 15.3.3).

Zur Schätzung der Parameter wird die CFA mit dem Programm *Mplus*, Version 8 (Muthén und Muthén 2017), unter Verwendung der robusten Maximum-Likelihood-Methode (MLR; vgl. ► Kap. 24) durchgeführt.

15.2.2.1 Reliabilitätsschätzung mit Cronbachs Alpha

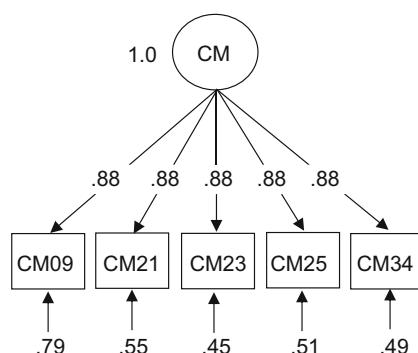
Zur Schätzung der Reliabilität der Skala CM anhand von Cronbachs Alpha wird ein eindimensionales Faktormodell spezifiziert, in dem alle fünf Faktorladungen im Modell (► Abb. 15.1) gleichgesetzt werden müssen, was der Annahme der essentiellen τ -Äquivalenz entspricht und somit der Annahme gleicher Faktorladungen der Items in der Population (die Mittelwerte der Items dürfen sich dagegen unterscheiden). Zur Identifikation des Modells wird die Varianz der latenten Variablen auf eins fixiert (zur Identifikation von CFA-Modellen ► Kap. 24). In diesem Modell sind aufgrund der essentiellen τ -Äquivalenz nur eine (über alle Items hinweg konstante) Faktorladung und fünf Fehlervarianzen zu schätzen. Das Pfaddiagramm mit den unstandardisierten Parameterschätzungen findet sich in ► Abb. 15.2.

Die Ergebnisse zeigen einen nicht zufriedenstellenden Modellfit mit $\chi^2(9) = 29.095, p = .001$, Root Mean Square Error of Approximation (RMSEA) = .095, Comparative Fit Index (CFI) = .956, Tucker Lewis Index (TLI) = .951, Standardized Root Mean Square Residual (SRMR) = .062 (zur Beurteilung der Gütekriterien s. ► Kap. 24; Schermelleh-Engel et al. 2003). Für ein gut passendes Modell wäre ein nicht signifikanter χ^2 -Wert ($p > .01$) erwartet worden. Somit passt das Modell mit identischen Faktorladungen der Itemvariablen nicht zu den Daten und die Parameterschätzungen dürfen nicht interpretiert werden.

Nur zum Vergleich mit den anderen Koeffizienten wird hier α über Gl. (15.1) in ► Tab. 15.2 berechnet:

$$\begin{aligned}\alpha &= \frac{(.88 + .88 + .88 + .88 + .88)^2 \cdot 1.00}{(.88 + .88 + .88 + .88 + .88)^2 \cdot 1.00 + (.79 + .55 + .45 + .51 + .49)} \\ &= \frac{19.36}{19.36 + 2.79} = .875\end{aligned}$$

Die Reliabilität der Skala würde anhand von Cronbachs Alpha auf .875 geschätzt werden (► Tab. 15.3). Aufgrund des schlechten Modellfits darf eine Schätzung der Reliabilität anhand des Koeffizienten α allerdings nicht vorgenommen werden. Ein Grund für den schlechten Modellfit könnte in der strengen Annahme der gleichen Faktorladungen liegen.



► Abb. 15.2 CFA-Modell der Skala *Concern over Mistakes* (CM), in dem alle unstandardisierten Faktorladungen gleichgesetzt wurden (Annahme der essentiellen τ -Äquivalenz)

Tabelle 15.3 Reliabilitätsschätzungen der auf fünf Items gekürzten Skala CM anhand eines eindimensionalen Faktormodells

Koeffizient	Reliabilitätsschätzung	95 %-Konfidenzintervall
Cronbachs Alpha	$\alpha = .875^a$	–
McDonalds Omega	$\omega = .874^a$	–
Bollens Omega	$\omega^* = .850$	[.806; .885]

^a Der Koeffizient darf nicht interpretiert werden, da das Modell nicht zu den Daten passt.

15.2.2.2 Reliabilitätsschätzung mit McDonalds Omega

Da die von Cronbachs Alpha implizierten Annahmen nicht zu den Daten passen, wird im Folgenden die Annahme der essentiellen τ -Äquivalenz zugunsten der weniger strengen Annahme der τ -Kongenerität aufgegeben. Dazu wird ein τ -kongenerisches Messmodell spezifiziert, in dem nun alle fünf Faktorladungen frei geschätzt werden. Mit diesem Modell werden die Annahme der Eindimensionalität der Indikatoren und damit implizit auch die Unkorreliertheit der Fehlervariablen überprüft. Zur Identifikation des Modells wird die Varianz der latenten Variablen wiederum auf eins fixiert. In diesem Modell sind fünf Faktorladungen und fünf Fehlervarianzen zu schätzen. Das Pfaddiagramm mit den unstandardisierten Parameterschätzungen findet sich in Abb. 15.3.

Die Ergebnisse zeigen wiederum einen nicht zufriedenstellenden Modellfit mit $\chi^2(5) = 17.840, p = .003, \text{RMSEA} = .101, \text{CFI} = .972, \text{TLI} = .943, \text{SRMR} = .030$. Für ein gut passendes Modell wäre ein nicht-signifikanter χ^2 -Wert ($p > .01$) zu erwarten. Somit passt auch dieses Modell nicht zu den Daten und die Parameter dürfen nicht interpretiert werden.

Nur zum Vergleich mit den anderen Koeffizienten wird hier ω über Gl. (15.2) in Tab. 15.2 berechnet:

$$\begin{aligned}\omega &= \frac{(.82 + .82 + .90 + .99 + .85)^2 \cdot 1.00}{(.82 + .82 + .90 + .99 + .85)^2 \cdot 1.00 + (.80 + .56 + .45 + .46 + .49)} \\ &= \frac{19.18}{19.18 + 2.76} = .874\end{aligned}$$

Die Reliabilität der Skala würde anhand von McDonalds Omega gemäß Tab. 15.2 Gl. (15.2) auf .874 geschätzt werden (Tab. 15.3). Aufgrund des schlechten Modellfits sollte die Reliabilität der Skala allerdings nicht anhand des Koeffizienten ω geschätzt werden. Die Ursache für den nicht zufriedenstellenden Modellfit wird nachfolgend untersucht.

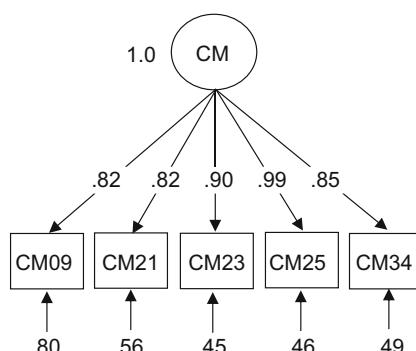


Abb. 15.3 CFA-Modell der Skala *Concern over Mistakes* (CM) mit unstandardisierten Faktorladungen, die alle frei geschätzt wurden (Annahme der τ -Kongenerität)

15.2.2.3 Reliabilitätsschätzung mit Bollens Omega

Sowohl das Modell essentiell τ -äquivalenter Variablen als auch das Modell τ -kongenerischer Variablen setzen die Unkorreliertheit der Fehlervariablen voraus. Ein Grund für den mangelnden Modellfit des essentiell τ -äquivalenten und des τ -kongenerischen Modells könnte darin liegen, dass diese Voraussetzung – zumindest teilweise – nicht erfüllt ist. Bei der Inspektion der Itemformulierungen fällt auf, dass die beiden Items CM09 und CM23 den Ausdruck „ich als Mensch“ beinhalten:

- CM09: „Wenn ich bei der Arbeit/beim Studium versage, dann bin ich auch als Mensch ein Versager.“
- CM23: „Wenn ich nicht so gut bin wie andere, dann heißt das, dass ich als Mensch weniger wert bin.“

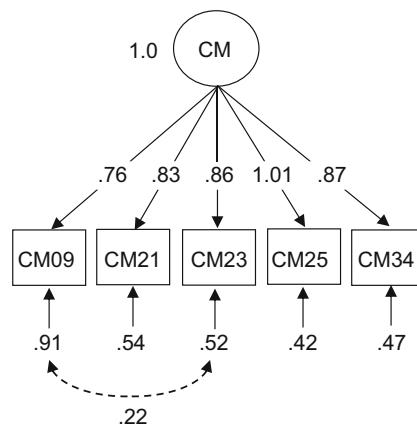
Diese ähnliche Formulierung der Items („ich als Mensch“) könnte auf einen Methodeneffekt schließen lassen. Ein Methodeneffekt (► Exkurs 15.1) kann über eine Fehlerkovarianz im Modell oder einen Methodenfaktor berücksichtigt und modelliert werden. Mit diesem modifizierten Modell kann die Reliabilität der Skala über die Erweiterung des Koeffizienten ω zu ω^* geschätzt werden, indem die Fehlerkovarianz in den Nenner von Gl. (15.3) (► Tab. 15.2) aufgenommen wird.

Modifikationsindizes (MI), die vom verwendeten Programm (z. B. von Mplus oder R) ausgegeben werden, geben an, um wie viel sich der χ^2 -Wert ungefähr verringern würde, wenn ein fixierter Parameter freigesetzt würde (vgl. ► Kap. 24; Sörbom 1989). Für das Beispiel zeigt sich, dass die Aufnahme einer Kovarianz zwischen den Fehlervariablen der Items CM09 und CM23 ins Modell zu einer deutlichen Verbesserung des Modellfits führen würde, da sich der χ^2 -Wert beträchtlich verringern würde ($MI = 14.292$). Die Kovarianz zwischen den Fehlervariablen der Items CM09 und CM23 wird folglich als frei zu schätzender Parameter ins Modell aufgenommen, sodass nun elf Parameter zu schätzen sind. Die unstandardisierten Parameterschätzungen des modifizierten Modells sind ► Abb. 15.4 zu entnehmen.

Die Aufnahme der Fehlerkovarianz zwischen den Items CM09 und CM23 ins Modell führt zu einer substantiellen Verbesserung des Modellfits. Da nun mit $\chi^2(4) = 4.564$, $p = .335$, RMSEA = .024, CFI = .999, TLI = .997, SRMR = .013 ein guter Modellfit vorliegt, kann die Reliabilität der Skala über den Koeffizienten ω^* (Bollen 1980) geschätzt werden.

Fehlerkovarianz

Modifikationsindizes



► Abb. 15.4 CFA-Modell der Skala *Concern over Mistakes* (CM) mit unstandardisierten, frei geschätzten Faktorladungen (Annahme der τ -Kongenerität) und Aufnahme einer zusätzlichen Kovarianz zwischen den Fehlervariablen der Items CM09 und CM23

Werden die geschätzten Werte der Parameter (Faktorladungen, Fehlervarianzen und die Fehlerkovarianz) in die Formel von Bollens Omega eingesetzt, so wird ω^* gemäß Gl. (15.3) in ▶ Tab. 15.2 wie folgt berechnet:

$$\begin{aligned}\omega^* &= \frac{(.76+.83+.86+1.01+.87)^2 \cdot 1.00}{(.76+.83+.86+1.01+.87)^2 \cdot 1.00 + (.91+.54+.52+.42+.47) + (2 \cdot .22)} \\ &= \frac{18.749}{18.749 + 2.86 + .44} = .850\end{aligned}$$

Wie dieses Ergebnis zeigt, weist der Reliabilitätskoeffizient ω^* (.850) erwartungsgemäß einen geringeren Wert auf als ω (.874), der wiederum erwartungsgemäß etwas geringer ist als α (.875). Die Werte der Koeffizienten α und ω unterscheiden sich in diesem Beispiel nur geringfügig. Der geringe Unterschied kann damit erklärt werden, dass sich die frei geschätzten Faktorladungen (zwischen .76 und 1.01) nur wenig von den gleichgesetzten Faktorladungen (alle .88) unterscheiden.

Mit der CFA wurde die Annahme der Eindimensionalität der Messungen geprüft und das Modell nach theoretischen Überlegungen durch Aufnahme einer Fehlerkovarianz modifiziert. Die Fehlerkovarianz wurde ins Modell aufgenommen, jedoch inhaltlich als nicht bedeutsam eingestuft und deshalb dem Messfehler zugeordnet. Die Skala kann somit auch weiterhin als im Wesentlichen eindimensional angesehen werden. Aufgrund der Fehlerkovarianz sollte jedoch nur ω^* als Reliabilitätsschätzung verwendet werden.

15.2.3 Asymmetrisches Konfidenzintervall

Neben der Punktschätzung des Reliabilitätskoeffizienten ist eine Intervallschätzung in Form eines 2-seitigen 95 %-Konfidenzintervalls nötig, um eine Aussage über die Präzision der Schätzung zu erhalten (Kelley und Pornprasertmanit 2016; Raykov 2002). Da der Wertebereich der Reliabilitätskoeffizienten auf die Grenzen null und eins beschränkt ist und somit das Konfidenzintervall – vor allem in der Nähe der Grenzwerte – nicht symmetrisch sein kann, wird ein adäquates *asymmetrisches Konfidenzintervall* verwendet (Eid und Schmidt 2014; Raykov 2002; Raykov und Marcoulides 2011; vgl. auch die ausführliche Erläuterung im ▶ Exkurs 15.3).

Unter Verwendung entsprechender Statistik-Software (z.B. Mplus oder R) kann ein geeigneter Standardfehler $SE(Re)$ für Reliabilitätskoeffizienten geschätzt werden. Dabei kommt meist entweder die Delta-Methode oder das Bootstrapping zur Anwendung (vgl. Raykov 2002; Raykov und Marcoulides 2004), wodurch eine im Vergleich zu herkömmlichen Methoden genauere Schätzung des Standardfehlers möglich wird.

Das Konfidenzintervall umfasst denjenigen Bereich der Reliabilität, in dem sich $(1 - \alpha) \cdot 100\%$ aller möglichen wahren Werte der Reliabilität befinden, die den in der Stichprobe beobachteten Wert des Reliabilitätskoeffizienten erzeugt haben können. Das Konfidenzintervall wird unter Verwendung des $z_{1-\alpha/2}$ -Wertes aus der Standardnormalverteilung (z-Verteilung, Anhang) ermittelt.

15 Asymmetrisches Konfidenzintervall nötig

Schätzung des Standardfehlers für Reliabilitätskoeffizienten

Exkurs 15.3**Schätzung des asymmetrischen Konfidenzintervalls für Reliabilitätskoeffizienten**

Der „Trick“ zur Bestimmung des asymmetrischen Konfidenzintervalls besteht darin, dass zunächst eine nichtlineare Logit-Transformation des geschätzten Reliabilitätskoeffizienten durchgeführt wird, $\ln [Reliabilität/(1 - Reliabilität)]$, dann für den transformierten Wert die Grenzen des Konfidenzintervalls bestimmt werden und anschließend eine Rücktransformation in die ursprüngliche Metrik unter Verwendung der Exponentialfunktion vorgenommen wird.

Durch die Logit-Transformation ist der Wertebereich der transformierten Variablen nicht mehr begrenzt, sondern geht von $-\infty$ bis $+\infty$. Der Logit wird hier mit L bezeichnet, die geschätzte Reliabilität mit Rel :

$$L = \ln \left(\frac{Rel}{1 - Rel} \right) \quad (15.10)$$

Nun wird die Tatsache berücksichtigt, dass die natürliche Logarithmusfunktion die Umkehrfunktion der e-Funktion ist. Die Anwendung der e-Funktion auf beiden Seiten von Gl. (15.10) führt zu:

$$e^L = \frac{Rel}{1 - Rel} \quad (15.11)$$

Auflösen von Gl. (15.11) nach Rel und Umformen ergibt:

$$Rel = \frac{1}{1 + e^{-L}} \quad (15.12)$$

! Anmerkung

Wie kommt man von Gl. (15.11) zu Gl. (15.12)?

Folgende zwei Umformungen müssen vorgenommen werden:

- Der Koeffizient Rel aus Gl. (15.11) wird auf die linke Seite der Gleichung gebracht:

$$\begin{aligned} Rel &= e^L (1 - Rel) = e^L - e^L Rel \\ \Rightarrow Rel + e^L Rel &= Rel (1 + e^L) = e^L \\ \Rightarrow Rel &= \frac{e^L}{1 + e^L} \end{aligned}$$

- Zähler und Nenner des Quotienten werden durch e^L geteilt, wobei $1/e^L = e^{-L}$ ist:

$$Rel = \frac{e^L}{1 + e^L} = \frac{e^L/e^L}{(1/e^L) + (e^L/e^L)} = \frac{1}{1 + e^{-L}}$$

Es resultiert Gl. (15.12).

Des Weiteren wird zur Berechnung des Konfidenzintervalls der Standardfehler des Logits, $SE(L)$, benötigt, der unter Verwendung des geschätzten Standardfehlers der Reliabilität, $SE(Rel)$, berechnet wird:

$$SE(L) = \frac{SE(Rel)}{Rel \cdot (1 - Rel)} \quad (15.13)$$

Für ein 95 %-Konfidenzintervall mit $z_{(1-\alpha/2)} = 1.96$ werden die Grenzen des Logits wie folgt bestimmt:

$$L - 1.96 \cdot SE(L); \quad L + 1.96 \cdot SE(L) \quad (15.14)$$

Die Grenzen des Logits aus Gl. (15.14) werden nun in Gl. (15.12) eingesetzt. Aus dieser Berechnung resultiert das 95 %-Konfidenzintervall (KI) für die auf den ursprünglichen Wertebereich rücktransformierte geschätzte Reliabilität:

$$\begin{aligned} KI_{\text{Untere Grenze}} &= \frac{1}{1 + e^{-L+1.96 \cdot SE(L)}} \\ KI_{\text{Obere Grenze}} &= \frac{1}{1 + e^{-L-1.96 \cdot SE(L)}} \end{aligned} \quad (15.15)$$

Gl. (15.15) ist identisch mit Gl. (15.16) und unterscheidet sich lediglich in der mathematischen Schreibweise.

Die Grenzen eines asymmetrischen 2-seitigen $(1 - \alpha) \cdot 100\% \text{-Konfidenzintervalls}$ können wie folgt berechnet werden (Raykov und Marcoulides 2011, S. 166; s. auch Eid und Schmidt 2014, S. 285):

$$\begin{aligned} KI_{\text{Untere Grenze}} &= \frac{1}{1 + \exp \left[-\ln \left(\frac{Rel}{1 - Rel} \right) + z_{(1-\alpha/2)} \cdot \frac{SE(Rel)}{Rel \cdot (1 - Rel)} \right]} \\ KI_{\text{Obere Grenze}} &= \frac{1}{1 + \exp \left[-\ln \left(\frac{Rel}{1 - Rel} \right) - z_{(1-\alpha/2)} \cdot \frac{SE(Rel)}{Rel \cdot (1 - Rel)} \right]} \end{aligned} \quad (15.16)$$

Grenzen des asymmetrischen Konfidenzintervalls

Hier bezeichnet KI das Konfidenzintervall, $\exp[\dots]$ die Exponentialfunktion mit der Basis e, Rel den geschätzten Reliabilitätskoeffizienten, \ln den Logarithmus naturalis und $SE(Rel)$ den geschätzten Standardfehler. Das ► Beispiel 15.1 ver-

anschaulicht die Berechnung für das empirische Beispiel aus ▶ Abschn. 15.2.2, der ▶ Exkurs 15.3 beschreibt das nähere Vorgehen.

Beispiel 15.1: Konfidenzintervallbestimmung – empirisches Beispiel

Für das empirische Beispiel (▶ Abschn. 15.2.2) wird das zweiseitige asymmetrische Konfidenzintervall (KI) um den Reliabilitätskoeffizienten $\omega^* = .85$ (Tab. 15.3) nach Gl. (15.16) geschätzt und beträgt $KI = [.806; .885]$. Damit weist die Skala CM eine zufriedenstellend hohe Reliabilität auf, die mit ausreichender Präzision geschätzt wird.

15.3 Mehrdimensionale Modelle

Mehrere Facetten eines übergeordneten Merkmals

Varianzaufteilung in mehrdimensionalen Modellen

Mehrdimensionale Modelle werden immer dann benötigt, wenn die Items eines Fragebogens oder Tests mehrere Facetten eines übergeordneten Merkmals erfassen (vgl. ▶ Kap. 3). Viele psychologische Konstrukte, z. B. Intelligenz, Selbstkonzept, Extraversion oder Perfektionismus, sind als mehrdimensionale Konstrukte angelegt. Sind nun für einen mehrdimensionalen Test sowohl Subskalenwerte der einzelnen Facetten als auch Gesamtwerte (Testsummenwerte, d. h. die Summe sämtlicher Itemwerte) vorgesehen, so sind Koeffizienten, die auf der Annahme der Eindimensionalität beruhen, zur Beurteilung der Reliabilität des Tests nicht angemessen. Für mehrdimensionale Konstrukte sollten daher geeignete modellbasierte Reliabilitätsmaße anstelle von klassischen Reliabilitätsmaßen verwendet werden.

Bei mehrdimensionalen Merkmalen wird – analog zu eindimensionalen Merkmalen – eine Unterteilung der Testwertvarianz in wahre Varianz und Fehlervarianz vorgenommen. Die wahre Varianz unterteilt sich nun aber zusätzlich in einen Anteil, der auf das gemeinsame übergeordnete Konstrukt (im Weiteren als Generalfaktor bezeichnet) zurückgeführt werden kann, und in einen Anteil, der die gesamte subskalenspezifische Varianz umfasst, die unabhängig vom gemeinsamen Konstrukt ist.

! Nur wenn der Generalfaktor einen hinreichend großen Anteil an der wahren Varianz der manifesten Itemvariablen erklärt, ist die Bildung von Gesamttestwerten durch Aufsummierung der Itemwerte über alle Subskalen hinweg sinnvoll. Andernfalls sollte von der Bildung von Gesamttestwerten abgesehen werden; stattdessen sollten die Itemwerte innerhalb jeder Subskala getrennt zu Subskalenwerten aufaddiert werden.

Ob es sinnvoll ist, einen gemeinsamen übergeordneten Generalfaktor anzunehmen, sodass eine Aufsummierung der Items zu einem Gesamttestwert gerechtfertigt ist, kann anhand einer CFA – z. B. im Rahmen eines Bifaktormodells – überprüft werden. Bei Gültigkeit des Modells können verschiedene Omega-Koeffizienten für den gesamten Test und für die einzelnen Subskalen im Rahmen des mehrdimensionalen Modells geschätzt werden.

15.3.1 Bifaktormodell

In den letzten Jahren hat sich das Bifaktormodell (vgl. Chen et al. 2006; Eid et al. 2017a; Gignac 2008; Holzinger und Swineford 1937), das auch als direkt hierarchisches Modell oder als „nested model“ (geschachteltes Modell) bezeichnet wird, zunehmend zur Untersuchung übergeordneter Merkmale mit mehreren Facetten

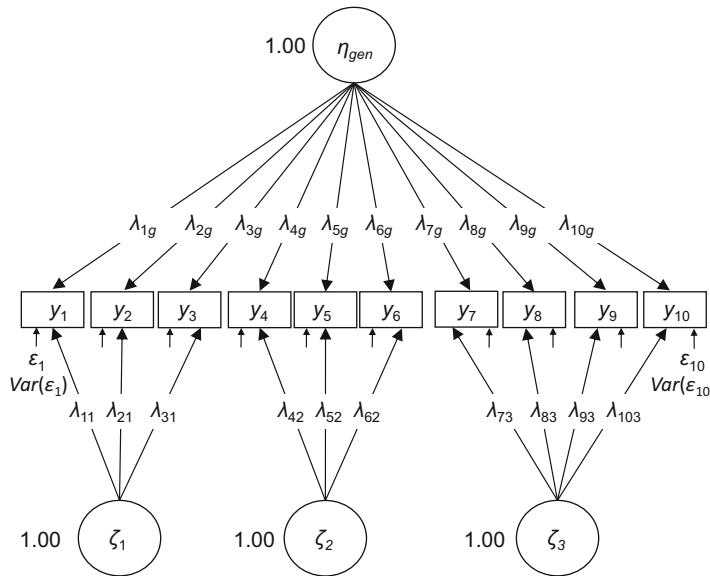


Abb. 15.5 Bifaktormodell mit zehn Items, einem Generalfaktor (η_{Gen}), drei spezifischen Residualfaktoren ($\zeta_1, \zeta_2, \zeta_3$) und allen zu schätzenden Parametern. Die Varianzen der latenten Variablen sind auf eins fixiert, die Fehlervariablen und -varianzen wurden zur Vereinfachung nur für die erste und die letzte Variable eingetragen

durchgesetzt, da es die hierarchische Struktur von latenten Merkmalen optimal abbildet.

Das Bifaktormodell (vgl. ▶ Kap. 24) ist ein mehrdimensionales konfirmatorisches Faktormodell, in dem jede Itemvariable auf einem gemeinsamen Faktor, dem Generalfaktor (η_{Gen}), lädt und zusätzlich jeweils noch auf einem subskalen-spezifischen Residualfaktor (ζ_1, ζ_2 oder ζ_3 ; vgl. □ Abb. 15.5). Der Generalfaktor repräsentiert ein breit angelegtes, in allen Subskalen enthaltenes Konstrukt, das mit dem Fragebogen oder Test primär gemessen werden soll; die spezifischen Residualfaktoren hingegen sollen konzeptuell enger definierte Teilespekte des Konstruktus erfassen, die unabhängig vom Generalfaktor sind (vgl. Rodriguez et al. 2016).

Im Bifaktormodell wird der Anteil des Generalfaktors aus den spezifischen Faktoren herauspartielliert, sodass die spezifischen Faktoren Residualfaktoren darstellen und sowohl unabhängig vom Generalfaktor als auch unabhängig voneinander sind. Die modellbasierte Reliabilitätsschätzung erfolgt für den Gesamttest auf Basis aller Items und für die Subskalen auf Basis der Items der einzelnen Subskalen.

Zur Illustration ist das Beispiel eines mehrdimensionalen Modells mit zehn Items, einem Generalfaktor und drei spezifischen Faktoren in □ Abb. 15.5 veranschaulicht.

Generalfaktor und spezifische Residualfaktoren

15.3.2 Koeffizienten zur Schätzung der Reliabilität mehrdimensionaler Tests

15.3.2.1 Omega-Koeffizienten des Gesamttests

Eine Erweiterung des Reliabilitätskonzepts für mehrdimensionale Konstrukte wurde von McDonald (1970, 1978, 1999) vorgenommen. Ungünstigerweise bezeichnete McDonald zwei verschiedene Reliabilitätskoeffizienten gleichermaßen mit ω , sodass später eine Umbenennung der Koeffizienten nötig wurde. Inzwischen wird einer der beiden Koeffizienten als ω_T (Omega-total) bezeichnet (Revelle und Zinbarg 2009), der andere als ω_H (Omega-hierarchisch; Zinbarg et al. 2005).

Omega-total

Der Koeffizient *Omega-total* (ω_T) bezeichnet den Anteil der gesamten wahren Varianz an der Gesamtvarianz der Testwerte eines mehrdimensionalen Konstrukt (vgl. □ Tab. 15.1). Er schätzt somit den Anteil an der Gesamtvarianz der Testwerte, der auf alle modellierten systematischen Varianzquellen im Rahmen eines mehrdimensionalen Faktormodells zurückgeführt werden kann. Die Bezeichnung „modelliert“ ist hier notwendig, da nicht berücksichtigte Varianzquellen auch nicht in den Koeffizienten eingehen können (vgl. Rodriguez et al. 2016). Der Koeffizient ω_T berücksichtigt somit als systematische Varianzquellen einerseits den übergeordneten Generalfaktor und andererseits alle subskalenspezifischen Faktoren gemeinsam. Er lässt sich additiv in zwei Koeffizienten aufteilen (□ Tab. 15.1), in ω_H (Omega-hierarchisch) und ω_S (Omega-spezifisch).

Omega-hierarchisch

Der Koeffizient *Omega-hierarchisch* (ω_H) ist ein Maß für den Anteil an der wahren Varianz an der Gesamtvarianz der mehrdimensionalen Testwertvariablen, der nur auf den Generalfaktor zurückgeführt wird und keine subskalenspezifischen Anteile beinhaltet. Anhand des Koeffizienten ω_H kann somit abgeschätzt werden, in welchem Ausmaß sich der Generalfaktor in den Testwerten widerspiegelt und ob die Testwerte damit im Wesentlichen eindimensional sind (Reise et al. 2013a; Rodriguez et al. 2016).

Omega-spezifisch

Der Koeffizient *Omega-spezifisch* (ω_S) ist ein Maß für den Anteil an der wahren Varianz an der Gesamtvarianz der mehrdimensionalen Testwertvariablen, der auf die subskalenspezifischen Faktoren zurückgeführt wird und unabhängig vom Generalfaktor ist. Anhand des Koeffizienten ω_S kann somit abgeschätzt werden, in welchem Ausmaß sich alle spezifischen Faktoren gemeinsam in den Testwerten widerspiegeln. In ▶ Exkurs 15.4 werden die einzelnen Formeln für ω_T , ω_H und ω_S hergeleitet.

Exkurs 15.4**Herleitung der Formeln für die Omega-Koeffizienten des Gesamttests**

Für zehn Items (vgl. □ Abb. 15.5) resultieren die folgenden Messmodellgleichungen:

$$\begin{aligned} y_1 &= \lambda_{1g} \cdot \eta_{\text{Gen}} + \lambda_{11} \cdot \xi_1 + \varepsilon_1 \\ &\vdots \\ y_4 &= \lambda_{4g} \cdot \eta_{\text{Gen}} + \lambda_{42} \cdot \xi_2 + \varepsilon_4 \\ &\vdots \\ y_{10} &= \lambda_{10g} \cdot \eta_{\text{Gen}} + \lambda_{103} \cdot \xi_3 + \varepsilon_{10} \end{aligned} \quad (15.17)$$

Die Variable der Gesamttestwerte Y ergibt sich, indem die Itemvariablen y_1 bis y_{10} , die η_{Gen} sowie ξ_1 , ξ_2 oder ξ_3 mit unterschiedlicher Genauigkeit messen, ungewichtet aufsummiert werden. Die zehn Itemvariablen können durch ihre Messmodellgleichungen ersetzt werden:

$$\begin{aligned} Y &= y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} \\ &= (\lambda_{1g} \cdot \eta_{\text{Gen}} + \lambda_{11} \cdot \xi_1 + \varepsilon_1) + \dots \\ &\quad + (\lambda_{4g} \cdot \eta_{\text{Gen}} + \lambda_{42} \cdot \xi_2 + \varepsilon_4) + \dots \\ &\quad + (\lambda_{10g} \cdot \eta_{\text{Gen}} + \lambda_{103} \cdot \xi_3 + \varepsilon_{10}) \end{aligned} \quad (15.18)$$

Umsortieren, Zusammenfassen und Ausklammern der latenten Variablen führen zu der folgenden Gleichung, wobei die

ersten vier Summanden dem vom Modell erklärten wahren Anteil entsprechen und die Summe der Fehlervariablen dem Fehleranteil:

$$\begin{aligned} Y &= \left(\sum_{i=1}^{10} \lambda_{ig} \right) \cdot \eta_{\text{Gen}} + \left(\sum_{i=1}^3 \lambda_{i1} \right) \cdot \xi_1 + \left(\sum_{i=4}^6 \lambda_{i2} \right) \cdot \xi_2 \\ &\quad + \left(\sum_{i=7}^{10} \lambda_{i3} \right) \cdot \xi_3 + \sum_{i=1}^{10} \varepsilon_i \end{aligned} \quad (15.19)$$

Die Varianz der Testwertvariablen Y (Gesamttest) lässt sich somit aufteilen in wahre Varianz und Fehlervarianz. Die wahre Varianz wiederum setzt sich aus der durch den Generalfaktor erklärten Varianz und die vom Generalfaktor unabhängigen drei subskalenspezifischen Varianzen (Residualvarianzen) zusammen:

$$\begin{aligned} \text{Var}(Y) &= \left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \\ &\quad + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\xi_2) + \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\xi_3) \\ &\quad + \sum_{i=1}^{10} \text{Var}(\varepsilon_i) \end{aligned} \quad (15.20)$$

Diese Varianzanteile werden für die Schätzung der verschiedenen Omega-Koeffizienten verwendet. *Omega-total* (ω_T) berechnet sich nun als Quotient aus totaler wahrer Varianz und Gesamtvarianz des Tests:

$$\begin{aligned} \omega_T = & \left[\left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\zeta_1) \right. \\ & + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\zeta_2) + \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\zeta_3) \Bigg] \Bigg/ \\ & \left[\left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\zeta_1) \right. \\ & + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\zeta_2) + \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\zeta_3) \\ & \left. + \sum_{i=1}^{10} \text{Var}(\varepsilon_i) \right] \end{aligned} \quad (15.21)$$

Omega-hierarchisch (ω_H) berechnet sich als Anteil der Varianz, die durch den Generalfaktor erklärt wird, an der Gesamtvarianz von Y :

$$\begin{aligned} \omega_H = & \left[\left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) \right] \Bigg/ \\ & \left[\left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\zeta_1) \right. \\ & + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\zeta_2) + \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\zeta_3) \\ & \left. + \sum_{i=1}^{10} \text{Var}(\varepsilon_i) \right] \end{aligned} \quad (15.22)$$

Omega-spezifisch (ω_S) berechnet sich als Anteil der durch alle Subskalen (Residualfaktoren) insgesamt erklärten Varianz an der Gesamtvarianz von Y und ist zugleich die Differenz zwischen Omega-total und Omega-hierarchisch:

$$\begin{aligned} \omega_S = & \left[\left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\zeta_1) + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\zeta_2) \right. \\ & + \left. \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\zeta_3) \right] \Bigg/ \\ & \left[\left(\sum_{i=1}^{10} \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\zeta_1) \right. \\ & + \left(\sum_{i=4}^6 \lambda_{i2} \right)^2 \cdot \text{Var}(\zeta_2) + \left(\sum_{i=7}^{10} \lambda_{i3} \right)^2 \cdot \text{Var}(\zeta_3) \\ & \left. + \sum_{i=1}^{10} \text{Var}(\varepsilon_i) \right] \end{aligned} \quad (15.23)$$

Anmerkung: Wenn aus theoretischen Überlegungen Fehlerkovarianzen ins Modell mit aufgenommen werden sollen, so erweitern sich die Formeln der verschiedenen Koeffizienten zu ω_T^* , ω_H^* und ω_S^* .

15.3.2.2 Omega-Koeffizienten der Subskalen

Ist man darüber hinaus daran interessiert, in welchem Ausmaß die Variablen der einzelnen Subskalenswerte des mehrdimensionalen Tests nicht nur durch den Generalfaktor determiniert sind, sondern auch spezifische Anteile enthalten, so müssen subskalenspezifische Koeffizienten bestimmt werden. Analog zur Aufteilung der Gesamtvarianz der Testwertvariablen des Gesamttests lässt sich auch für die Summenwerte jeder Subskala eine Aufteilung der Gesamtvarianz in einen Anteil erklärter Varianz, der auf den Generalfaktor zurückgeht, einen subskalenspezifischen Anteil und einen nicht erklärten Fehleranteil vornehmen.

Subskalenspezifische Koeffizienten im Rahmen eines mehrfaktoriellen Modells werden für jede Subskala getrennt geschätzt (vgl. Reise et al. 2013a). Hierbei handelt es sich um den Koeffizienten *Omega-Subskala-total* ($\omega_{\text{Skala-T}}$) zur Schätzung der gesamten wahren Varianz an der totalen Varianz der Werte einer Subskala, wobei sich $\omega_{\text{Skala-T}}$ wiederum additiv aufteilt in zwei Koeffizienten: *Omega-Subskala-hierarchisch* ($\omega_{\text{Skala-H}}$) schätzt den durch den Generalfaktor erklärten wahren Varianzanteil an der totalen Varianz der Subskala; *Omega-Subskala-spezifisch* ($\omega_{\text{Skala-S}}$) schätzt den durch den spezifischen Faktor erklärten wahren Varianzanteil an der totalen Varianz der Subskala unabhängig vom Generalfaktor (► Tab. 15.1). Zu beachten ist, dass diese drei Reliabilitätskoeffizienten für jede Subskala geschätzt werden müssen (► Exkurs 15.5).

**Subskalenspezifische
Omega-Koeffizienten**

Exkurs 15.5

Herleitung der Formeln für die Omega-Koeffizienten einer Subskala

Die erste Subskala (Skala 1) soll hier die ersten drei Items in Abb. 15.5 umfassen. Für diese Items resultieren die folgenden Messmodellgleichungen (► Kap. 24):

$$\begin{aligned} y_1 &= \lambda_{1g} \cdot \eta_{\text{Gen}} + \lambda_{11} \cdot \xi_1 + \varepsilon_1 \\ y_2 &= \lambda_{2g} \cdot \eta_{\text{Gen}} + \lambda_{21} \cdot \xi_1 + \varepsilon_2 \\ y_3 &= \lambda_{3g} \cdot \eta_{\text{Gen}} + \lambda_{31} \cdot \xi_1 + \varepsilon_3 \end{aligned} \quad (15.24)$$

Die Variable der Skalensummenwerte Y_{Skala1} erhält man, indem die drei Itemvariablen ungewichtet aufsummiert werden. Die drei Itemvariablen können durch ihre Messmodellgleichungen ersetzt werden:

$$\begin{aligned} Y_{\text{Skala1}} &= y_1 + y_2 + y_3 \\ &= (\lambda_{1g} \cdot \eta_{\text{Gen}} + \lambda_{11} \cdot \xi_1 + \varepsilon_1) \\ &\quad + (\lambda_{2g} \cdot \eta_{\text{Gen}} + \lambda_{21} \cdot \xi_1 + \varepsilon_2) \\ &\quad + (\lambda_{3g} \cdot \eta_{\text{Gen}} + \lambda_{31} \cdot \xi_1 + \varepsilon_3) \end{aligned} \quad (15.25)$$

Umsortieren, Zusammenfassen und Ausklammern der latenten Variablen führen zu der folgenden Gleichung, wobei die ersten zwei Summanden den vom Modell erklärten wahren Anteil bezeichnen und die Summe der Fehlervariablen den Fehleranteil:

$$Y_{\text{Skala1}} = \left(\sum_{i=1}^3 \lambda_{ig} \right) \cdot \eta_{\text{Gen}} + \left(\sum_{i=1}^3 \lambda_{i1} \right) \cdot \xi_1 + \sum_{i=1}^3 \varepsilon_i \quad (15.26)$$

Die Varianz der Subskala Y_{Skala1} lässt sich somit aufteilen in wahre Varianz und Fehlervarianz. Die wahre Varianz wiederum setzt sich aus der durch den Generalfaktor erklärten Varianz und der vom Generalfaktor unabhängigen subskalenspezifischen Varianz (Residualvarianz) zusammen:

$$\begin{aligned} \text{Var}(Y_{\text{Skala1}}) &= \left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) \\ &\quad + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) + \sum_{i=1}^3 \text{Var}(\varepsilon_i) \end{aligned} \quad (15.27)$$

Diese Varianzanteile werden für die Schätzung der subskalenspezifischen Omega-Koeffizienten verwendet.

Omega-Subskala-total ($\omega_{\text{Skala1-T}}$) berechnet sich nun als Quotient aus der gesamten wahren Varianz und der Gesamtvarianz der Subskala im Rahmen des Bifaktormodells:

$$\begin{aligned} \omega_{\text{Skala1-T}} &= \frac{\left[\left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \right]}{\left[\left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \right.} \\ &\quad \left. + \sum_{i=1}^3 \text{Var}(\varepsilon_i) \right]} \end{aligned} \quad (15.28)$$

Omega-Subskala-hierarchisch ($\omega_{\text{Skala1-H}}$) berechnet sich als Anteil der durch den Generalfaktor erklärten Varianz an der Gesamtvarianz der Subskala:

$$\begin{aligned} \omega_{\text{Skala1-H}} &= \frac{\left[\left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) \right]}{\left[\left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \right.} \\ &\quad \left. + \sum_{i=1}^3 \text{Var}(\varepsilon_i) \right]} \end{aligned} \quad (15.29)$$

Omega-Subskala-spezifisch ($\omega_{\text{Skala1-S}}$) berechnet sich als Anteil der durch den spezifischen Faktor erklärten Varianz an der Gesamtvarianz der Subskala und ist zugleich die Differenz zwischen Omega-Subskala-total und Omega-Subskala-hierarchisch:

$$\begin{aligned} \omega_{\text{Skala1-S}} &= \frac{\left[\left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \right]}{\left[\left(\sum_{i=1}^3 \lambda_{ig} \right)^2 \cdot \text{Var}(\eta_{\text{Gen}}) + \left(\sum_{i=1}^3 \lambda_{i1} \right)^2 \cdot \text{Var}(\xi_1) \right.} \\ &\quad \left. + \sum_{i=1}^3 \text{Var}(\varepsilon_i) \right]} \end{aligned} \quad (15.30)$$

Anmerkung: Wenn aus theoretischen Überlegungen Fehlerkovarianzen ins Modell mit aufgenommen werden sollen, so erweitern sich die Formeln der verschiedenen Koeffizienten zu $\omega_{\text{Skala1-T}}^*$, $\omega_{\text{Skala1-H}}^*$ und $\omega_{\text{Skala1-S}}^*$.

15.3.3 Empirisches Beispiel

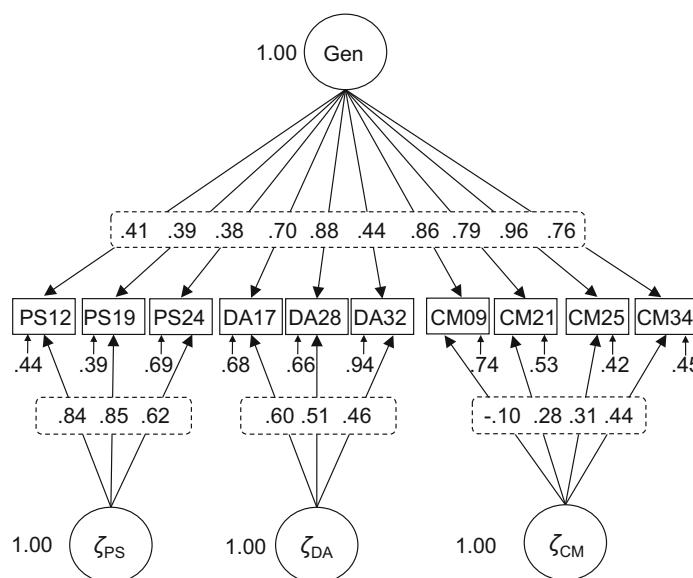
Das Konstrukt Perfektionismus ist mehrdimensional konzipiert. Mehrere Studien fanden z. B. übereinstimmend zwei Perfektionismusfaktoren, die den sechs Skalen der MPS-F zugrunde liegen und die positiv miteinander korrelieren. Diese beiden Faktoren werden als *Perfectionistic Strivings* und *Perfectionistic Concerns* bezeichnet.

Nachfolgend soll das Bifaktormodell aus □ Abb. 15.6 am Beispiel des mehrdimensionalen Konstrukts Perfektionismus zur Anwendung kommen. Zur Schätzung der Omega-Koeffizienten wird hier der zu didaktischen Zwecken verkürzte Test des Beispiels in ► Abschn. 15.2.2 verwendet. Die zehn ausgewählten Items laden im Bifaktormodell einerseits auf dem Generalfaktor Perfektionismus und andererseits zusätzlich auf einem der drei spezifischen Faktoren der Subskalen *Personal Standards* (PS, 3 Items), *Doubts about Actions* (DA, 3 Items) und *Concern over Mistakes* (CM, 4 Items). Inhaltlich werden die Subskalen DA und CM der perfektionistischen Sorge (*Perfectionistic Concerns*) und die Subskala PS dem perfektionistischen Streben (*Perfectionistic Strivings*) zugeordnet (vgl. u. a. Bieling et al. 2004; Cox et al. 2002; Frost et al. 1990; Stoeber und Damian 2014).

Die Schätzung der Omega-Koeffizienten für die drei Subskalen folgt in ► Abschn. 15.3.3.2. Im Unterschied zu ► Abschn. 15.2.2 wird das Item CM23 zur Vereinfachung hier nicht mehr verwendet, sodass keine korrelierten Messfehler berücksichtigt werden müssen. Zur Identifikation des Modells wurden die Varianzen der latenten Variablen jeweils auf den Wert eins fixiert. Die spezifischen Residualfaktoren ξ_{PS} , ξ_{DA} und ξ_{CM} beinhalten jeweils den Anteil an der jeweiligen Subskala PS, DA oder CM, der unabhängig vom Generalfaktor ist. Sie sollten deshalb nicht verwechselt werden mit den Subskalen, die sowohl Anteile des Generalfaktors als auch jeweils einen subskalenspezifischen Anteil erfassen.

15.3.3.1 Omega-Koeffizienten des Gesamttests

Die CFA ergab einen guten Modellfit für das Bifaktormodell mit $\chi^2(25) = 32.435$, $p = .146$, RMSEA = .034, CFI = .991, TLI = .984 und SRMR = .032. Die Reliabilitätskoeffizienten können nun anhand der unstandardisierten Parameter



■ Abb. 15.6 Bifaktormodell mit unstandardisierten Parameterschätzungen, in dem die Faktorladungen auf dem Generalfaktor (Gen) und die Faktorladungen auf den drei spezifischen Residualfaktoren ξ_{PS} , ξ_{DA} und ξ_{CM} hervorgehoben sind

Perfektionismus

Tabelle 15.4 Reliabilitätsschätzungen im Rahmen des Bifaktormodells mit Konfidenzintervallen und Anteilen an der wahren Varianz des Gesamttests für die gekürzten Subskalen PS, DA und CM der MPS-F

Modell	Reliabilitäts-schätzung	95 %-Konfidenz-intervall	Anteil an der wahren Varianz
Gesamttest im Rahmen des Bifaktormodells	$\omega_T = .897$	[.876; .915]	
	$\omega_H = .748$	[.687; .801]	.748/.897 = 83.39 %
	$\omega_S = .149$	[.109; .200]	.149/.897 = 16.61 %

ω_T = Omega-total, ω_H = Omega-hierarchisch, ω_S = Omega-spezifisch

(Abb. 15.6) berechnet werden (► Beispiel 15.2). Unter Verwendung der geschätzten Standardfehler der Koeffizienten lassen sich zusätzlich auch deren Konfidenzintervalle berechnen. Außerdem werden die prozentualen Anteile von ω_H und ω_S an ω_T bestimmt. Diese Quotienten geben jeweils an, wie groß der Anteil der wahren Varianz der Koeffizienten an der gesamten wahren Varianz ist (Tab. 15.4).

Beispiel 15.2: Empirisches Beispiel

Berechnung der Omega-Koeffizienten für den Gesamttest

Für die nachfolgenden Berechnungen der Omega-Koeffizienten werden die unstandardisierten Parameter verwendet:

Omega-total

Omega-total wird berechnet als Quotient aus der wahren Varianz, die sowohl auf den Generalfaktor als auch auf die drei spezifischen Faktoren zurückzuführen ist, und der Gesamtvarianz.

$$\begin{aligned} \omega_T &= \frac{(.41 + .39 + .38 + .70 + .88 + .44 + .86 + .79 + .96 + .76)^2 \cdot 1.0}{(.41 + .39 + .38 + .70 + .88 + .44 + .86 + .79 + .96 + .76)^2 \cdot 1.0} \\ &\quad + (.84 + .85 + .62)^2 \cdot 1.0 + (.60 + .51 + .46)^2 \cdot 1.0 \\ &\quad + (-.10 + .28 + .31 + .44)^2 \cdot 1.0 \\ &= \frac{43.165 + 5.336 + 2.465 + .865}{43.165 + 5.336 + 2.465 + .865 + 5.94} = .897 \end{aligned} \quad (15.31)$$

Omega-hierarchisch

Omega-hierarchisch wird berechnet als Quotient aus der wahren Varianz, die auf den Generalfaktor zurückzuführen ist, und der Gesamtvarianz:

$$\omega_H = \frac{43.165}{43.165 + 5.336 + 2.465 + .865 + 5.94} = .748 \quad (15.32)$$

Omega-spezifisch

Omega-spezifisch wird berechnet als Quotient aus der Summe der wahren Varianzanteile, die auf alle drei spezifischen Residualfaktoren zurückzuführen sind, und der Gesamtvarianz:

$$\omega_S = \frac{5.336 + 2.465 + .865}{43.165 + 5.336 + 2.465 + .865 + 5.94} = .149 \quad (15.33)$$

Um Aussagen über die Präzision der Punktschätzungen der Reliabilitätskoeffizienten zu erhalten, werden zusätzlich Intervallschätzungen in Form von 2-seitigen asymmetrischen 95 %-Konfidenzintervallen vorgenommen. Die Reliabilitätsschätzungen, die 2-seitigen 95 %-Konfidenzintervalle und die Anteile an der wahren Varianz sind in □ Tab. 15.4 aufgelistet.

Wie die Omega-Koeffizienten für das Gesamtmodell zeigen, ist ω_H (.748) deutlich höher als ω_S (.149): Circa 83 % der wahren Varianz der über alle Items aufsummierten Testwerte können auf den Generalfaktor zurückgeführt werden und nur ca. 17 % auf alle spezifischen Residualfaktoren zusammen. Die 2-seitigen 95 %-Konfidenzintervalle sind relativ schmal und zeigen, dass die Schätzungen vertrauenswürdig sind.

15.3.3.2 Omega-Koeffizienten der Subskalen

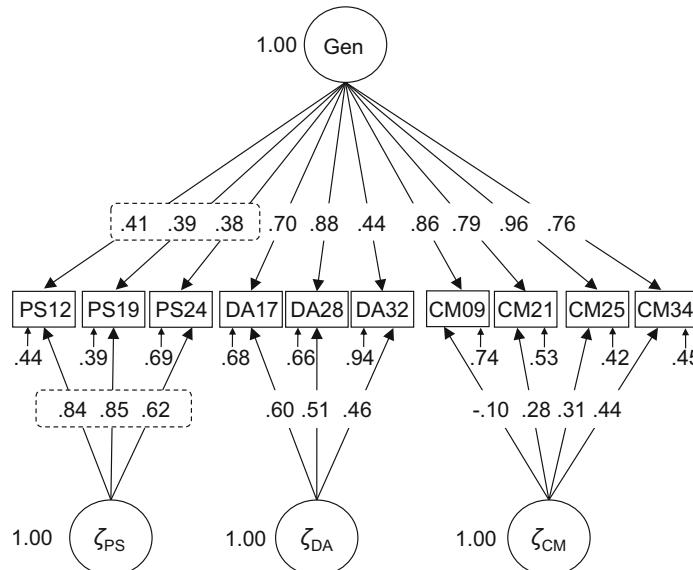
Um nun zu überprüfen, ob der relativ niedrige Wert von Omega-spezifisch des Gesamttests mit $\omega_S = .149$ [.109; .200] durch alle Subskalen gleichermaßen oder im Wesentlichen nur durch eine Subskala bedingt ist, können die subskalenspezifischen Omega-Koeffizienten geschätzt werden. In ► Beispiel 15.3 soll deren Berechnung exemplarisch anhand der Subskala PS demonstriert werden; die Berechnungen für die Subskalen DA und CM können in analoger Weise erfolgen. Die unstandardisierten Parameterschätzungen für die Subskala PS sind □ Abb. 15.7 zu entnehmen.

Subskala Personal Standards (PS)

Beispiel 15.3: Empirisches Beispiel

Berechnung der Omega-Koeffizienten für die Subskala Personal Standards (PS)

Für das empirische Beispiel sollen exemplarisch nur die Omega-Koeffizienten für die Subskala PS im Rahmen des Bifaktormodells bestimmt werden. Hierzu wird zunächst die Summenvariable Y_{PS} der Subskala PS als ungewichtete Summe der Antwortvariablen der drei Items PS12, PS19 und PS24 (y_{PS12} , y_{PS19} und y_{PS24}) ge-



■ Abb. 15.7 Bifaktormodell mit unstandardisierten Parameterschätzungen, in dem die Faktorladungen der drei Items der Subskala *Personal Standards* (PS) auf dem Generalfaktor (Gen) und die Faktorladungen auf dem spezifischen Residualfaktor ζ_{PS} hervorgehoben sind

bildet wird:

$$Y_{PS} = y_{PS12} + y_{PS19} + y_{PS24} \quad (15.34)$$

Die Messmodellgleichungen mit den geschätzten Parametern aus Abb. 15.7 lauten dann wie folgt:

$$\begin{aligned} y_{PS12} &= .41 \cdot \eta_{Gen} + .84 \cdot \zeta_{PS} + \varepsilon_{PS12} \\ y_{PS19} &= .39 \cdot \eta_{Gen} + .85 \cdot \zeta_{PS} + \varepsilon_{PS19} \\ y_{PS24} &= .38 \cdot \eta_{Gen} + .62 \cdot \zeta_{PS} + \varepsilon_{PS24} \end{aligned} \quad (15.35)$$

Einsetzen der Messmodellgleichungen in Gl. (15.34) führt zu:

$$\begin{aligned} Y_{PS} &= (.41 \cdot \eta_{Gen} + .84 \cdot \zeta_{PS} + \varepsilon_{PS12}) + (.39 \cdot \eta_{Gen} + .85 \cdot \zeta_{PS} + \varepsilon_{PS19}) \\ &\quad + (.38 \cdot \eta_{Gen} + .62 \cdot \zeta_{PS} + \varepsilon_{PS24}) \end{aligned} \quad (15.36)$$

Umsortieren, Zusammenfassen und Ausklammern der latenten Variablen führen zu der folgenden Gleichung, wobei die ersten beiden Summanden den wahren Anteil bezeichnen und der dritte Summand den Fehleranteil:

$$Y_{PS} = (.41 + .39 + .38) \cdot \eta_{Gen} + (.84 + .85 + .62) \cdot \zeta_{PS} + (\varepsilon_{PS12} + \varepsilon_{PS19} + \varepsilon_{PS24}) \quad (15.37)$$

Zur Berechnung der Varianz der Subskalenwerte müssen die Faktorladungen auf dem Generalfaktor aufaddiert und quadriert, die Faktorladungen auf dem jeweiligen spezifischen Residualfaktor aufaddiert und quadriert und die resultierenden Werte mit den Varianzen der latenten Variablen (hier jeweils 1.0) multipliziert werden. Des Weiteren muss noch die Summe der Fehlervarianzen der Items hinzugefügt werden. Die Varianz der Variablen Y_{PS} der Subskala PS setzt sich nun aus den folgenden drei Varianzanteilen zusammen: durch den Generalfaktor erklärter Anteil (1.392), durch den spezifischen Residualfaktor erklärter Anteil (5.336) und unerklärter Anteil (1.52):

$$\begin{aligned} Var(Y_{PS}) &= (.41 + .39 + .38)^2 \cdot 1.0 + (.84 + .85 + .62)^2 \cdot 1.0 \\ &\quad + (.44 + .39 + .69) \\ &= 1.392 + 5.336 + 1.52 \end{aligned} \quad (15.38)$$

Omega-PS-total

Omega-PS-total wird berechnet als Quotient aus der gesamten wahren Varianz der Subskala PS (Summe der durch den Generalfaktor erklärten Varianz und der durch den spezifischen Residualfaktor erklärten Varianz) und der Gesamtvarianz der Subskala:

$$\omega_{PS-T} = \frac{1.392 + 5.336}{1.392 + 5.336 + 1.52} = .816 \quad (15.39)$$

Omega-PS-hierarchisch

Omega-PS-hierarchisch wird berechnet als Quotient aus der wahren Varianz der Subskala PS, die auf den Generalfaktor zurückgeführt werden kann, und der Gesamtvarianz der Subskala:

$$\omega_{PS-H} = \frac{1.392}{1.392 + 5.336 + 1.52} = .169 \quad (15.40)$$

Omega-PS-spezifisch

Omega-PS-spezifisch wird berechnet als Quotient aus der wahren Varianz der Subskala PS, die auf den spezifischen Faktor zurückgeführt werden kann, und der Gesamtvarianz der Subskala:

$$\omega_{PS-S} = \frac{5.336}{1.392 + 5.336 + 1.52} = .647 \quad (15.41)$$

■ **Tabelle 15.5** Reliabilitäts schätzungen der Subskalen PS, DA und CM im Rahmen des Bifaktormodells mit Konfidenzintervallen und Anteilen an der wahren Varianz der jeweiligen Subskala

Modell	Reliabilitäts schätzung	95 %-Konfidenz intervall	Anteil an der wahren Varianz der Subskala
Subskala PS im Rahmen des Bifaktormodells	$\omega_{PS-T} = .814$	[.765; .855]	
	$\omega_{PS-H} = .168$	[.088; .298]	.168/.814 = 20.64 %
	$\omega_{PS-S} = .646$	[.535; .744]	.646/.814 = 79.36 %
Subskala DA im Rahmen des Bifaktormodells	$\omega_{DA-T} = .742$	[.683; .793]	
	$\omega_{DA-H} = .464$	[.322; .612]	.464/.742 = 62.53 %
	$\omega_{DA-S} = .278$	[.156; .445]	.278/.742 = 37.47 %
Subskala CM im Rahmen des Bifaktormodells	$\omega_{CM-T} = .851$	[.805; .888]	
	$\omega_{CM-H} = .791$	[.504; .934]	.791/.851 = 92.95 %
	$\omega_{CM-S} = .060$ n. s.	[.002; .681]	.060/.851 = 7.05 %

n. s. = nicht signifikant

Die Ergebnisse der Subskala PS sowie die analog berechneten Ergebnisse der Subskalen DA und CM sind in ■ Tab. 15.5 aufgeführt (die geringfügigen Unterschiede zu den Ergebnissen in Gln. (15.39) bis (15.41) beruhen auf Rundungsfehlern, da bei der ausführlichen Berechnung nur zwei Dezimalstellen für die Schätzungen der Faktorladungen und Fehlervarianzen verwendet wurden). Wie die Ergebnisse zeigen (■ Tab. 15.5), hängt die Reliabilität der drei Subskalen (Omega-PS-total, Omega-DA-total und Omega-CM-total) in unterschiedlichem Ausmaß vom Generalfaktor ab. Die wahre Varianz der Subskalen CM und DA kann fast ausschließlich (CM) oder zu einem bedeutenden Anteil (DA) auf den Generalfaktor zurückgeführt werden, wie anhand der Koeffizienten $\omega_{CM-H} = .791$ und $\omega_{DA-H} = .464$ zu erkennen ist: Der Generalfaktor erklärt ca. 93 % ($.791/.851 = .92.95\%$) der gesamten wahren Varianz der Subskala CM und ca. 63 % ($.464/.742 = .62.53\%$) der gesamten wahren Varianz der Subskala DA.

Im Gegensatz dazu beruht die wahre Varianz der Subskala PS fast ausschließlich auf der subskalenspezifischen Varianz, während der Generalfaktor sich nur relativ wenig auswirkt. Dies zeigt sich am Koeffizienten $\omega_{PS-H} = .168$ [.088; .298]: Der Generalfaktor erklärt nur ca. 21 % der gesamten wahren Varianz der Subskala ($.168/.814 = 20.64\%$). Der subskalenspezifische Faktor erklärt dagegen mit einem Koeffizienten von $\omega_{PS-S} = .646$ [.535; .744] ca. 79 % der wahren Varianz von PS ($.646/.814 = 79.36\%$).

Der geringe, durch alle spezifischen Faktoren gemeinsam erklärte Anteil an der Varianz des Gesamttests von ca. 17 % ($.149/.897 = 16.61\%$; vgl. ■ Tab. 15.4) resultiert somit aus den geringen subskalenspezifischen Varianzanteilen der Subskalen DA und CM. Bei der Subskala DA mit $\omega_{DA-S} = .278$ werden nur 37 % ($.278/.742 = 37.47\%$) der wahren Varianz dieser Subskala durch die subskalenspezifische Varianz erklärt, und bei der Subskala CM sind dies nur ca. 7 % ($.060/.851 = 7.05\%$). Der Koeffizient ω_{CM-S} mit einem Wert von .060 [.002; .681] ist nicht signifikant von null verschieden ($p = .55$), sodass die gesamte wahre Varianz dieser Subskala durch den Generalfaktor bestimmt wird.

Das breite Konfidenzintervall von $\omega_{CM-S} = .060$ ist darauf zurückzuführen, dass die Schätzung mit einer großen Unsicherheit behaftet ist: Alle Faktorladungen der Itemvariablen auf dem spezifischen Residualfaktor ζ_{CM} , die in die Gleichung des Koeffizienten ω_{CM-S} eingehen, sind nicht signifikant. Damit handelt es sich bei

Vanishing factor/collapsing factor

diesem Faktor um einen sog. „vanishing factor“ (Eid et al. 2017a) oder „collapsing factor“ (Geiser et al. 2015). Solche anomalen Ergebnisse können in Bifaktormodellen z. B. erwartet werden, wenn zwei Faktoren sehr hoch miteinander korrelieren, sodass einer der beiden Faktoren über die gemeinsame Varianz, die durch den Generalfaktor erklärt wird, hinausgehend keine spezifische Varianz mehr aufweist (s. auch Eid et al. 2008). Als Alternative könnte das Bifaktor-($S - 1$)-Modell (► Abschn. 15.5.2) spezifiziert werden, bei dem ein spezifischer Residualfaktor S weniger modelliert wird, als Subskalen vorhanden sind (Eid et al. 2017a). Für das Beispiel bedeutet das Ergebnis, dass die Berechnung des Koeffizienten ω_{CM-S} sowie des Konfidenzintervalls in diesem Fall nicht sinnvoll ist.

Bildung eines Gesamttestwertes

Anhand der vorliegenden Ergebnisse kann nun die Frage beantwortet werden, ob ein Gesamttestwert über alle drei Subskalen hinweg gebildet werden sollte. Das ist hier nicht der Fall, da der Generalfaktor nur wenig zum Subskalenwert von PS beiträgt, zu den Subskalenwerten von DA und CM hingegen viel, wobei CM praktisch nur aus dem Generalfaktor besteht. Somit würde eine Verwendung des Gesamttestwertes über alle drei Subskalen hinweg insoweit irreführend sein, als mit diesem Wert im Wesentlichen die perfektionistische Sorge erfasst würde (vgl. Gädé et al. 2017). Die Aufteilung in perfektionistisches Streben (Skala PS) und perfektionistische Sorge (Skalen DA und CM) kann somit auch durch die Ergebnisse dieser sehr verkürzten MPS-F gestützt werden.

15.4 Omega-Koeffizienten im Rahmen weiterer Faktormodelle

Neben dem Bifaktormodell werden auch weitere CFA-Modelle häufig angewandt, u. a. das Faktormodell höherer Ordnung, in dem die Kovarianzen zwischen den Faktoren erster Ordnung durch einen Faktor zweiter Ordnung erklärt werden (s. dazu ► Kap. 24), und das Modell mit korrelierten Faktoren, das so viele Faktoren umfasst, wie es Subskalen gibt. Weitere Modelle, die häufig angewandt werden, sind das Mehrebenenmodell und das latente Klassenmodell. Alle Modelle ermöglichen eine modellbasierte Schätzung der Reliabilität.

- !** Sind in diesen Modellen Fehlerkovarianzen enthalten, die theoretisch begründbar, aber inhaltlich nicht relevant sind, so sollte jeweils die Gleichung zur Berechnung von Bollens Omega verwendet werden, in der die Fehlerkovarianzen in den Nenner aufgenommen werden.

■ ■ Faktormodell höherer Ordnung

In diesem Modell werden die Kovarianzen zwischen den Faktoren erster Ordnung durch einen übergeordneten gemeinsamen Faktor zweiter Ordnung erklärt (vgl. ► Kap. 24). Omega-Koeffizienten können in diesem Fall analog zum Bifaktormodell geschätzt werden, wobei Omega-total den Anteil an der wahren Varianz an der Gesamtvarianz der Items erklärt, der sowohl auf dem gemeinsamen Faktor zweiter Ordnung als auch auf den spezifischen Anteilen (Residuen) der Faktoren erster Ordnung beruht. Omega-hierarchisch beinhaltet dann den Anteil der wahren Varianz an der Gesamtvarianz, der auf den gemeinsamen übergeordneten Faktor zurückgeführt werden kann, und Omega-spezifisch den Anteil, der auf alle spezifischen Faktoren gemeinsam zurückgeführt werden kann. Ebenso wie im Bifaktormodell können die Reliabilitätskoeffizienten auch für jede Subskala im Rahmen des Modells geschätzt werden.

■ ■ Faktormodell mit korrelierten Faktoren

Sind die Faktoren eines Modells mit korrelierten Faktoren (vgl. ► Kap. 24) nur gering miteinander korreliert oder ist bei höher korrelierten Faktoren aufgrund theoretischer Überlegungen ein übergeordneter gemeinsamer Faktor nicht begründbar

Omega-Koeffizienten für ein Faktormodell höherer Ordnung**Omega-Koeffizienten für ein korreliertes Faktormodell**

15.5 · Bewertung der modellbasierten Reliabilitätsschätzung

(z. B. bei perfektionistischem Streben und perfektionistischer Sorge), so sollte die Reliabilität der einzelnen Skalen getrennt anhand von McDonalds Omega für ein-dimensionale Modelle geschätzt werden. Ansonsten sollte die Reliabilität anhand eines Bifaktormodells oder eines Faktormodells höherer Ordnung bestimmt werden.

■■ Mehrebenenmodell

Geldhof et al. (2014) konnten zeigen, dass Omega-Koeffizienten auch in Mehr-ebenenmodellen auf beiden Ebenen geschätzt werden können. Allerdings müssen die Stichproben sowohl auf Ebene 1 als auch auf Ebene 2 ausreichend groß und die Intraklassenkorrelation $\geq .05$ sowie die Faktorladungen auf beiden Ebenen hinreichend hoch sein, um unverzerrte ebenenspezifische Omega-Koeffizienten zu erhalten.

Ebenenspezifische Omega-Koeffizienten

■■ Latentes Klassenmodell

Omega-Koeffizienten können auch in einem Mischverteilungsmodell mit einer unbekannten Anzahl latenter Klassen geschätzt werden (Raykov und Marcoulides 2015). Dafür wird zunächst die Anzahl der latenten Klassen bestimmt und danach die klassenspezifischen Reliabilitätskoeffizienten der Testwertvariablen geschätzt.

Klassenspezifische Omega-Koeffizienten

15.5 Bewertung der modellbasierten Reliabilitätsschätzung

Die modellbasierte Reliabilitätsschätzung weist einige Vorteile gegenüber der klassischen Reliabilitätsschätzung auf, von denen nachfolgend einige aufgeführt sind (► Abschn. 15.5.1). Probleme, die sich auf die Reliabilitätsschätzung auswirken können, werden ebenfalls thematisiert (► Abschn. 15.5.2).

15.5.1 Vorteile gegenüber der klassischen Reliabilitätsschätzung

1. *Ein- und mehrdimensionale Modelle:* Die modellbasierte Reliabilitätsschätzung liefert differenzierte Informationen bezüglich der psychometrischen Güte eines Gesamttests und seiner Subskalen. Dies ist ein besonderer Vorteil gegenüber der klassischen Reliabilitätsschätzung, die in der Regel auf eindimensionalen Modellen mit relativ strengen Annahmen beruht. Anhand der Omega-Koeffizienten mehrdimensionaler Modelle kann geprüft werden, ob alle Itemwerte zu einem Testwert aufsummiert werden können und ob die Subskalen eines mehrdimensionalen Tests über den Generalfaktor hinausgehend genügend eigenständige Varianz aufweisen, um die Bildung von Subskalenwerten zu rechtfertigen.
2. *Überprüfbarkeit der Modellannahmen:* Ein weiterer Vorteil gegenüber der klassischen Reliabilitätsschätzung liegt darin, dass das den Koeffizienten zugrunde liegende Modell mit den jeweils implizierten Modellannahmen anhand der CFA explizit überprüft wird. Da zur Beurteilung der Modellgüte klare Entscheidungsregeln existieren, kann ein Modell somit auch verworfen werden. Eine solche Überprüfung der Voraussetzungen wird bei den klassischen Methoden zumeist leider nicht durchgeführt, obwohl dies möglich wäre.
3. *Verwendung unterschiedlicher Schätzmethoden:* Des Weiteren können zur Schätzung der Parameter eines CFA-Modells unterschiedliche Schätzmethoden verwendet und dabei auch fehlende Werte berücksichtigt werden. Die Wahl der Schätzmethode hängt insbesondere von der Anzahl der Antwortkategorien sowie von der Verteilungsform der Itemvariablen ab (vgl. ► Kap. 24).

Nutzen bei ein- und mehrdimensionalen Tests

Klare Entscheidungsregeln

Wahl einer geeigneten Schätzmethode

Adäquate Schätzung der Konfidenzintervalle

4. *Adäquate Schätzung der Konfidenzintervalle:* Ein weiterer Vorteil liegt in der adäquaten Schätzung der Konfidenzintervalle. Für klassische Maße sind diese in der Regel nicht zuverlässig schätzbar, da keine präzisen Standardfehlerschätzungen vorliegen. Die Punktschätzungen der Reliabilitätskoeffizienten besitzen jedoch nur eine begrenzte Aussagekraft über die Populationswerte, daher sind Konfidenzintervalle unverzichtbar, die plausible Bereiche der Populationswerte liefern.
5. *Berücksichtigung von Methodeneffekten:* Die modellbasierten Omega-Koeffizienten können Methodeneffekte berücksichtigen, indem diese als Fehlerkovarianzen ins Modell aufgenommen und als Messfehler behandelt werden, wenn sie inhaltlich nicht bedeutsam sind (Bollens Omega, □ Tab. 15.2). Sind die Fehlerkovarianzen aber inhaltlich relevant oder wesentlich für die Messungen, so können die Kovarianzen zwischen den Fehlervariablen als weitere erklärte Varianz in den Zähler der Formel aufgenommen werden (vgl. Bollen 1989, S. 218; Eid et al. 1994; Geiser und Lockhart 2012; Steyer et al. 2015).

Berücksichtigung von Methodeneffekten

Hinzunahme weiterer Items

Stichprobengröße

Abhängigkeit vom verwendeten Modell

15.5.2 Probleme der modellbasierten Reliabilitätsschätzung

1. *Abhängigkeit von der Itemanzahl:* Alle Omega-Koeffizienten sind ebenso wie Cronbachs Alpha abhängig von der Anzahl der Items (vgl. u. a. Rodriguez et al. 2016): Mit zunehmender Itemanzahl erhöht sich die Reliabilität der Testwertvariablen. Die Hinzunahme weiterer Items zugunsten einer höheren Reliabilität sollte jedoch nicht unkritisch erfolgen, da sich dadurch die Bedeutung des gemessenen Konstrukts verändern könnte und die Konstruktion eindimensionaler Tests oder Subskalen möglicherweise erschwert wird.
2. *Notwendigkeit relativ großer Stichproben:* Bei Verwendung der CFA werden – wie bei allen latenten Modellen – relativ große Stichproben benötigt, um die Parameter und damit auch die Reliabilitätskoeffizienten zuverlässig schätzen zu können. Eine präzise Schätzung ist nötig, um ein korrektes Konfidenzintervall bestimmen zu können. Die benötigte Stichprobengröße hängt von mehreren Einflussgrößen ab, sodass keine klaren Richtlinien dafür gegeben werden können, wie groß eine Stichprobe generell sein sollte. Es ist aber bekannt, dass bei einer größeren Anzahl von Items, bei hohen Faktorladungen, bei höheren Faktorkorrelationen und bei normalverteilten Variablen schon anhand kleinerer Stichproben ($N < 200$) zuverlässige Schätzungen möglich sind (vgl. Wolf et al. 2013).
3. *Unplausible Ergebnisse:* Bifaktormodelle und Faktormodelle höherer Ordnung werden oftmals zur Schätzung der Reliabilitätskoeffizienten mehrdimensionaler Modelle verwendet. Voraussetzung hierfür – ebenso wie bei eindimensionalen Modellen – ist, dass das Modell zu den Daten passt. Trotz guten Modellfits können jedoch unplausible Ergebnisse auftreten, wenn z. B. die Varianz eines spezifischen Faktors nicht signifikant von null verschieden ist oder die Itemvariablen nicht auf dem intendierten spezifischen Faktor oder dem Generalfaktor laden. Da sich diese Probleme auf die Reliabilitätskoeffizienten auswirken, sollten unter Umständen alternative Modelle in Betracht gezogen werden, z. B. das Bifaktor-($S - 1$)-Modell, bei dem ein spezifischer Residualfaktor S weniger modelliert wird, als Subskalen vorhanden sind. In diesem Modell laden die zu diesem Faktor gehörigen Itemvariablen nur noch auf dem Generalfaktor und bestimmen damit die Bedeutung dieses Faktors (Eid et al. 2017a; Geiser et al. 2015). Damit wird deutlich, dass die Reliabilitätskoeffizienten nur bezogen auf das verwendete, passende und plausible Modell interpretiert werden dürfen.

15.6 Reliabilitätsschätzung ordinalskalierter Variablen

Die bisher beschriebenen Methoden der modellbasierten Reliabilitätsschätzung setzen streng genommen voraus, dass die Antworten auf die Items, d. h. die Itemvariablen, metrisch (in der Regel intervallskaliert) sind. In der Praxis ist dies aber oftmals nicht der Fall. Itemvariablen sind häufig ordinalskaliert, d.h., es liegen kategoriale Variablen mit geordneten Antwortkategorien vor, die – wenn sie fünf oder mehr Abstufungen aufweisen – als näherungsweise kontinuierlich angenommen werden können. Doch auch für die Reliabilitätsschätzung ordinalskalierter Variablen mit zwei oder mehr Antwortkategorien stehen verschiedene Methoden zur Verfügung, die nachfolgend kurz dargestellt werden sollen.

15.6.1 Variablen mit fünf oder mehr Antwortkategorien

Itemvariablen mit wenigen geordneten Antwortkategorien können nicht als kontinuierlich eingestuft werden, sondern als kategorial mit geordneten Antwortkategorien. Ordinalskalierte Itemvariablen werden aber oftmals als näherungsweise kontinuierlich angesehen, wenn sie mindestens fünf Abstufungen aufweisen (Raykov und Marcoulides 2011; Rhemtulla et al. 2012). In diesem Fall kann eine CFA durchgeführt werden, wobei die Parameter anhand von Schätzmethoden für kontinuierliche Variablen (z. B. Maximum-Likelihood- [ML-] oder robuste Maximum-Likelihood-Methode [MLR-Methode]; vgl. ▶ Kap. 24) und die Omega-Koeffizienten unter Verwendung dieser Methoden geschätzt werden.

Annähernd kontinuierliche Itemvariablen

15.6.2 Variablen mit zwei bis vier Antwortkategorien

Itemvariablen mit zwei bis vier geordneten Antwortkategorien können streng genommen nicht mehr als kontinuierlich angesehen werden. In diesem Fall kann eine von Green und Yang (2009) vorgestellte Methode zur Reliabilitätsschätzung verwendet werden, die auf einem nichtlinearen Faktormodell beruht. Diese Methode führt – auch bei moderat schieverteilten Variablen – zu zuverlässigen Ergebnissen (Yang und Green 2015). Sie benötigt jedoch relativ große Stichproben, wenn die zugrunde liegende Verteilung nicht normal ist oder die Schwellenwerte heterogen sind (eine Erläuterung der Schwellenwerte findet sich z. B. in Bollen 1989, S. 439 ff., oder in Reinecke 2014, S. 36 ff.).

Methoden auf Basis nichtlinearer Faktormodelle

Eine Alternative für Itemvariablen mit wenigen Antwortkategorien basierend auf der Bayes'schen Schätzmethode wurde von Yang und Xia (2019) entwickelt. Diese Methode ist bereits für kleine Stichproben von $N = 100$ geeignet. Diese neueren Methoden sind allerdings (noch) recht kompliziert in der Anwendung und somit für die Praxis momentan nur bedingt geeignet.

Bayes'sche Schätzmethode

15.6.3 Variablen mit zwei Antwortkategorien

Weisen die beobachtbaren Itemvariablen nur zwei geordnete Antwortkategorien auf, sind sie also dichotom, so kann eine Methode von Raykov et al. (2010) verwendet werden, die auf dem 1PL- oder dem 2PL-Modell der Item-Response-Theorie (IRT, einparametrisches bzw. zweiparametrisches logistisches Modell; vgl. ▶ Kap. 16) beruht. Anhand eines Modelltests kann die Struktur des Modells überprüft werden. Modellbasierte Reliabilitätsschätzungen sind anhand dieser Methode für einzelne Itemvariablen und für die Testwertvariable sowohl als Punkt- als auch als Intervallschätzung möglich.

Methode von Raykov et al. (2010)

15.6.4 Item-Parcels

Methode von Raykov und Marcoulides (2011)

Raykov und Marcoulides (2011, S. 176 ff.) schlagen für Itemvariablen mit geordneten Antwortkategorien eine vereinfachte Methode vor, die auf der Aufsummierung einzelner Itemvariablen zu mehreren „Päckchen“ (Teilskalen) beruht, die auch als „Parcels“ oder „Item-Parcels“ bezeichnet werden (vgl. Bandalos 2002, 2008; Little et al. 2013; s. auch ► Kap. 26). Anstelle der einzelnen kategorialen Itemvariablen können diese Parcels in der CFA als näherungsweise kontinuierliche Indikatoren verwendet werden. Der Vorteil der Parcels besteht u. a. darin, dass sie im Vergleich zu Itemvariablen eine größere Anzahl an Abstufungen aufweisen und zudem eher normalverteilt sind. Die Grundidee ist hier, dass ein Testwert nicht nur über die Addition aller Itemwerte gebildet werden kann, sondern ebenso über die Aufsummierung der Parcel-Werte. Auch bei Verwendung der Parcels müssen die Voraussetzungen und Modellannahmen der Reliabilitätskoeffizienten zunächst geprüft werden. Sind die Annahmen erfüllt, erfolgt die Reliabilitätsschätzung wie beschrieben über die Formeln der verschiedenen Omega-Koeffizienten. Eine systematische Überprüfung dieser Methode steht aber noch aus.

15.6.5 IRT-Modelle

Reliabilität latenter Personenwerte

Zur Schätzung der Reliabilität anhand von Modellen mit kategorialen, insbesondere zweiwertigen Antwortvariablen eignen sich vor allem IRT-Modelle. Im Rahmen der IRT wird die Reliabilität der geschätzten latenten Personenwerte modellbasiert bestimmt und nicht die Reliabilität der messfehlerbehafteten Testwertvariablen, sodass die Reliabilität in der IRT etwas anders definiert ist als in der Klassischen Testtheorie (KTT, vgl. ► Kap. 19; Eid und Schmidt 2014). Das Reliabilitätsmaß der IRT gibt an, inwieweit sich Unterschiede der geschätzten Personenwerte auf wahre Unterschiede zwischen Personen zurückführen lassen; dieses Reliabilitätsmaß kann übrigens auch im Rahmen der KTT bestimmt werden (► Kap. 12; Eid und Schmidt 2014, S. 292). Dafür wird die Testinformationsfunktion verwendet, die auf den einzelnen Iteminformationen beruht. Die Iteminformationen geben jeweils an, welchen Beitrag ein Item zur Schätzgenauigkeit eines latenten Personenwertes leistet (vgl. z. B. Bandalos 2018, S. 429; ► Kap. 16 und 19). Da die Reliabilität in der IRT in Abhängigkeit von den Ausprägungen der latenten Variablen variiert, wurden auch sog. „marginale Reliabilitätskoeffizienten“ entwickelt, die als Kennwerte der durchschnittlichen Messgenauigkeit eines Tests dienen können (vgl. ► Kap. 19).

15.7 Erste Empfehlungen zur Beurteilung der Omega-Koeffizienten

Omega-Koeffizienten $\geq .70$, besser $\geq .80$ für eindimensionale Tests

Für eindimensionale Tests können hinsichtlich der erstrebenswerten Höhe der Omega-Koeffizienten die Empfehlungen für die klassischen Reliabilitätsmaße verwendet werden, indem die Koeffizienten ω und ω^* – abhängig von der Heterogenität bzw. Homogenität der Items und der diagnostischen Zielsetzung – mindestens einen Wert von .70, besser .80 oder .90 aufweisen sollten (vgl. Eid und Schmidt 2014, S. 286; ► Kap. 14). Zusätzlich sollten die Koeffizienten ein kleines Konfidenzintervall aufweisen.

Keine verbindlichen Richtlinien für mehrdimensionale Tests

Für mehrdimensionale Tests liegen bisher noch keine verbindlichen Richtlinien zur Beurteilung der Omega-Koeffizienten vor. Somit ist nicht klar, wie groß ein Omega-Koeffizient sein sollte, um das gemessene Konstrukt als „im Wesentlichen“ eindimensional einstuften zu können, bzw. wie groß die über den Generalfaktor hinausgehende reliable spezifische Varianz der Subskalen sein muss, damit die

15.8 · Zusammenfassung

Subskalen als bedeutsame eigenständige Facetten des gemeinsamen Konstrukts angesehen werden können.

Aufgrund der Angaben in der Literatur und eigener Erfahrungen wollen wir erste Empfehlungen zur Beurteilung der Omega-Koeffizienten für mehrdimensionale Tests formulieren. Die Koeffizienten ω_T und $\omega_{Skala-T}$ sollten nach gängiger Praxis mindestens einen Wert von .80 aufweisen mit einem schmalen Konfidenzintervall, wenn der untersuchte Test relativ homogene Items enthält.

Soll der Test im Wesentlichen ein eindimensionales Konstrukt erfassen, soll also ein Gesamttestwert über alle Items und über alle Subskalen hinweg gebildet werden, so sollte ω_H bzw. $\omega_{Skala-H}$ mindestens einen Wert von .50, besser Werte von .70 oder .75 aufweisen (vgl. Reise et al. 2013a; Reise et al. 2013b).

Um von einer ausreichend hohen subskalenspezifischen Varianz unabhängig vom Generalfaktor sprechen zu können, sollte die Punktschätzung von $\omega_{Skala-S}$ mindestens .30 (Smits et al. 2015) und signifikant von null verschieden sein. Die untere Grenze des Konfidenzintervalls sollte dabei nicht zu nahe am Wert null liegen, wobei ein Wert von etwa .10 in der Regel nicht unterschritten werden sollte. Die untere Grenze des Konfidenzintervalls kann anhand des verwendeten Programms (z. B. Mplus oder R) auf Signifikanz getestet werden.

$$\omega_T \text{ und } \omega_{Skala-T} \geq .80$$

$$\omega_H \text{ und } \omega_{Skala-H} \geq .50, \text{ besser } \geq .70$$

$$\omega_{Skala-S} \geq .30$$

15.8 Zusammenfassung

Die modellbasierten Methoden der Reliabilitätschätzung beruhen im Vergleich zu den klassischen Methoden der Reliabilitätschätzung auf weniger strengen Annahmen und haben den Vorteil, dass zusammen mit der Parameterschätzung die Voraussetzungen und Modellannahmen der jeweiligen Reliabilitätskoeffizienten anhand der CFA explizit überprüft und Modelle somit auch als nicht passend verworfen werden können. Die Modellannahmen beziehen sich neben der grundsätzlichen Frage der Dimensionalität eines Tests auf die Stufe der Messäquivalenz sowie auf die Unkorreliertheit der Fehlervariablen. „Modellbasiert“ bedeutet somit zum einen, dass die Modelle und Annahmen explizit überprüft werden, und zum anderen, dass die Reliabilitätskoeffizienten im Rahmen der CFA anhand der Modellparameter geschätzt werden.

Für eindimensionale Tests wurden Cronbachs Alpha, McDonalds Omega und Bollens Omega sowie für mehrdimensionale Tests verschiedene Omega-Koeffizienten vorgestellt, die auf einer ungewichteten Aufsummierung der Itemwerte des gesamten Tests oder der Itemwerte der einzelnen Subskalen eines mehrdimensionalen Tests beruhen.

Für die Testwerte mehrdimensionaler Tests können drei Reliabilitätskoeffizienten anhand der Parameter eines Bifaktormodells berechnet werden. Omega-total (ω_T) gibt Auskunft darüber, wie hoch der Anteil der totalen wahren Varianz an der gesamten Varianz eines Tests ist, während sich Omega-hierarchisch (ω_H) nur auf den Anteil der wahren Varianz bezieht, der durch den Generalfaktor bedingt ist. Die Summe aller subskalenspezifischen Varianzanteile bezogen auf die Gesamtvarianz wird als Omega-spezifisch (ω_S) bezeichnet.

Für die Subskalenvwerte im Rahmen eines mehrfaktoriellen Modells lassen sich ebenfalls jeweils drei Reliabilitätskoeffizienten berechnen. Subskalenspezifische Koeffizienten sind folgende:

- Omega-Subskala-total ($\omega_{Skala-T}$) zur Schätzung der totalen wahren Varianz an der Gesamtvarianz der Subskala
- Omega-Subskala-hierarchisch ($\omega_{Skala-H}$) zur Schätzung des erklärten Varianzanteils an der Gesamtvarianz der Subskala, der nur auf den Generalfaktor zurückgeht

- Omega-Subskala-spezifisch ($\omega_{\text{Skala-S}}$) zur Schätzung des erklärten Varianzanteils an der Gesamtvarianz einer Subskala, der nur auf den spezifischen Faktor zurückgeht

Alle modellbasierten Reliabilitätsschätzungen können als Punktschätzungen vorteilhaft durch Intervallschätzungen ergänzt werden.

Als empirisches Anwendungsbeispiel wurde das mehrdimensionale Persönlichkeitsmerkmal Perfektionismus, gemessen mit der MPS-F, verwendet. Die Schätzung der Varianzkomponenten anhand der CFA und die Berechnung der verschiedenen Omega-Koeffizienten wurden sowohl für die verkürzte eindimensionale Skala (*Concern over Mistakes*) als auch für den verkürzten mehrdimensionalen Test mit den drei Subskalen *Personal Standards* (PS), *Doubts about Actions* (DA) und *Concern over Mistakes* (CM) demonstriert.

15.9 EDV-Hinweise

Alle in diesem Kapitel behandelten CFA-Modelle können mit gängigen EDV-Programmen zur Analyse von Strukturgleichungsmodellen oder mit dem Programm R analysiert werden. Die Omega-Koeffizienten und die Konfidenzintervalle lassen sich am einfachsten mit den Programmen *Mplus* (Muthén und Muthén 2017) oder dem R-Paket *lavaan* (Rosseel 2012) schätzen. *Mplus*- und R-Syntax für Omega-Koeffizienten der eindimensionalen Modelle und der mehrdimensionalen Modelle finden sich in den Zusatzmaterialien unter ► <http://www.lehrbuch-psychologie.springer.com>.

15.10 Kontrollfragen

?(?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Welche Vorteile hat die modellbasierte Reliabilitätsschätzung im Vergleich zur klassischen Reliabilitätsschätzung?
2. Worin besteht der Unterschied zwischen den Reliabilitätskoeffizienten ω und ω^* ?
3. Welche Omega-Koeffizienten werden bei mehrdimensionalen Tests unterschieden?
4. Warum ist der Modellfit wesentlich für modellbasierte Reliabilitätsschätzungen?
5. Warum kann das übliche symmetrische Konfidenzintervall für Reliabilitätschätzungen nicht verwendet werden?
6. Wie würden Sie die Reliabilität eines mehrdimensionalen Tests mit den Reliabilitätsschätzungen $\omega_H = .55$ und $\omega_S = .30$ beurteilen?

Literatur

- Altstötter-Gleich, C. & Bergemann, N. (2006). Testgüte einer deutschsprachigen Version der Mehrdimensionalen Perfektionismus Skala von Frost, Marten, Lahart und Rosenblatt (MPS-F). *Diagnostica*, 52, 105–118.
- Amend, N. (2015). *Who's perfect? Pilotstudie zur Untersuchung potenzieller Korrelate des Merkmals Perfektionismus*. Unveröffentlichte Bachelorarbeit, Institut für Psychologie, Goethe Universität, Frankfurt am Main.
- Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, 9, 78–102.
- Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, 15, 211–240.

- Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York, NY: The Guilford Press.
- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74, 137–143.
- Bieling, P. J., Israeli, A. L. & Anthony, M. M. (2004). Is perfectionism good, bad, or both? Examining models of the perfectionism construct. *Personality and Individual Differences*, 36, 1373–1385.
- Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review* 45, 370–390.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley and Sons.
- Chen, F. F., West, S. G. & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189–225.
- Cox, B. J., Enns, M. W. & Clara, I. P. (2002). The multidimensional structure of perfectionism in clinically distressed and college student samples. *Psychological Assessment*, 14, 365–373.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eid, M., Notz, P., Steyer, R. & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analyses. *Personality and Individual Differences*, 16, 63–76.
- 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., Geiser, C. & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science*, 25, 275–280.
- Eid, M., Geiser, C., Koch, T. & Heene, M. (2017a). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods*, 22, 541–562.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017b). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
- Frost, R. O., Marten, P., Lahart, C. & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research*, 14, 449–468.
- Gäde, J. C., Schermelleh-Engel, K. & Klein, A. G. (2017). Disentangling the common variance of perfectionistic strivings and perfectionistic concerns: A bifactor model of perfectionism. *Frontiers in Psychology*, 8, 160. <https://doi.org/10.3389/fpsyg.2017.00160>
- Geiser, C. & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, 17, 255–283.
- Geiser, C., Bishop, J. & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology*, 6:946.
- Geldhof, G. J., Preacher, K. J. & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19, 72–91.
- Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: G as superordinate or breadth factor? *Psychology Science*, 50, 21–43.
- Green, S. B. & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74, 155–167.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Holzinger, K. J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41–54.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Kelley, K. & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for homogeneous composite measures. *Psychological Methods*, 21, 69–92.
- Little, T. D., Rhett, M., Gibson, K. & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300.
- McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1–21.
- McDonald, R. P. (1978). A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 31, 59–72.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Psychological Methods*, 88, 879–903.
- Podsakoff, P. M., MacKenzie, S. B. & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569.
- Rauch, W. A. & Moosbrugger, H. (2011). Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B Methodologie und Methoden, Serie II Psychologische Diagnostik, Band 2, Methoden der psychologischen Diagnostik* (S. 1–87). Göttingen: Hogrefe.

- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184.
- Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, 37, 89–103.
- Raykov, T. (2004). Point and interval estimation of reliability for multiple component measuring instruments via linear constraint covariance structure modeling. *Structural Equation Modeling*, 11, 452–483.
- Raykov, T. & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, 11, 621–637.
- Raykov, T. & Marcoulides, G. A. (2011). *Psychometric Theory*. New York: Routledge.
- Raykov, T. & Marcoulides, G. A. (2015). Scale reliability evaluation with heterogeneous populations. *Educational and Psychological Measurement*, 75, 146–156.
- Raykov, T. & Marcoulides, G. A. (2016). Scale reliability evaluation under multiple assumption violations. *Structural Equation Modeling*, 23, 302–313.
- Raykov, T., Dimitrov, D. M. & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling*, 17, 122–132.
- Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften* (2. Aufl.). München: Oldenbourg Wissenschaftsverlag.
- Reise, S. P., Bonifay, W. E. & Haviland, M. G. (2013a). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95, 129–140.
- Reise, S. P., Scheines, R., Widaman, K. F. & Haviland, M. G. (2013b). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73, 5–26.
- Revelle, W. & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth & D. Hughes (Eds.), *The Wiley-Blackwell Handbook of Psychometric Testing*. West Sussex, UK: Blackwell Publishing Ltd.
- Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 145–154.
- Rhemtulla, M., Brosseau-Liard, P. E. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373.
- Rodriguez, A., Reise, S. P. & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21, 137–150.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48, 1–36.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research-Online*, 8, 23–74.
- Sijtsma, K. (2009). The use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120.
- Smith, M. M. & Saklofske, D. H. (2017). The structure of multidimensional perfectionism: Support for a bifactor model with a dominant general factor. *Journal of Personality Assessment*, 99, 297–303.
- Smits, I. A. M., Timmerman, M. E., Barelds, D. P. H. & Meijer, R. R. (2015). The Dutch Symptom Checklist-90-Revised: Is the use of the subscales justified? *European Journal of Psychological Assessment*, 31, 263–271.
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54, 371–384.
- Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin, Heidelberg: Springer.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A Theory of States and Traits – Revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- Stöber, J. (1995). Frost Multidimensional Perfectionism Scale-Deutsch (FMPS-D). Unveröff. Manuskript. Freie Universität Berlin, Institut für Psychologie.
- Stoeber, J. (2014). Perfectionism. In R. C. Eklund & G. Tenenbaum (Eds.), *Encyclopedia of sport and exercise psychology* (Vol. 2, pp. 527–530). Thousand Oaks, CA: Sage.
- Stoeber, J. & Damian, L. (2014). The clinical perfectionism questionnaire: further evidence for two factors capturing perfectionistic strivings and perfectionistic concerns. *Personality and Individual Differences*, 61–62, 38–42.
- Werts, C. E., Linn, R. L. & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34, 25–33.
- Wolf, E. J., Harrington, K. M., Clark, S. L. & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73, 913–934.
- Yang, Y. & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology*, 11, 23–34.
- Yang, Y. & Xia, Y. (2019). Categorical omega with small sample sizes via Bayesian estimation: an alternative to Frequentist estimators. *Educational and Psychological Measurement*, 79, 19–39.
- Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω_H : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 1–11.

Einführung in die Item-Response-Theorie (IRT)

Augustin Kelava und Helfried Moosbrugger

Inhaltsverzeichnis

16.1	Grundüberlegungen zur IRT – 371
16.1.1	Dichotomes Antwortformat – 371
16.1.2	Zusammenhänge dichotomer Antwortvariablen – 372
16.2	Latent-Trait-Modelle – 372
16.3	Dichotomes Rasch-Modell (1PL-Modell) – 373
16.3.1	Vorüberlegungen – 373
16.3.2	Rasch-Homogenität – 373
16.3.3	Logistische IC-Funktion – 373
16.3.4	Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit – 374
16.3.5	Beziehung des Item- und Personenparameters – 376
16.3.5.1	Joint Scale (gemeinsame Skala) – 376
16.3.5.2	Personenparameter bei Konstanthaltung des Itemparameters – 377
16.3.5.3	Itemparameter bei Konstanthaltung des Personenparameters – 378
16.3.5.4	Parameternormierung – 379
16.3.6	Spezifische Objektivität der Vergleiche – 380
16.3.7	Lokale stochastische Unabhängigkeit – 381
16.3.7.1	Paarweise Betrachtung der Antworten auf Items – 381
16.3.7.2	Formale Definition – 382
16.3.7.3	Veranschaulichung über Korrelationen – 382
16.3.7.4	Empirische Nutzungsmöglichkeit – 384
16.3.8	Parameterschätzung – 387
16.3.8.1	Joint Maximum Likelihood (JML) zur Schätzung der Item- und Personenparameter – 387
16.3.8.2	Conditional Maximum Likelihood (CML) – 390
16.3.8.3	Marginal Maximum Likelihood (MML) – 391
16.3.8.4	Bayes'sche Schätzung der Itemparameter – 392
16.3.8.5	Unconditional und Weighted Maximum Likelihood zur Schätzung der Personenparameter – 393
16.3.8.6	Weitere Verfahren zur Personenparameterschätzung – 393
16.3.9	Iteminformationsfunktion – 394
16.3.10	Testinformation und Konfidenzintervall für η_v – 395

16.3.11 Überprüfung der Modellpassung/Modellkonformität – 396
16.3.11.1 Empirische Modellkontrollen und Itemselektion – 396
16.3.11.2 Personenselektion – 398

16.4 2PL-Modell nach Birnbaum – 399

16.4.1 Charakteristika – 399
16.4.2 Modellgleichung – 400
16.4.3 Parameterschätzung – 401

16.5 3PL-Modell nach Birnbaum – 401

16.6 Weitere IRT-Modelle – 402

16.6.1 Polytome Latent-Trait-Modelle – 403
16.6.2 Mixed-Rasch-Modelle – 404
16.6.3 Linear-logistische Modelle – 405

16.7 Zusammenfassung – 406

16.8 EDV-Hinweise – 407

16.9 Kontrollfragen – 407

Literatur – 407

i Bei der Modellierung menschlichen Erlebens und Verhaltens wird oft die Annahme gemacht, dass Unterschiede im beobachtbaren Verhalten auf die Ausprägungen von latenten Personenvariablen (hypothetische Konstrukte, latent Traits) zurückgeführt werden können. Kennzeichnend für latente Merkmale ist, dass man sie nicht direkt beobachten kann; vielmehr wird die Merkmalsausprägung aus einem Muster von beobachtbarem Antwortverhalten auf Testitems („Item Response“, z. B. Zustimmung/Lösung/Richtigantwort oder Ablehnung/Nichtlösung/Falschantwort) erschlossen. Vereinfacht schließt man beispielsweise aus der Zustimmung auf eine Aussage wie „Ich stehe bei einer Party gerne im Mittelpunkt der Aufmerksamkeit“ auf eine hohe Ausprägung des latenten Merkmals „Extraversion“.

16.1 Grundüberlegungen zur IRT

In diesem Kapitel erfolgt eine Einführung in die sog. „Item-Response-Theorie“ (IRT; z. B. Lord und Nowick 1968). Die grundlegende testtheoretische Idee der IRT besteht darin, die Wahrscheinlichkeit eines gezeigten Antwortverhaltens (*Response*) einer Person bei einem Item (z. B. das Bejahren/Nichtbejahren einer Aussage in einem Einstellungstest bzw. das Lösen/Nichtlösen einer Aufgabe/eines Items in einem Leistungstest) in Form einer (zumeist einfachen) Wahrscheinlichkeitsfunktion zu beschreiben. Die Wahrscheinlichkeit, mit der die Person ein konkretes Antwortverhalten zeigt, wird dabei einerseits von den Eigenschaften des Items (z. B. der Anforderung/Aufgabenschwierigkeit) und andererseits von der Ausprägung der Person im interessierenden latenten Merkmal (z. B. der Einstellung oder Leistungsfähigkeit) bestimmt.

Antwortverhalten (Response) hängt von Eigenschaften des Items und der Merkmalsausprägung der Person ab

16.1.1 Dichotomes Antwortformat

Im einfachsten Fall gehen Modelle der IRT davon aus, dass das Antwortverhalten auf zwei distinkte Ausprägungen reduziert ist, denen zwei numerische Werte der Antwortvariablen/Itemvariablen zugeordnet werden können. Typische Werte bei diesem dichotomen Antwortformat (vgl. ► Kap. 5) sind die Werte „1“ und „0“ (es sind aber auch zwei beliebige andere, unterschiedliche Werte denkbar), wobei dem Bejahren/Lösen zumeist der Wert „1“ und dem Nichtbejahren/Nichtlösen zumeist der Wert „0“ zugeordnet wird.

Zweiwertige Antwortvariablen beim dichotomen Antwortformat

Anmerkung: Vom dichotomen Antwortformat zu unterscheiden ist das sog. „polytomous Antwortformat“, bei dem mehreren disjunkten Ausprägungen des Antwortverhaltens entsprechende Werte der Antwortvariablen zugeordnet werden. Beispielsweise könnten bei Situationen, in denen eine Aufgabe „nicht“, „teilweise“ oder „vollständig“ gelöst wird, die Werte „0“, „1“ oder „2“ vergeben werden. Modelle für solche mehrwertige Antwortvariablen in Partial-Credit-Situationen werden von Kelava, Robitzsch und Noventa in ► Kap. 18 besprochen.

Mehrwertige Antwortvariablen beim polytomous Antwortformat

Im weiteren Verlauf des Kapitels werden wir uns mit IRT-Modellen beschäftigen, die einen Rückschluss von zweiwertigem Antwortverhalten auf dahinterliegende latente Personenmerkmale ermöglichen. Als Grundvoraussetzung ist hierbei wichtig, dass es sich bei allen verwendeten Items (genauer: bei den Antwortvariablen der Personen auf die Items) um Indikatoren desselben dahinterliegenden latenten Merkmals handelt. Einen ersten deskriptiven Hinweis darauf liefern die Korrelationen zwischen den Antwortvariablen: Sofern diese nicht null sind, kann es sich bei den Items um Indikatoren eines gemeinsamen dahinterliegenden latenten Merkmals handeln (vgl. auch ► Kap. 7 zur deskriptivstatistischen Itemanalyse).

■ **Tabelle 16.1** Korrelationsmatrix der Antwortvariablen von fünf Items aus Abschnitt 6 des Law School Admission Test (LSAT)

	Item 11	Item 12	Item 13	Item 14	Item 15
Item 11	1.00				
Item 12	.12	1.00			
Item 13	.16	.26	1.00		
Item 14	.16	.13	.16	1.00	
Item 15	.15	.07	.14	.09	1.00

16.1.2 Zusammenhänge dichotomer Antwortvariablen

ϕ-Koeffizient als Maß des Zusammenhangs zweier dichotomer Variablen

Zur Bestimmung der Enge des empirischen Zusammenhangs zwischen dichotomen Antwortvariablen werden üblicherweise sog. „ ϕ -Koeffizienten“ berechnet (vgl. Bortz und Schuster 2010, S. 174). Basierend auf den Antwortvariablen von n Personen in p Items werden dabei vier Häufigkeiten („keines der beiden Items gelöst“, „das eine Item gelöst, das andere nicht“, „das eine Item nicht gelöst, das andere schon“ und „beide Items gelöst“) miteinander in Beziehung gesetzt. Zur Veranschaulichung sei als Beispiel auf einige Items aus dem *Law School Admission Test* (LSAT) zurückgegriffen, einem Eignungstest zum Jurastudium im angelsächsischen Raum (Items 11 bis 15 aus Abschnitt 6, s. Bock und Lieberman 1970, S. 188). Die Korrelationsmatrix (ϕ -Koeffizienten) mit den $5 \times (5 + 1)/2$ nicht redundanten Elementen ist in ■ Tab. 16.1 gegeben. Wie man erkennen kann, korrelieren die Items 11 und 12 zu .12, die Items 12 und 13 zu .26 usw.

Nun gilt es, die Frage zu klären, ob sich die Korrelationen zwischen je zwei Items auf ein oder mehrere dahinterliegende Merkmale zurückführen lassen. In diesem Kapitel werden wir uns darauf konzentrieren, dass die verschiedenen Items genau *ein* latentes Merkmal messen, auf das die Korrelationen zwischen den Antwortvariablen zurückgeführt werden können. In diesem Fall spricht man von eindimensionalen IRT-Modellen. (Multidimensionale IRT-Modelle sind Gegenstand von ► Kap. 18.)

Die wissenschaftliche Aufgabe besteht nun darin, mit geeigneten IRT-Modellen den (empirischen) Zusammenhang zwischen den dichotomen Antwortvariablen zu erklären, damit ein Rückschluss auf das interessierende dahinterliegende latente Personenmerkmal plausibel erscheint.

16.2 Latent-Trait-Modelle

Ein- vs. mehrdimensionale IRT-Modelle

Kontinuierlicher latenter Trait als Personenvariable

Das Konzept der Lösungswahrscheinlichkeit

Bei sog. „Latent-Trait-Modellen“ nimmt man – im Unterschied zu Latent-Class-Modellen (► Kap. 22) – an, dass dem gezeigten Verhalten (d. h. den konkreten „Responses“ auf ein Item) ein kontinuierliches latentes Merkmal (Trait) zugrunde liegt, das von Person zu Person eine unterschiedliche Ausprägung aufweist. Diese latente Personenvariable wird im weiteren Verlauf dieses Kapitels mit η bezeichnet, die individuelle Ausprägung einer Person v mit η_v . In den beobachteten Antwortvariablen werden die zustimmenden Antworten auf ein Item i mit $Y_i = 1$ gekennzeichnet, die ablehnenden Antworten entsprechend mit $Y_i = 0$.

Ausgehend von der Annahme einer latenten Personenvariable wird – im Unterschied zu Modellen der Klassischen Testtheorie (KTT, ► Kap. 13) – im Rahmen der IRT die bedingte Wahrscheinlichkeit $P(Y_i = 1 | \eta)$ modelliert, die angibt, mit welcher Wahrscheinlichkeit bei gegebener Ausprägung der Personenvariable η eine richtige/zustimmende Antwort auf Item i gegeben wird („Lösungswahrscheinlichkeit“).

Hierzu wird eine Funktion, die sog. „Itemcharakteristische Funktion“ (IC-Funktion), definiert, die beschreibt, wie sich die Lösungswahrscheinlichkeit in Abhängigkeit von der Ausprägung der latenten Personenvariable η (z. B. einer Fähigkeit) verändert. Sinnvollerweise wird ein monotoner Funktionstypus gewählt, der bei einer höheren latenten Merkmalsausprägung (z. B. höherer Fähigkeit) eine höhere Lösungswahrscheinlichkeit impliziert. Diese *Monotonie* der Beziehung ist eine zentrale Anforderung bei der Auswahl des Funktionstypus.

Bedingte Wahrscheinlichkeit der Lösung eines Items i in Abhängigkeit von der Personenvariable η

Monotone IC-Funktion

16.3 Dichotomes Rasch-Modell (1PL-Modell)

16.3.1 Vorüberlegungen

Das Rasch-Modell, auch einparametrisches logistisches Modell (1PL-Modell) genannt, gehört zu den am häufigsten angewandten Modellen aus der Gruppe der Latent-Trait-Modelle. Es wird insbesondere in der Leistungsdagnostik angewendet, in der auf Grundlage des Lösen bzw. Nichtlösen von Aufgaben eine Bestimmung der latenten Merkmalsausprägung (Leistungsfähigkeit) vorgenommen wird.

16.3.2 Rasch-Homogenität

Eine wesentliche Annahme der Rasch-Modellierung der Itemantworten ist – als spezielle Form von Eindimensionalität – die sog. „Rasch-Homogenität“ der Items. Unter der Rasch-Homogenität versteht man, dass den Antworten auf alle Items eines Tests genau *eine* latente Variable η (nämlich das interessierende Merkmal) zugrunde liegt und dass – abgesehen von den variierenden Itemschwierigkeiten β_i – genau diese eine latente Personenvariable die Unterschiede im Antwortverhalten der verschiedenen Personen erzeugt (und in gewisser Weise auch erklärt). Weitere systematische Einflüsse (in Form weiterer Merkmale oder Parameter, z. B. eines Diskriminationsparameters, ► Abschn. 16.4) auf die Lösungswahrscheinlichkeit werden als nicht existent angenommen. Alle Items messen somit dasselbe latente Merkmal; sie stellen aber – in Form unterschiedlicher „Itemschwierigkeiten“ – unterschiedlich hohe Anforderungen an die Testpersonen. Darüber hinaus sind es nur die verschiedenen Merkmalsausprägungen auf der latenten Variable η , die als „ursächlich“ für die jeweilige Lösungswahrscheinlichkeit eines Items $P(Y_i = 1 | \eta)$ angesehen werden.

Rasch-homogene Items messen *eine* latente Variable

16.3.3 Logistische IC-Funktion

Als IC-Funktion, d. h. als funktionale Modellierung des Zusammenhangs zwischen der Lösungswahrscheinlichkeit eines Items und dem interessierenden latenten Merkmal, wird im Rasch-Modell eine „logistische“ Funktion angenommen. Diese Beziehung („Modellgleichung“) ist monoton und lässt sich mathematisch-statistisch wie folgt ausdrücken:

$$P(Y_i = 1 | \eta) = \frac{e^{\eta - \beta_i}}{1 + e^{\eta - \beta_i}} \quad (16.1)$$

Für eine konkrete Person v ergibt sich bei einem Item i :

$$P(Y_i = 1 | \eta_v) = \frac{e^{\eta_v - \beta_i}}{1 + e^{\eta_v - \beta_i}} \quad (16.2)$$

Logistische IC-Funktion, Modellgleichung

Modellgleichung, Fähigkeits- und Schwierigkeitsparameter, Joint Scale

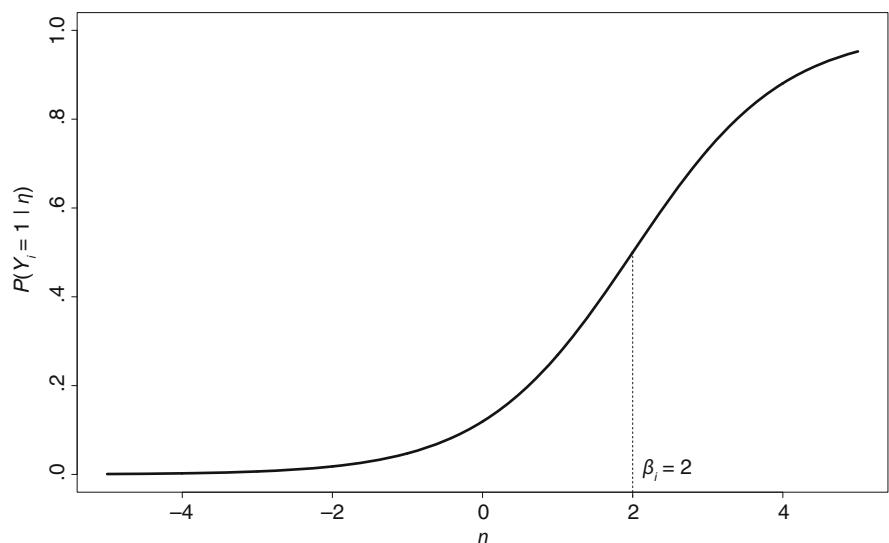
In der Modellgleichung (Gl. 16.1) wird die Wahrscheinlichkeit, auf Item i eine richtige Antwort zu geben, in Abhängigkeit von einem Itemschwierigkeitsparameter β_i sowie in Abhängigkeit von der Personenvariablen η (bzw. in Gl. 16.2 individualisiert für eine konkreten Person v in Abhängigkeit vom individuellen Personenparameter η_v) beschrieben. Dabei wird der Personenparameter (zumeist Fähigkeitsparameter) η_v , d. h. die interessierende Merkmalsausprägung einer konkreten Person v , ebenso wie der Itemschwierigkeitsparameter β_i üblicherweise auf einer gemeinsamen Skala („Joint Scale“) von η und β (► Abschn. 16.3.5.1) abgetragen.

Anmerkung: Anstelle der logistischen Funktion wären natürlich auch andere funktionale Zusammenhänge denkbar. Es erweist sich aber als forschungsoökonomisch sehr sinnvoll, die logistische IC-Funktion zu wählen, da das Rasch-Modell bei entsprechender Passung von Modell und Daten (Modellkonformität, ► Abschn. 16.3.11) sowohl statistisch als auch inhaltlich einige sehr vorteilhafte Eigenschaften aufweist (► Abschn. 16.3.6).

16.3.4 Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit

Steigung und Wendepunkt der IC-Funktion

Für ein sicheres Verständnis wird im Folgenden die IC-Funktion hinsichtlich ihrer Eigenschaften genauer erläutert. Dazu wird ein „schwieriges“ Item betrachtet (s. dazu die Parameternormierung in ► Abschn. 16.3.5.4), das überdurchschnittliche Anforderungen an die Testpersonen stellt und eine Itemschwierigkeit von $\beta_i = 2$ hat. In □ Abb. 16.1 ist der Verlauf der IC-Funktion veranschaulicht. Wie man erkennen kann, beginnt die liegend s-förmige Kurve für das Item i auf der linken Seite mit einer Lösungswahrscheinlichkeit nahe null. Mit zunehmender Personenfähigkeit η nimmt die Lösungswahrscheinlichkeit zwar zunächst gering, so dann aber immer stärker zu, wie man an der Steigung der IC-Funktion erkennen kann. Im Punkt $\eta = \beta_i = 2$ ist die Steigung maximal; hier hat die IC-Funktion ihren Wendepunkt. Mit noch größerer Personenfähigkeit nimmt die Lösungswahrscheinlichkeit zwar weiter zu, allerdings mit geringerer Veränderung, d. h. mit ab-



□ Abb. 16.1 Lösungswahrscheinlichkeit (IC-Funktion) im Rasch-Modell. Gezeigt wird der Verlauf der logistischen IC-Funktion des Rasch-Modells bei einem Item i mit dem Schwierigkeitsparameter $\beta_i = 2$. Abszisse: gemeinsame Skala (Joint Scale) von η und β ; Ordinate: bedingte Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ in Abhängigkeit von η

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

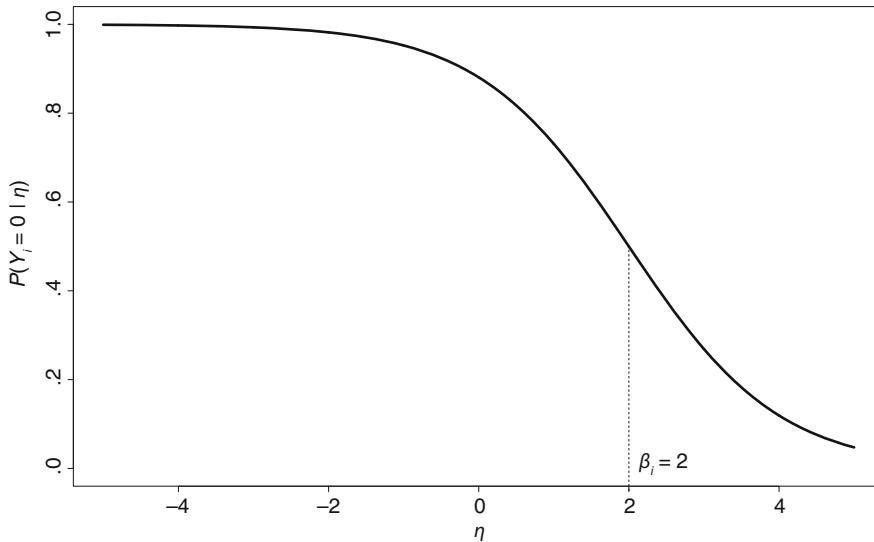


Abb. 16.2 Gegenwahrscheinlichkeit im Rasch-Modell. Gezeigt wird der Verlauf der logistischen IC-Funktion für die Gegenwahrscheinlichkeit bei einem Item i mit dem Schwierigkeitsparameter $\beta_i = 2$. Die Gegenwahrscheinlichkeit bezeichnet die bedingte Wahrscheinlichkeit $P(Y_i = 0 | \eta)$, ein Item nicht zu lösen/nicht zu bejahen in Abhängigkeit von η

nehmender Steigung; bei sehr großer Ausprägung der Personenfähigkeit strebt die Lösungswahrscheinlichkeit schließlich dem Wert 1 entgegen.

Formal nimmt die Lösungswahrscheinlichkeit aus Gl. (16.1) für $\beta_i = 2$ die folgende Form an:

$$P(Y_i = 1 | \eta) = \frac{e^{\eta-2}}{1 + e^{\eta-2}} \quad (16.3)$$

Nun stelle man sich eine (durchschnittliche) Person mit $\eta_v = 0$ vor (zur Parameter-normierung s. auch ► Abschn. 16.3.5.4). Nach Einsetzen ihres Wertes in Gl. (16.2) erhält man als Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta_v) = .119$.

In □ Abb. 16.2 ist die Gegenwahrscheinlichkeit der Lösungswahrscheinlichkeit abgetragen, d. h. die bedingte Wahrscheinlichkeit, das Item nicht zu lösen ($P(Y_i = 0 | \eta)$). Man sieht, dass die Wahrscheinlichkeit, das Item nicht lösen zu können, anfänglich (d. h. bei geringer Merkmalsausprägung η) sehr groß ist und dann einem verkehrt s-förmigen Verlauf folgend immer schneller abnimmt, bis bei $\eta = \beta_i = 2$ der Wendepunkt erreicht ist; mit größer werdendem η geht die Gegenwahrscheinlichkeit langsam gegen 0.

□ Abb. 16.3 veranschaulicht die Lösungs- und Gegenwahrscheinlichkeit für ein Item i mit einem Schwierigkeitsparameter von $\beta_i = 2$ gemeinsam. Wie man erkennen kann, sind beide Wahrscheinlichkeiten gegenläufig und addieren sich stets zu eins.

Da sich die Lösungswahrscheinlichkeit und die Gegenwahrscheinlichkeit stets zu eins ergänzen, lässt sich die Gegenwahrscheinlichkeit $P(Y_i = 0 | \eta)$ allgemein ausdrücken als:

$$\begin{aligned} P(Y_i = 0 | \eta) &= 1 - P(Y_i = 1 | \eta) = 1 - \frac{e^{\eta-\beta_i}}{1 + e^{\eta-\beta_i}} \\ &= \frac{1 + e^{\eta-\beta_i}}{1 + e^{\eta-\beta_i}} - \frac{e^{\eta-\beta_i}}{1 + e^{\eta-\beta_i}} = \frac{1}{1 + e^{\eta-\beta_i}} \end{aligned} \quad (16.4)$$

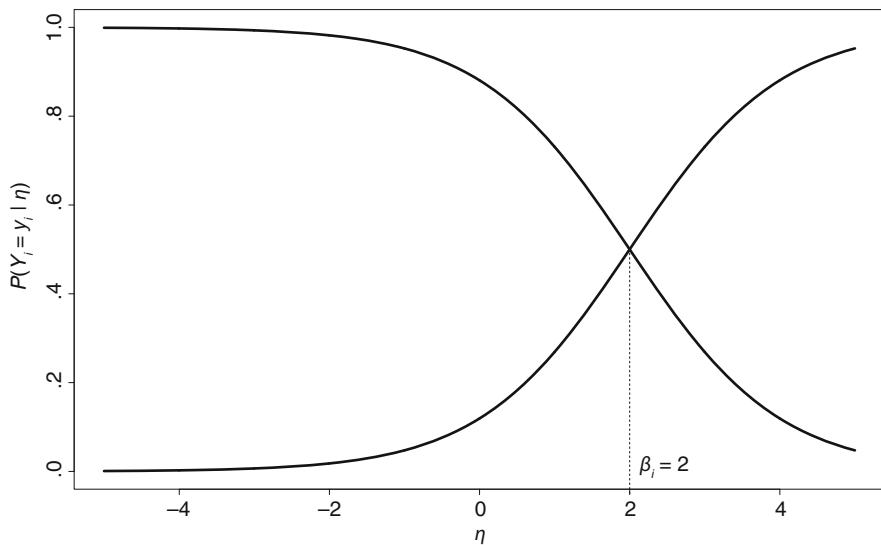


Abb. 16.3 Lösungs- und Gegenwahrscheinlichkeit im Rasch-Modell. Gezeigt wird die Gegenläufigkeit der Lösungs- und Gegenwahrscheinlichkeit eines Items i mit einem Schwierigkeitsparameter von $\beta_i = 2$ im Rasch-Modell. Beide Wahrscheinlichkeiten addieren sich stets zu eins

Formal nimmt die Gegenwahrscheinlichkeit (Gl. 16.4) für $\beta_i = 2$ folgende Form an:

$$P(Y_i = 0 | \eta) = \frac{1}{1 + e^{\eta-2}} \quad (16.5)$$

Nach Einsetzen für eine (durchschnittliche) Person mit $\eta_v = 0$ (zur Personennormierung s. auch ► Abschn. 16.3.5.4) in Gl. (16.5) erhielte man als Gegenwahrscheinlichkeit $P(Y_i = 0 | \eta_v) = .881$.

Die Formeln für die Lösungswahrscheinlichkeit und die Gegenwahrscheinlichkeit lassen sich auch in eine gemeinsame Formel zusammenziehen:

$$P(Y_i = y_i | \eta) = \frac{e^{y_i(\eta-\beta_i)}}{1 + e^{\eta-\beta_i}} \quad (16.6)$$

Für $y_i = 1$ vereinfacht sich Gl. (16.6) zu Gl. (16.1) und für $y_i = 0$ zu Gl. (16.4).

16.3.5 Beziehung des Item- und Personenparameters

16.3.5.1 Joint Scale (gemeinsame Skala)

In der Modellgleichung des Rasch-Modells (16.1) ist die Differenz von η und β_i (sog. „Logit-Wert“ der Logit-Variablen, Näheres ► Abschn. 16.3.6) von besonderer Bedeutung. Sie kann auf der „Joint Scale“, d. h. auf der gemeinsamen Skala von η und β , abgetragen werden. Ganz allgemein hat die Verortung der Differenz $\eta - \beta_i$ in der IC-Funktion (vgl. Gl. 16.1) auf einer gemeinsamen Skala den Vorteil, dass ein unmittelbarer Bezug zwischen der Personenfähigkeit η und dem Itemparameter β_i hergestellt wird, wie □ Abb. 16.4 zu entnehmen ist.

Wie man in □ Abb. 16.4 erkennen kann, hat die logistische IC-Funktion an der Stelle $\eta - \beta_i = 0$ ihren Wendepunkt. Der Itemschwierigkeitsparameter β_i ist dafür entscheidend, welche Anforderung das Item i an die Merkmalsausprägung der Personen in der latenten Variablen η stellt. Je stärker β_i von η übertroffen wird, desto größer ist die Lösungs-/Bejahungswahrscheinlichkeit $P(Y_i = 1 | \eta)$; je stärker η hinter β_i zurückbleibt, desto kleiner ist $P(Y_i = 1 | \eta)$ und desto größer ist die Gegenwahrscheinlichkeit $P(Y_i = 0 | \eta)$, das Item nicht zu lösen bzw. zu verneinen.

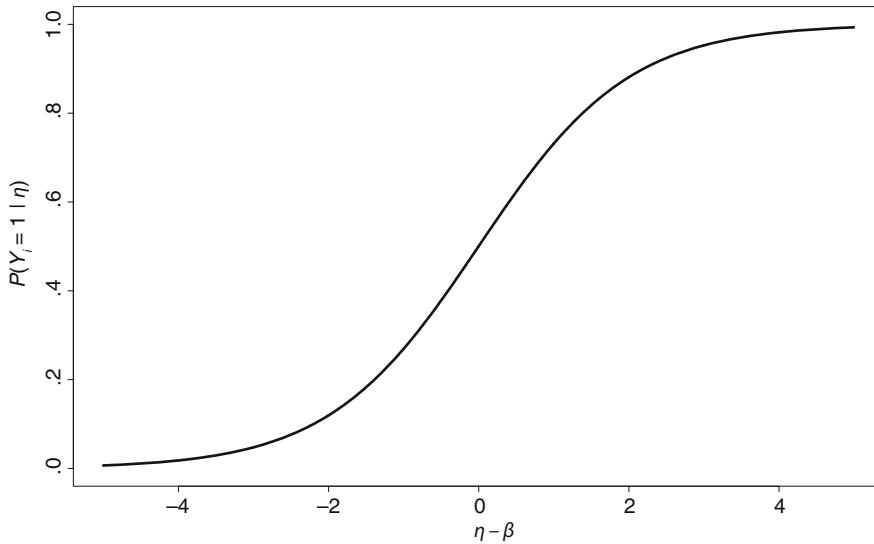


Abb. 16.4 Lösungswahrscheinlichkeit (IC-Funktion) im Rasch-Modell. Gezeigt wird der Verlauf der logistischen IC-Funktion des Rasch-Modells, wenn auf der Joint Scale (Abszisse) die Logit-Differenz $(\eta - \beta_i)$ abgetragen wird

Wie man später in **Abb. 16.5** anhand der variierten Itemschwierigkeiten sehen kann, hängt die Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ sowohl von der Personenfähigkeit η als auch vom jeweiligen Itemparameter β_i ab. Konkret stellt die Differenz $\eta_v - \beta_i$ zwischen der individuellen Merkmalsausprägung der Person v und der Anforderung des jeweiligen Items i , die entscheidende Größe für die individuelle Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta = \eta_v)$ dar. Eine Fallunterscheidung mit bestimmten Werten von β_i soll die Verbindung zwischen den Abbildungen und der Modellgleichung Gl. (16.1) herstellen und das Verständnis erleichtern:

- Für $\eta_v = \beta_i$ entspricht die Fähigkeit der Person v genau der Schwierigkeit des Items und mit $e^0/(1 + e^0) = 1/2$ ist die Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta = \eta_v)$ genau .5.
- Für $\eta_v > \beta_i$ ist die Fähigkeit der Person v größer als die Schwierigkeit des Items; die Lösungswahrscheinlichkeit steigt an ($P(Y_i = 1 | \eta = \eta_v) > .5$) und geht bei entsprechend großer Fähigkeit asymptotisch gegen 1.
- Für $\eta_v < \beta_i$ ist die Fähigkeit der Person v kleiner als die Schwierigkeit des Items; die Lösungswahrscheinlichkeit fällt ab ($P(Y_i = 1 | \eta = \eta_v) < .5$) und geht bei entsprechend geringer Fähigkeit asymptotisch gegen 0.

Als Konsequenz ergibt sich, dass sich IC-Funktionen für Rasch-homogene Items beliebiger Schwierigkeit β_i (also solcher Items, die der IC-Funktion aus Gl. 16.1 folgen) in einer einzigen Abbildung zusammenfassen lassen, wenn man auf der Joint Scale jeweils die Differenz zwischen η und β_i abträgt. Diese Skala wird auch als „Logit-Skala“ (Näheres hierzu in ► Abschn. 16.3.6) bezeichnet. Unabhängig von den konkreten Ausprägungen η_v und β_i ergibt sich für $(\eta_v - \beta_i) = 0$ jeweils die Lösungswahrscheinlichkeit von .5; für $(\eta - \beta_i) > 0$ erhält man hingegen eine Lösungswahrscheinlichkeit zwischen .5 und 1 sowie für $(\eta - \beta_i) < 0$ eine zwischen 0 und .5 (**Abb. 16.4**).

16.3.5.2 Personenparameter bei Konstanthaltung des Itemparameters

Um die Bedeutung der Differenz $\eta - \beta_i$ für die Lösungswahrscheinlichkeit in Abhängigkeit vom Personenparameter η_v detailliert zu veranschaulichen, halten wir zunächst den Itemparameter konstant und betrachten drei Personen mit den Personenparametern $\eta_1 = -2$, $\eta_2 = 0$, $\eta_3 = 1$. Die Lösungswahrscheinlichkeit für ein

Fallunterscheidung bezüglich der Größen von η_v und β_i

Zusammenfassung der IC-Funktionen auf der Logit-Skala

Bedeutung der Differenz $\eta - \beta_i$ bei Konstanthaltung des Itemparameters

Item i mit konstant gehaltener Itemschwierigkeit $\beta_i = 0$ beträgt dann:

$$\begin{aligned} P(Y_i = 1 | \eta = -2, \beta_i = 0) &= \frac{e^{(-2)-0}}{1 + e^{(-2)-0}} = \frac{e^{(-2)}}{1 + e^{(-2)}} \approx \frac{.135}{1 + .135} \approx .119 \\ P(Y_i = 1 | \eta = 0, \beta_i = 0) &= \frac{e^{0-0}}{1 + e^{0-0}} = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = .5 \\ P(Y_i = 1 | \eta = 1, \beta_i = 0) &= \frac{e^{1-0}}{1 + e^{1-0}} = \frac{e^1}{1 + e^1} \approx \frac{2.718}{1 + 2.718} \approx .731 \end{aligned}$$

Analog zu Abb. 16.4 sieht man, dass für Personen mit geringer Merkmalsausprägung/Fähigkeit und einem entsprechend niedrigen Personenparameterwert ($\eta = -2$) eine niedrigere Lösungswahrscheinlichkeit erwartet wird als bei mittlerem Fähigkeitsniveau ($\eta = 0$). Für Personen mit einem höheren Fähigkeitsparameter ($\eta = 1$) wird eine hohe Lösungswahrscheinlichkeit erwartet.

16.3.5.3 Itemparameter bei Konstanthaltung des Personenparameters

Bedeutung der Differenz $\eta - \beta_i$ bei Konstanthaltung des Personenparameters

Um die Bedeutung der Differenz $\eta - \beta_i$ für die Lösungswahrscheinlichkeit in Abhängigkeit vom Itemparameter β_i detailliert zu veranschaulichen, halten wir nun den Personenparameter konstant und betrachten drei Items mit den Itemparametern $\beta_1 = -1$, $\beta_2 = 0$ und $\beta_3 = 2$. Die Lösungswahrscheinlichkeiten für einen konstant gehaltenen Personenparameter $\eta = 0$ betragen dann:

$$\begin{aligned} P(Y_i = 1 | \eta = 0, \beta_i = -1) &= \frac{e^{0-(-1)}}{1 + e^{0-(-1)}} = \frac{e^1}{1 + e^1} \approx \frac{2.718}{1 + 2.718} \approx .731 \\ P(Y_i = 1 | \eta = 0, \beta_i = 0) &= \frac{e^{0-0}}{1 + e^{0-0}} = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = .5 \\ P(Y_i = 1 | \eta = 0, \beta_i = 2) &= \frac{e^{0-2}}{1 + e^{0-2}} = \frac{e^{-2}}{1 + e^{-2}} \approx \frac{0.135}{1 + 0.135} \approx .119 \end{aligned}$$

Abb. 16.5 veranschaulicht drei IC-Funktionen mit den Lösungswahrscheinlichkeiten für drei Items mit variierenden Itemschwierigkeitsparametern. Man sieht,

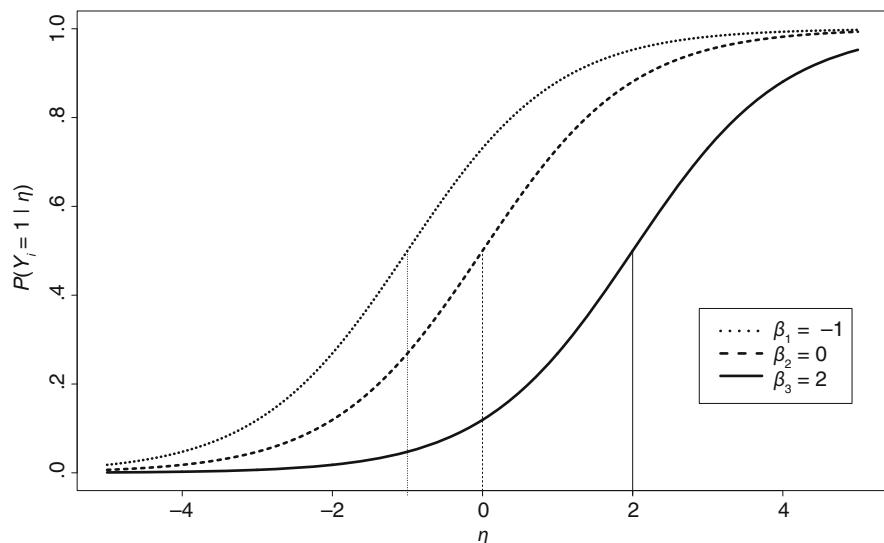


Abb. 16.5 Logistische IC-Funktionen für drei Rasch-homogenen Items mit unterschiedlichen Itemschwierigkeitsparametern. Item 1 ist das leichteste mit $\beta_1 = -1$; Item 2 ist schwieriger mit $\beta_2 = 0$; Item 3 ist das schwierigste mit $\beta_3 = 2$. Die Wendepunkte der drei IC-Funktionen liegen bei den jeweiligen Ausprägungen der Schwierigkeitsparameter

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

dass für das Item mit dem niedrigsten Schwierigkeitsparameter („leichtes Item“) höhere Lösungswahrscheinlichkeiten erwartet werden als für das mittlere Item. Für das Item mit dem höchsten Schwierigkeitsparameter („schwieriges Item“) werden die niedrigsten Lösungswahrscheinlichkeiten erwartet.

Wie man darüber hinaus erkennen kann, sind die Kurven der IC-Funktion entlang der Abszisse (Joint Scale) parallel verschoben. Diese Eigenschaft wird im weiteren Verlauf noch eine wichtige Rolle spielen (► Abschn. 16.3.6).

Der Itemschwierigkeitsparameter β_i in der IRT entspricht in der IC-Funktion exakt jener Merkmalsausprägung η_v einer Person v , bei der die Lösungswahrscheinlichkeit für das Item i (und ebenso die Gegenwahrscheinlichkeit; vgl. □ Abb. 16.2) genau .5 beträgt $P(Y_i = 1 | \eta_v = \beta_i) = P(Y_i = 0 | \eta_v = \beta_i) = .5$. Nimmt also der Personenparameter η für eine Person v den Wert η_v an, der gleich groß ist wie β_i (sprich $\eta_v = \beta_i$), so ergibt sich:

$$P(Y_i = 1 | \eta = \beta_i) = \frac{e^{\eta - \beta_i}}{1 + e^{\eta - \beta_i}} = \frac{e^{\beta_i - \beta_i}}{1 + e^{\beta_i - \beta_i}} = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = .5 \quad (16.7)$$

Anmerkung: Der Schwierigkeitsparameter (Itemparameter) β_i unterscheidet sich sowohl konzeptuell als auch interpretativ vom Schwierigkeitsindex P_i der Itemanalyse (vgl. ► Kap. 7) wie auch vom Interzept α_i in den Messmodellen der KTT (► Kap. 14). Während der Schwierigkeitsindex P_i die „Leichtigkeit“ der Aufgabe/des Items – konkret im dichotomen Fall die mit 100 multiplizierte relative Lösungshäufigkeit – beschreibt, so gibt die Itemschwierigkeit β_i im Kontext der IRT dabei tatsächlich die Schwierigkeit im inhaltlichen Sinne an, der Leichtigkeitsparameter im Kontext der KTT (Interzept α_i) hingegen die inhaltliche Leichtigkeit.

Lösungswahrscheinlichkeit von .5

Unterschied zwischen Schwierigkeitsparameter (IRT), Leichtigkeitsparameter (KTT) und Schwierigkeitsindex der Itemanalyse

16.3.5.4 Parameternormierung

Wie man den vorherigen Ausführungen entnehmen kann, hängt die Lösungswahrscheinlichkeit eines Items für eine konkrete Person v lediglich von der Differenz $\eta_v - \beta_i$ ab. Diese Differenz ist in empirischen Situationen (d. h., bei konkret vorliegenden Itemantworten) nur dann eindeutig bestimmbar, wenn man weitere Annahmen trifft. Um die sog. „Bestimmbarkeit“ der Parameter (auch „Identifikation“) zu gewährleisten, kann man verschiedene zusätzliche Normierungsannahmen treffen:

1. Die einfachste Form der Normierung ist die Festsetzung (auch Fixierung) eines Itemparameters auf einen konkreten Wert. In der Regel wird dann der Wert 0 gewählt. Beispielsweise kann man den Itemschwierigkeitsparameter des ersten Items 1 auf den Wert null fixieren, sodass $\beta_1 = 0$ ist. Im weiteren Verlauf der Parameterschätzung (vgl. ► Abschn. 16.3.8 und 16.3.11) wird hierdurch ein Bezugssystem geschaffen und die oben beschriebene Differenz ist eindeutig bestimmbar.
2. Eine andere Normierungsannahme besteht darin, den Erwartungswert $E(\eta)$ der Personenparameter auf einen konkreten Wert festzulegen. Zumeist wird dann der Wert 0 gewählt, sodass $E(\eta) = 0$ ist.
3. Eine weitere Möglichkeit ist die Annahme, dass die Summe der Itemparameter einem vorher festgelegten Wert entspricht. Auch hier nimmt man zumeist an, dass die Summe über alle p Items null sei: $\sum_{i=1}^p \beta_i = 0$. Bei dieser Form der Annahme handelt es sich um die sog. „Summennormierung“.

Zusatzzannahmen zur Bestimmbarkeit („Identifikation“) der Parameter

Fixierung eines Itemparameters

Festlegung des Erwartungswertes der Personenparameter

Summennormierung

Alle drei Arten von Normierungsannahmen führen zu eindeutigen Lösungen der zu schätzenden Parameter. Welche Form der Normierung in einem konkreten Anwendungsfall gewählt wird, hängt von den inhaltlich zu treffenden Aussagen bezüglich der Parameter ab. Im Fall der sog. „PISA-Studien“ (Programme for International Student Assessment; OECD 2017) wurde beispielsweise für die teilnehmenden Staaten ein ursprüngliches Bezugssystem von 500 (als Mittelwert aller Staaten) und eine Standardabweichung von 100 Punkten für die Leistungsfähigkeitspara-

Weitere Möglichkeiten der Parameternormierung

IC-Funktionen sind im Rasch-Modell parallel verschoben

Bedeutung der Spezifischen Objektivität

meter der Personen vereinbart. Auch diese abweichende Form der Normierung ist somit denkbar (und ebenso andere, die der Joint Scale eine Metrik geben; für ausführliche Ausführungen vgl. ► Kap. 17 und 19).

16.3.6 Spezifische Objektivität der Vergleiche

Im Falle von Modellkonformität (zur Überprüfung ► Abschn. 16.3.11) wird für das Rasch-Modell davon ausgegangen, dass die IC-Funktionen aller Items die gleiche Form aufweisen und lediglich horizontal entlang der Abszisse (Joint Scale) parallel verschoben sind (vgl. □ Abb. 16.4). Inhaltlich bedeutet dies, dass alle Items dasselbe Personenmerkmal gleichermaßen messen, aber auf verschiedenen Schwierigkeitsstufen.

Dieser Sachverhalt ermöglicht die sog. „Spezifische Objektivität der Vergleichs“, die bedeutet, dass ein Vergleich der

1. Fähigkeitsparameterausprägungen $\eta_w - \eta_v$ zweier Personen w und v unabhängig davon erfolgen kann, ob einfache oder schwierige Items verwendet werden;
2. Schwierigkeitsparameter $\beta_j - \beta_i$ zweier Items j und i unabhängig davon erfolgen kann, ob Personen mit niedrigen oder hohen Ausprägungen der Personenvariable η untersucht wurden.

Um die Modelleigenschaft der spezifischen Objektivität genauer zeigen zu können, wählen wir die *Logit-Schreibweise* des Rasch-Modells (► Unter der Lupe).

Unter der Lupe

Logit-Schreibweise des Rasch-Modells

Die Logit-Schreibweise des Rasch-Modells (vgl. Eid und Schmidt 2014) nimmt ihren Ausgang vom Quotienten („Wettquotient“) aus der Lösungswahrscheinlichkeit (Gl. 16.1) und der Gegenwahrscheinlichkeit (Gl. 16.4). Durch Logarithmieren der Wettquotientvariablen erhält man die sog. „Logit-Variable“:

$$\begin{aligned} \ln\left(\frac{P(Y_i = 1 | \eta)}{P(Y_i = 0 | \eta)}\right) &= \ln\left(\frac{e^{\eta - \beta_i}}{1 + e^{\eta - \beta_i}} / \frac{1}{1 + e^{\eta - \beta_i}}\right) \\ &= \ln\left(\frac{e^{\eta - \beta_i}}{1 + e^{\eta - \beta_i}} \cdot \frac{1 + e^{\eta - \beta_i}}{1}\right) \\ &= \ln\left(\frac{e^{\eta - \beta_i}}{1}\right) \\ &= \ln(e^{\eta - \beta_i}) = \eta - \beta_i \end{aligned} \quad (16.8)$$

Gl. (16.8) zeigt, dass die einzelnen Werte der Logit-Variablen, d. h. die logarithmierten Werte der Relation aus Lösungs- und Gegenwahrscheinlichkeit gerade so groß sind wie die Differenz $\eta - \beta_i$ (vgl. ► Abschn. 16.3.5.1).

Logit-Differenz zweier Personen

Zur Veranschaulichung der Spezifischen Objektivität der Vergleiche bilden wir unter Verwendung von Gl. (16.8) für zwei

Personen w und v bei Vorlage eines beliebigen Items i die Differenz ihrer Logit-Werte:

$$\begin{aligned} \ln\left(\frac{P(Y_i = 1 | \eta = \eta_w)}{P(Y_i = 0 | \eta = \eta_w)}\right) - \ln\left(\frac{P(Y_i = 1 | \eta = \eta_v)}{P(Y_i = 0 | \eta = \eta_v)}\right) \\ = (\eta_w - \beta_i) - (\eta_v - \beta_i) = \eta_w - \eta_v \end{aligned} \quad (16.9)$$

Man sieht, dass die Differenz $\eta_w - \eta_v$ unabhängig von der Itemschwierigkeit β_i ist, die durch Kürzen in Gl. (16.9) entfällt. Somit ist gezeigt, dass der Vergleich der Fähigkeitsparameter $\eta_w - \eta_v$ zweier Personen w und v unabhängig davon festgestellt werden kann, ob die zwei Personen ein leichtes oder ein schwieriges Item bearbeitet haben. Der Vergleich ist somit spezifisch objektiv („itemunabhängig“).

Logit-Differenz zweier Items

In gleicher Weise kann unter erneuter Verwendung von Gl. (16.8) nun für zwei Items j und i die Differenz ihrer Logit-Werte bei Bearbeitung durch beliebige Personen bestimmt werden:

$$\begin{aligned} \ln\left(\frac{P(Y_i = 1 | \eta)}{P(Y_i = 0 | \eta)}\right) - \ln\left(\frac{P(Y_j = 1 | \eta)}{P(Y_j = 0 | \eta)}\right) \\ = (\eta - \beta_i) - (\eta - \beta_j) = \beta_j - \beta_i \end{aligned} \quad (16.10)$$

Man sieht, dass die Differenz $\beta_j - \beta_i$ unabhängig von der Personenvariable η ist, die durch Kürzen in Gl. (16.10) entfällt. Somit ist gezeigt, dass der Vergleich der Schwierigkeitsparameter $\beta_j - \beta_i$ zweier Items j und i unabhängig davon festgestellt werden kann, ob die zwei Items j und i von einer Person mit einer niedrigen oder hohen Ausprägung der Personenvariablen η bearbeitet wurden. Der Vergleich ist somit spezifisch objektiv „personenunabhängig“.

Anmerkung:

Wichtig ist aber, dass nicht zwei Personen mit unterschiedlichen Merkmalsausprägungen für den Vergleich herangezogen werden, sondern nur eine Person. Diese darf jede beliebige Merkmalsausprägung haben!

16.3.7 Lokale stochastische Unabhängigkeit

Eine wesentliche Bedingung, die bei Itemhomogenität (= alle Items messen das gleiche eindimensionale latente Merkmal) erfüllt sein muss, ist die sog. „lokale stochastische Unabhängigkeit“ der Antworten von Personen auf die Items. Die lokale (auch bedingte) stochastische Unabhängigkeit bildet die Basis sowohl für die in ▶ Abschn. 16.3.8 beschriebene Parameterschätzung wie auch für die in ▶ Abschn. 16.3.11 beschriebenen empirischen Kontrollen der Modellkonformität.

16.3.7.1 Paarweise Betrachtung der Antworten auf Items

Die lokale stochastische Unabhängigkeit erfordert, dass die Antworten auf zwei beliebige Rasch-homogene Items i und j bei gegebener Personenvariable η paarweise voneinander unabhängig sind. Dies ist dann der Fall, wenn die Verbundwahrscheinlichkeit der Antworten der Items i und j ebenso groß ist wie das Produkt der Einzelwahrscheinlichkeiten gemäß Gl. (16.1) bzw. (16.4) (▶ Abschn. 16.3.7.2).

Die geforderte lokale stochastische Unabhängigkeit Rasch-homogener Items kann nur unter folgenden Voraussetzungen erfüllt sein:

- Die Wahrscheinlichkeit einer konkreten Antwort auf Item i darf nicht von einer konkreten Antwort auf ein anderes Item j abhängen. Dieser Aspekt wäre beispielsweise dann verletzt, wenn die Lösung eines Items i die Lösungswahrscheinlichkeit eines anderen Items j – gegenüber der gemäß der Ausprägung von η zu erwartenden Lösungswahrscheinlichkeit – anhebt/absenkt. Solche Mängel untereinander abhängiger Items sollten möglichst schon bei der Itemgenerierung vermieden werden (▶ Kap. 4).
- Ebenso darf die Wahrscheinlichkeit einer konkreten Antwort von Person v nicht von der konkreten Antwort einer anderen Person w abhängen (vgl. hierzu auch die Annahme unkorrelierter Residuen in der KTT, ▶ Kap. 13). Zur Erfüllung dieses Aspekts ist durch geeignete Maßnahmen der Testdurchführung sicherzustellen, dass eine Person v nach Beantwortung eines Items i bei gleich gebliebener Fähigkeit η_v nicht plötzlich eine höhere Wahrscheinlichkeit einer Richtigantwort auf das nächste Item j hätte, wie es z. B. durch Abschreiben beim leistungsfähigeren Sitznachbarn der Fall wäre.

Lokale stochastische Unabhängigkeit als Basis für die Parameterschätzung und Modellkontrollen

Verbundwahrscheinlichkeit

Bei Itemgenerierung auf Unabhängigkeit der Items achten

Bei Testdurchführung auf Unabhängigkeit der Personen achten

Zusammenfassend: Über die Fähigkeitsvariable η hinausgehend darf die Beantwortung von Item i (Antwort $Y_i = 0$ bzw. $Y_i = 1$) bei gegebener lokaler stochastischer Unabhängigkeit keinen Einfluss auf die Antwortwahrscheinlichkeit eines anderen Items j (Antwort $Y_j = 0$ bzw. $Y_j = 1$) haben.

Die Spezifische Objektivität der Vergleiche ist – bei Modellgültigkeit! – eine besonders hervorzuhebende vorteilhafte Eigenschaft des Rasch-Modells (und bildet auch eine Grundlage für das sog. „adaptive Testen“, ▶ Kap. 20).

Multiplikationstheorem für unabhängige Ereignisse

Formale Bedingung für lokale stochastische Unabhängigkeit

Likelihood der Datenmatrix

Betrachtung der Korrelationen auf lokalen Stufen von η

16.3.7.2 Formale Definition

Durch Anwendung des mathematischen Multiplikationstheorems für unabhängige Ereignisse, demgemäß die Verbundwahrscheinlichkeit mehrerer unabhängiger Ereignisse so groß ist wie das Produkt der Wahrscheinlichkeiten der einzelnen Ereignisse, lässt sich die Forderung nach paarweiser lokaler stochastischer Unabhängigkeit von zwei beliebigen Items i und j formal wie folgt ausdrücken:

$$P(Y_i = y_i, Y_j = y_j | \eta) = P(Y_i = y_i | \eta) \cdot P(Y_j = y_j | \eta) \quad (16.11)$$

Durch sequentielle Anwendung des Multiplikationstheorems kann die Bedingung der lokalen stochastischen Unabhängigkeit unter Benutzung des Produktoperators $\prod_{i=1}^p P(Y_i = y_i | \eta)$ für die Verbundwahrscheinlichkeit eines Antwortmusters von p Items bei gegebenem η (d. h. für eine beliebige Zeile aus einer Datenmatrix) als

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_p = y_p | \eta) = \prod_{i=1}^p P(Y_i = y_i | \eta) \quad (16.12)$$

sowie durch Einbezug aller Zeilen der $v = 1, \dots, n$ Personen für die Verbundwahrscheinlichkeit der ganzen $(n \times p)$ Datenmatrix (*Likelihood L der Datenmatrix*; s. auch ► Abschn. 16.3.8.1) wie folgt formuliert werden:

$$L = \prod_{v=1}^n \prod_{i=1}^p P(Y_i = y_{vi} | \eta_v) \quad (16.13)$$

Die jeweiligen Verbundwahrscheinlichkeiten auf der linken Seite von Gln. (16.11) und (16.12) müssen empirisch als beobachtete (relative) Häufigkeiten aus einer Datenmatrix ermittelt werden; die Einzelwahrscheinlichkeiten auf der rechten Seite findet man für die Richtigantworten ($Y_i = 1$) gemäß Gl. (16.1) und für die Falschantworten ($Y_i = 0$) gemäß Gl. (16.14). Die Annahme der lokalen stochastischen Unabhängigkeit wäre verletzt, wenn die Verbundwahrscheinlichkeiten auf der linken Seite *nicht* dem Produkt der Einzelwahrscheinlichkeiten auf der rechten Seite entsprechen.

16.3.7.3 Veranschaulichung über Korrelationen

Die lokale stochastische Unabhängigkeit der Antworten lässt sich auch anhand der Betrachtung der Korrelationen $\text{Corr}(Y_i, Y_j)$ zwischen beliebigen Itemvariablen Y_i und Y_j veranschaulichen. Dazu hält man gedanklich die latente Personenvariable η auf einem bestimmten Wert (auf einer „lokalen Stufe“) konstant und untersucht die Korrelationen der Itemvariablen nur bei Personen mit dieser Ausprägung des interessierenden Persönlichkeitsmerkmals η . Sofern die Korrelationen zwischen den Itemvariablen bei konstant gehaltenem η auf null zurückgehen, ist die Bedingung der lokalen stochastischen Unabhängigkeit erfüllt. Bei Korrelationen ungleich null wäre die Bedingung der lokalen stochastischen Unabhängigkeit hingegen verletzt. Diese Überlegungen werden im Folgenden zunächst für eine *Situation A* mit einer niedrigeren/durchschnittlichen lokalen Stufe η_v und sodann für eine *Situation B* mit einer höheren lokalen Stufe η_w illustriert.

■ ■ Situation A: Niedrigere lokale Stufe η_v

Man stelle sich zunächst eine Stichprobe von Personen vor, die hinsichtlich eines latenten Merkmals (z. B. Reasoning-Intelligenz) alle die gleiche – und zwar im Vergleich zu Situation B (s. u.) niedrigere – Merkmalsausprägung η_v aufweisen, und betrachte das Antwortverhalten auf zwei beliebige Items i und j aus einem Pool von homogenen Items, die alle Intelligenz erfassen. Ohne Beschränkung der Allgemeinheit gehen wir für beide Items von gleichen Itemschwierigkeiten aus.

Beispielsweise liege die (bedingte) Lösungswahrscheinlichkeit auf der Stufe η_v für die Items i und j jeweils bei .60. Wenn nun die Antworten auf die beiden

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

Items i und j voneinander lokal stochastisch unabhängig sind, müssen theoretisch alle Personen der Stichprobe das Item i mit gleicher (bedingter) Wahrscheinlichkeit beantworten können (Gl. 16.1) und ebenso das Item j , da die antwortbedingende Variable η auf der lokalen Stufe η_v konstant gehalten wird:

		Item j		
		$Y_j = 1$	$Y_j = 0$	
Item i	$Y_i = 1$.60 · .60 = .36	.60 · .40 = .24	.60
	$Y_i = 0$.40 · .60 = .24	.40 · .40 = .16	.40
		.60	.40	$\text{Corr}(Y_i, Y_j) = .00$

Personen mit gleicher Ausprägung η_v werden also gemäß Multiplikationstheorem entweder beide Items mit einer Wahrscheinlichkeit von .36 lösen (wegen $.60 \cdot .60 = .36$) oder nur eines der beiden und das andere nicht mit einer Wahrscheinlichkeit von .24 (wegen $.60 \cdot .40 = .24$ bzw. $.40 \cdot .60 = .24$) oder keines von beiden mit einer Wahrscheinlichkeit von .16 (wegen $.40 \cdot .40 = .16$). Die Berechnung des ϕ -Koeffizienten zur Bestimmung des Zusammenhangs der Itemvariablen zeigt, dass die Korrelation $\text{Corr}(Y_i, Y_j)$ bei lokaler stochastischer Unabhängigkeit gleich null ist.

Wäre innerhalb der Gruppe mit gleicher Intelligenz $\eta = \eta_v$ die Korrelation von null verschieden, so wäre die Forderung nach lokaler stochastischer Unabhängigkeit verletzt. Es wäre dann naheliegend, dass die Korrelation zwischen den beiden Itemvariablen von einem anderen/weiteren als dem interessierenden Merkmal erzeugt wurde.

Nullkorrelation in den Gruppen mit niedriger Merkmalsausprägung

■ ■ Situation B: Höhere lokale Stufe η_w

Nun stelle man sich eine andere (der Einfachheit halber gleich große) Stichprobe von Personen aus der gleichen Population vor, die alle eine höhere Merkmalsausprägung η_w aufweisen. Für diese Gruppe mit höherer Intelligenz gelte eine wesentlich höhere (bedingte) Lösungswahrscheinlichkeit von .90 für jeweils beide Items i und j .

Wenn nun die beiden Items i und j voneinander lokal stochastisch unabhängig sind, müssen theoretisch alle Personen der Stichprobe das Item i mit gleicher (bedingter) Wahrscheinlichkeit beantworten (Gl. 16.1) und ebenso das Item j , da die antwortbedingende Variable η auf der lokalen Stufe η_w konstant gehalten wird:

		Item j		
		$Y_j = 1$	$Y_j = 0$	
Item i	$Y_i = 1$.90 · .90 = .81	.90 · .10 = .09	.90
	$Y_i = 0$.10 · .90 = .09	.10 · .10 = .01	.10
		.90	.10	$\text{Corr}(Y_i, Y_j) = .00$

Personen mit gleicher Ausprägung η_w werden also gemäß Multiplikationstheorem entweder beide Items mit einer Wahrscheinlichkeit von .81 lösen (wegen $.90 \cdot .90 = .81$) oder nur eines der beiden und das andere nicht mit einer Wahrscheinlichkeit von .09 (wegen $.90 \cdot .10 = .09$ bzw. $.10 \cdot .90 = .09$) oder keines von beiden mit einer Wahrscheinlichkeit von .01 (wegen $.10 \cdot .10 = .01$). Die Berechnung des ϕ -Koeffizienten zur Bestimmung des Zusammenhangs der Itemvariablen zeigt, dass die Korrelation $\text{Corr}(Y_i, Y_j)$ bei lokaler stochastischer Unabhängigkeit gleich null ist (Situation c).

Nullkorrelation in den Gruppen mit hoher Merkmalsausprägung

Wäre innerhalb der Gruppe mit gleicher Intelligenz $\eta = \eta_w$ die Korrelation von null verschieden, so wäre die lokale stochastische Unabhängigkeit verletzt. Es wäre dann naheliegend, dass die Korrelation zwischen den beiden Itemvariablen von einem anderen/weiteren als dem interessierenden Merkmal erzeugt wurde.

■ ■ Situation C: Keine Trennung nach lokalen Stufen

Werden hingegen keine nach lokalen Stufen des latenten Merkmals getrennte Stichproben betrachtet, sondern beide Stichproben zusammengefasst in einer gemeinsamen mit den verschiedenen Merkmalsausprägungen η_v und η_w , so zeigt sich ein anderes Bild:

		Item j			
		$Y_j = 1$	$Y_j = 0$		
Item i	$Y_i = 1$	$(.36+.81)/2 = .585$	$(.24+.09)/2 = .165$.75	
	$Y_i = 0$	$(.24+.09)/2 = .165$	$(.16+.09)/2 = .085$.25	
		.75	.25	$Corr(Y_i, Y_j) = .13$	

Als Zeilenrandverteilungen erhalten wir nun die unbedingten Lösungs- und Nichtlösungs wahrscheinlichkeiten von Item i aus dem Durchschnitt der bedingten Lösungs- bzw. Nichtlösungs wahrscheinlichkeiten, also $(.60 + .90)/2 = .75$ sowie $(.40 + .10)/2 = .25$, und als Spaltenrandverteilung jene von Item j , also $(.60 + .90)/2 = .75$ sowie $(.40 + .10)/2 = .25$. Als Wahrscheinlichkeiten in den vier Zellen erhalten wir den Durchschnitt der Wahrscheinlichkeiten aus Situation A und Situation B. Die Korrelation $Corr(Y_i, Y_j)$ ist mit .13 nun ungleich null; über die Gruppen hinweg steht das Lösen von Item i mit dem Lösen von Item j in Zusammenhang.

Mit den Einzeldarstellungen in Situation A und Situation B sowie deren Zusammenlegung in Situation C ist gezeigt, dass die Korrelation der Itemvariablen durch die unterschiedlichen Merkmalsausprägungen von η hervorgerufen wurde. Folglich kann man auch den umgekehrten Weg gehen und bei korrelierenden Itemvariablen prüfen, ob die Korrelationen auf lokalen Stufen einer – hypothetisch – dahinterliegenden latenten Personenvariablen verschwinden (► Abschn. 16.3.7.4).

16.3.7.4 Empirische Nutzungsmöglichkeit

Bei empirischen Fragestellungen kann – im Unterschied zu der gedanklichen Veranschaulichung in ► Abschn. 16.3.7.3 – nicht von separierten Teilgruppen auf konstant gehaltenen Stufen eines latenten Personenmerkmals ausgegangen werden; vielmehr muss man das betreffende Merkmal, das die Korrelationen zwischen den Itemantworten erzeugt haben kann, erst identifizieren. Hierbei erweist sich die empirisch prüfbare Eigenschaft der lokalen stochastischen Unabhängigkeit als sehr nützlich, wie ► Beispiel 16.1 zeigt: Man hält das vermutete Personenmerkmal η auf lokalen Stufen konstant und überprüft, ob die Korrelationen null werden. Wenn dies der Fall ist, kann man davon ausgehen, dass die Items das vermutete Merkmal tatsächlich messen; andernfalls messen die Items etwas anderes.

Beispiel 16.1: Messung des Persönlichkeitsmerkmals „Emotionale Labilität“

Mit der Absicht, das Persönlichkeitsmerkmal „Emotionale Labilität“ zu messen, werden verschiedene Items mit dichotomem Beantwortungsmodus (1 für „stimmt“, 0 für „stimmt nicht“) konstruiert, so z. B. ein Item i „Termindruck und Hektik lösen bei mir körperliche Beschwerden aus“ und ein Item j „Es gibt Zeiten, in denen ich ganz traurig und niedergedrückt bin“ (Quelle: Items 49 und 106 der revidierten

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

Fassung des Freiburger Persönlichkeitsinventars, FPI-R; Fahrenberg et al. 2001). Gesucht werden belastbare Anhaltspunkte, ob beide Items das vermutete Merkmal (hier „Emotionale Labilität“) messen.

Hierzu werden z. B. aus dem Antwortverhalten von $N = 100$ Personen die zwei Zustimmungswahrscheinlichkeiten $P(Y_i = 1)$ und $P(Y_j = 1)$, die zwei Ablehnungswahrscheinlichkeiten $P(Y_i = 0)$, $P(Y_j = 0)$ sowie die vier Verbundwahrscheinlichkeiten $P(Y_i = 1, Y_j = 1)$, $P(Y_i = 1, Y_j = 0)$, $P(Y_i = 0, Y_j = 1)$ und $P(Y_i = 0, Y_j = 0)$ sowie die Korrelationen $\text{Corr}(y_i, y_j)$ festgestellt, zunächst undifferenziert (A) und sodann probeweise geteilt nach niedriger (B) bzw. hoher Ausprägung (C) des interessierenden Merkmals (im Kontext von Persönlichkeitfragebogen bedeutet die Zustimmungswahrscheinlichkeit den Anteil der Personen, die symptomatisch im Sinne einer höheren Merkmalsausprägung geantwortet haben, vgl. ► Kap. 7).

A: Betrachten wir zunächst die undifferenzierten Randwahrscheinlichkeiten der beiden Items:

		Item <i>j</i>		
		$Y_j = 1$	$Y_j = 0$	
Item <i>i</i>	$Y_i = 1$.33	.27	.60
	$Y_i = 0$.07	.33	.40
		.40	.60	$\text{Corr}(Y_i, Y_j) = .38$

Man erkennt, dass das Item *i* das leichtere Item ist (höhere Zustimmungswahrscheinlichkeit $P(Y_i = 1) = .60$), das Item *j* hingegen das schwierigere Item (niedrigere Zustimmungswahrscheinlichkeit $P(Y_j = 1) = .40$). Außerdem erkennt man an $P(Y_i = 1, Y_j = 1) \neq P(Y_i = 1) \cdot P(Y_j = 1)$ (in Zahlen $.33 \neq .60 \cdot .40$), dass die Antwortvariablen nicht unabhängig sind, sondern korrelieren ($\text{Corr}(y_i, y_j) = .38$), was auf eine möglicherweise dahinterliegende Personenvariable η schließen lässt. Die Beantwortung beider Items in symptomatischer Richtung tritt also mit $P(Y_i = 1, Y_j = 1) = .33$ häufiger auf als nach dem Multiplikationstheorem für unabhängige Ereignisse (vgl. Gl. 16.11) mit $.6 \cdot .4 = .24$ zu erwarten wäre.

Zur Kontrolle, ob die vermutete latente Personenvariable η die beobachtete Korrelation erklären kann, teilen wir die 100 Personen in zwei Gruppen gleichen Umfangs, wobei die eine Gruppe hinsichtlich der vermuteten Personenvariablen η eine niedrigere Ausprägung η_v , die andere hingegen eine höhere Ausprägung η_w habe (man nimmt also eine Betrachtung auf zwei lokalen Stufen von η vor, ► Abschn. 16.3.7.3). In den zwei Gruppen seien die in B (für $\eta = \eta_v$) bzw. C (für $\eta = \eta_w$) aufgeführten bedingten Zustimmungs-, Ablehnungs- und Verbundwahrscheinlichkeiten beobachtbar.

B: Probeweise Teilung nach niedriger Ausprägung:

		Item <i>j</i>		
		$Y_j = 1$	$Y_j = 0$	
Item <i>i</i>	$Y_i = 1$.03	.27	.30
	$Y_i = 0$.07	.63	.70
		.10	.90	$\text{Corr}(Y_i, Y_j) = .00$

C: Probeweise Teilung nach hoher Ausprägung:

		Item j		
		$Y_j = 1$	$Y_j = 0$	
Item i	$Y_i = 1$.63	.27	.90
	$Y_i = 0$.07	.03	.10
		.70	.30	$\text{Corr}(Y_i, Y_j) = .00$

An den Randwahrscheinlichkeiten sieht man nun, dass – jeweils im Vergleich zu den unbedingten Randwahrscheinlichkeiten – die bedingte Wahrscheinlichkeit, dem Item i bzw. j zuzustimmen, für Personen mit $\eta = \eta_v$ auf $P(Y_i = 1 | \eta_v) = .30$ bzw. $P(Y_j = 1 | \eta_v) = .10$ gefallen ist; andererseits ist die bedingte Wahrscheinlichkeit, dem Item i bzw. j zuzustimmen, für Personen mit $\eta = \eta_w$ auf $P(Y_i = 1 | \eta_w) = .90$ bzw. $P(Y_j = 1 | \eta_w) = .70$ angestiegen, was wegen $\eta_v < \eta_w$ zu erwarten war.

Die Forderung nach lokaler stochastischer Unabhängigkeit wäre erfüllt, wenn die Itemantworten auf den lokalen Stufen nun voneinander unabhängig sind, da die vermutete „Verursachung“, d. h. der Einfluss der latenten Personenvariablen, durch Konstanthaltung eliminiert wurde. Die Überprüfung, ob bei Konstanthaltung von η die Forderung nach lokaler stochastischer Unabhängigkeit erfüllt ist, ergibt gemäß Gl. (16.11)

– für die Stufe $\eta = \eta_v$:

$$\begin{aligned} P(Y_i = 1, Y_j = 1 | \eta_v) &= P(Y_i = 1 | \eta_v) \cdot P(Y_j = 1 | \eta_v) \\ &.03 = .30 \cdot .10 \\ \text{Corr}(y_i, y_j) &= 0 \end{aligned}$$

– und für die Stufe $\eta = \eta_w$:

$$\begin{aligned} P(Y_i = 1, Y_j = 1 | \eta_w) &= P(Y_i = 1 | \eta_w) \cdot P(Y_j = 1 | \eta_w) \\ &.63 = .90 \cdot .70 \\ \text{Corr}(y_i, y_j) &= 0 \end{aligned}$$

Sowohl auf der lokalen Stufe η_v (niedrige Merkmalsausprägung) als auch auf der lokalen Stufe η_w (hohe Merkmalsausprägung) erfüllen die bedingten Verbundwahrscheinlichkeiten nun das Multiplikationstheorem für unabhängige Ereignisse, $\text{Corr}(Y_i, Y_j)$ ist jeweils null. Somit liegt eine lokale stochastische Unabhängigkeit vor und man darf annehmen, dass beide Items dasselbe latente Merkmal „Emotionale Labilität“ messen.

Nullkorrelationen (im Sinne der lokalen stochastischen Unabhängigkeit) sind nicht selbstverständlich.

Anmerkung: Gründlichkeitshalber sei hier nochmals erwähnt, dass (näherungsweise) Nullkorrelationen nur dann zu erwarten sind, wenn die Aufteilung in zwei Gruppen nach einer für den Zusammenhang verantwortlichen latenten Personenvariable erfolgt. Hätte man die Stichprobe aus obigem Beispiel nicht nach dem richtigen latenten Merkmal „Emotionalität“ geteilt, sondern nach einem anderen, fälschlich vermuteten Merkmal (z. B. „Extraversion“), das die Korrelationen nicht verursacht, so wären die Korrelationen in den beiden Gruppen gleich oder zumindest ähnlich geblieben wie in der ungeteilten Stichprobe, da die beiden Items eben nicht Extraversion, sondern ein anderes latentes Merkmal messen.

16.3.8 Parameterschätzung

Da alle Modelle der IRT und hier insbesondere das Rasch-Modell die Wahrscheinlichkeit einer Datenmatrix und der in ihr enthaltenen Antwortmuster in Abhängigkeit von Item- und Personenparametern beschreiben, müssen diese Parameter geschätzt werden. Hierbei werden die Parameter so bestimmt, dass die empirisch beobachteten Werte der Antwortvariable Y (genau genommen das gesamte Antwortmuster) in Abhängigkeit von β und η in der Modellgleichung Gl. (16.6) eine maximale Wahrscheinlichkeit aufweisen. Als Optimierungskriterium dient grundsätzlich zumeist die Likelihood der Datenmatrix, die unter der Bedingung der lokalen stochastischen Unabhängigkeit gemäß Gl. (16.13) kalkuliert wird.

Zur Schätzung der Item- und Personenparameter haben sich in den vergangenen Jahrzehnten unterschiedliche Herangehensweisen etabliert, die jeweils Vor- und Nachteile mit sich bringen. Allen gemein ist, dass für die Schätzung die Annahme der lokalen stochastischen Unabhängigkeit benötigt wird (► Abschn. 16.3.7).

Innerhalb der Maximum-Likelihood-Verfahren (ML-Verfahren) lassen sich folgende Schätzverfahren unterscheiden:

- Joint Maximum Likelihood (JML, ► Abschn. 16.3.8.1)
- Conditional Maximum Likelihood (CML, ► Abschn. 16.3.8.2)
- Marginal Maximum Likelihood (MML), das in 2PL- (► Abschn. 16.4) und 3PL-Modellen (► Abschn. 16.5) zur Anwendung kommt

Ferner ist es möglich, eine Schätzung nach dem Bayes'schen Ansatz vorzunehmen (► Abschn. 16.3.8.4). Im weiteren Verlauf sollen diese und weitere Herangehensweisen beschrieben werden (für einen zusammenfassenden Überblick s. Cai und Thissen 2014; Näheres s. ► Kap. 19).

16.3.8.1 Joint Maximum Likelihood (JML) zur Schätzung der Item- und Personenparameter

Das JML-Schätzverfahren ist eine Herangehensweise, bei der gleichzeitig (daher die Bezeichnung „joint“) die Item- und die Personenparameter geschätzt werden. (Die beiden anderen ML-basierten Verfahrensweisen, CML und MML, schätzen zunächst nur die Itemparameter.) Beim JML-Schätzverfahren werden die Personenparameter als den Personen zugehörige, d. h. fixe Größen aufgefasst und nicht als zufällige Realisierungen einer (latenten) Variable, die einer bestimmten Verteilung folgen (z. B. der Normalverteilung). Bei der Schätzung werden daher die Item- und Personenparameter gleich behandelt.

■■ Parameterschätzung

Die Schätzung der unbekannten Item- und Personenparameter nimmt ihren Ausgang bei den einzelnen Reaktionen y_{vi} aller Personen auf alle Items. Sie werden in einer Datenmatrix Y gesammelt, in der die $v = 1, \dots, n$ Personen die Zeilen und die $i = 1, \dots, p$ Items die Spalten bilden.

Unter Benutzung der Modellgleichung (Gl. 16.6) für die Wahrscheinlichkeiten der einzelnen Itemantworten y_{vi} (► Abschn. 16.3.3) ergibt sich die Wahrscheinlichkeit für die gesamte Datenmatrix $P(y_{vi})$ wegen der postulierten lokalen stochastischen Unabhängigkeit durch systematisch wiederholte Anwendung des Multiplikationstheorems für unabhängige Ereignisse. Hierbei werden sämtliche $P(y_{vi})$ zeilen- oder spaltenweise miteinander multipliziert, was mit zwei Produktoperatoren für die Zeilen und für die Spalten ausgedrückt werden kann. Die Wahrscheinlichkeit der beobachteten Daten $P(Y = y_{vi})$ in Abhängigkeit von den unbekannten Parametern η_v und β_i wird als sog. „Likelihood-Funktion“ L bezeichnet:

$$L = \prod_{v=1}^n \prod_{i=1}^p P(Y_i = y_{iv} | \eta_v, \beta_i), \quad (16.14)$$

Optimierung der Likelihood der Datenmatrix

Datenmatrix

Likelihood-Funktion

wobei gemäß Gl. (16.6)

$$P(Y_i = y_{vi} \mid \eta_v, \beta_i) = \frac{e^{y_{vi}(\eta_v - \beta_i)}}{1 + e^{\eta_v - \beta_i}}$$

Prinzipiell kann die Likelihood Werte zwischen 0 und 1 annehmen. Das Ziel besteht vereinfacht gesprochen darin, durch die Wahl von geeigneten Werten für die unbekannten Parameter eine hohe Likelihood zu erzielen (Beispiel 16.2).

Beispiel 16.2: Likelihood-Funktion

Zur Illustration der Parameterschätzung und des Verhaltens der Likelihood-Funktion in Abhängigkeit von verschiedenen Parameterwerten nehmen wir an, es hätten drei Personen zwei dichotome Items bearbeitet und dabei folgendes Antwortverhalten (Datenmatrix (Y)) gezeigt. Hierbei bedeutet ($y_{vi} = 1$), dass Person v das Item i bejaht bzw. gelöst hat; ($y_{vi} = 0$) bedeutet, dass das Item i nicht bejaht bzw. nicht gelöst wurde:

		Item		
		1	2	Zeilensumme
Person	1	$y_{11} = 1$	$y_{12} = 1$	$y_{1\cdot} = 2$
	2	$y_{21} = 1$	$y_{22} = 0$	$y_{2\cdot} = 1$
	3	$y_{31} = 0$	$y_{32} = 0$	$y_{3\cdot} = 0$
Spaltensumme		$y_{\cdot 1} = 2$	$y_{\cdot 2} = 1$	

Es stellt sich nun die Frage, welche Werte der dahinterliegenden Item- und Personenparameter β_i und η_v diese Datenmatrix am wahrscheinlichsten erzeugt haben. Hierfür untersuchen wir die Likelihood der Datenmatrix, indem wir verschiedene Werte für die Item- und Personenparameter auswählen: günstige Werte werden zu einer höheren, weniger günstige hingegen nur zu einer niedrigen Likelihood für die beobachtete Datenmatrix führen. Zur Veranschaulichung wählen wir zunächst günstige Parameterwerte und vergleichen die resultierende Likelihood nachfolgend mit der Likelihood bei ungünstig gewählten Parameterwerten.

Um günstige Parameterwerte zu finden, stellen wir zunächst an den Spaltensummen fest, dass Item 1 offensichtlich leichter zu bejahen ist als Item 2. Deshalb wählen wir für Item 1 einen niedrigen Schwierigkeitsparameter, z. B. $\beta_1 = -1$, und für Item 2 einen höheren, z. B. $\beta_2 = +1$.

Darüber hinaus stellen wir an den Zeilensummen fest, dass Person 1 offensichtlich eine höhere Merkmalsausprägung als Person 2 und Person 2 eine höhere als Person 3 aufweist. Deshalb wählen wir für Person 1 einen hohen Personenparameter, z. B. $\eta_1 = 2$, für Person 2 einen mittleren, z. B. $\eta_2 = 0$, und für Person 3 einen niedrigen, z. B. $\eta_3 = -2$.

Durch Einsetzen der beobachteten Daten y_{vi} und der gewählten Parameter für η_v und β_i in die Modellgleichung (Gl. 16.6) und weiter in die Likelihood-Funktion L (Gl. 16.14) können die Wahrscheinlichkeiten für die sechs Zellen und die Likelihood für die gesamte Datenmatrix $Y = (y_{vi})$ berechnet werden:

16 Likelihood bei günstigen Parameterwerten

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

		Item	
		1	2
Person	1	.953	.731
	2	.731	.731
	3	.731	.953

Die Likelihood für die gesamte Datenmatrix ergibt sich aus dem Produkt der Wahrscheinlichkeiten $P(y_{vi})$ für die empirisch beobachteten Antworten y_{vi} in den sechs Zellen:

$$\begin{aligned}
 L &= \prod_{v=1}^n \prod_{i=1}^p P(Y_i = y_{iv} | \eta_v, \beta_i) = \frac{e^{(y_{v1}(\eta_v - \beta_1))}}{1 + e^{\eta_v - \beta_1}} \\
 L &= \frac{\exp(y_{11}(\eta_1 - \beta_1))}{1 + \exp(\eta_1 - \beta_1)} \cdot \frac{\exp(y_{12}(\eta_1 - \beta_2))}{1 + \exp(\eta_1 - \beta_2)} \cdot \frac{\exp(y_{21}(\eta_2 - \beta_1))}{1 + \exp(\eta_2 - \beta_1)} \\
 &\quad \cdot \frac{\exp(y_{22}(\eta_2 - \beta_2))}{1 + \exp(\eta_2 - \beta_2)} \cdot \frac{\exp(y_{31}(\eta_3 - \beta_1))}{1 + \exp(\eta_3 - \beta_1)} \cdot \frac{\exp(y_{32}(\eta_3 - \beta_2))}{1 + \exp(\eta_3 - \beta_2)} \\
 &= \frac{\exp(1(2 - (-1)))}{1 + \exp(2 - (-1))} \cdot \frac{\exp(1(2 - 1))}{1 + \exp(2 - 1)} \cdot \frac{\exp(1(0 - (-1)))}{1 + \exp(0 - (-1))} \\
 &\quad \cdot \frac{\exp(0(0 - 1))}{1 + \exp(0 - 1)} \cdot \frac{\exp(0((-2) - (-1)))}{1 + \exp((-2) - (-1))} \cdot \frac{\exp(0((-2) - 1))}{1 + \exp((-2) - 1)} \\
 &\approx .953 \cdot .731 \cdot .731 \cdot .731 \cdot .953 \approx .259
 \end{aligned}$$

Die Likelihood für die gesamte Datenmatrix ist mit $L = .259$ hier relativ hoch und nach Gl. (16.6) erhält man für die empirisch beobachteten Daten y_{vi} in den sechs Zellen relativ hohe Wahrscheinlichkeiten; man kann also davon ausgehen, dass es sich bei den gewählten Werten um eher günstige Personenparameterwerte handelt.

Bei ungünstig gewählten Parameterwerten ist die Likelihood hingegen deutlich niedriger: Hätten wir z. B. für die beste Person den niedrigsten Personenparameterwert und umgekehrt gewählt, also $\eta_1 = -2$, $\eta_2 = 0$ und $\eta_3 = 2$, so würden wir bei unveränderten Werten von β_i nur folgende Wahrscheinlichkeiten $P(y_{vi})$ für die sechs Zellen der Datenmatrix erhalten:

		Item	
		1	2
Person	1	.269	.047
	2	.731	.731
	3	.047	.269

Man erkennt, dass ungünstig gewählte Personenparameterwerte zu deutlich niedrigeren Wahrscheinlichkeiten für die sechs Zellen führen, weshalb sie als beste Schätzer für die unbekannten Parameter ausgeschlossen werden können. Die resultierende Likelihood wäre lediglich:

$$L = .269 \cdot .731 \cdot .047 \cdot .047 \cdot .731 \cdot .269 = .00009$$

Die Höhe der Likelihood variiert also in Abhängigkeit von den gewählten Parameterwerten. Sie erreicht das für eine gegebene Datenmatrix mögliche Maximum dann, wenn bei der Parameterschätzung jeweils optimale Werte sowohl für die Personenparameter als auch für die Itemparameter gefunden wurden.

Likelihood bei ungünstig gewählten Parameterwerten

Simultane Schätzung der Item- und Personenparameter

Score-Funktion

Summenscore als suffiziente Statistik

Separierbarkeit der Parameter und Stichprobenunabhängigkeit

Anmerkung: Bei der geringen Anzahl von Daten kann für die hier gewählte Veranschaulichung keine optimale Personenparameterschätzung bestimmt werden, da der Datensatz zu klein ist.

Je höher die Likelihood für die empirisch beobachtete Datenmatrix in Abhängigkeit der gewählten Werte für η_v und β_i ausfällt, desto wahrscheinlicher ist es, die richtigen Werte für die Parameter gefunden zu haben. Durch systematische Veränderung der Werte versucht man, das erzielbare Maximum der Likelihood zu finden (*ML-Verfahren*). Als beste Schätzer für die unbekannten Parameter gelten jene Werte, bei denen die Likelihood unter der Annahme lokaler stochastischer Unabhängigkeit ihren relativen Maximalwert erreicht.

Prozedural wird also das Maximum der Likelihood-Funktion aller beobachteter Itemantworten über alle Personen hinweg gesucht:

$$\varphi_{ML} = \arg \max L(\varphi)$$

Dabei ist „*arg max*“ eine Funktion zur Bestimmung der Stelle φ_{ML} , an der die Likelihood-Funktion $L(\varphi)$ für die möglichen Parameterausprägungen, die im Parametervektor φ zusammengefasst sind, maximal wird; mit diesen Parameterausprägungen sind die beobachteten Daten bei gegebenem Gesamtmodell maximal wahrscheinlich.

Um die Stelle φ_{ML} zu finden, wird beim JML-Schätzer zunächst die erste Ableitung der logarithmierten Funktion L gebildet. Daraus resultiert eine sog. „Score-Funktion“ l' . Die Nullstelle der Score-Funktion l' , d.h. der resultierende Parametervektor, wird dann numerisch (z.B. mit einem Newton-artigen Algorithmus) bestimmt.

Um Standardfehlerschätzungen zu erhalten, kann man bei ML-Schätzungen ganz allgemein die zweite Ableitung l'' der logarithmierten Likelihood bilden. Aus der resultierenden Matrix, auch *Hesse-Matrix* genannt, lässt sich – nachdem man sie mit (-1) multipliziert hat – die Inverse bilden, die sog. „beobachtete Fisher-Informationsmatrix“. Ihre Diagonale enthält die Varianzen der Parameterschätzungen, die Wurzel dieser Varianzen sind die jeweiligen Standardfehlerschätzungen.

Alternativ kann man auch den Erwartungswert der Ableitung der logarithmierten Likelihood-Funktion anstelle der ML-Parameterschätzung bilden. Die resultierende sog. „erwartete Fisher-Informationsmatrix“ liefert ebenfalls Standardfehlerschätzungen.

16.3.8.2 Conditional Maximum Likelihood (CML)

Das CML-Schätzverfahren ist ein Vorgehen, bei dem keine simultane Schätzung der Personenparameter nötig ist. So können zunächst nur die Itemparameter geschätzt werden. Mit diesem häufig im Rasch-Modell angewandten Verfahren umgeht man die Problematik, dass die simultane Schätzung von Personen- und Itemparametern zu verzerrten (inkonsistenten) Schätzungen der Parameter führen kann. Dabei ist der jeweilige Summenscore, d.h. die Zahl der beantworteten Aufgaben einer Testperson, eine suffiziente Statistik, die Eingang in die Schätzung findet.

Der Vorteil dieser Vorgehensweise besteht nicht nur darin, dass die Inkonsistenzproblematik umgangen wird, sondern vor allem auch darin, dass die Itemparameter separiert von den Personenparametern geschätzt werden können. Diese *Separierbarkeit der Parameter* wird mit dem etwas missverständlichen Ausdruck der *Stichprobenunabhängigkeit* bezeichnet, die für die Überprüfung der Modellkonformität (► Abschn. 16.3.11) eine wesentliche Rolle spielt. Das Faktum, dass die Personenparameter in einem zweiten Schritt (► Abschn. 16.3.8.5) geschätzt werden müssen, stellt einen vernachlässigbaren Nachteil dar. Dabei lässt sich das CML-Verfahren nur beim Rasch-Modell anwenden, nicht hingegen beim

sog. „2PL-Modell“ (► Abschn. 16.4), das neben dem Itemschwierigkeitsparameter auch einen Diskriminationsparameter aufweist.

Prozedural wird ähnlich wie bei der JML-Schätzung vorgegangen. Es wird aber im Unterschied zur JML-Schätzung eine bedingte („conditional“) Log-Likelihood maximiert. Wie auch zuvor wird die Score-Funktion gebildet und die Nullstellen werden anhand eines Newton-artigen Algorithmus gesucht. Die Standardfehlerschätzungen der Parameterschätzungen werden anhand der beobachteten oder erwarteten Fisher-Informationsmatrix gebildet. Der wesentliche Unterschied ist, dass über die Personenparameter keine Aussagen (im Sinne von Schätzungen) vorgenommen werden.

■■ Veranschaulichung

Bezogen auf die in ► Abschn. 16.1 verwendeten Items des *Law School Admission Test* (LSAT) seien die Ergebnisse einer CML-Schätzung veranschaulicht. Die Analysen wurden mit dem „eRm-Package“ der R-Project-Software durchgeführt. Sie ergaben Folgendes:

```
Item (Category) Difficulty Parameters (eta): with 0.95 CI:
  Estimate Std. Error lower CI upper CI
Item11 -1.256
Item12  0.475   0.070  0.338  0.612
Item13  1.236   0.069  1.101  1.371
Item14  0.168   0.073  0.026  0.311
Item15 -0.623   0.086 -0.792 -0.455
```

Hinweis: Im Software-Output werden entgegen unserer Notation die Itemparameter mit „eta“ statt mit „beta“ (β) gekennzeichnet.

Hierbei ist zu beachten, dass der Itemschwierigkeitsparameter des 11. Items nicht geschätzt wurde; vielmehr hat eine (Summen-)Normierung stattgefunden (vgl. ► Abschn. 16.3.5.4), wonach dessen Itemschwierigkeitsparameter (-1.256) die negative Summe der geschätzten Itemschwierigkeiten aller anderen Items ist, sodass die Summe der Itemschwierigkeiten insgesamt null ergibt. Die verbleibenden Items wurden unter dieser Normierungsbedingung/Restriktion geschätzt. Zusätzlich zur eigentlichen Schätzung („Estimate“) wurden Standardfehlerschätzungen für die Punktschätzungen der Itemschwierigkeitsparameter („Std. Error“) und dazu gehörige Konfidenzintervallgrenzen (5 %- und 95 %-Grenzen; „lower CI“ und „upper CI“ als β_i plus/minus 2 Std. Error) ausgegeben.

Die Schätzung der Itemparameter erfolgt bei der CML-Schätzung ohne Schätzung der Personenparameter. Die □ Abb. 16.6 veranschaulicht die IC-Funktionen der fünf Items. Wie man der Abbildung entnehmen kann, unterscheiden sich die Items voneinander: Item 11 ist das leichteste, es kann schon bei einer unterdurchschnittlichen Merkmalsausprägung von -1.256 mit 50 %iger Wahrscheinlichkeit gelöst werden. Es folgen die Items 15, 14 und 12. Das schwierigste Item ist Item 13, das erst bei einer überdurchschnittlichen Merkmalsausprägung von $+1.236$ mit 50 %iger Wahrscheinlichkeit gelöst werden kann (in den Zusatzmaterialien zum Buch ist die Syntax für die Berechnungen und die Grafik nachvollziehbar wiedergegeben, ► Abschn. 16.8).

16.3.8.3 Marginal Maximum Likelihood (MML)

Da das CML-Schätzverfahren auf der einen Seite die Inkonsistenzproblematik adressiert, auf der anderen Seite aber auf Item-Response-Modelle aus der Rasch-Familie beschränkt ist, wurde als Alternative (z. B. für sog. „mehrparametrische Item-Response-Modelle“ wie das 2PL- oder das 3PL-Modell; vgl. ► Abschn. 16.4 und 16.5) das MML-Schätzverfahren entwickelt. Bei diesem Verfahren werden die Personenparameter als zufällige Realisierungen einer latenten Variable η aufgefasst, die einer zuvor festgelegten Verteilung(-sform), z. B. der Normalverteilung folgen. Durch die Annahme/Festlegung der Verteilung der latenten Variable η re-

Annahme einer (Normal-)Verteilung der Personenparameter

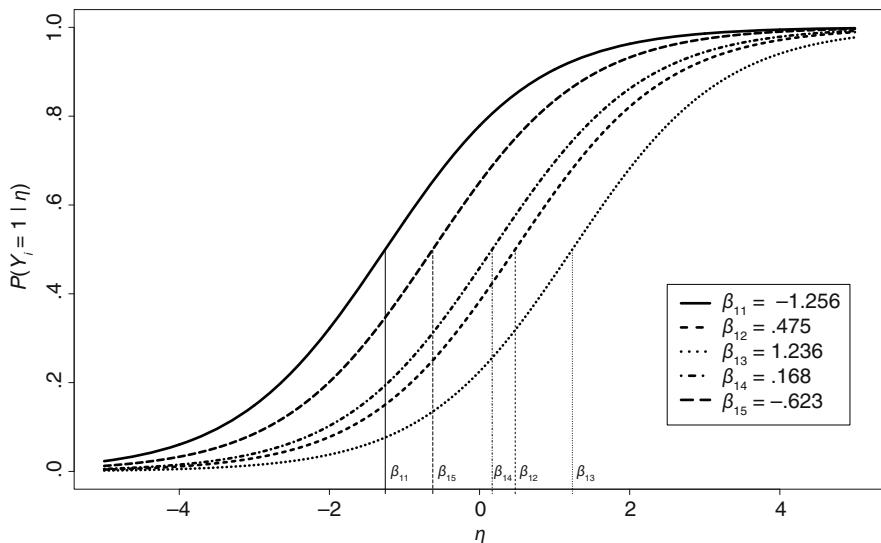


Abb. 16.6 IC-Funktionen der Items aus dem LSAT. Veranschaulicht werden die anhand des CML-Schätzverfahrens erhaltenen Itemparameterschätzung der fünf LSAT-Items

duziert sich die Zahl der zu schätzenden Parameter ganz erheblich, da nicht mehr für jede Person einer Testung ein Personenparameter geschätzt wird, sondern vielmehr nur noch die Verteilungsparameter (z. B. Mittelwert und Varianz im Falle der Normalverteilung) im zu schätzenden Modell enthalten sind. Selbst diese lassen sich durch gezielte Annahmen (z. B. eine standardnormalverteilte latente Variable η mit einem Mittelwert 0 und einer Varianz von 1) reduzieren.

Prozedural werden, wie der Name „marginal“ suggeriert, die unbedingten Antwortmusterwahrscheinlichkeiten (also jene, die nicht von den Personenparametern abhängen, sondern bei denen die Personenparameter „herausintegriert“ wurden) maximiert (vgl. hierzu auch Latent-Class-Analyse, LCA, ▶ Kap. 22). Erneut werden nur die Itemparameter geschätzt. Bei der Maximierung selbst werden sog. „Quadraturverfahren“ für die numerische Integration innerhalb der Likelihood-Funktion benötigt (zu technischen Details ▶ Kap. 19). Aufwendig ist zudem die Bestimmung der Standardfehlerschätzungen für die Itemparameter (vgl. auch Bock und Aitkin 1981).

Maximierung der unbedingten Antwortmusterwahrscheinlichkeiten

16

A-posteriori-Verteilung der Parameter

16.3.8.4 Bayes'sche Schätzung der Itemparameter

Einen anderen Weg der Bestimmung der Itemparameter stellt die Schätzung nach dem Bayes'schen Ansatz dar. Im Unterschied zu den ML-Schätzverfahren, die zuvor dargestellt wurden, werden im Rahmen dieser Konzeption die Itemparameter nicht als fixe Größen angenommen, sondern vielmehr als Variablen mit einer eigenen Verteilung. Im Rahmen dieses Schätzverfahrens ist daher die sog. „A-posteriori-Verteilung“ der/sämtlicher Itemparameter $f(\Theta|Y = y)$ von Interesse. Um diese zu erhalten, werden einerseits A-priori-Annahmen über die Itemparameterverteilungen (z. B. Normalverteilung) getroffen, andererseits werden im Rahmen eines Samplings (z. B. Gibbs-Sampler) die spezifizierten Verteilungen generiert. Auf technische Details soll an dieser Stelle nicht eingegangen werden, hierzu sei auf ▶ Kap. 19 verwiesen. Wesentlich ist jedoch, dass die Konzeption der Bayes'schen Schätzung unmittelbar aus der Generierung der A-posteriori-Verteilung eine Herleitung der Konfidenzintervalle für die Itemparameter und damit ihrer „Unreliabilität“ ermöglicht.

16.3.8.5 Unconditional und Weighted Maximum Likelihood zur Schätzung der Personenparameter

Neben der bereits erwähnten gemeinsamen Schätzung der Personen- und Itemparameter anhand einer JML-Schätzung (► Abschn. 16.3.8.1) kann die Personenparameterschätzung auch auf Grundlage bereits geschätzter Itemparameter erfolgen, da es bei vorliegenden CML-Schätzwerten der Itemparameter möglich ist, diese in die Likelihood-Grundgleichung einzusetzen und die Personenparameter daraus zu schätzen. Diese Herangehensweise wird auch als Schätzung der *Unconditional Maximum Likelihood* (UML) bezeichnet. Hierbei besteht das Problem, dass die Itemparameterschätzungen nicht zwingend mit den realen Ausprägungen der Itemparameter in der Population übereinstimmen, sondern mit einer gewissen Unreliabilität behaftet sind. Abhilfe kann man durch die Verwendung möglichst großer Stichproben schaffen, sodass die Itemparameterschätzungen eine kleinere Schwankung erfahren (im Sinne kleinerer Konfidenzintervalle).

Wegen der vorteilhaften Eigenschaft spezifischer objektiver Vergleiche im Rasch-Modell (vgl. ► Abschn. 16.3.6) und der dadurch möglichen bedingten Likelihood-Schätzung (CML-Schätzung) der Itemparameter bilden die „Summenscores“ eine suffiziente („erschöpfende“) Statistik für die Schätzung der Personenparameter η_v . Als Summenscore wird die Anzahl der von einer Testperson gelösten/bejahten Items bezeichnet. Aus der Höhe des Summenscores kann unmittelbar auf die Merkmalsausprägung η_v geschlossen werden, weil sich bei gleicher Anzahl gelöster Aufgaben trotz unterschiedlicher Antwortmuster die gleichen Personenparameterschätzungen ergeben. Es ist im Rasch-Modell also nicht entscheidend, welche Aufgaben gelöst/bejaht wurden, sondern lediglich deren Anzahl (also der Summenscore).

Beziiglich des Summenscores ist festzuhalten, dass es bei beispielsweise fünf Items zwar 32 Antwortmuster, aber nur sechs verschiedene Summenscores gibt, da die Summe (Anzahl) der mit „1“ kodierten gelösten/bejahten Aufgaben nur zwischen null und fünf variieren kann.

Ein Problem für die Parameterschätzung besteht darin, dass man für Personen, die keine Aufgabe (null Mal die „1“) oder alle Aufgaben (fünf Mal die „1“) gelöst haben, keine realistischen Schätzungen der Personenparameter erhält, da diese gegen minus unendlich bzw. plus unendlich tendieren. Um auch für solche Personen Schätzungen der Personenparameter zu erhalten, kann man alternativ eine Schätzung der *Weighted Maximum Likelihood* (WML) durchführen (Warm 1989; s. auch Rost 2004, S. 313 f.), die sich dadurch auszeichnet, dass sie insgesamt weniger verzerrte Personenparameterschätzungen für endliche Stichproben liefert.

Summenscore als erschöpfende Statistik

16.3.8.6 Weitere Verfahren zur Personenparameterschätzung

Neben den zuvor beschriebenen Verfahrensweisen zur Schätzung von Personenparametern sind auch Bayes'sche Verfahrensweisen üblich. Dazu gehören die Expected-a-posteriori-Schätzung (EAP) und Maximum-a-posteriori-Schätzung (MAP).

Bei der EAP-Schätzung des Personenparameters wird der Erwartungswert der A-posteriori-Verteilung als Schätzer des Personenparameters verwendet. Analog wird bei der MAP-Schätzung das Maximum, d. h. der Modalwert der A-posteriori-Verteilung als (Punkt-)Schätzer des Personenparameters verwendet.

In beide Verfahrensweisen fließen die zuvor geschätzten Itemparameter ein. Daher sind diese Verfahrensweisen sog. „empirical Bayes-Schätzer“, bei denen ein Teil der Parameter (hier Itemparameter) aus den Daten geschätzt wird und dann in einen zweiten Schritt (hier bei der Personenparameterschätzung) einfließen. Der Nachteil ist, dass sich die Verzerrtheit der Itemparameter bei endlicher Itemzahl und endlichen Stichprobengrößen auf die Schätzung im zweiten Schritt auswirkt. Außerdem fließen in beide Schätzverfahren Annahmen über die Personenparameterverteilung (z. B. Normalverteilung) ein. Sofern die Annahmen nicht

Expected-a-posteriori-Schätzung (EAP) und Maximum-a-posteriori-Schätzung (MAP)

korrekt sind, resultieren daraus falsche Schlüsse in Bezug auf die Merkmalsausprägungen der Personen.

Der Vergleich der ML- und der Bayes-Schätzer ergibt, dass ML-Schätzungen die geringste Verzerrung (Bias) aufweisen, während der MAP-Schätzer die geringste Variabilität aufweist (im Sinne eines Konfidenzintervalls). Beide Verfahrenstypen (ML- und Bayes-Schätzung) finden ihre Anwendung, einen perfekten Schätzer gibt es gegenwärtig nicht.

16.3.9 Iteminformationsfunktion

Breite der Konfidenzintervalle abhängig von verwendeten Items

Nach erfolgter Schätzung der Personenparameter lassen sich für diese auf Grundlage der ML-Methode Konfidenzintervalle berechnen. Die Breite der Konfidenzintervalle differiert in Abhängigkeit der verwendeten Items und deren jeweiliger Iteminformation.

Warum ein Item unterschiedlich viel Information liefern kann, wollen wir wie folgt verdeutlichen: Grundsätzlich erscheinen, wie man an den Personenvergleichen sehen kann (vgl. ► Abschn. 16.3.6), alle Items eines homogenen Itempools zur Erfassung von Unterschieden der Merkmalsausprägungen geeignet. Dennoch darf aber nicht der Eindruck entstehen, dass jedes Item über die Merkmalsausprägungen verschiedener Personen gleich viel Information liefert. Vielmehr macht die logistische IC-Funktion deutlich, dass die Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ ihren stärksten Zuwachs gerade dann aufweist, wenn die Itemschwierigkeit β_i mit der Merkmalsausprägung η_v übereinstimmt.

Untersucht man die Unterschiede der Lösungswahrscheinlichkeit systematisch für immer kleiner werdende Merkmalsdifferenzen, so erhält man als Grenzfall den Differenzenquotienten, der die Steigung der IC-Funktion angibt (im dichotomen Rasch-Modell variiert die Steigung bei gegebener Fähigkeit η_v in Abhängigkeit von der Differenz zwischen Fähigkeit und Itemschwierigkeit). Je größer die Steigung der IC-Funktion ist, desto höher ist der Gewinn an Information durch Anwendung des Items i bei Person v und desto größer ist die Iteminformationsfunktion, die mit I_i bezeichnet wird.

Die numerische Ausprägung der Iteminformationsfunktion $I_i(\eta)$ eines bestimmten Items i ist festgelegt durch (vgl. Fischer 1974, S. 295):

$$I_i(\eta) = P(Y_i = 1 | \eta) \cdot P(Y_i = 0 | \eta) \quad (16.15)$$

Iteminformation I_i

Iteminformationsfunktion $I_i(\eta)$

Sie entspricht dem Produkt aus Lösungs- und Nichtlösungswahrscheinlichkeit des Items bei gegebener Personenfähigkeit η . Die Iteminformationsfunktion $I_i(\eta)$ für ein Item mit $\beta_i = 0$ ist in □ Abb. 16.7 (unten) für variierendes η abgetragen. Wie man sieht, erreicht die Iteminformationsfunktion I_i an der Stelle $\eta = \beta_i$ ihr Maximum; die Steigung der IC-Funktion im Punkt $\beta_i = 0$ beträgt .25 und entspricht dem Wert der Iteminformationsfunktion im Punkt $I_i(\eta = 0)$. Mit zunehmender Differenz zwischen η und β_i fällt die Iteminformationsfunktion nach beiden Seiten zunächst langsam, dann schnell und bei sehr großer Differenz wieder langsam asymptotisch gegen null ab.

Bezogen auf das Item mit $\beta_i = 0$ (vgl. □ Abb. 16.7) erhält man für die vier eingetragenen Merkmalsausprägungen $\eta_1 = -1, \eta_2 = 1, \eta_3 = 2, \eta_4 = 4$ folgende numerische Ausprägungen der Iteminformationsfunktion:

$$\begin{aligned} I_i(\eta = -1) &= .27 \cdot .73 = .197 \\ I_i(\eta = 1) &= .73 \cdot .27 = .197 \\ I_i(\eta = 2) &= .88 \cdot .12 = .105 \\ I_i(\eta = 4) &= .98 \cdot .02 = .018 \end{aligned}$$

16.3 · Dichotomes Rasch-Modell (1PL-Modell)

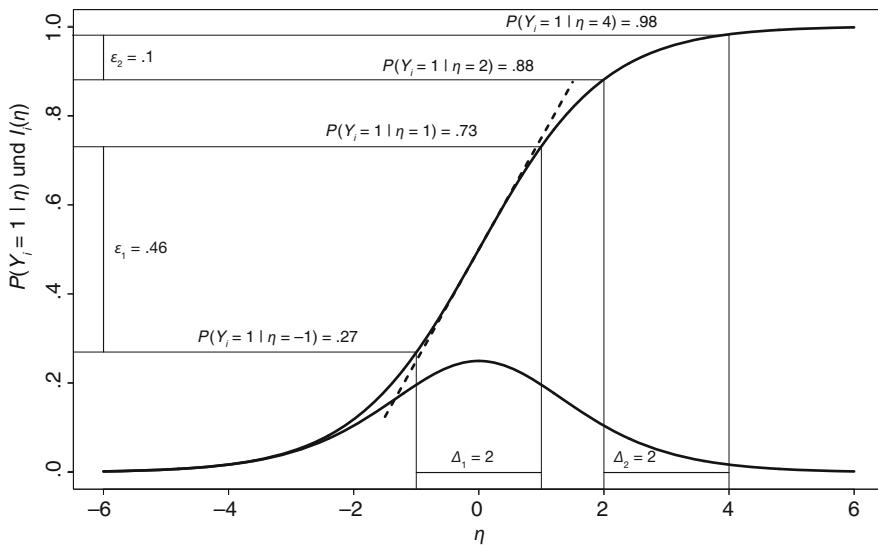


Abb. 16.7 Vergleich der Itemformation für verschiedene Merkmalsausprägungen. Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ und (Item-)Informationsfunktion $I_i(\eta)$ eines Rasch-homogenen Items i mit durchschnittlicher Itemschwierigkeit $\beta_i = 0$ in Abhängigkeit von η . Die Iteminformationsfunktion $I_i(\eta)$ variiert in Abhängigkeit von der Fähigkeit η : Liegen die Personenparameter wie im Fall von $\eta_1 = -1$ und $\eta_2 = 1$ nahe bei $\beta_i = 0$, so führt die Fähigkeitsdifferenz $\Delta_1 = \eta_2 - \eta_1 = 2$ der Personen 1 und 2 zu einem großen Unterschied ε_1 der Lösungswahrscheinlichkeit ($\varepsilon_1 = .73 - .27 = .46$); liegen die Personenparameter wie im Fall von $\eta_3 = 2$ und $\eta_4 = 4$ nicht nahe bei $\beta_i = 0$, so führt die Differenz $\Delta_2 = \eta_4 - \eta_3 = 2$ der Testpersonen 3 und 4 trotz gleicher Größe zu einem geringen Unterschied ε_2 in der Lösungswahrscheinlichkeit ($\varepsilon_2 = .98 - .88 = .10$). Die Steigung der IC-Funktion im Punkt $\beta_i = 0$ beträgt .25 (gestrichelte Linie) und entspricht dem Wert der Iteminformationsfunktion im Punkt $I_i(\eta = 0)$

Will man also Vergleiche zwischen zwei Personen mit der Merkmalsdifferenz $\Delta = \eta_v - \eta_w$ vornehmen, so sind nur dann deutliche Unterschiede der Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ zu erwarten, wenn man Items verwendet, deren Itemschwierigkeiten im Bereich der jeweiligen Personenfähigkeiten liegen. Weichen die Itemschwierigkeiten von den Fähigkeiten hingegen deutlich ab, so fallen die Unterschiede im Lösungsverhalten viel geringer aus, wie Abb. 16.7 zeigt.

16.3.10 Testinformation und Konfidenzintervall für η_v

Für einen aus p Items bestehenden Test und vorliegender Rasch-Modellkonformität (► Abschn. 16.3.11) lässt sich ebenfalls die Testinformation $I(\eta_v)$ berechnen, und zwar für eine bestimmte Testperson v mit dem Personenparameter η_v durch Addition der einzelnen Iteminformationsbeträge $I_i(\eta_v)$ (vgl. Kubinger 2003, S. 4). Die Testinformation variiert interindividuell von Testperson zu Testperson, je nach Ausprägung des Personenparameters η_v

$$I(\eta_v) = \sum_{i=1}^p I_i(\eta_v) \quad (16.16)$$

Mithilfe der Testinformation $I(\eta_v)$ kann die interindividuell variierende Genauigkeit der Personenparameterschätzung η_v als asymptotisches 95 %-Konfidenzintervall (z. B. mit $z_{\alpha/2} = 1.96$) berechnet werden (vgl. Fischer 1983, S. 609):

$$\hat{\eta}_v - \frac{1.96}{\sqrt{I(\hat{\eta}_v)}} \leq \eta_v \leq \hat{\eta}_v + \frac{1.96}{\sqrt{I(\hat{\eta}_v)}} \quad (16.17)$$

Testinformation $I(\eta)$ variiert interindividuell

Individuelles Konfidenzintervall für η_v

Die individuelle Testgenauigkeit wird umso größer, je höher die Testinformation $I(\eta_v)$ für die jeweilige Testperson v ausfällt. Die Testinformation kann durch Vermehrung der Itemanzahl und/oder durch Vergrößerung der einzelnen additiven Iteminformationsbeträge $I_i(\eta_v)$ gesteigert werden. Letztere Überlegung findet beim adaptiven Testen Verwendung (s. ▶ Kap. 20).

16.3.11 Überprüfung der Modellpassung/Modellkonformität

**Modellkonformität vs.
Modellinkonformität**

Wie bereits ausgeführt, muss für die Parameterschätzung eine IC-Funktion festgelegt werden, die in Form einer mathematischen Gleichung angibt, welche Annahmen über den Zusammenhang zwischen den manifesten Antwortvariablen und der latenten Personenfähigkeit im jeweiligen Modell (hier Rasch-Modell) getroffen werden. Die Likelihood-Schätzung selbst sagt nichts darüber aus, ob die getroffenen Modellannahmen auch zutreffen. Um die oben genannten vorteilhaften Modelleigenschaften des Rasch-Modells in Anspruch nehmen zu können, muss die Passung zwischen den Modellannahmen und den empirisch gefundenen Daten dahingehend geprüft/getestet werden, ob *Modellkonformität* besteht. Es könnte beispielsweise vorkommen, dass den bestmöglich geschätzten Parametern ein nicht zutreffendes Modell zugrunde lag, welches die empirische Realität nicht angemessen repräsentiert; in diesem Fall läge eine *Modellinkonformität* vor (also eine mangelnde Modellpassung).

**Itemrevision bzw.
Modellmodifikation**

Wenn für einen Test ein Mangel an Modellkonformität aufgedeckt wird, kann einerseits eine Itemselektion/-revision durchgeführt werden. Es kann andererseits aber auch angezeigt sein, die Modellannahmen des 1PL- zugunsten des 2PL- oder des 3PL-Modells (► Abschn. 16.4 und 16.5) zu lockern/relaxieren bzw. zu ändern; wenn auch das zu keiner Modellkonformität führt, müssen die Modellannahmen vollends verworfen werden.

16.3.11.1 Empirische Modellkontrollen und Itemselektion

**Teilung der Stichprobe nach
relevantem Kriterium**

Die Modellkonformität kann mittels *empirischer Modellkontrollen* überprüft werden. Das einfachste Vorgehen für das Rasch-Modell besteht darin, die postulierte *Stichprobenunabhängigkeit* zu hinterfragen und die Personenstichprobe nach einem relevanten Kriterium (z. B. Alter, Geschlecht, Sozialisation) oder nach dem untersuchten Persönlichkeitsmerkmal selbst (vgl. dazu auch ► Kap. 22) in zwei oder mehrere Substichproben zu unterteilen und in jeder der Substichproben getrennte Itemparameterschätzungen vorzunehmen. Auf diese Weise gewinnt man für jeden Parameter jeweils zwei Werte, die sich bei Modellkonformität nicht bzw. nur zufällig unterscheiden sollten.

Grafischer Modelltest

Einen ersten Überblick über die Modellpassung/-konformität der Items kann man sich mit dem *grafischen Modelltest* verschaffen, bei dem die beiden Itemparameterschätzungen in einem bivariaten Streudiagramm gegeneinander abgetragen werden (s. Lord 1980, S. 37). Je näher die Itemparameter an der Hauptdiagonalen zu liegen kommen, desto größer ist die Stichprobenunabhängigkeit und desto eindeutiger die Rasch-Homogenität. Systematische Abweichungen würden hingegen Hinweise auf modellinkonforme Wechselwirkungen zwischen der Itemschwierigkeit und jenem Kriterium liefern, das für die Teilung der Stichprobe verwendet wurde. Ein gelungenes Beispiel für modellkonforme Daten zeigt □ Abb. 16.8.

Itemselektion

Liegt ein bedeutsamer Unterschied zwischen den Parameterschätzungen vor, so können/sollten die *modellinkonformen Items*, d. h. jene mit auffälligen Unterschieden, entfernt oder auch nachgebessert werden (s. dazu die Überlegungen zur Itemrevision in ► Kap. 7). Weitere Optimierungsmöglichkeiten bestehen durch Itemselektion oder auf Residualmaßen beruhende Item-Fit-Indizes (s. Rost 2004, S. 369 ff.; Strobl 2012; ► Kap. 22).

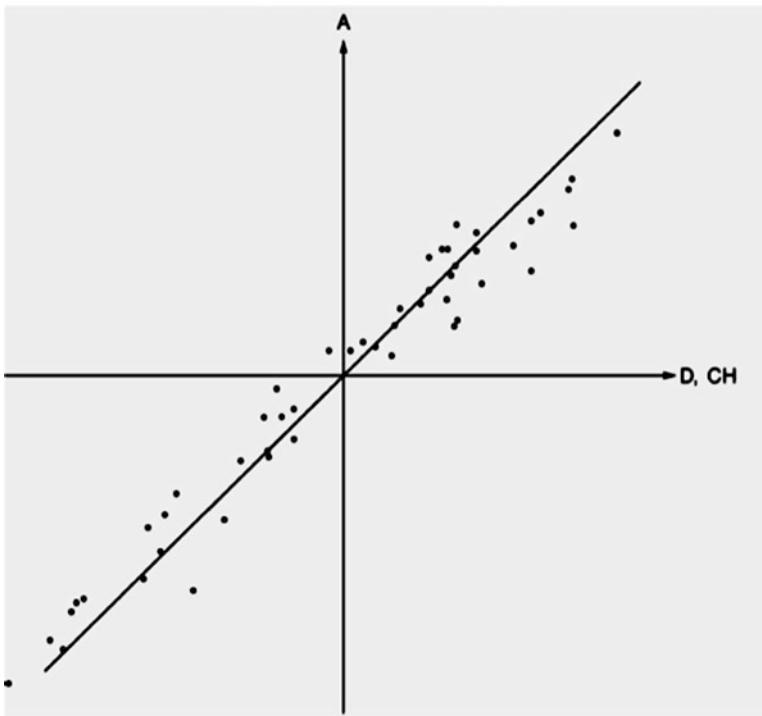


Abb. 16.8 Grafischer Modelltest: Gegenüberstellung der nach dem Rasch-Modell geschätzten Itemparameter der Testskala „Alltagswissen“ aus dem Adaptiven Intelligenz Diagnostikum (AID; Kubinger und Wurst 2000), einerseits für Kinder aus Deutschland und der Schweiz (Abszisse, Stichprobe 1), andererseits für Kinder aus Österreich (Ordinate, Stichprobe 2). (Nach Kubinger 1995, S. 70)

Will man sich nicht mit der grafischen Kontrolle begnügen, sondern die Modellkonformität numerisch beurteilen, so wird häufig der *bedingte Likelihood-Quotienten-Test* (auch Conditional Likelihood-Ratio-Test) von Andersen (1973) eingesetzt, bei dem für beide Teilstichproben CML-Schätzungen durchgeführt und diese mittels Signifikanztest auf Unterschiedlichkeit geprüft werden. Kann die Nullhypothese beibehalten werden, so spricht das für Modellkonformität; das Verwerfen der Nullhypothese spräche gegen die Modellkonformität. Sofern in diesem Fall nur bei einzelnen Items Differenzen auftreten, kann nach Aussonderung oder Überarbeitung des/der betroffenen Items erneut überprüft werden, ob nunmehr Modellkonformität vorliegt. Dazu sollten möglichst neue Daten herangezogen werden.

Likelihood-Quotienten-Test

Eine weitere Möglichkeit zur Testung der Modellkonformität bietet der *Wald-Test* (vgl. Strobl 2012). Mit diesem ist es möglich, auf Itemebene die Gleichheit von Itemparametern über zwei Subpopulationen (Stichproben) hinweg zu testen. Dazu wird die aus den zwei Subpopulationen erhaltene Differenz zweier Itemparameterschätzungen an der Summe der quadrierten Standardfehler der Itemparameterschätzungen relativiert. Die resultierende Prüfstatistik ist standardnormalverteilt, sodass ein Signifikanztest vorgenommen werden kann.

Wald-Test

Ist der erhaltene z -Wert signifikant, so ist die Abweichung der Itemparameterschätzungen nicht vernachlässigbar und die Items funktionieren über die Subpopulationen hinweg unterschiedlich. Diese spezifischen Items weisen ein sog. „Differential Item Functioning“ (DIF), also eine differenzielle Itemfunktionsweise, auf (Holland und Wainer 1993). Diese äußert sich darin, dass manche Items hinsichtlich ihrer Anforderungen an das interessierende Merkmal über Subpopulationen oder Populationen hinweg variieren. Diejenige Variable, die für die Variation verantwortlich ist, kann man als sog. „Moderatorvariable“ bei der Modellierung berücksichtigen. Wenn z. B. manche Leistungstestaufgaben regional aufgrund einer Sprachgebundenheit hinsichtlich ihrer Schwierigkeit variieren, dann kann man

Differential Item Functioning (DIF)

Globale Modellkonformität: χ^2 -Test nach Pearson und Bootstrap

Weitere Testmöglichkeiten

Unangemessene Bearbeitungsstile

Auffällige Antwortmuster

Person-Fit-Indizes

dies in Form eines erweiterten Modells berücksichtigen, indem man die Region als kategoriale Moderatorvariable in das Modell aufnimmt.

Um eine globale Modellkonformität zu beurteilen, bietet es sich an, die Verteilung der Antwortmuster in Betracht zu ziehen. Weichen die beobachteten Häufigkeiten der Antwortmuster (Zeilen der Datenmatrix) von den erwarteten Häufigkeiten der Antwortmuster bedeutsam ab, so ist von einer schlechten Modellpassung auszugehen. Dies kann mit dem χ^2 -Test nach Pearson getestet werden. Dazu werden über alle Antwortmuster hinweg die quadrierten Abweichungen der beobachteten und erwarteten Antwortmusterhäufigkeiten an den erwarteten Antwortmusterhäufigkeiten relativiert und summiert. Ist die resultierende Prüfstatistik signifikant, so muss die globale Passung verworfen werden. Da die Prüfgröße nur bei großen Stichproben χ^2 -verteilt ist, bietet es sich bei kleinen Stichproben an, anhand eines Bootstrap-Verfahrens eine für den jeweiligen Fall spezifische Prüfverteilung zu generieren (vgl. Rost 2004), sodass die Prüfstatistik an dieser spezifischen Prüfverteilung hinsichtlich ihrer Signifikanz beurteilt werden kann.

In einer weiteren Abwandlung des *Likelihood-Quotienten-Tests* kann man die Likelihood mit den beobachteten Antwortmusterwahrscheinlichkeiten in Beziehung setzen. Diese Prüfstatistik ist ebenfalls asymptotisch χ^2 -verteilt, sodass bei kleinen Stichproben mittels Bootstrap eine Prüfverteilung zur Modellkonformitätsbeurteilung generiert werden kann. Kommt es auch hier zu bedeutsamen Abweichungen, so ist das Gesamtmodell global zurückzuweisen. Eine weitere Variante des *Wald-Tests* (vgl. Strobl 2012) kann auch genutzt werden, um nicht nur einzelne Itemparameterschätzungen zu vergleichen, sondern um die globale Modellkonformität (Modellpassung) zu beurteilen. Dieser ist asymptotisch äquivalent zum Likelihood-Quotienten-Test.

16.3.11.2 Personenselektion

Mängel eines Tests hinsichtlich der Modellkonformität können aber nicht nur auf die Items oder die Modelleigenschaften zurückzuführen sein, sondern auch darauf, dass einzelne Personen auf die Testitems nicht in angemessener Weise reagieren, sondern vielmehr unangemessene Bearbeitungsstile zeigen: Akquieszenz, Schwundeln, Raten, soziale Desirabilität und arbiträres Verhalten wären hier als mögliche Gründe ebenso aufzuführen wie Probleme beim Instruktionsverständnis oder allgemeine Sprachschwierigkeiten (s. dazu auch ► Kap. 4). Personen mit unangemessenen Bearbeitungsstilen, welche möglichst auch transsituativ in anderen (Sub-)Tests abgesichert sein sollten, müssen gegebenenfalls ausgesondert werden, um die Personenstichprobe hinsichtlich ihres Bearbeitungsstiles zu homogenisieren.

Die Personenselektion nutzt die Tatsache, dass sich inadäquate Bearbeitungsstile in der Regel in auffälligen Antwortmustern („aberrant response patterns“) manifestieren, denen unter Modellgültigkeit nur eine sehr geringe Auftretenswahrscheinlichkeit zukommt. Ein deutlich abweichendes Antwortmuster läge beispielsweise vor, wenn eine Person grundsätzlich alle Items bejaht oder ein alternierendes „Ja-Nein-Ja-Nein-...“-Muster erzeugt. Die Antworten dieser Personen würden dann zu fehlerhaften Schlussfolgerungen hinsichtlich der latenten Fähigkeit führen.

Bei der Testanwendung sollte im diagnostischen Einzelfall stets geprüft werden, ob sich die einzelne Person „modellkonform“ verhält oder nicht. Dazu wurden sog. „Person-Fit-Indizes“ (auch „Caution-Indizes“) entwickelt, die auf der Basis der Antwortmuster eine Beurteilung erlauben, ob es sich um plausible oder um unplausible Testergebnisse handelt. Während etliche Verfahren aus verschiedenen Gründen nur eingeschränkt empfohlen werden können (s. Fischer 1996, S. 692), erweisen sich die auf der Likelihood-Funktion basierenden Ansätze (beispielsweise von Klauer 1991; Molenaar und Hoijtink 1990; Tarnai und Rost 1990) als wissenschaftlich gut fundiert. Für eine Diskussion von auf der Likelihood-Funktion basierenden Ansätzen sei auf Snijders (2001) verwiesen. Fällt ein Person-Fit-Index zu ungünstig aus, so ist bei dem jeweiligen Testergebnis Vorsicht angezeigt; die Testinterpretation sollte dann entweder unterlassen oder nur mit entsprechender

16.4 · 2PL-Modell nach Birnbaum

Umsicht vorgenommen werden (für weitere Informationen zu Person-Fit-Indizes s. Klauer 1995, für Optimierungsmöglichkeiten durch Personenselektion s. auch Rost 2004, S. 363 ff.).

Im Falle von unterschiedlich funktionierenden Itemparametern (DIF) kann es auch lohnend sein, das Rasch-Modell mit der Latent-Class-Analyse (LCA, vgl. ▶ Kap. 22) zu kombinieren und – neben dem interessierenden latenten Merkmal – eine Heterogenität anhand latenter Klassen (Subpopulationen) zu suchen, die für die unterschiedlichen Itemparameterschätzungen verantwortlich sind. Dazu berechnet man eine steigende Anzahl von Klassenlösungen (d. h. erst eine, dann zwei, dann drei etc.) und erhält für jede latente Klasse Itemparameterschätzungen (über alle Items oder Teilmengen der Items hinweg). Basierend auf den Likelihood-Lösungen lassen sich unter Einbezug sog. „Informationsmaße“ wie dem *Akaike-Informationskriterium* (engl. „Akaike Information Criterion“, AIC) oder dem *Bayes-Informationskriterium* (engl. „Bayesian Information Criterion“, BIC) und ihren Weiterentwicklungen Entscheidungen zugunsten einer Anzahl von Klassen treffen. Wenn die Zahl der Klassen bestimmt wurde, liegt der nächste Schritt darin, mögliche Erklärungsansätze für die Unterschiede in Bezug auf die Funktionsweise der Items zwischen den Klassen zu finden (vgl. dazu ▶ Kap. 22) oder nach sorgfältigen Erwägungen ggf. Items (oder auch Personen) zu selektieren.

Anstelle einer vorschnellen Item- bzw. Personenselektion sollte aber auch überlegt werden, ob das modellinkonforme Verhalten eine relevante Information im Sinne der differenziellen Psychologie darstellt. So können gerade niedrige Person-Fit-Indizes auch ein Hinweis dafür sein, dass man es mit Personen zu tun hat, deren Arbeitsstil anders ist als derjenige der Mehrheit. Diese Überlegung findet beispielsweise in der Sportpsychologie Anwendung zur Identifikation von Personen, die über die Gabe verfügen, ihre Leistung unter Belastung zu steigern (s. z. B. Guttmann und Ettlinger 1991).

Kombination von Rasch-Modell und LCA

Keine vorschnelle Personenselektion

16.4 2PL-Modell nach Birnbaum

16.4.1 Charakteristika

Die bisherigen Darstellungen beschränkten sich auf ein sehr prominentes Modell der IRT, das Rasch-Modell, auch 1PL-Modell genannt, da es ein logistisches Modell ist und es nur einen Itemparameter, den Itemschwierigkeitsparameter, enthält.

Wenn für die strengen Annahmen des 1PL-Modells keine Modellkonformität erzielbar ist, kann auf Modelle mit relaxierten Annahmen zurückgegriffen werden. Ein solches Modell ist beispielsweise das *dichotome Birnbaum-Modell* (zwei-parametrisches logistisches Modell bzw. 2PL-Modell nach Birnbaum). Es verwendet ebenfalls die logistische IC-Funktion. Im Unterschied zum Rasch-Modell, das für alle Items dieselbe Steigung der IC-Funktion annimmt, erlaubt das 2PL-Modell unterschiedliche Steigungen, indem es neben der Personenfähigkeit η zwei Itemparameter, und zwar den Itemschwierigkeitsparameter β_i und zusätzlich den sog. „Diskriminationsparameter“ λ_i enthält. Auf diese Weise trägt das 2PL-Modell dem Umstand Rechnung, dass verschiedene Items unterschiedlich gut zwischen Personen mit schwächeren bzw. stärkeren Merkmalsausprägungen trennen können.

Zusätzlicher Itemdiskriminationsparameter

16.4.2 Modellgleichung

Modellgleichung des 2PL-Modells

Formal lässt sich das 2PL-Modell unter Verwendung der logistischen Funktion durch folgende Gleichung beschreiben:

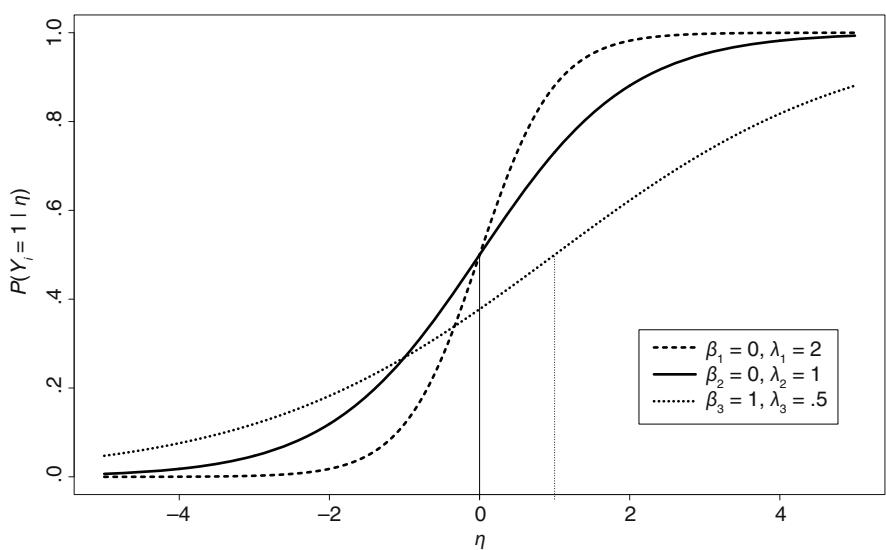
$$P(Y_i = 1 | \eta) = \frac{e^{\lambda_i(\eta - \beta_i)}}{1 + e^{\lambda_i(\eta - \beta_i)}} \quad (16.18)$$

Wie auch im Rasch-Modell ist η die Personenfähigkeit. Der Itemschwierigkeitsparameter β_i gibt an, wie weit links (leichte Items) bzw. wie weit rechts (schwierige Items) die IC-Funktion des Items i auf der gemeinsamen Skala von η und β_i zu liegen kommt. Vom Diskriminationsparameter λ_i hängt für jedes Item die Steilheit der IC-Funktion ab, die wie im Rasch-Modell im Wendepunkt ihr Maximum erreicht.

Das 2PL-Modell lässt also für die verschiedenen Items logistische IC-Funktionen mit verschiedenen Steigungen (charakterisiert durch die jeweiligen Diskriminationsparameter λ_i) zu. Im Unterschied dazu waren im dichotomen Rasch-Modell (► Abschn. 16.3) alle Diskriminationsparameter λ_i auf den konstanten Wert 1 gesetzt, sodass das Rasch-Modell als Spezialfall des 2PL-Modells angesehen werden kann.

Die □ Abb. 16.9 veranschaulicht die logistischen IC-Funktionen von drei Items (1, 2, 3) mit unterschiedlichen Diskriminationsparametern in absteigender Größe $\lambda_1 = 2 > \lambda_2 = 1 > \lambda_3 = .5$. Die Diskriminationsparameter geben an, wie stark sich die Lösungswahrscheinlichkeiten $P(Y_i = 1 | \eta)$ in Abhängigkeit von der Personenfähigkeit η verändern. Sie stellen ein Maß der Sensitivität der Items für Merkmalsunterschiede dar und entsprechen in gewisser Weise den Trennschärfen der Itemanalyse (► Kap. 7). Die Itemdiskriminationsparameter λ_i charakterisieren die Steigungen der IC-Funktionen an ihrem jeweiligen Wendepunkt. Wie im Rasch-Modell ist dies gleichzeitig der Punkt im Personenfähigkeitskontinuum, an dem das Item am stärksten zwischen Personen mit niedrigeren bzw. höheren Merkmalsausprägungen diskriminiert. Je kleiner λ_i ausfällt, desto flacher wird die IC-Funktion und desto weniger gut kann das Item Personen mit höherer von Personen mit niedriger Merkmalsausprägung trennen (s. □ Abb. 16.9).

Sensitivität der Items für Merkmalsunterschiede



□ Abb. 16.9 IC-Funktionen mit unterschiedlichen Steigungen für drei Items des 2PL-Modells. Die Schwierigkeitsparameter sind $\beta_1 = \beta_2 = 0$ und $\beta_3 = 1$. Für die Diskriminationsparameter gilt $\lambda_1 = 2 > \lambda_2 = 1 > \lambda_3 = .5$. Das zweite Item entspricht wegen $\lambda_2 = 1$ der IC-Funktion des Rasch-Modells

16.5 · 3PL-Modell nach Birnbaum

Wie man in der Abbildung sieht, ist Item 3 jenes, das am flachsten verläuft und den niedrigsten Itemdiskriminationsparameter $\lambda_3 = .5$ aufweist. Andererseits ist – vor allem im Vergleich zu Item 1, aber auch zu Item 2 – ein Zugewinn an Iteminformation im oberen und unteren Bereich des Merkmals η zu verzeichnen. Die Schwierigkeitsparameter sind bei Item 1 und 2 gleich ($\beta_1 = \beta_2 = 0$) und bei Item 3 größer ($\beta_3 = 1$).

16.4.3 Parameterschätzung

Wegen der unterschiedlichen Steigungen der IC-Funktionen muss im 2PL-Modell auf spezifisch objektiv Vergleiche als vorteilhafte Eigenschaft des Rasch-Modells (vgl. ► Abschn. 16.3.6) explizit verzichtet werden. Im Unterschied zum Rasch-Modell ist in der Folge keine bedingte Likelihood-Schätzung (CML-Schätzung) in der bisherigen Form möglich und der Summenscore ist im 2PL-Modell keine erschöpfende (suffiziente) Statistik für den Personenparameter η_v . Letzteres heißt, dass aus dem Summenscore nicht unmittelbar auf die Merkmalsausprägung η_v geschlossen werden kann. Es kommt bei gleicher Anzahl gelöster Aufgaben, aber unterschiedlichen Antwortmustern (Zeilen in der Datenmatrix) zu unterschiedlichen Personenparameterschätzungen (was im Rasch-Modell nicht der Fall war). Im Unterschied zum Rasch-Modell ist es also entscheidend, welche Aufgaben bearbeitet und gelöst wurden (dieser Umstand hat Implikationen für das computerisierte adaptive Testen, ► Kap. 20).

Da das CML-Schätzverfahren auf der einen Seite die Inkonsistenzproblematik adressiert, auf der anderen Seite aber auf Item-Response-Modelle aus der Rasch-Familie beschränkt ist, kann für mehrparametrische Item-Response-Modelle wie dem 2PL- oder dem 3PL-Modell (► Abschn. 16.5) als Alternative das *MML-Schätzverfahren* (► Abschn. 16.3.8.3) eingesetzt werden. Bei diesem Verfahren werden die Personenparameter als zufällige Realisierungen einer latenten Variable η aufgefasst, die einer zuvor festgelegten Verteilung(-sform), insbesondere der Normalverteilung, folgen. Durch die Festlegung der Verteilung der latenten Variablen η reduziert sich die Zahl der zu schätzenden Parameter erheblich, da nur noch die Verteilungsparameter (z. B. Mittelwert und Varianz im Fall der Normalverteilung) im zu schätzenden Modell enthalten sind, sodass nicht mehr für jede Testperson ein eigener Personenparameter geschätzt wird. Durch gezielte Annahmen (z. B. eine standardnormalverteilte latente Variable η mit einem Mittelwert 0 und einer Varianz von 1) lässt sich die Zahl der Parameter noch weiter reduzieren.

Prozedural werden die unbedingten Antwortmusterwahrscheinlichkeiten (also jene, die nicht von den Personenparametern abhängen, sondern bei denen die Personenparameter „herausintegriert“ wurden) maximiert (vgl. hierzu auch ► Kap. 22). Erneut werden nur die Itemparameter geschätzt. Bei der Maximierung selbst werden sog. „Quadraturverfahren“ für die numerische Integration innerhalb der Likelihoodfunktion benötigt (für technische Details vgl. ► Kap. 19). Aufwendig ist zudem die Bestimmung der Standardfehlerschätzungen für die Itemparameter (vgl. auch Bock und Aitkin 1981).

16.5 3PL-Modell nach Birnbaum

Das „Dichotome Rate-Modell von Birnbaum“ (dreiparametrisches logistisches Modell, 3PL-Modell) verwendet zur Beschreibung des Antwortverhaltens der Testpersonen ein zusätzlich zum Itemschwierigkeitsparameter β_i und dem Diskriminationsparameter λ_i einen dritten Itemparameter, der als *Rateparameter* γ_i (Gamma) bezeichnet wird. Dieser Rateparameter γ_i wird dem Umstand gerecht, dass z. B. bei vierkategorialen Multiple-Choice-Items (s. dazu ► Kap. 5) in einem

Keine spezifisch objektiven Vergleiche möglich

MML-Schätzverfahren

Maximierung der unbedingten Antwortmusterwahrscheinlichkeiten

Zusätzlicher Rateparameter im 3PL-Modell

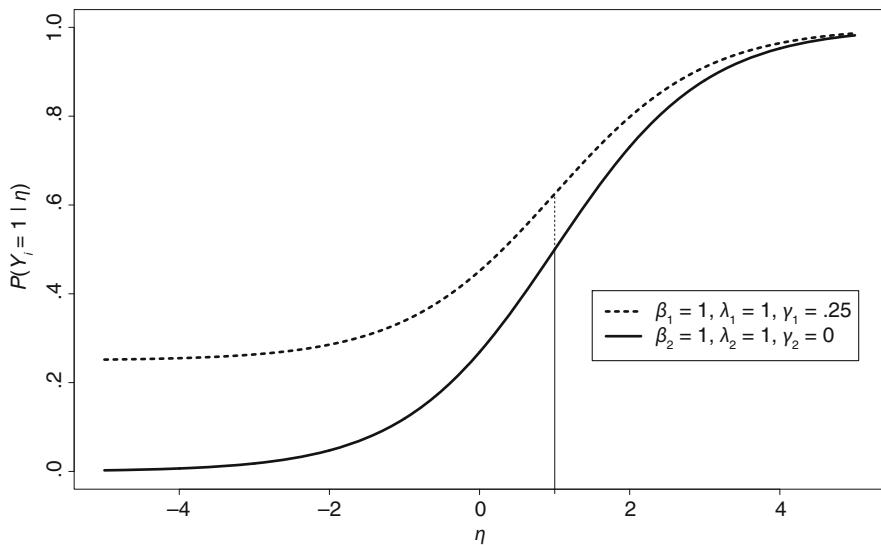


Abb. 16.10 Zwei IC-Funktionen des 3PL-Modells mit den Ratewahrscheinlichkeiten $\gamma_1 = .25$ und $\gamma_2 = 0$. Die IC-Funktion des ersten Items entspricht einem Birnbaum-Modell mit Rateparameter. Das zweite Item folgt dem 2PL-Modell und wegen $\lambda_2 = 1$ auch dem 1PL-Modell (Rasch-Modell) ohne Rateparameter. Erkennbar ist, dass die IC-Funktion beim ersten Item eine Ratewahrscheinlichkeit von .25 berücksichtigt; auch Personen mit extrem niedriger Fähigkeit haben noch eine Lösungswahrscheinlichkeit von .25, wenn sie raten

Leistungstest mit einer 25 %igen Ratewahrscheinlichkeit zu rechnen ist, da auch ohne merkmalspezifische Fähigkeit die Richtigantwort mit einer Wahrscheinlichkeit von .25 erraten werden kann. Die Parametrisierung der Möglichkeit, dass eine richtige Lösung durch Raten zustande kommt, ist Kernmerkmal der Erweiterung der IC-Funktion des 2PL-Modell zur IC-Funktion des 3PL-Modells.

Modellgleichung

Die IC-Funktion des 3PL-Modells lässt sich formal durch folgende Gleichung beschreiben:

$$P(Y_i = 1 | \eta) = \gamma_i + (1 - \gamma_i) \frac{e^{\lambda_i(\eta - \beta_i)}}{1 + e^{\lambda_i(\eta - \beta_i)}} \quad (16.19)$$

Abb. 16.10 zeigt zwei IC-Funktionen des 3PL-Modells. Man sieht für $\gamma_1 = .25$, dass die Lösungswahrscheinlichkeit $P(Y_i = 1 | \eta)$ bei sehr geringer Merkmalsausprägung nicht gegen 0, sondern gegen .25 geht. Für $\gamma_2 = 0$ und $\lambda_2 \neq 1$ reduziert sich das 3PL-Modell auf das 2PL-Modell und bei $\lambda_2 = 1$ auf das Rasch-Modell. Sind $\gamma_1 = .25$ und $\lambda_1 = 1$, so erhält man als Spezialfall des 3PL-Modells ein dichotomes Rasch-Modell mit einem Rateparameter.

Obwohl mit dem 3PL-Modell und dem 2PL-Modell eine „genauere“ Modellierung des Personenverhaltens als mit dem 1PL-Modell möglich ist, bleibt aus testtheoretischer Sicht anzumerken, dass nur das 1PL-Modell (Rasch-Modell) die oben beschriebenen besonderen Vorteile (erschöpfende Statistiken, Spezifische Objektivität der Vergleiche, Schätzverfahren) auf sich vereinen kann. Es ist zudem auch sparsamer, da weniger Parameter benötigt werden. Somit weist das 1PL-Modell letztlich die vorteilhafteren Modelleigenschaften auf.

Vergleich der Modelleigenschaften

16.6 Weitere IRT-Modelle

Neben den genannten dichotomen 1PL-, 2PL- und 3PL-Modellen umfasst das Gebiet der IRT heute eine Vielzahl weiterer Modelle, denen ein eigenes Kapitel gewidmet ist (► Kap. 18). Sie sind ebenfalls probabilistisch, unterscheiden sich aber

u. a. durch die Art der manifesten und/oder latenten Variablen und die Art der verwendeten Modellparameter. Die in der IRT zentrale Annahme der lokalen stochastischen Unabhängigkeit gilt sinngemäß auch hier. Die meisten der im Folgenden skizzierten Modelle lassen sich als Weiterentwicklungen und Verallgemeinerungen der genannten Modelltypen interpretieren, andere haben ihre eigene Geschichte. Die folgende Darstellung zeigt beispielhaft grundlegende Ansätze auf und erhebt keinen Anspruch auf Vollständigkeit.

16.6.1 Polytome Latent-Trait-Modelle

Rasch (1961) hat sein dichotomes Modell auf den Fall polytomer Items (d. h. Items mit mehrkategoriellem Antwortmodus) erweitert. Da es sich um Items mit nominalen Kategorien handeln kann (z. B. Signierungen bei Fragen mit freier Beantwortung), ist das polytome Rasch-Modell im allgemeinsten Fall mehrdimensional: Abgesehen von einer Referenzkategorie wird für jede Kategorie ein eigener Personen- und ein eigener Itemparameter eingeführt. Auch hier sind wieder spezifisch objektive Vergleiche möglich und es existieren Verfahren zur Parameterschätzung und Modellkontrolle (z. B. Fischer 1974, 1983; Fischer und Molenaar 1995). Ein Anwendungsproblem besteht allerdings darin, dass bei vielen Personen bestimmte Kategorien gar nicht vorkommen (Rost 2004).

Von großer praktischer Bedeutung ist der eindimensionale Spezialfall des polytomen Rasch-Modells (vgl. hierzu ▶ Kap. 18), in dem sich die Antwortkategorien im Sinne einer Rangskala ordnen lassen. Das zugehörige Modell enthält nur einen Personen- und einen Itemparameter; wie im dichotomen Fall sind die Parameter z. B. als Fähigkeit (allgemeiner: Merkmalsausprägung) bzw. als Schwierigkeit interpretierbar. Zusätzlich gibt es für jede Kategorie eine Gewichtszahl und einen Parameter, der als Aufforderungscharakter der jeweiligen Kategorie bezeichnet werden kann (Fischer 1974, 1983). Sofern sich im Einklang mit der Rangordnung der Kategorien „gleichabständige“ Gewichtszahlen ergeben, sind auch hier spezifisch objektive Vergleiche möglich (Andersen 1995).

Anmerkung: Gleichabständige Gewichtungen der Form 0, 1, 2, … o. Ä. für Stufenantwortaufgaben und Ratingskalen finden auch ohne Bezug auf die IRT Verwendung, jedoch fehlt dort häufig ihre Legitimation mangels Einbettung in ein empirisch prüftbares Modell (vgl. ▶ Kap. 5; vgl. dazu aber auch die Überlegungen zur Verwendung von Stufenantwortaufgaben im Rahmen der KTT, ▶ Kap. 13).

Andrich (1978) gelang es, das eindimensionale polytome Rasch-Modell auf der Basis dichotomer Latent-Trait-Modelle zu interpretieren (für eine ausführlichere Behandlung s. ▶ Kap. 18). In Andrichs Darstellung, dem Rating-Scale-Modell (RSM), werden die manifesten Kategoriengrenzen durch sog. „Schwellen“ auf der latenten Variablen repräsentiert, die sich ähnlich wie dichotome Items durch Diskriminations- und Schwierigkeitsparameter beschreiben lassen. Dabei zeigte sich, dass die oben hervorgehobene gleichabständige Gewichtung nur dann resultiert, wenn man gleich diskriminierende Schwellen annimmt. Folglich werden im RSM alle Diskriminationsparameter gleich eins gesetzt und die Kategorien mit fortlaufenden ganzen Zahlen (0, 1, 2, …) gewichtet (s. auch Rost 2004).

■ Abb. 16.11 kann als Illustration des RSM für den Fall von vier Antwortkategorien dienen. Der Aufforderungscharakter der einzelnen Kategorien hängt von den relativen Positionen der Schwellen auf dem latenten Kontinuum ab, die sich aus den Schnittpunkten der Kurven benachbarter Kategorien ergeben.

Anmerkung: Das dichotome Rasch-Modell ist als Spezialfall im RSM enthalten: Allgemein ist die IC-Funktion bei dichotomen Latent-Trait-Modellen nichts ande-

Mehrdimensionales polytomes Rasch-Modell

Eindimensionales polytomes Rasch-Modell

Schwellen im RSM

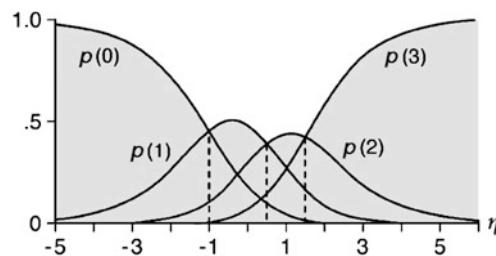


Abb. 16.11 Kategoriencharakteristiken eines vierkategorialen Items mit Kategorie 0 (z. B. „lehne ich ab“) bis Kategorie 3 („stimme völlig zu“). (Nach Rost 2004, S. 216)

res als die Kategoriencharakteristik der positiven oder symptomatischen Kategorie (hier Kategorie 3).

Kategoriencharakteristiken

Wahrscheinlichste Antwort

Eindimensionale polytome Modelle lassen sich durch *Kategoriencharakteristiken* veranschaulichen, die die Wahrscheinlichkeiten für alle möglichen Antworten als Funktion des Personenparameters zeigen.

Wird im RSM der Personenparameter variiert, ergibt sich die jeweils wahrscheinlichste Antwort unter Berücksichtigung der Kategorienfunktionen der gesamten Ratingskala. Demnach wäre für Personen mit einem Summenscore < -1 die Kategorie 0 die wahrscheinlichste Antwort; bei $-1 \leq \text{Summenscore} < .5$ die Kategorie 1, bei $.5 \leq \text{Summenscore} < 1.5$ die Kategorie 2 und bei $\text{Summenscore} \geq 1.5$ die Kategorie 3 (Näheres s. Rost 2006).

Die in **Abb. 16.11** gezeigte Kurvenschar wäre bei einem leichteren Item lediglich insgesamt nach links, bei einem schwereren Item nach rechts verschoben. Die wesentlichen Merkmale von Rasch-Modellen (z. B. Summenwerte als erschöpfende Statistiken für die Modellparameter, Existenz konsistenter Schätzverfahren) bleiben jedoch erhalten, wenn auch sog. „Interaktionseffekte“ zugelassen werden, und zwar dergestalt, dass die relativen Positionen der Schwellen und sogar die Anzahl der Kategorien von Item zu Item schwanken können. Masters (1982) konzipierte dieses sehr allgemeine Modell zunächst für Leistungstests mit abgestufter Bewertung der Antworten (z. B. bei Teil- bzw. Gesamtlösung) und nannte es dementsprechend *Partial-Credit-Modell* (PCM, für eine ausführliche Behandlung ► Kap. 18). Es eignet sich aber auch als Bezugsrahmen für eine Reihe spezieller Rasch-Modelle mit geordneten Kategorien (Masters und Wright 1984; Rost 1988; Wright und Masters 1982), sodass die neutrale Bezeichnung „ordinales Rasch-Modell“ (Rost 2004) angemessener erscheint.

Eine Verallgemeinerung auf ein Rasch-Modell für kontinuierliche Ratingskalen entwickelte Müller (1987). Für nähere Einzelheiten der vorgeschlagenen Spezialfälle und mögliche Anwendungen sei insbesondere auf Müller (1999) und Rost (2004) verwiesen.

Partial-Credit-Modell (PCM)

Rasch-Modell für kontinuierliche Ratingskalen

16.6.2 Mixed-Rasch-Modelle

Wie bereits in ► Abschn. 16.3 dargestellt, setzen herkömmliche Rasch-Modelle Stichprobenunabhängigkeit (z. B. van den Wollenberg 1988) bzw. Rasch-Homogenität (► Abschn. 16.3.2) in dem Sinne voraus, dass die Items bei allen getesteten Personen dasselbe Merkmal erfassen sollen, was sich darin äußert, dass in verschiedenen Teilstichproben gleiche Itemparameter vorliegen. Ob Stichprobenabhängigkeit gegeben ist, beurteilt man mit Modellgeltungstests wie dem Likelihood-Quotienten-Test von Andersen (1973), der die Gleichheit der Itemparameter des dichotomen Rasch-Modells in manifesten Teilstichproben der Personen überprüft (► Abschn. 16.3.11). Solche Modellkontrollen sind im Allgemeinen gut interpretierbar, beinhalten aber die Gefahr, dass relevante Teilungskriterien übersehen

werden; seit Kurzem kann man sich allerdings mit einem neuen Testverfahren von Strobl et al. (2010) gegen dieses Problem absichern.

Sofern es sich im Fall von Stichprobenabhängigkeit und insbesondere von DIF (s. hierzu auch ► Abschn. 16.3.11.1) als nicht haltbar erweist, für die gesamte Personenstichprobe dieselben Itemparameterwerte anzunehmen, können für verschiedene Teilstichproben unterschiedliche Itemparameter zugelassen werden. Diese Möglichkeit bieten sog. „Mixed-Rasch-Modelle“ (Rost 1990, 2004), die auch als *Mischverteilungsmodelle* bezeichnet werden. Sie beruhen sowohl auf der LCA (► Kap. 22) als auch auf der IRT und lassen dementsprechend die Möglichkeit zu, dass nur innerhalb von verschiedenen, zunächst nicht bekannten latenten Klassen, die mithilfe der LCA identifiziert werden, Rasch-Homogenität gegeben ist.

Spricht in einer empirischen Anwendung Vieles für das Vorliegen mehrerer latenter Klassen, kann dies z. B. auf unterschiedliche Lösungsstrategien oder Antwortstile der Personen hindeuten und eine Modifikation der inhaltlichen Modellvorstellungen nahelegen, z. B. in der Weise, dass Personen mit zuvor mäßigem „Personen-Fit“ nunmehr als eigenständige Klasse mit homogenem Antwortverhalten identifiziert werden können (s. Köller 1993, vgl. ► Kap. 22). So gesehen lassen sich Mixed-Rasch-Modelle auch als Modelltests zur Überprüfung herkömmlicher Rasch-Modelle nutzen.

Aus der Sicht der LCA ist an Mischverteilungsmodelle zu denken, wenn in einer Typologie bestimmte Typen als polar (s. z. B. Amelang et al. 2006) konzipiert sind. Als konkretes Anwendungsbeispiel sei der Vergleich zweier Geschlechtsrollytypologien durch Strauß et al. (1996) genannt, bei dem ordinale, Latent-Class- und Mixed-Rasch-Modelle zum Einsatz kamen, also fast alle bisher skizzierten Arten komplexerer IRT-Modelle.

Mischverteilungsmodelle mit klassenspezifischen Itemcharakteristiken

16.6.3 Linear-logistische Modelle

Die Grundidee linear-logistischer Modelle besteht darin, die Itemparameter in IRT-Modellen näher zu erklären, indem sie als Linearkombination, d. h. als gewichtete Summe einer geringeren Anzahl von Basisparametern, aufgefasst werden.

Zerlegung der Itemparameter in eine Linearkombination von Basisparametern

In psychologisch-inhaltlicher Hinsicht ermöglichen linear-logistische Modelle Erweiterungen gewöhnlicher IRT-Modelle, weil sich die linear kombinierten Basisparameter z. B. auf die einzelnen Schritte kognitiver Operationen beziehen können, die hypothetisch zur Bearbeitung der Testitems erforderlich sind. Mit welchem Gewicht die jeweilige Operation für die Lösung eines Items erforderlich ist (z. B. einmal, zweimal, oder auch gar nicht), muss inhaltlich begründet vorab festgelegt werden. Ein in dieser Weise spezifiziertes linear-logistisches Modell kann wegen der geringeren Parameteranzahl nur gültig sein, wenn als notwendige (aber nicht hinreichende) Bedingung auch für das zugehörige logistische IRT-Modell ohne die lineare Zerlegung Modellkonformität besteht. In formaler Hinsicht sind linear-logistische Modelle also Spezialfälle von IRT-Modellen. Sie zwingen zu einer gründlichen Analyse der Anforderungsstruktur von Testaufgaben, was sowohl für Konstruktvalidierungen als auch für das Assessment von Kompetenzen von großer Bedeutung ist (Hartig 2007). Sie eröffnen damit die Möglichkeit, die Aufgabenschwierigkeiten (Itemparameter) zum Gegenstand der Hypothesenbildung und Modellierung zu machen.

Scheiblechner (1972) und Fischer (1995b) haben das dichotome Rasch-Modell zum linear-logistischen Testmodell (LLTM) erweitert, indem sie die Schwierigkeitsparameter als Linearkombination von Basisparametern darstellen. Als Anwendungsbeispiel für das LLTM sei ein Test zur Messung des räumlichen Vorstellungsvermögens von Gittler (1990) angeführt, der das Prinzip der aus dem Intelligenz-Struktur-Test 70 (IST 70) bekannten Würfelaufgaben (Amthauer 1970) aufgreift und diese verbessert. Als relevante Merkmale der Anforderungsstruktur

Linear-logistisches Testmodell (LLTM)

erwiesen sich hier u. a. die Anzahl der (mental) Dreh- oder Kippbewegungen, die Symmetrieeigenschaften der Muster auf den Würfelflächen und die Position des Lösungswürfels im Multiple-Choice-Antwortformat. Zusätzlich spielt der Lernzuwachs während des Tests eine Rolle, was vor allem beim adaptiven Testen zu beachten ist (Fischer 1983; Gittler und Wild 1988). Insbesondere durch die Large-Scale-Assessments im Rahmen von TIMSS (Trends in International Mathematics and Science Study; Klieme et al. 2000) und PISA (Carstensen et al. 2007; Hartig et al. 2008) erlangten solche IRT-Modelle eine größere öffentliche Beachtung.

Die Zerlegung der Itemparameter in eine Linearkombination von Basisparametern ist auch bei erweiterten Rasch-Modellen sowie bei Latent-Class-Modellen möglich. Das lineare RSM (Fischer und Parzer 1991) und das lineare PCM (Glas und Verhelst 1989; Fischer und Ponocny 1995) basieren auf entsprechenden ordinalen Rasch-Modellen. Bei der linear-logistischen LCA für dichotome Items (Formann 1984) werden die Itemparameter, d. h. die klassenspezifischen Lösungswahrscheinlichkeiten, erst nach einer logistischen Transformation zerlegt, um einen anschaulicherem Wertebereich als den zwischen null und eins zu erzielen. Der Fall polytomer Items wird z. B. von Formann (1993) behandelt.

Linear-logistische Modelle sind insgesamt flexibler, als hier dargestellt werden kann. Insbesondere sind sie auch im Fall mehrerer Messzeitpunkte einsetzbar, sodass sich im Rahmen der IRT auch Fragestellungen der Veränderungsmessung untersuchen lassen (z. B. Fischer 1974, 1995a; Fischer und Ponocny 1995). Hierbei ist es nötig, zunächst zwischen verschiedenen Arten von Veränderungshypothesen zu unterscheiden (Rost 2004; Rost und Spada 1983). Geht es beispielsweise um den Nachweis „globaler“ Veränderungen aufgrund einer pädagogischen oder therapeutischen Intervention, so stellt dies insofern eine strenge Form einer Veränderungshypothese dar, als für alle Personen und bei allen Items der gleiche Effekt erwartet wird. Da hierdurch der differenziell-psychologische Aspekt in den Hintergrund tritt, erscheint die Forderung nach „spezifisch objektiven Vergleichen“ zwischen Personen in einem solchen Fall entbehrlich. Hier kann das von Fischer (z. B. 1983, 1995a) vorgeschlagene *linear-logistische Modell mit relaxierten Annahmen* bzw. „linear logistic model with relaxed assumptions“ (LLRA) eingesetzt werden, das ohne die für Rasch-Modelle charakteristische Annahme der Eindimensionalität bzw. Homogenität der Items auskommt.

16.7 Zusammenfassung

In diesem Kapitel wurde in die IRT eingeführt. Die grundlegende testtheoretische Idee der IRT besteht darin, die Wahrscheinlichkeit eines gezeigten Antwortverhaltens („Response“) einer Person bei einem Item (z. B. das Bejahren/Nichtbejahren einer Aussage in einem Einstellungstest bzw. das Lösen/Nichtlösen einer Aufgabe in einem Leistungstest) in Form einer (zumeist einfachen) Wahrscheinlichkeitsfunktion zu beschreiben. Das Kapitel stellte zunächst verschiedene Grundüberlegungen zu dichotomen Itemformaten und ihren Zusammenhängen dar. Danach erfolgt – in Abgrenzung zu Latent-Class-Modellen – eine Einführung in Latent-Trait Modelle. Das dichotome Rasch-Modell (1PL-Modell) als sehr grundlegendes Modell der IRT wurde vorgestellt. Dies umfasste die Modellgleichung und ihre Bestandteile wie Personenparameter und Itemparameter sowie die Funktionsweise der IC-Funktion. In diesem Abschnitt wurden außerdem Konzepte der sog. Rasch-Homogenität, Joint Scale, Interpretationen von Item- und Personenparametern, Parameternormierung, Spezifische Objektivität, Parameterschätzung, Item- und Testinformation sowie Modellpassung ausführlich behandelt. Das sog. 2PL-Modell und 3PL-Modell nach Birnbaum und ihre Eigenschaften wurden danach vorgestellt. Insbesondere die Eigenschaften variierender Diskriminationsparame-

16.8 EDV-Hinweise

ter und Rateparameter wurden beschrieben. Abschließend wurde ein Ausblick auf weitere IRT-Modelle gegeben.

16.8 EDV-Hinweise

Die Syntax für Berechnungen des LSAT-Beispiels ist in den Zusatzmaterialien zum Buch auf ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion) wiedergegeben.

Häufig eingesetzte Softwarepakete für die IRT-Modellierung sind:

- ConQuest (► <https://www.acer.org/conquest>)
- IRTPRO (► <http://www.ssicentral.com>)
- Mplus (► <http://www.statmodel.com>)
- R-project (► <https://www.r-project.org>): Hier gibt es fast unüberschaubare Zahl von Paketen für uni- und mehrdimensionale IRT-Modelle. Ein Überblick ist auf ► <https://cran.r-project.org/web/views/Psychometrics.html> gegeben.
- flexMIRT (► <https://www.vpgcentral.com/software/irt-software>)
- WINMIRA 2001 (► <http://www.von-davier.com>)

Es gilt zu beachten, dass die Entwicklung von Softwarepaketen ein sehr dynamisches Feld ist und entsprechende Produkte „kommen und gehen“. Deshalb ist die Wahrscheinlichkeit hoch, dass die obige Aufzählung in wenigen Jahren nicht mehr aktuell ist. Einen hilfreichen Überblick kann man sich ebenfalls unter ► <http://www.rasch.org/software.htm> oder ► https://en.wikipedia.org/wiki/Psychometric_software verschaffen.

16.9 Kontrollfragen

❓ Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was wird in der Item-Response-Theorie (IRT) im Unterschied zur Klassischen Testtheorie (KTT) modelliert?
2. Was versteht man unter Rasch-Homogenität?
3. Was beschreibt eine IC-Funktion?
4. In welcher Beziehung stehen Lösungswahrscheinlichkeit, Nichtlösungswahrscheinlichkeit und Iteminformationsfunktion im Rasch-Modell?
5. Erläutern Sie den Begriff „Spezifische Objektivität“.
6. Was versteht man unter „lokaler stochastischer Unabhängigkeit“?
7. Was versteht man unter „adaptivem Testen“?
8. Welche Fälle können im polytomous Rasch-Modell unterschieden werden?
9. Worin unterscheiden sich Latent-Class-Modelle von Latent-Trait-Modellen?
10. Worin besteht die Grundidee linear-logistischer Modelle?

Literatur

- Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung* (6. Aufl.). Stuttgart: Kohlhammer.
- Amthauer, R. (1970). *Intelligenz-Struktur-Test (I-S-T 70)*. Göttingen: Hogrefe.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123–140.
- Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 271–291). New York: Springer.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

- Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D. & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika*, 35, 179–197.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler Springer-Lehrbuch* (7. Aufl.). Berlin, Heidelberg: Springer.
- Cai, L. & Thissen, D. (2014). Modern approaches to parameter estimation in item response theory. In S. P. Reise & D. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York, NY: Taylor & Francis.
- Carstensen, C. H., Frey, A., Walter, O. & Knoll, S. (2007). Technische Grundlagen des dritten internationalen Vergleichs. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 367–390). Münster: Waxmann.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Fahrenberg, J., Hampel, R. & Selg, H. (2001). *Das Freiburger Persönlichkeitsinventar FPI-R mit neuer Normierung. Handanweisung* (7. Aufl.). Göttingen: Hogrefe.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Fischer, G. H. (1983). Neuere Testtheorie. In J. Bredenkamp & H. Feger (Hrsg.), *Messen und Testen* (S. 604–692). Göttingen: Hogrefe.
- Fischer, G. H. (1995a). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 157–180). New York: Springer.
- Fischer, G. H. (1995b). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). New York: Springer.
- Fischer, G. H. (1996). IRT-Modelle als Forschungsinstrumente der Differentiellen Psychologie. In K. Pawlik (Hrsg.), *Grundlagen und Methoden der Differentiellen Psychologie* (S. 673–729). Göttingen: Hogrefe.
- Fischer, G. H. & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.
- Fischer, G. H. & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of treatment effects. *Psychometrika*, 56, 637–651.
- Fischer, G. H. & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer, I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 353–370). New York: Springer.
- Formann, A. K. (1984). *Die Latent-Class-Analyse*. Weinheim: Beltz.
- Formann, A. K. (1993). Some simple latent class models for attitudinal scaling in the presence of polytomous items. *Methodika*, 7, 62–78.
- Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW). Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvorvermögens*. Weinheim: Beltz.
- Gittler, G. & Wild, B. (1988). Der Einsatz des LLTM bei der Konstruktion eines Itempools für das adaptive Testen. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie* (S. 115–139). Weinheim: Psychologie Verlags Union.
- Glas, C. A. W. & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, 54, 635–659.
- Guttmann, G. & Ettlinger, S. C. (1991). Susceptibility to stress and anxiety in relation to performance, emotion, and personality: The ergopsychometric approach. In C. D. Spielberger, I. G. Sarason, J. Strelau & J. M. T. Brebner (Eds.), *Stress and anxiety* (Vol. 13, pp. 23–52). New York: Hemisphere Publishing Corporation.
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (Hrsg.), *2007. Sprachliche Kompetenzen – Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Weinheim: Beltz.
- Hartig, J., Klieme, E. & Leutner, D. (Eds.). (2008). *Assessment of competencies in educational contexts*. Göttingen: Hogrefe.
- Holland, P. & Wainer, H. (1993). Differential item functioning. New York: Erlbaum.
- Klauer, K. C. (1991). An exact and optimal standardized person fit test for assessing consistency with the Rasch model. *Psychometrika*, 56, 213–228.
- Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97–110). New York: Springer.
- Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. H. Lehmann (Hrsg.) *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit*. Opladen: Leske + Budrich.
- Köller, O. (1993). Die Identifikation von Ratern bei Leistungstests mit Hilfe des Mixed-Rasch-Modells. Vortrag auf der 1. Tagung der Fachgruppe Methoden der Deutschen Gesellschaft für Psychologie in Kiel. Empirische Pädagogik (o. A.).

Literatur

- Kubinger, K. D. (1995). *Einführung in die Diagnostik*. Weinheim: Psychologie Verlags Union.
- Kubinger, K. D. (2003). Adaptives Testen. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz PVU.
- Kubinger, K. D. & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum (AID 2)*. Göttingen: Hogrefe.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.
- Lord, F. N. & Nowick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Masters, G. N. & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, 49, 529–544.
- Molenaar, I. W. & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75–106.
- Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, 52, 165–181.
- Müller, H. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen*. Bern: Huber.
- OECD (2017). *PISA 2015 Technical Report*. Paris: OECD.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Berkeley, CA: University of California Press.
- Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie*. Bern: Huber.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
- Rost, J. (2006). Item-Response-Theorie. In F. Petermann & M. Eid (Hrsg.), *Handbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.
- Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 4, 29–49.
- Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für experimentelle und angewandte Psychologie*, 19, 476–506.
- Snijders, T. A. B. (2001). Asymptotic Null Distribution of Person Fit Statistics with Estimated Person Parameter. *Psychometrika*, 66, 331–342.
- Strauß, B., Köller, O. & Möller, J. (1996). Geschlechtsrollentypologien – eine empirische Prüfung des additiven und des balancierten Modells. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 17, 67–83.
- Strobl, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis (Sozialwissenschaftliche Forschungsmethoden)*. Hampp, Mering.
- Strobl, C., Kopf, J. & Zeileis, A. (2010). Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben. In S. Trepte & M. Verbeet (Hrsg.), *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studenten-pisa-Test* (S. 255–272). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Tarnai, C. & Rost, J. (1990). Identifying aberrant response patterns in the Rasch model. *The Q Index. Sozialwissenschaftliche Forschungsdokumentation*. Münster: Institut für sozialwissenschaftliche Forschung e. V.
- van den Wollenberg, A. L. (1988). Testing a latent trait model. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 31–50). New York: Plenum.
- Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Interpretation von Testwerten in der Item-Response-Theorie (IRT)

Dominique Rauch und Johannes Hartig

Inhaltsverzeichnis

- 17.1 Vorbemerkungen – 412
- 17.2 Grundlagen kriteriumsorientierter Testwertinterpretation in IRT-Modellen – 414
 - 17.3 Definition von Kompetenzniveaus zur kriteriumsorientierten Testwertinterpretation – 417
 - 17.3.1 Grundidee von Kompetenzniveaus – 417
 - 17.3.2 Methoden zur Bestimmung von Schwellen zwischen den Kompetenzniveaus – 417
 - 17.4 Verwendung von Post-hoc-Analysen und A-priori-Merkmalen zur Testwertbeschreibung – 418
 - 17.4.1 Post-hoc-Analyse von Iteminhalten – 418
 - 17.4.2 Verwendung von A-priori-Aufgabenmerkmalen zur Testwertbeschreibung – 420
- 17.5 Zusammenfassung – 422
- 17.6 EDV-Hinweise – 423
- 17.7 Kontrollfragen – 423
- Literatur – 423

i Die IRT hat eine Reihe von anwendungsbezogenen Vorteilen, die vor allem im Rahmen von Large-Scale-Assessments in der empirischen Bildungsforschung genutzt werden. Sie ermöglicht das Matrix-Sampling von Testaufgaben, die Erstellung paralleler Testformen und die Entwicklung computerisierter adaptiver Tests. Vor allem aber ist es bei Gültigkeit des Raschmodells durch die Joint Scale von Itemschwierigkeiten und Personenfähigkeiten möglich, individuelle Testwerte durch ihre Abstände zu Itemschwierigkeiten zu interpretieren. Durch die Unterteilung der gemeinsamen Skala in sog. „Kompetenzniveaus“ können Kompetenzen von Schülerinnen und Schülern in einem bestimmten Abschnitt der Kompetenzskala beschrieben werden. Alternative Verfahren zur Erstellung dieser Kompetenzniveaus werden an einem gemeinsamen Beispiel aus der dritten internationale Mathematik- und Naturwissenschaftsstudie „Trends in International Mathematics and Science Study“ (TIMSS/III) und der Deutsch Englisch Schülerleistungen International (DESI) vorgestellt.

17.1 Vorbemerkungen

Indirekte Messung

Im Gegensatz zur deskriptivstatistischen Itemanalyse (► Kap. 7) setzt die Item-Response-Theorie (IRT, ► Kap. 16) bei der Testwertebildung die Antworten von Personen auf die Items eines Tests nicht mit der Messung des im Test erfassenen Konstrukts gleich, sondern konzipiert die Messung des Konstrukt als indirekt: IRT-Modelle postulieren, dass dem im Test gezeigten Verhalten, also den Antworten auf die Items („item responses“) des Tests, eine Fähigkeit oder Eigenschaft zugrunde liegt, die das Testverhalten „verursacht“. Das beobachtete Verhalten (erfasst als manifeste Variable) stellt lediglich einen Indikator für das dahinterliegende Konstrukt (latente Variable) dar, dessen Ausprägung erschlossen werden muss. Bei der Anwendung von IRT-Modellen ist es möglich, für jede Person eine Schätzung ihrer individuellen Ausprägung η_v auf der latenten Variable η vorzunehmen (für die Schätzung im Rasch-Modell s. ► Abschn. 16.3.8). Diese individuelle Schätzung des Personenparameters η_v stellt den IRT-basierten Testwert einer Person v dar.

Einen prominenten Anwendungsbereich hat die IRT-basierte Schätzung von Personenmerkmalen u. a. in der empirischen Bildungsforschung. Groß angelegte Erhebungen von Schülerleistungen (*Large-Scale-Assessments*) wie das „Programme for International Student Assessment“ (PISA, vgl. Baumert et al. 2001; Klieme et al. 2010; OECD 2001, 2004a, 2004b; PISA-Konsortium Deutschland 2004; Reiss et al. 2016) oder die internationale Mathematik- und Naturwissenschaftsstudie „Trends in International Mathematics and Science Study“ (TIMSS; vgl. Klieme et al. 2000) verwenden Modelle der IRT bei der Auswertung von Leistungstests. Dabei werden spezifische Vorteile der IRT genutzt: So wird es durch IRT-Analysen möglich, jede/n Schüler/in nur eine Stichprobe aus einer Gesamtheit homogener Testaufgaben bearbeiten zu lassen, die zur Erfassung einer spezifischen Kompetenz eingesetzt werden (*Matrix-Sampling*). IRT-basierte Schätzungen der zu erfassenden Fähigkeiten erlauben es, trotz unterschiedlicher bearbeiteter Itemmengen für alle SchülerInnen Testwerte auf einer gemeinsamen Skala zu bestimmen.

Matrix-Sampling von Testaufgaben

Ebenfalls Gebrauch von dieser Möglichkeit kann bei der Erstellung paralleler Testformen mithilfe von IRT-Modellen gemacht werden. *Parallele Testformen* (vgl. auch ► Kap. 13) werden u. a. eingesetzt, um bei wiederholten Messungen Erinnerungseffekte ausschließen zu können. Items eines IRT-skalierten Tests können auf zwei oder mehr Testformen aufgeteilt werden. Sofern das IRT-Modell gilt, können mit jeder Testform Ausprägungen auf derselben Variablen (oder im Fall eines mehrdimensionalen Modells auf mehreren Variablen, s. dazu ► Kap. 18) gemessen werden. Dabei dienen sog. „Ankeritems“ dazu, die Items der Testformen auf einer Skala mit einer gemeinsamen Metrik zu positionieren (zu verankern). Individuelle Testwerte, die für Personen nach Bearbeitung unterschiedlicher Test-

17.1 · Vorbemerkungen

formen geschätzt werden, können so miteinander verglichen werden. Ein Beispiel für die Nutzung einer IRT-basierten Testwertschätzung für die Auswertung paralleler Testformen ist der Test für mathematisches Fachwissen von Mathematiklehrern von Hill et al. (2004).

Ein weiterer Anwendungsbereich von IRT-Modellen ist das sog. „computerisierte adaptive Testen“. Bei computerisierten adaptiven Tests wird im Verlauf des Testens auf Basis sich wiederholender Personenfähigkeitschätzungen aus einer großen Anzahl *kalibrierter Items* immer dasjenige Item vorgegeben, das für die jeweilige Schätzung der Personenfähigkeit die höchste Iteminformation (s. Iteminformationsfunktion, ► Abschn. 16.3.9) aufweist. So werden nur solche Items vorgegeben, die für das Fähigkeitsniveau η_v der getesteten Person eine hohe Messgenauigkeit aufweisen, also weder zu leicht noch zu schwierig sind. Aufgrund dieser maßgeschneiderten Vorgabe von Items spricht man auch von „*Tailored Testing*“. Diese Form des adaptiven Testens macht von der Möglichkeit Gebrauch, die Personenfähigkeit im Verlauf des Testens ständig neu zu schätzen und diese Schätzung zur Grundlage der weiteren Itemauswahl zu machen. Die spezifische Auswahl von vorgegebenen Items dient einer möglichst genauen Messung in möglichst kurzer Zeit. Grundlagen und Anwendungen des computerisierten adaptiven Testens werden von Frey in ► Kap. 20 beschrieben.

Personenfähigkeiten und Itemschwierigkeiten werden in der IRT auf einer gemeinsamen Skala (*Joint Scale*) verortet. Diese gemeinsame Skala wird in IRT-Modellen mit logistischen itemcharakteristischen Funktionen (IC-Funktionen) auch als *Logit-Skala* (z. B. Rost 2004) bezeichnet. Ein *Logit-Wert* ist der Logarithmus des Wettquotienten (*Odds*) aus Lösungswahrscheinlichkeit $P(y_{vi} = 1)$ und Gegenwahrscheinlichkeit $P(y_{vi} = 0)$. Die Metrik der gemeinsamen Skala ist abhängig vom gewählten IRT-Modell und den Restriktionen der Parameterschätzung. Um diese gemeinsame Skala von Personenfähigkeiten und Itemschwierigkeiten zu definieren, muss zunächst der Nullpunkt der Skala festgelegt werden. Die Metrik der gemeinsamen Skala wird festgelegt, indem entweder die durchschnittliche Itemschwierigkeit oder die durchschnittliche Personenfähigkeit auf null restriktiert wird. In Abhängigkeit vom gewählten Nullpunkt der Skala können die Testwerte somit bezogen auf die durchschnittliche Itemschwierigkeit oder die durchschnittliche Personenfähigkeit interpretiert werden (vgl. dazu ► Abschn. 16.3.5 und 16.3.6).

Trotz dieser Referenz auf mittlere Fähigkeit oder mittlere Schwierigkeit bleibt die Metrik von Testwerten aus IRT-Modellen unhandlich, typischerweise resultieren Werte in einem numerisch relativ kleinen Wertebereich um null (z. B. -3 Logits bis +3 Logits). Um anschaulichere Werte zu gewinnen, können geschätzte IRT-basierte Testwerte nach den gleichen Regeln normiert werden wie Testwerte, die auf Basis der KTT gewonnen wurden (► Kap. 9). In der *PISA-Studie* beispielsweise wurden die Testwerte so normiert, dass der Leistungsmittelwert über alle teilnehmenden OECD-Staaten 500 und die Standardabweichung 100 Punkte beträgt (z. B. OECD 2001, 2004a). Für deutsche Fünfzehnjährige ergab sich 2003 auf der PISA-Gesamtskala für Lesekompetenz ein Mittelwert von 491 Punkten bei einer Standardabweichung von 109 Punkten (OECD 2004a). Bei normorientierter Interpretation lag die Leseleistung deutscher Schüler demnach knapp unter dem OECD-Mittelwert, die Streuung der Testergebnisse ist jedoch etwas größer als der internationale Durchschnitt.

In Hinblick auf Unterrichtsverbesserungen und Möglichkeiten zur gezielten Förderung von Schülergruppen wird der normorientierte Vergleich von Leistungswerten mit Bezugspopulationen oder der Vergleich von Subpopulationen untereinander (z. B. die Ländervergleiche in der PISA-Studie) oftmals nicht als ausreichend erachtet (z. B. Helmke und Hosenfeld 2004). Es besteht vielmehr der Bedarf nach einer *kriteriumsorientierten Interpretation* der Schülertestwerte (vgl. ► Kap. 9): So interessiert beispielsweise, über welche spezifischen Kompetenzen bestimmte Schülergruppen verfügen und welche fachbezogenen Leistungsanforderungen sie

Computerisierte adaptive Tests

Kalibrierte Items

Joint Scale für Personenfähigkeiten und Itemschwierigkeiten, Logit-Werte

Normorientierte Interpretation IRT-basierter Testwerte am Beispiel von PISA

Kriteriumsorientierte Interpretation IRT-basierter Testwerte

mit ausreichender Sicherheit bewältigen können. Hierzu müssen die Testwerte auf der Fähigkeitsskala zu konkreten, fachbezogenen Anforderungen in Bezug gesetzt werden.

Normorientierte und kriteriumsorientierte Interpretationen von Leistungsmessungen müssen nicht als konkurrierend verstanden werden; unter bestimmten Voraussetzungen können Testergebnisse sowohl norm- als auch kriteriumsorientiert interpretiert werden.

17.2 Grundlagen kriteriumsorientierter Testwertinterpretation in IRT-Modellen

Grundvoraussetzung für eine kriteriumsorientierte Interpretation individueller Testwerte ist die Abbildung von Itemschwierigkeiten und Personenfähigkeiten auf einer gemeinsamen Skala. Im Rahmen der deskriptivstatistischen Itemanalyse (► Kap. 7) wird zwischen der individuellen Leistung einer Person (z. B. Prozent gelöster Items) und der Schwierigkeit eines Items (z. B. Prozent der Personen, die das Item gelöst haben) kein expliziter Bezug hergestellt. In IRT-Modellen dagegen werden individuelle Fähigkeitsschätzungen und Itemschwierigkeiten auf einer gemeinsamen Skala abgebildet. Dadurch ist es möglich, individuelle Testwerte durch ihre Abstände zu Itemschwierigkeiten zu interpretieren (Embreton 2006). Eine eindeutige relative Lokalisation von Personenfähigkeit und Itemschwierigkeit ist allerdings nur dann möglich, wenn die IC-Funktionen aller Items parallel verlaufen (Spezifische Objektivität, s. ► Kap. 16, ► Abschn. 16.3.6). Dies ist im Rasch-Modell, auch einparametrisches logistisches Modell (1PL-Modell) genannt, der Fall: Die IC-Funktion eines Items ist hier durch einen einzigen Parameter, und zwar die Itemschwierigkeit, vollständig determiniert.

Im Unterschied zum Rasch-Modell haben mehrparametrische Modelle wie das 2PL-Modell oder das 3PL-Modell gegenüber dem einparametrischen Rasch-Modell den für die Interpretierbarkeit der Skala schwerwiegenden Nachteil, dass sich Differenzen zwischen den Lösungswahrscheinlichkeiten mehrerer Items in Abhängigkeit von der Personenfähigkeit verändern, d. h., die IC-Funktionen verschiedener Items schneiden sich (s. ► Kap. 16, ► Abschn. 16.4, 16.5 und □ Abb. 16.9). Dies kann zu dem paradoxen Ergebnis führen, dass ein Item dem Modell zufolge für eine bestimmte Person leichter ist als ein anderes und sich dieses Verhältnis für eine andere Person umkehrt. Angesichts der Vorteile des Rasch-Modells für die kriteriumsorientierte Testwertinterpretation (vgl. Wilson 2003) wird im Folgenden ausschließlich auf dieses Modell Bezug genommen.

Im Rasch-Modell ist die Schwierigkeit eines Items definiert als jene Ausprägung auf der Fähigkeitsskala, die erforderlich ist, um das Item mit einer Wahrscheinlichkeit von 50 % zu lösen (□ Abb. 17.1). Über die itemcharakteristische Funktion können spezifische Ausprägungen der Personenfähigkeit in Lösungswahrscheinlichkeiten für Items mit bestimmten Schwierigkeiten übertragen werden. Betrachtet man z. B. die IC-Funktionen der drei in □ Abb. 17.1 abgetragenen Items, so ist zu erwarten, dass Personen, deren geschätzte Fähigkeit $\eta_v = 0$ den gleichen Wert auf der Skala aufweist wie die Schwierigkeit von Item 2 ($\beta_2 = 0$), dieses Item zu 50 % lösen können. Item 1 ($\beta_1 = -1$) hingegen sollten etwa 70 % dieser Personen lösen können, während Item 3 ($\beta_3 = 2$) von etwa 10 % der Personen bewältigt werden dürfte.

Personen, deren individueller Testwert die 50 %-Schwelle eines Items übersteigt, können dieses Item mit „hinreichender“ Wahrscheinlichkeit (also mit mindestens 50%iger Lösungswahrscheinlichkeit) korrekt lösen, Personen mit niedrigeren Testwerten hingegen nicht. Auf Basis derjenigen Items, die mit einer Wahrscheinlichkeit von 50 % oder mehr richtig gelöst werden können, lässt sich kri-

Vorteile des Rasch-Modells

Nachteile mehrparametrischer Modelle

17.2 · Grundlagen kriteriumsorientierter Testwertinterpretation in IRT-Modellen

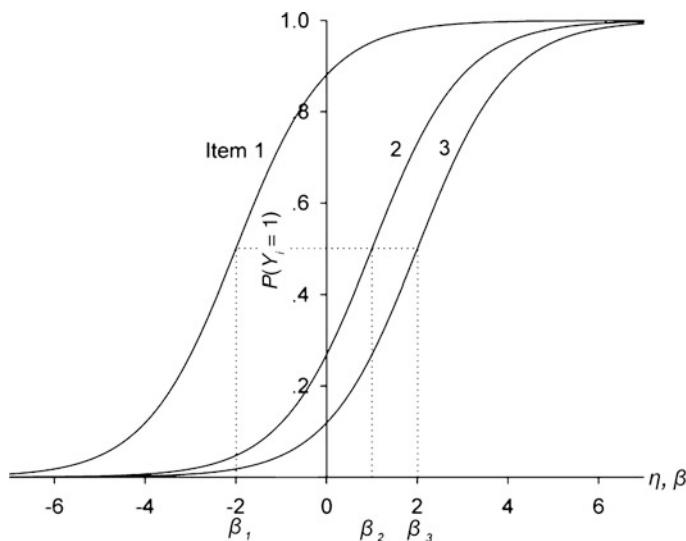


Abb. 17.1 Logistische IC-Funktionen für drei Rasch-homogene Items mit unterschiedlichen Schwierigkeitsparametern. Item 1 ist das leichteste mit $\beta_1 = -1$; Item 2 ist schwieriger mit $\beta_2 = 0$; Item 3 ist das schwierigste mit $\beta_3 = 2$

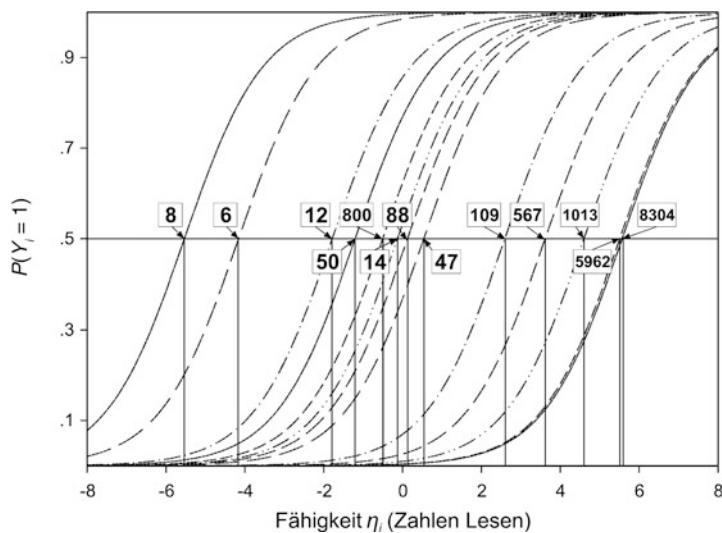
teriumsorientiert beschreiben, welche Anforderungen diese Personen bewältigen können (► Beispiel 17.1).

Beispiel 17.1: Kriteriumsorientierte Interpretation von individuellen Testwerten

Als Veranschaulichung für die kriteriumsorientierte Interpretation von individuellen Fähigkeitsschätzungen soll hier der Subtest „Zahlen lesen“ (Graf et al. 2011) aus dem Projekt „Persönlichkeits- und Lernentwicklung an sächsischen Grundschulen“ (PERLE) dienen. Die Studie PERLE untersucht die Lernentwicklung und Persönlichkeitsentwicklung von Grundschulkindern, im Mittelpunkt stehen die Bereiche Schriftspracherwerb, Mathematik und bildende Kunst. Im Bereich Mathematik wurden an Erstklässlern unmittelbar nach Schuleintritt die mathematischen Vorläuferfähigkeiten für spätere Leistungen im Mathematikunterricht erhoben. Ein Subtest erfasst die Fähigkeit der Kinder, Zahlen zu lesen. Hierbei wurden den Kindern 13 Zahlen vorgelegt, die sie laut benennen sollten. Der Test beginnt mit einstelligen Zahlen, dann folgen zwei-, drei- und vierstellige Zahlen. Es wird nur zwischen richtigen und falschen Antworten unterschieden. Die Skalierung der Daten erfolgte auf Basis des dichotomen Rasch-Modells (Greb 2007). □ Abb. 17.2 zeigt die IC-Funktionen der 13 Items.

Am leichtesten richtig zu benennen sind die einstelligen Zahlen „8“ und „6“; sie erfordern nur eine geringe Lesefähigkeit und liegen somit auf der Joint Scale ganz links. Schwerer und somit weiter rechts sind die zweistelligen Zahlen, wobei entsprechend den theoretischen Annahmen Zahlen ohne Inversion („50“) leichter zu benennen sind als solche, bei denen die zweite Ziffer zuerst genannt werden muss („47“). Die Items, bei denen dreistellige Zahlen zu lesen sind, finden sich mit deutlichem Abstand noch weiter rechts auf der Joint Scale, d. h., sie erfordern eine deutlich höhere Fähigkeit beim Lesen von Zahlen. Eine Ausnahme bildet die Zahl 800, die aufgrund ihres einfachen Aufbaus im Schwierigkeitsbereich der zweistelligen Zahlen liegt. Am schwierigsten richtig zu benennen sind erwartungsgemäß die vierstelligen Zahlen.

Kinder, deren Fähigkeit ungefähr bei $\eta_v = 1$ auf der gemeinsamen Skala von Itemschwierigkeit und Personenfähigkeit zu verorten ist, werden alle Items, die

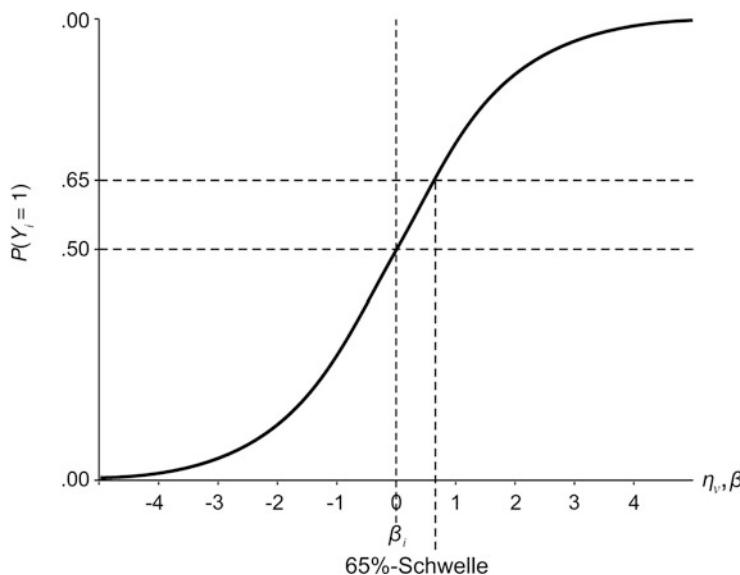


■ Abb. 17.2 Auf Basis des dichotomen Rasch-Modells ermittelte IC-Funktionen der 13 Items mit ein-, zwei-, drei- und vierstelligen Zahlen des PERLE-Subtests „Zahlen lesen“. (Nach Greb 2007)

leichter sind ($\beta_i < 1$), mit einer Wahrscheinlichkeit von über 50 % lösen. Sie beherrschen mit sehr großer Wahrscheinlichkeit (> 95 %) die einstelligen Zahlen, deren Schwierigkeiten am niedrigsten liegen, und können auch alle zweistelligen Zahlen schon zuverlässig benennen (> 70 %). Bei drei- und vierstelligen Zahlen kann man jedoch nicht mit hinreichender Sicherheit davon ausgehen, dass diese Kinder sie bereits lesen können. Eine kriteriumsbezogene Interpretation der numerischen Werte von Testergebnissen wird durch den Vergleich der individuellen Fähigkeitsschätzungen mit den Itemschwierigkeiten möglich, da die Kompetenzen der Kinder auf Anforderungen der Items bezogen werden können.

Eine Lösungswahrscheinlichkeit von 50 % erscheint als relativ niedrig, um darauf zu schließen, dass Personen mit der Fähigkeit $\eta_v = \beta_i$ die Anforderungen von Item i hinreichend sicher bewältigen können. Daher werden in Schulleistungsstudien, in denen das Vorhandensein spezifischer Kompetenzen untersucht werden soll, oft auch höhere Lösungswahrscheinlichkeiten als 50 % gewählt, um einzelne Items auf der Kompetenzskala zu verorten. Anhand des im Rasch-Modell angenommenen Zusammenhangs zwischen Personenfähigkeit und Lösungswahrscheinlichkeit lassen sich leicht auch Punkte auf der Kompetenzskala bestimmen, an denen die Lösungswahrscheinlichkeit für ein spezifisches Item einen beliebigen anderen Wert als 50 % annimmt. In ■ Abb. 17.3 ist dies am Beispiel der „65 %-Schwelle“, wie sie z. B. in der internationale Mathematik- und Naturwissenschaftsstudie TIMSS (Klieme et al. 2000) und in der DESI-Studie zu sprachlichen Leistungen in Deutsch und Englisch (Beck und Klieme 2003; Hartig 2007; Klieme und Beck 2007) verwendet wurde, dargestellt. Personen, deren individueller Testwert die 65 %-Schwelle eines Items übersteigt, können dieses Item mit „hinreichender“ Wahrscheinlichkeit (also mit mindestens 65 %iger Lösungswahrscheinlichkeit) korrekt lösen, Personen mit niedrigeren Testwerten hingegen nicht. Auf Basis derjenigen Items, die mit einer Wahrscheinlichkeit von 65 % oder mehr richtig gelöst werden können, lässt sich kriteriumsorientiert beschreiben, welche Anforderungen diese Personen bewältigen können (■ Abb. 17.3).

17.3 · Definition von Kompetenzniveaus zur kriteriumsorientierten Testwertinterpretation



■ Abb. 17.3 Veranschaulichung der Bildung der „65 %-Schwelle“ für ein Item mit $\beta_i = 0$

17.3 Definition von Kompetenzniveaus zur kriteriumsorientierten Testwertinterpretation

17.3.1 Grundidee von Kompetenzniveaus

Um die quantitativen Werte einer Kompetenzskala kriteriumsorientiert zu beschreiben, wird in der empirischen Bildungsforschung ein pragmatisches Vorgehen gewählt: Die kontinuierliche Skala wird in Abschnitte unterteilt, die als Kompetenzniveaus bezeichnet werden (vgl. Hartig und Klieme 2006). Die kriteriumsorientierte Beschreibung erfolgt für jeden Skalenabschnitt; innerhalb der gebildeten Kompetenzniveaus wird keine weitere inhaltliche Differenzierung der erfassten Kompetenz vorgenommen. Dieses Vorgehen wird nicht zuletzt damit begründet, dass es in der Praxis nicht realisierbar ist, jeden einzelnen Punkt auf einer quantitativen Skala anhand konkreter, fachbezogener Kompetenzen inhaltlich zu beschreiben (Beaton und Allen 1992).

17.3.2 Methoden zur Bestimmung von Schwellen zwischen den Kompetenzniveaus

Um für die Skala eines existierenden Tests Kompetenzniveaus zu bilden, stehen verschiedene Methoden zur Verfügung. Die Grundlage für die Definition von Kompetenzniveaus liefern bei jeder dieser möglichen Vorgehensweisen zum einen die Items selbst, genauer gesagt ihre fachbezogenen Anforderungen, und zum anderen die im Rahmen der IRT-Skalierung ermittelten Itemschwierigkeiten. Entscheidend bei der Bildung von Niveaus ist die Methode zur Bestimmung der *Schwellen zwischen den Niveaus*, wobei verschiedene Methoden unterschieden werden können. Im einfachsten Fall werden die Schwellen zwischen den Abschnitten auf der Kompetenzskala willkürlich gesetzt, z. B. in gleichen Abständen oder bezogen auf die Mittelwerte bestimmter Bezugsgruppen. Anschließend wird nach Items gesucht, deren Schwierigkeiten für die gesetzten Schwellen charakteristisch sind. Die inhaltliche Beschreibung der Skalenabschnitte erfolgt dann anhand einer an-

Post-hoc-Analyse der Iteminhalte

schließenden *Post-hoc-Analyse* der Inhalte dieser Items (Beaton und Allen 1992; ► Abschn. 17.4).

A-priori-Aufgabenmerkmale

Häufig können jedoch schon während der Testentwicklung Annahmen darüber formuliert werden, aus welchem Grund bestimmte Items leichter oder schwerer sind. Liegen theoretisch begründete *a priori definierte Annahmen* darüber vor, welche spezifischen Merkmale des Items seine Schwierigkeit bedingen, so können diese Merkmale herangezogen werden, um die Schwellen zwischen den Kompetenzniveaus festzulegen. Man spricht in diesem Zusammenhang von A-priori-Aufgabenmerkmalen, da sich diese beispielsweise bei einem Leseverständnistest aus den Anforderungen des Textes und den Aufgaben zum Text (d. h. den Items) ergeben.

17.4 Verwendung von Post-hoc-Analysen und A-priori-Merkmalen zur Testwertbeschreibung

Im Folgenden (► Abschn. 17.4.1) wird die Definition und Beschreibung von Kompetenzniveaus anhand eines Vorgehens mit Post-hoc-Analysen der Items expliziert (► Beispiel 17.2). In ► Abschn. 17.4.2 folgt die Veranschaulichung für die Verwendung von A-priori-Aufgabenmerkmalen anhand desselben Beispiels.

17.4.1 Post-hoc-Analyse von Iteminhalten

Ein Post-hoc-Verfahren zur Definition und inhaltlichen Beschreibung von Kompetenzniveaus soll beispielhaft (► Beispiel 17.2) am TIMSS/III-Test (Klieme et al. 2000) zur naturwissenschaftlichen Grundbildung dargestellt werden. TIMSS/III ist eine international vergleichende Schulleistungsuntersuchung, in der Mathematik- und Naturwissenschaftsleistungen von Schülerinnen und Schülern der Sekundarstufe II untersucht wurden.

Beispiel 17.2: Bestimmung von Kompetenzniveaus in der TIMSS/III

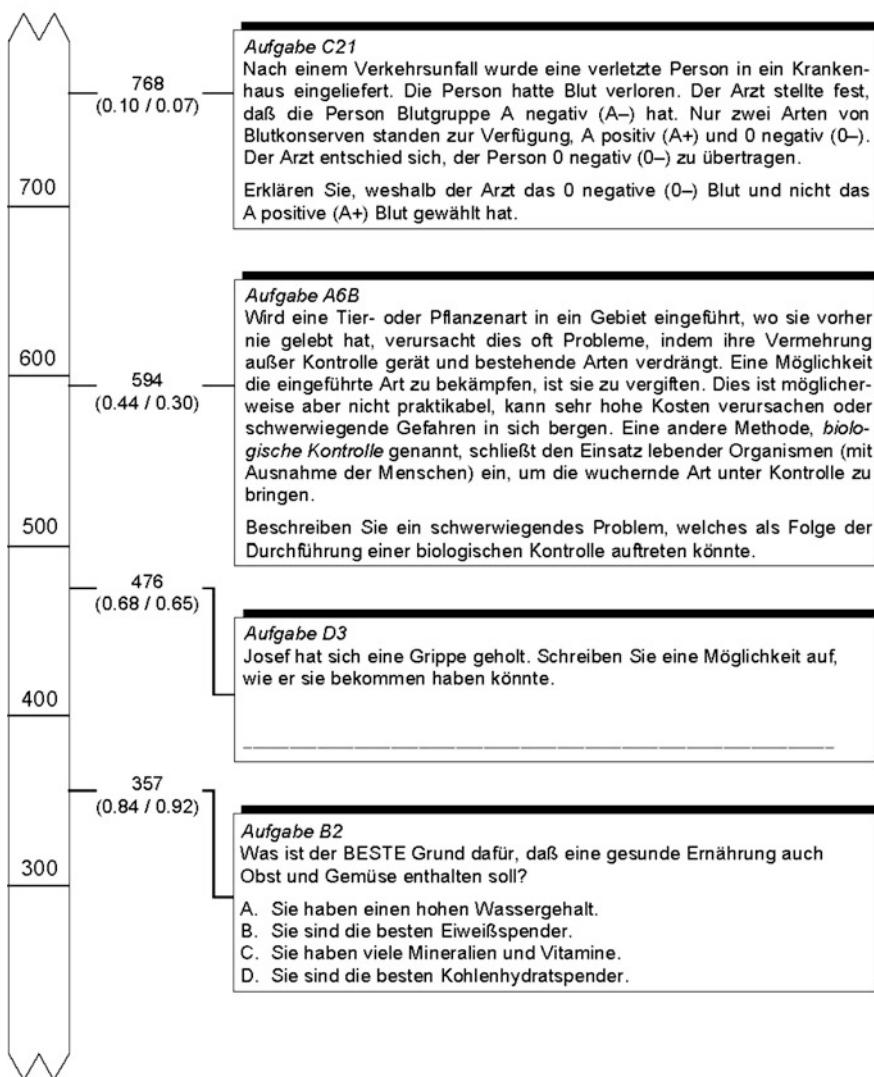
Die Bestimmung von Kompetenzniveaus in der TIMSS erfolgt nach einem von Beaton und Allen (1992) vorgestellten Verfahren, bei dem zuerst nach einer Sichtung der Testitems durch Experten/-innen Ankerpunkte auf der Kompetenzskala gesetzt und anschließend diejenigen Items identifiziert und inhaltlich betrachtet werden, die zur Beschreibung dieser Ankerpunkte geeignet sind. Im Folgenden werden die in der Terminologie von Beaton und Allen (1992) als Ankerpunkte bezeichneten Punkte auf der Kompetenzskala als Schwellen zwischen Kompetenzstufen betrachtet: Ein Schüler, der eine solche Schwelle erreicht, wird als kompetent diagnostiziert, die mit diesem Punkt verbundenen inhaltlichen Anforderungen zu bewältigen.

Auch für die TIMSS/III wurden die Daten aus den Leistungstests mit dem Rasch-Modell ausgewertet, Grundlage für diese Skalierung bildete der internationale Datensatz. Anschließend wurden die auf Basis des Rasch-Modells gebildeten Testwerte so normiert, dass der internationale Mittelwert der Skala 500 Punkte und die Standardabweichung 100 Punkte beträgt (zum Verfahren s. ► Kap. 9, ► Abschn. 9.2.2). Auf Grundlage einer ersten Inspektion der Aufgaben wurden vier Schwellen gesetzt: Zunächst wurde eine Schwelle auf den internationalen Mittelwert (500) gelegt; eine weitere Schwelle wurde eine Standardabweichung unterhalb des Mittelwerts (400) gesetzt, zwei weitere eine und zwei Standardabweichungen oberhalb des Mittelwerts (600 und 700). Klieme et al. (2000, S. 118) betonen, dass „die primäre Festlegung der Zahl der Kompetenzstufen und deren Abstände in gewissem Maße arbiträr“ erfolgte.

17.4 · Verwendung von Post-hoc-Analysen und A-priori-Merkmalen zur Testwertbeschreibung

Zur inhaltlichen Bestimmung der Schwellen wurde zunächst für jede Schwelle die Menge derjenigen Items gebildet, deren auf Basis der IC-Funktion erwartete Lösungswahrscheinlichkeit an diesem Punkt hinreichend hoch ($> 65\%$) und an der nächsten tieferen Schwelle hinreichend niedrig ($< 50\%$) ist. Items, die diese Kriterien für eine Schwelle erfüllen, werden als charakteristisch für die Kompetenz an diesem Punkt der Kompetenzskala betrachtet. Die genannten Auswahlkriterien führen dazu, dass jedes Item maximal einer Schwelle zugeordnet wird. Es werden allerdings nicht alle Items einer Schwelle zugeordnet. Items, die die Auswahlkriterien für keinen Punkt erfüllen, werden für die inhaltliche Spezifikation der Kompetenzniveaus nicht mehr berücksichtigt. Die Items, die einer Schwelle zugeordnet wurden, wurden dann hinsichtlich ihrer Inhalte und Anforderungen analysiert. Auf diese Weise kann pro Kompetenzniveau eine charakteristische Itemmenge gebildet werden, deren inhaltliche Analyse Aufschluss über die ihnen gemeinsa-

Fähigkeit



■ Abb. 17.4 Kompetenzniveaudefinition in TIMSS/III: Zur inhaltlichen Beschreibung der Schwellen wurden Aufgaben unterhalb des Schwellenwertes herangezogen. Kompetenzniveau 1 ab 400 Punkte, 2 ab 500 Punkte, 3 ab 600 Punkte und 4 ab 700 Punkte. Unterhalb von Kompetenzniveau 1 ist kein weiteres Niveau definiert, lediglich die Skala weiter fortgeführt. (Aus Klieme et al. 2000, S. 128)

men Anforderungen gibt. Die Beschreibungen dieser für jedes Niveau spezifischen Anforderungen dienen dann als inhaltliche Spezifikation der Kompetenzniveaus (s. □ Abb. 17.4).

□ Abb. 17.4 aus dem ersten TIMSS/III-Band (Klieme et al. 2000) zeigt Kompetenzniveaus der TIMSS/III-Skala „Naturwissenschaftliche Grundbildung“ mit den für die Schwellen charakteristischen Items (in □ Abb. 17.4 mit „Aufgabe“ bezeichnet). Zur inhaltlichen Bestimmung der Kompetenzniveaus wurden Items knapp unter dem jeweiligen Schwellenwert herangezogen.

Aufgabe A6B in □ Abb. 17.4 liegt beispielsweise mit einer Schwierigkeit von 594 Punkten sehr knapp unter der Schwelle zum Kompetenzniveau 3 (in TIMSS/III mit „Stufe“ bezeichnet), die auf 600 Punkte gelegt wurde. Diese Aufgabe eignet sich wegen der unmittelbaren Nähe zur Schwelle gut, um zu beschreiben, was Schülerinnen und Schüler auf Kompetenzniveau 3 „Anwendung elementarer naturwissenschaftlicher Modellvorstellungen“ bereits mit hinreichender Sicherheit können: Sie sind in der Lage, Modellvorstellungen, die im naturwissenschaftlichen Unterricht der Mittelstufe vermittelt werden, auf ein konkretes Problem anzuwenden. Dementsprechend können sie erklären, dass die Einführung eines neuen Organismus in ein Ökosystem unerwünschte Folgen haben kann, beispielsweise eine unkontrollierte Vermehrung dieses Organismus, wenn dieser keine natürlichen Feinde hat.

17.4.2 Verwendung von A-priori-Aufgabenmerkmalen zur Testwertbeschreibung

Oft können bereits vor der Testanwendung Annahmen über Aufgabenmerkmale, die sich auf die Schwierigkeiten der Items auswirken, formuliert werden. Derartige a priori begründete Aufgabenmerkmale können verwendet werden, um IRT-basierte Testwerte kriterienorientiert zu beschreiben. Mit einer auf Aufgabenmerkmale bezogenen Beschreibung wird es leichter, das zu erfassende Merkmal η in generalisierter Weise, d. h. unabhängig von den einzelnen Testaufgaben, zu beschreiben.

Relevante Aufgabenmerkmale für die Testwertbeschreibung beziehen sich auf Anforderungen, die beim Bearbeiten und Lösen der Items bewältigt werden müssen. Aufgabenmerkmale können sich auf kognitive Prozesse beim Lösen, auf Eigenschaften des Aufgabenmaterials, aber auch auf technische Oberflächencharakteristika der Aufgaben beziehen. Hartig und Klieme (2006, S. 136) nennen folgende Beispiele für schwierigkeitsrelevante Aufgabenmerkmale:

- Zum Lösen der Aufgabe auszuführende kognitive Operationen (z. B. Bilden eines mentalen Modells beim Lesen)
- Schwierigkeit hinsichtlich spezifischer Kriterien (z. B. Wortschatz eines Lese-tektes)
- Spezifische Phänomene im jeweiligen Leistungsbereich (z. B. Bilden von Konjunktivformen)
- Aufgabenformate (z. B. geschlossene vs. offene Antworten)

Der erste Schritt auf dem Weg zu einer Beschreibung von Testwerten mittels A-priori-Aufgabenmerkmalen ist es, den Zusammenhang zwischen Merkmalen und Itemschwierigkeiten zu untersuchen. Nur solche Aufgabenmerkmale, die tatsächlich einen Effekt auf die Itemschwierigkeiten eines Tests haben, können zur Beschreibung der Testwerte verwendet werden. Lässt sich der Effekt der Aufgabenmerkmale auf die Itemschwierigkeiten empirisch nachweisen, können die Testwerte auf Itemschwierigkeiten β_i bezogen werden, die sich für spezifische Konfigurationen von Aufgabenmerkmalen ergeben.

Relevante Aufgabenmerkmale

17.4 · Verwendung von Post-hoc-Analysen und A-priori-Merkmalen zur Testwertbeschreibung

Es gibt verschiedene Methoden, die Aufgabenmerkmale mit den Itemschwierigkeiten in Beziehung zu setzen und zur Testwertbeschreibung zu verwenden. Hier soll kurz das Vorgehen skizziert werden, das in der DESI-Studie zur Anwendung kam (vgl. Hartig 2007): Jedes Item der in der DESI-Studie verwendeten Sprachtests wurde hinsichtlich mehrerer Aufgabenmerkmale eingestuft, die sich auf die Itemschwierigkeit auswirken sollten. Die empirischen Itemschwierigkeiten β_i aller Items wurden auf Basis des Rasch-Modells (1PL) geschätzt. In einer anschließenden linearen Regressionsanalyse (s. z. B. Moosbrugger 2011) wurden die $k = 1, \dots, K$ Aufgabenmerkmale von Item i als Prädiktoren y_{ik} und die Itemschwierigkeiten β_i als Kriterium verwendet. Bei diesem Verfahren werden die Schwierigkeiten β_i als gewichtete Linearkombination von K Aufgabenmerkmalen modelliert:

$$\beta_i = \alpha_0 + \sum_{k=1}^K \alpha_k y_{ik} + \varepsilon_i \quad (17.1)$$

Hierbei ist α_k der Effekt von Merkmal k auf die Schwierigkeiten β_i , y_{ik} die Ausprägung von Merkmal k für Item i , und ε_i das Regressionsresiduum. Die Regressionskonstante α_0 entspricht der erwarteten Schwierigkeit einer Aufgabe, die in allen Aufgabenmerkmalen die Ausprägung $y_{ik} = 0$ aufweist. Diese Analysen liefern zunächst Aufschluss darüber, welche Aufgabenmerkmale tatsächlich in Zusammenhang mit den Itemschwierigkeiten stehen, sodass auf dieser Basis eine Auswahl von relevanten Merkmalen getroffen werden kann.

Auf Basis der Ergebnisse der Regressionsanalyse ist es möglich, die *erwarteten Schwierigkeiten* $\hat{\beta}_i$ für bestimmte Konfigurationen von Aufgabenmerkmalen zu ermitteln:

$$\hat{\beta}_i = \alpha_0 + \sum_{k=1}^K \alpha_k y_{ik} \quad (17.2)$$

Diese erwarteten Schwierigkeiten können auf derselben gemeinsamen Skala wie Personenfähigkeiten η_v und Itemschwierigkeiten β_i verortet werden. Die erwarteten Schwierigkeiten können daher verwendet werden, um Schwellen zwischen Kompetenzniveaus zu definieren (► Beispiel 17.3).

Verwendung von erwarteten Aufgabenschwierigkeiten

Beispiel 17.3: Festlegung der Aufgabenmerkmale für „Englisch Hörverstehen“

Aufgabenmerkmale für den Test in Englisch Hörverstehen (Nold und Rossa 2007) waren beispielsweise die Sprechgeschwindigkeit sowie die Komplexität der aus dem gehörten Text zu erschließenden Information. Das unterste Kompetenzniveau A in Hörverstehen, das mit Bezug auf diese Aufgabenmerkmale definiert wurde, ist u. a. dadurch charakterisiert, dass Schülerinnen und Schüler „konkrete Einzelinformationen [...] hörend verstehen [können], wenn diese Informationen langsam, deutlich gesprochen und in einfacher Sprache explizit präsentiert werden“ (Nold und Rossa 2007, S. 191). Die Kompetenzen von Schülerinnen und Schülern auf dem höchsten Niveau C hingegen werden wie folgt beschrieben: „Kann abstrakte Informationen [...] verstehen, indem implizite Informationen erschlossen oder inhaltlich komplexe Einzelinformationen interpretiert werden, auch wenn diese sprachlich komplex und in partiell schneller Sprechgeschwindigkeit präsentiert werden, wie Muttersprachler dies in natürlicher Interaktion tun“ (Nold und Rossa 2007, S. 192). Diese inhaltlichen Beschreibungen der Kompetenzniveaus erfolgen also mit Bezug auf die zur Beschreibung der Items verwendeten Aufgabenmerkmale. Dieser Bezug wird nicht durch eine Post-hoc-Analyse der Inhalte im Nachhinein hergestellt, sondern a priori über den empirischen Zusammenhang der Aufgabenmerkmale mit den Aufgabenschwierigkeiten.

Alternativ können Aufgabenmerkmale auch direkt im IRT-Modell berücksichtigt werden, wobei hierfür insbesondere das *linear-logistische Testmodell* (LLTM; Fischer 1973, 1995; ► Abschn. 16.6.3) zu nennen wäre. Modelle, die den Einbezug von Prädiktoren für die Itemschwierigkeiten erlauben, werden auch als *erklärende Item-Response-Modelle* (Explanatory Item Response Models; Wilson und De Boeck 2004) bezeichnet. Zusätzlich zum Nutzen für die Testwertbeschreibung kann die Erklärung von IRT-basierten Itemschwierigkeiten durch Aufgabenmerkmale auch als eine Prüfung der Konstruktvalidität betrachtet werden (Embretson 1983, 1998; vgl. ► Kap. 21). Die Definition von Aufgabenmerkmalen setzt nämlich gerichtete Annahmen darüber voraus, welche Aufgaben höhere oder niedrigere Anforderungen an die zu messende Personenfähigkeit η_v stellen. Können die Itemschwierigkeiten β_i durch relevante Aufgabenmerkmale erklärt werden, kann dies als ein Hinweis darauf betrachtet werden, dass tatsächlich die interessierende Fähigkeit erfasst wurde.

17.5 Zusammenfassung

Im vorliegenden Kapitel stand die Anwendung von IRT-Modellen im Rahmen der empirischen Bildungsforschung im Fokus. Bei großen Schulleistungsstudien werden spezifische Vorteile der IRT genutzt. Das Matrix-Sampling von Testaufgaben ermöglicht es, jeden Schüler nur eine Stichprobe aus einer Gesamtheit homogener Testaufgaben bearbeiten zu lassen. Die IRT wird auch genutzt, um parallele Testformen zu erstellen, indem Items eines IRT-skalierten Tests auf mehrere Testformen aufgeteilt werden. Ankeritems dienen dazu, die Items der Testformen auf einer Skala mit einer gemeinsamen Metrik zu verankern. Testwerte, die für Personen nach Bearbeitung unterschiedlicher Testformen geschätzt werden, können so miteinander verglichen werden. Computerisierte adaptive Tests legen einer Person aus einer großen Anzahl kalibrierter Items im Verlauf des Testens immer dasjenige Item vor, das für die jeweilige Schätzung der Personenfähigkeit die höchste Iteminformation aufweist. Auf diese Weise können die Messgenauigkeit maximiert und der Zeitaufwand minimiert werden.

Ein wesentlicher Vorteil von IRT-Modellen ist die Möglichkeit der kriteriumsorientierten Interpretation IRT-basierter Testwerte. Diese wird durch die gemeinsame Verortung von Itemschwierigkeiten und Personenfähigkeiten auf einer Joint Scale durchführbar. Dadurch ist es möglich, individuelle Testwerte durch ihre Abstände zu Itemschwierigkeiten zu interpretieren. Eine eindeutige relative Lokalisation von Personenfähigkeit und Itemschwierigkeit ist allerdings nur im Rasch-Modell möglich. Auf dieser zentralen Eigenschaft von Rasch-Modellen bauen auch sog. „Kompetenzniveaus“ auf. Zur leichteren Interpretation wird die kontinuierliche Skala in Abschnitte (Kompetenzniveaus) unterteilt, die dann als Ganzes kriteriumsorientiert beschrieben werden. Es wurden zwei Vorgehensweisen zur Erstellung von Kompetenzniveaus beispielhaft anhand von Daten aus der TIMSS und DESI-Studie vorgestellt: Post-hoc-Analysen der Items und Verwendung von A-priori-Aufgabenmerkmalen. Bei Post-hoc-Analysen der Items werden durch Experten Ankerpunkte auf der Kompetenzskala gesetzt und diejenigen Items identifiziert und inhaltlich betrachtet, die zur Beschreibung dieser Ankerpunkte geeignet sind. Verfahren zur Verwendung von A-priori-Aufgabenmerkmalen setzen bereits vor der Testanwendung an, indem Annahmen über Aufgabenmerkmale, die sich auf die Schwierigkeiten der Items auswirken, formuliert werden. Derartige a priori begründete Aufgabenmerkmale können verwendet werden, um IRT-basierte Testwerte kriterienorientiert zu beschreiben und die Schwellen zwischen Kompetenzniveaus festzulegen.

17.6 EDV-Hinweise

Bitte beachten Sie die EDV-Hinweise in ► Kap. 15.

17.7 Kontrollfragen

❓ Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <https://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was ist der zentrale Unterschied zwischen der Klassischen Testtheorie (KTT) und der Item-Response-Theorie (IRT)?
2. Was ist die Joint Scale von Itemschwierigkeiten und Personenfähigkeiten und welchen anwendungsbezogenen Vorteil bietet sie?
3. Warum ist eine kriteriumsorientierte Interpretation von Personenfähigkeiten aus mehrparametrischen Modellen schwierig?
4. Wie unterscheiden sich Methoden zur Definition von Kompetenzniveaus (Verwendung von A-priori-Aufgabenmerkmalen vs. Post-hoc-Analysen der Items)?

Literatur

- Baumert, J., Artelt, C., Klieme, E. & Stanat, P. (2001). PISA. Programme for International Student Assessment. Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen*. Weinheim: Beltz.
- Beaton, E. & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17, 191–204.
- Beck, B. & Klieme, E. (2003). DESI – Eine Large scale-Studie zur Untersuchung des Sprachunterrichts in deutschen Schulen. *Zeitschrift für empirische Pädagogik*, 17, 380–395.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Embretson, S. E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods*, 3, 300–396.
- Embretson, S. E. (2006). The Continued Search for nonarbitrary metrics in psychology. *American Psychologist*, 61, 50–55.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). New York: Springer.
- Graf, M., Greb, K., Jeising, E. & Lipowski, F. (2011). Mathematiktest Eingangsuntersuchung. In G. Faust & F. Lipowsky (Hrsg.), *Dokumentation der Erhebungsinstrumente zur Eingangsuntersuchung im Projekt „Persönlichkeits- und Lernentwicklung von Grundschulkindern (PERLE)“* (S. 41–54). Frankfurt am Main: GFPF.
- Greb, K. (2007). *Measuring number reading skills of students entering elementary school*. Poster präsentiert auf der Summer Academy 2007on Educational Measurement. Berlin.
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Weinheim: Beltz.
- Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Berlin, Heidelberg: Springer.
- Helmke, A. & Hosenfeld, I. (2004). Vergleichsarbeiten – Standards – Kompetenzstufen: Begriffliche Klärungen und Perspektiven. In R. S. Jäger & A. Frey (Hrsg.), *Lernprozesse, Lernumgebung und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert*. Landau: Verlag Empirische Pädagogik.
- Hill, C. H., Schilling, S. G., Loewenberg Ball, D. (2004). Developing Measures of Teachers' Mathematics Knowledge for Teaching. *The Elementary School Journal*, 105, 11–30.
- Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W. & Stanat, P. (Hrsg.). (2010). *PISA 2009: Bilanz nach einem Jahrzehnt*. Münster: Waxmann.
- Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. H. Lehmann (Hrsg.), *TIMSS/III. Dritte internationale Mathematik- und*

- Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit.* Opladen: Leske + Budrich.
- Klieme, E. & Beck, B. (Hrsg.). (2007). *Sprachliche Kompetenzen – Konzepte und Messung*. DESI-Studie (Deutsch Englisch Schuelerleistungen International) Weinheim: Beltz.
- Moosbrugger, H. (2011). *Lineare Modelle. Regressions- und Varianzanalysen* (4. Aufl., unter Mitarbeit von J. Engel, S. Etzler, K. Fischer und M. Weigand). Bern: Huber.
- Nold, G. & Rossa, H. (2007). Hörverstehen. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung. DESI-Studie (Deutsch Englisch Schuelerleistungen International)* (S. 178–196). Weinheim: Beltz.
- Organisation for Economic Co-operation and Development (OECD). (2001). *Lernen für das Leben. Erste Ergebnisse der internationalen Schulleistungsstudie PISA 2000*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD). (2004a). *Lernen für die Welt von morgen. Erste Ergebnisse von PISA 2003*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD). (2004b). *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Skills from PISA 2003*. Paris: OECD.
- Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.) (2016). *PISA 2015: Eine Studie zwischen Kontinuität und Innovation*. Münster: Waxmann.
- PISA-Konsortium Deutschland (Hrsg.). (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
- Wilson, M. R. (2003). On choosing a model for measuring. *Methods of Psychological Research Online*, 8, 1–22.
- Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 43–74). New York: Springer.



Überblick über Modelle der Item-Response-Theorie (IRT)

Augustin Kelava, Stefano Noventa und Alexander Robitzsch

Inhaltsverzeichnis

- 18.1 Modelle mit eindimensionalen latenten Merkmalen – 426**
 - 18.1.1 Generalized Partial-Credit-Modell (GPCM) nach Muraki – 426
 - 18.1.2 Rating-Scale-Modell (RSM) nach Andrich – 430
 - 18.1.3 Graded-Response-Modell (GRM) nach Samejima – 432
 - 18.1.4 Sequential Models für geordnet kategoriale Variablen – 436
- 18.2 Modelle mit mehrdimensionalen latenten Merkmalen – 438**
 - 18.2.1 Einführung in die multidimensionalen Modelle der IRT – 438
 - 18.2.2 Multidimensionales Generalized Partial-Credit-Modell (mGPCM) – 440
 - 18.2.3 Multidimensionales Graded-Response-Modell (mGRM) – 441
 - 18.2.4 Multidimensionales Graded-Rating-Scale-Modell (mGRSM) – 442
 - 18.2.5 Bifaktormodelle – 442
- 18.3 Ausblick auf weitere Modelle – 443**
- 18.4 Weiterführende Literatur – 444**
- 18.5 EDV-Hinweise – 444**
- 18.6 Kontrollfragen – 445**
- Literatur – 445**

i Nachdem in ▶ Kap. 16 eine einführende Darstellung von ausgewählten Modellen der IRT für dichotome Itemantworten vorgenommen wurde, widmet sich dieses Kapitel dem Überblick über einige exemplarische Modelle (vor allem bei polytomen, d. h. mehrkategorialen Itemantworten), die häufig Anwendung finden. Ziel dieses Kapitels ist es aufzuzeigen, dass die vorgestellten Modelle eine gewisse Verwandtschaft zueinander aufweisen und dass sich durch eine bestimmte Parametrisierung, d. h. die spezifische Ausgestaltung der sog. „Category Response Functions“, Spezialfälle ergeben, die unterschiedliche Modelltypen definieren. Da in den vergangenen Jahrzehnten eine unüberschaubare Zahl von IRT-Modellen entwickelt wurde, sind die Darstellungen dieses Kapitels nicht erschöpfend, sondern bieten nur einen exemplarischen Überblick über Modelle polytomer Itemantworten (bei eindimensionalen Merkmalen) und über Modelle multidimensionaler Merkmale. Die konkrete Schätzung der Modelle wird hier nicht beschrieben. Dazu sei auf das ▶ Kap. 19 verwiesen.

18.1 Modelle mit eindimensionalen latenten Merkmalen

18.1.1 Generalized Partial-Credit-Modell (GPCM) nach Muraki

Ein bekanntes Modell für polytome Antworten ist das GPCM nach Muraki (1992). Dieses wurde auf Grundlage des Partial-Credit-Modells (PCM) von Masters (1982) entwickelt. In diesem Abschnitt wird das GPCM beschrieben und danach der Spezialfall des (älteren) PCM betrachtet.

■■ Modellgleichungen des GPCM

Das GPCM nach Muraki (1992) beschreibt Items $i = 1, \dots, m$ mit mehreren Kategorien $k = 0, 1, \dots, K_i$, wobei die Kategorienanzahl K_i von Item zu Item verschieden sein kann. Für die Wahrscheinlichkeit, eine der Kategorien zu wählen, definieren wir als Schreibweise Folgendes:

$$P_{i,k}(\eta) := P(Y_i = k | \eta) \quad (18.1)$$

Modellierung der Wahl einer Kategorie k , die über Kategorie $k-1$ hinausgeht

Ausgangspunkt des GPCM ist die Modellierung der Wahl einer Kategorie k , die über die vorhergehende/benachbarte Kategorie $k-1$ hinausgeht. Dabei wird die Wahrscheinlichkeit $C_{i,k}$, die über $k-1$ hinausgehende Kategorie k zu erreichen, durch eine logistische Antwortfunktion definiert, die diese dichotome Wahl beschreibt (vgl. hierzu die Darstellungen aus Muraki und Muraki 2017):

$$C_{i,k} = P_{i,k|k-1,k}(\eta) = \frac{P_{i,k}(\eta)}{P_{i,k-1}(\eta) + P_{i,k}(\eta)} = \frac{\exp[Z_{i,k}(\eta)]}{1 + \exp[Z_{i,k}(\eta)]} \quad (18.2)$$

Unbedingte Wahrscheinlichkeit für Kategorie k

$P_{i,k|k-1,k}(\eta)$ ist somit die (bedingte) Wahrscheinlichkeit, Kategorie k zu wählen, wenn man zwischen k und $k-1$ wählen kann. Die (unbedingte) Wahrscheinlichkeit, Kategorie k zu wählen, entspricht:

$$P_{i,k}(\eta) = \frac{C_{i,k}}{1 - C_{i,k}} P_{i,k-1}(\eta) = \exp(Z_{i,k}(\eta)) P_{i,k-1}(\eta) \quad (18.3)$$

Dabei ist $C_{i,k}/(1 - C_{i,k})$ das Verhältnis zweier bedingter Wahrscheinlichkeiten. Es beschreibt das Verhältnis der bedingten Wahrscheinlichkeiten, k anstelle von $k-1$ und $k-1$ anstelle von k zu wählen. $Z_{i,k}(\eta)$ ist der sog. „Logit“.

Item Category Response Function (ICRF) $P_{i,k}(\eta)$

Wird für alle $P_{i,k}(\eta)$ eine Normierung in der Weise durchgeführt, dass die Summe der Kategorienwahrscheinlichkeiten eins ist ($\sum_{k=0}^{K_i} P_{i,k}(\eta) = 1$), so lässt sich die Kategorienwahrscheinlichkeit $P_{i,k}(\eta)$, die auch „Item Category Response

18.1 · Modelle mit eindimensionalen latenten Merkmalen

Function“ (ICRF) genannt wird, im GPCM wie folgt darstellen:

$$P_{i,k}(\eta) = \frac{\exp \left[\sum_{v=0}^k Z_{i,v}(\eta) \right]}{\sum_{c=0}^{K_i} \exp \left[\sum_{v=0}^c Z_{i,v}(\eta) \right]} \quad (18.4)$$

mit

$$Z_{i,k}(\eta) = D\lambda_i(\eta - \beta_{ik}) = D\lambda_i(\eta - (\beta_i - \tau_{ik})) \quad (18.5)$$

Dabei ist $D = 1.702$, das als Konstante eingefügt wird, um eine Vergleichbarkeit des logistischen Modells mit einem Modell nach der Normal-Ogiven-Repräsentation herzustellen.¹ Ferner sind λ_i ein Trennschärfenparameter (oder Diskriminationsparameter; vgl. ► Kap. 16), β_{ik} ein Itemkategorienparameter, β_i ein Itemlokationsparameter (s. hierzu auch ► Kap. 16) und τ_{ik} ein Kategorienparameter, der die Abweichung vom Itemlokationsparameter β_i beschreibt.

Bezüglich der Identifikation des GPCM ist wie auch bei anderen Modellen der IRT eine Normierung der Parameter notwendig. Da für jedes Item i , mit β_i Kategorien, $K_i - 1$ Kategorienparameter identifiziert werden können, lässt sich entweder ein Kategorienparameter auf einen bestimmten Wert fixieren (z. B. null, $\beta_{i1} := 0$) oder die Summe der Kategorienparameter entspricht einem fixierten Wert (in der Regel null).

Parametrisierung des GPCM

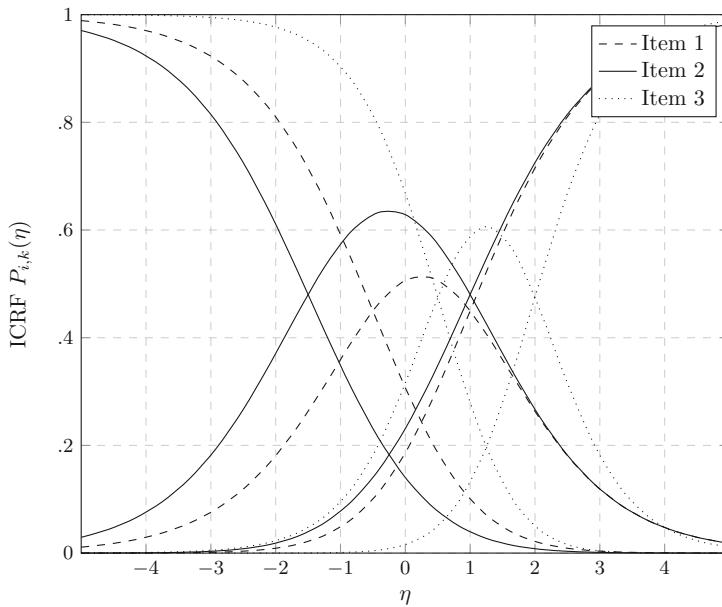
Normierung im GPCM

Fallunterscheidung

■■ Eigenschaften des GPCM

Die □ Abb. 18.1 veranschaulicht drei ICRF $P_{i,k}(\eta)$ von dreikategorialen Items.

Wie man der □ Abb. 18.1 (und den vorangehenden Gleichungen) entnehmen kann, sind die Itemkategorienparameter β_{ik} jene Stellen auf der Joint Scale, an



□ Abb. 18.1 Beispiel eines GPCM mit drei Items (mit je drei Kategorien). Abgebildet sind die insgesamt neun ICRF. Die Parameter betragen: Item 1: $\lambda_1 = 1$, $\beta_{11} = -.5$, $\beta_{12} = 1$; Item 2: $\lambda_2 = 1$, $\beta_{21} = -1.5$, $\beta_{22} = 1$; Item 3: $\lambda_3 = 1.5$, $\beta_{31} = .5$, $\beta_{32} = 2$

¹ Bei der Normal-Ogiven-Repräsentation wird die kumulative Normalverteilung statt der logistischen Funktion verwendet.

denen sich die beiden Wahrscheinlichkeiten $P_{i,k-1}(\eta)$ und $P_{i,k}(\eta)$ schneiden. Für positive λ_i (Trennschärfen) gilt daher:

$$\text{falls } \eta = \beta_{ik}, \text{ dann } P_{i,k}(\eta) = P_{i,k-1}(\eta), \quad (18.6)$$

$$\text{falls } \eta > \beta_{ik}, \text{ dann } P_{i,k}(\eta) > P_{i,k-1}(\eta), \quad (18.7)$$

$$\text{falls } \eta < \beta_{ik}, \text{ dann } P_{i,k}(\eta) < P_{i,k-1}(\eta). \quad (18.8)$$

Die β_{ik} Parameter in Abb. 18.1 betragen für Item 1: $\lambda_1 = 1$, $\beta_{11} = -.5$, $\beta_{12} = 1$. Für Item 1 schneiden sich entsprechend die jeweiligen Kategorienwahrscheinlichkeiten an den Stellen -5 (für Kategorie 1 und 2) und 1 (für Kategorie 2 und 3). Für Item 2 betragen die Parameter: $\lambda_2 = 1$, $\beta_{21} = -1.5$, $\beta_{22} = 1$. Im Fall von Item 3 betragen sie: $\lambda_3 = 1.5$, $\beta_{31} = .5$, $\beta_{32} = 2$. Wie man erkennen kann, kommt es hier zu einem größeren Intervall, in dem die Kategorienwahrscheinlichkeit für Kategorie 1 größer ist als für Kategorie 2. Wenn sich nun die Trennschärfe verändert (hier vermindert), verlaufen nicht nur die Kategorienwahrscheinlichkeiten flacher, sondern es kann sogar zu dem Phänomen kommen, dass für ein Item (hier Item 3) eine Kategorienwahrscheinlichkeit (hier Kategorie 2) stets kleiner bleibt als konkurrierende Kategorien (hier Kategorien 1 und 3).

Keine zwingende Ordnung der Parameter

Wichtig ist festzuhalten, dass die Parameter β_{ik} nicht geordnet sein müssen. Hierzu wurden keine weiteren Annahmen getroffen. Sie repräsentieren lediglich die relative Größe angrenzender Kategorien zueinander.

■ ■ PCM nach Masters

Ein Modell, das eine besondere Bedeutung hat, ist das ursprüngliche PCM nach Masters (1982). Dieses wird für geordnete kategoriale Variablen verwendet, d. h., wenn Teillösungen denkbar sind. Es stellt eine Erweiterung des Rasch-Modells, auch einparametrisches logistisches Modell (1PL-Modell) genannt, dar (vgl. hierzu ► Kap. 16). Analog zum allgemeinen Fall des GPCM wird die Wahrscheinlichkeit modelliert, Kategorie k anstelle von Kategorie $k-1$ zu erreichen. Im Unterschied zum allgemeinen GPCM wird für diese dichotome Entscheidung ein Rasch-Modell verwendet. Aus dieser Annahme resultieren die bekannten wünschenswerten Eigenschaften für Rasch und PCM, die im Zuge der Separierbarkeit der Personen- und Itemparameter den Einsatz des CML-Schätzers (Conditional Maximum Likelihood) ermöglicht (was im Falle anderer Modelle nicht möglich ist). Die Itemwerte (d. h., konkretes Antwortverhalten) sind erschöpfende (suffiziente) Statistiken für die Personenparameter.

Betrachtet man also erneut die Wahrscheinlichkeit, Kategorie k anstelle von $k-1$ zu wählen (oder in einem Leistungstest Stufe k zu erreichen), so stellt sich das Modell wie folgt dar:

$$\begin{aligned} C_{i,k} = P_{i,k|k-1,k}(\eta) &= \frac{P_{i,k}(\eta)}{P_{i,k-1}(\eta) + P_{i,k}(\eta)} = \frac{\exp(Z_{i,k}(\eta))}{1 + \exp(Z_{i,k}(\eta))} \\ &= \frac{\exp(\eta - \beta_{ik})}{1 + \exp(\eta - \beta_{ik})} \end{aligned} \quad (18.9)$$

Wahl der Kategorie k anstelle von Kategorie $k-1$

Unbedingte Kategorienwahrscheinlichkeit für Kategorie k

Erneut ist η die Fähigkeit; β_{ik} ist ein Itemparameter, der die Wahrscheinlichkeit, Kategorie k anstelle von $k-1$ zu wählen, bestimmt. Die unbedingte Kategorienwahrscheinlichkeit für Kategorie k ist damit:

$$P_{i,k}(\eta) = \frac{\exp \left[\sum_{v=0}^k (\eta - \beta_{iv}) \right]}{\sum_{c=0}^{K_i} \exp \left[\sum_{v=0}^c (\eta - \beta_{iv}) \right]} \quad (18.10)$$

18.1 · Modelle mit eindimensionalen latenten Merkmalen

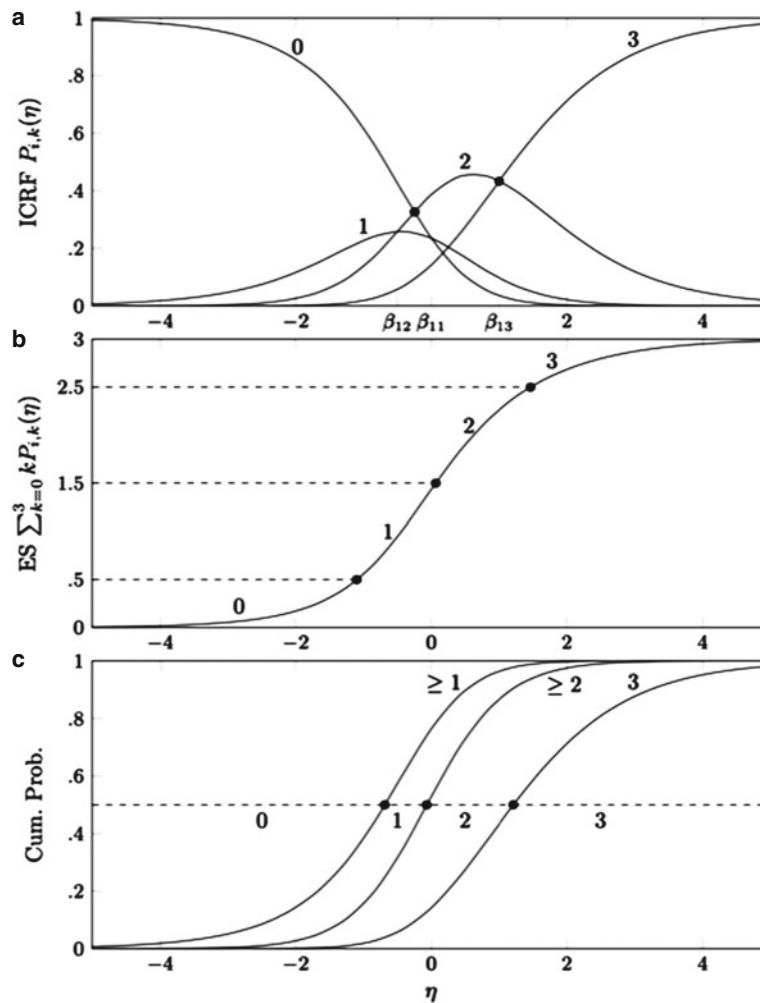


Abb. 18.2 Beispielitem für das PCM mit vier Kategorien (0, 1, 2, 3). **a** Item Category Response Function (ICRF) $P_{ik}(\eta)$. **b** Expected Score (ES), der die erwartete Antwortkategorie abbildet. **c** Kumulative Wahrscheinlichkeit (Cum. Prob.) für Kategorienwerte ≥ 1 , ≥ 2 und den Kategorienwert 3. Die Item(kategorien)parameter des abgebildeten Items lauten: $\beta_{11} = 0$, $\beta_{12} = -0.5$, $\beta_{13} = 1$

Dies lässt sich auch wie folgt formulieren:

$$P_{i,k}(\eta) = \frac{\exp(a_k \eta - \beta_{ik})}{\sum_{c=0}^{K_i} \exp(a_c \eta - \beta_{ic})} \quad (18.11)$$

Dabei entspricht der Kategorienscore $a_k = \sum_{v=1}^k 1 = k$ und $\beta_{ik} = \sum_{v=1}^k \beta_{iv}$. Aus Konventionsgründen wird ferner angenommen, dass $\eta - \beta_{i0}$ gleich null ist.

In der **Abb. 18.2** sind u. a. die Kategorienwahrscheinlichkeiten für ein PCM ersichtlich. Im Folgenden soll genauer auf **Abb. 18.2** eingegangen werden.

Wie man in **Abb. 18.2a** erkennen kann, sind die Modalwerte der ICRF aufsteigend geordnet (analog zu $1, \dots, k$). Dies ist eine grundlegende Eigenschaft des PCM. Die Item(kategorien)parameter β_{ik} haben dabei die Eigenschaft, dass sie jene Stelle anzeigen, für die eine Person identische Antwortwahrscheinlichkeiten für Kategorie k und $k-1$ aufweist. Dort schneiden sich die Kurven. Zur besseren Unterscheidung sind auf der Abszisse die β_{ik} markiert. Die dicken Punkte markieren in **Abb. 18.2a** jene Punkte, in denen sich Kategorie 0 und 2 sowie 2 und 3 schneiden. Wie man erkennen kann, ist die Wahrscheinlichkeit, eine andere Kategorie als

Aufsteigende Ordnung der Modalwerte der ICRF

Kategorie 1 zu wählen, stets größer. In □ Abb. 18.2b ist der *Expected Score* (ES), der erwartete beobachtete Wert, abgetragen. Dabei sind jene Punkte auf der Kurve markiert, die die Kategoriengrenzen markieren. Auf der Abszisse findet man die korrespondierenden Werte auf der Joint Scale. In □ Abb. 18.2c beschreibt in Abhängigkeit von der Merkmalsausprägung η die kumulativen Wahrscheinlichkeiten (Cum. Prob.) für Kategorienwerte ≥ 1 , ≥ 2 und 3. Wie man erkennen kann, ist die Wahlwahrscheinlichkeit für Kategorie 1 sehr gering. Dies äußert sich durch die Nähe zur Kurve für ≥ 2 .

Interpretation der Itemparameter

Die Item(kategorien)parameter nehmen hinsichtlich ihrer Interpretierbarkeit die Bedeutung und Skala der ursprünglichen beobachteten Variable (des Items) an. Die Personenparameter lassen sich daher zu ihnen in Beziehung setzen. Allerdings können – wie auch im GPCM – die Item(kategorien)parameter eine beliebige Reihenfolge einnehmen. Ist das der Fall, dann gibt es Kategorien, deren Wahrscheinlichkeit stets unter mindestens einer anderen Kategorienwahrscheinlichkeit liegt (vgl. □ Abb. 18.2, Kategorie 1). Ferner wird im Unterschied zum folgenden Rating-Scale-Modell (RSM) keine weitere Annahme in Form einer Restriktion bzw. der Distanz der Kategorien gemacht. Die Abstände können also variieren. Die Items eines Tests haben in der Regel daher nicht zwingend die gleichen Kategorienabstände. Die Abstände zwischen zwei Kategorien (z. B. Kategorie 1 und 2) über verschiedene Items hinweg sind in der Regel ungleich.

18.1.2 Rating-Scale-Modell (RSM) nach Andrich

Äquidistante Kategorien

Ein empirisch selten passendes, aber konzeptuell interessantes Modell, das einen Spezialfall des PCM darstellt, ist das Rating-Scale Modell (RSM) nach Andrich (1978, 2005; Rasch 1961). Das RSM ist ein Mitglied der Rasch-Familie und vor allem dann von Interesse, wenn Itemantworten in gleichabständig geordneten Kategorien denkbar sind. Dies gilt für typische Ratingskalenformate (z. B. Likert-Skalen, ▶ Kap. 5), für die dieses Modell durch zusätzliche Annahmen gegenüber dem PCM zu einem konzeptuell interessanten Modell wird.

Invarianz der Kategorienparameter

Werden nämlich im Fall von k Kategorien den jeweiligen Kategorien Scorewerte beginnend mit 0 und endend mit $k - 1$ zugewiesen und ist das Modell gültig, so sind die Summenwerte jeder Person für die Personenparameterschätzung suffizient. Das Modell erlaubt damit eine Überprüfung, ob die Kategorien bzw. Itemparameter die sinnhafte Reihenfolge einhalten (z. B. dass Kategorien auch wirklich geordnet sind). Im RSM geht man von der Annahme aus, dass die Kategorienparameter über Items hinweg invariant sind. Die Beurteilungsskala der Ratings wird letztlich über die Items hinweg als gleich angenommen, was für andere Modelle (ohne diese Restriktion) üblicherweise nicht gilt.

■■ Modellgleichung des RSM

Kategorienwahrscheinlichkeit $P_{i,k}(\eta)$

Im RSM werden für jedes Item Kategorienwerte angenommen mit $k \in \{0, 1, 2, \dots, K_i\}$. Die Modellgleichung sieht dann wie folgt aus²:

$$P_{i,k}(\eta) = \frac{\exp[a_k(\eta - \beta_i) + \kappa_k]}{1 + \sum_{c=1}^{K_i} \exp[a_c(\eta - \beta_i) + \kappa_c]} \quad (18.12)$$

Dabei sind a_k sog. „Kategorienscores“ (a_1, \dots, a_{K_i} und $a_0 = 0$). Diese folgen in der bekanntesten (und hier behandelten) Form der Score-Funktion $a_k = k \cdot 1$

² Das RSM wurde verschiedentlich definiert (vgl. Embretson und Reise 2013a). Obgleich eine erste Variante auf Rasch selbst zurückgeht, folgen wir an dieser Stelle jüngeren Darstellungen, wie sie sich z. B. bei Andrich (2005) finden lassen.

18.1 · Modelle mit eindimensionalen latenten Merkmalen

(sprich: $1, 2, 3, \dots, k, \dots, K_i$). Es werden damit äquidistante Antwortkategorien definiert, sodass daraus suffiziente Statistiken resultieren³.

Den Darstellungen von Andersen (1977) folgend lässt sich das Modell vereinfachen:

$$P_{i,k}(\eta) = \frac{\exp(a_k\eta - \beta'_{ik})}{1 + \sum_{c=1}^{K_i} \exp(a_c\eta - \beta'_{ic})} \quad (18.13)$$

Dabei ist $\beta'_{ik} = a_k\beta_i - \kappa_k$. Aus dieser Darstellung lässt sich erkennen, dass das RSM und das PCM zunächst äquivalent sind (vgl. hierzu Gl. 18.11; s. Anmerkung unten).

Wenn man alternativ den Darstellungen von Masters folgend mit $a_k = \sum_{v=1}^k 1 = k$ und $\beta'_{ik} = \sum_{v=1}^k \beta_{iv}$ eine andere Parametrisierung wählt, so erhält man:

$$P_{i,k}(\eta) = \frac{\exp\left[\sum_{v=0}^k (\eta - \beta_{iv})\right]}{1 + \sum_{c=1}^{K_i} \exp\left[\sum_{v=0}^c (\eta - \beta_{ic})\right]} \quad (18.14)$$

Eine andere Darstellungsform findet sich bei Andrich. Wenn man $\kappa_k = \sum_{v=1}^k \tau_v$ und $a_k = \sum_{v=1}^k 1 = k$ oder äquivalent $\beta'_{ik} = \sum_{v=0}^k (\beta_i - \tau_v)$ hat, so erhält man:

$$P_{i,k}(\eta) = \frac{\exp\left[\sum_{v=0}^k \eta - (\beta_i - \tau_v)\right]}{\sum_{c=0}^{K_i} \exp\left[\sum_{v=0}^c \eta - (\beta_i - \tau_v)\right]} \quad (18.15)$$

Dabei ist $\kappa_k = \sum_{v=1}^k \tau_v$, wobei $\tau_1, \tau_2, \dots, \tau_k$ Schwellenparameter sind, für die die Wahrscheinlichkeiten der Antworten auf zwei Kategorien k und $k-1$ gleich sind. Aus Gründen der Identifizierbarkeit wird angenommen, dass $\sum_{v=1}^{K_i} \tau_v = 0$ ist.

Abb. 18.3 soll die Beziehungen der Parametrisierungen veranschaulichen. Wie man erkennen kann, sind die Kategorienwahrscheinlichkeiten eines Items mit vier Kategorien (0, 1, 2, 3) abgetragen. Auf der Abszisse (unten) sind die Item(kategorien)parameter der Andersen'schen Parametrisierung abgetragen (β'_{ik}). Die Parametrisierung nach Masters ergibt sich auf der Abszisse (oben) (β_{ik}). Die Parametrisierung nach Andrich ergibt sich über den Lokationsparameter (β_i) und die Abweichungen davon (τ_k). Abb. 18.3 kann man entnehmen, dass die Parametrisierung nach Masters die Schnittpunkte angrenzender Kategorien markieren (als dicke Punkte angedeutet).

An dieser Stelle sollen noch einige Eigenschaften der Parameter und ihre Beziehungen betont werden (vgl. auch die Darstellungen bei Adams et al. 2012):

- Der Schwierigkeitsparameter in Andersens Darstellung ist die Summe der Itemparameter in dem Modell nach Masters: $\beta'_{ik} = \sum_{v=1}^k \beta_{iv}$.
- Der Schwierigkeitsparameter in Andrichs Darstellung ist der Durchschnitt der Parameter im Modell nach Masters: $\beta_i = 1/(k+1) \sum_{v=1}^k \beta_{iv}$. Ferner entspricht β_i dem Schnittpunkt der kleinsten und größten Kategorie (hier Kategorie 0 und 3).

Veranschaulichung und Parametrisierungen

Eigenschaften der Parameter

³ An dieser Stelle sei darauf hingewiesen, dass man in anderen Darstellungen (etwa bei van der Linden und Hambleton 1997) auch die Definition $a_k = K_i - k$ findet. Dies bedeutet, dass man alle Parameter entsprechend reskalieren muss.

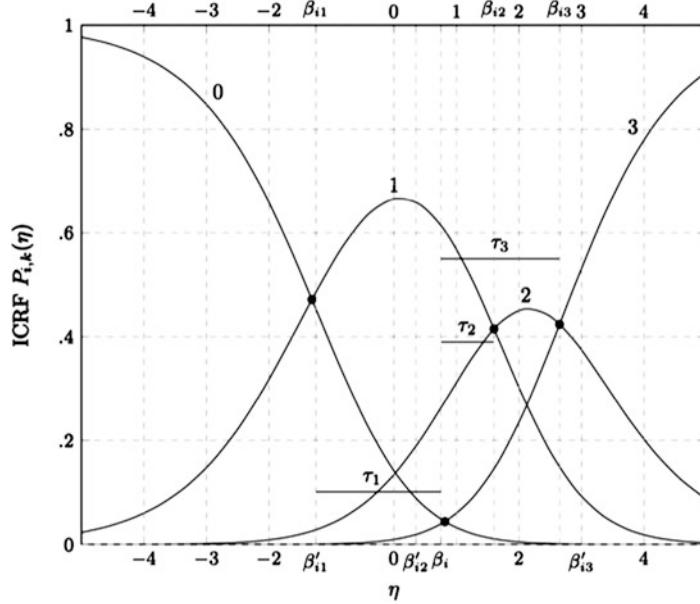


Abb. 18.3 Beispiel für ein RSM. Es werden drei (äquivalente) Parametrisierung beschrieben. Die Parametrisierung nach Andersen findet man auf der Abszisse unten: $\beta'_{i1} = -1.25$, $\beta'_{i2} = .35$, $\beta'_{i3} = 3$. Die Parametrisierung nach Masters findet man auf der Abszisse oben: $\beta_{i1} = -1.25$, $\beta_{i2} = 1.60$, $\beta_{i3} = 2.65$. Andrichs Parametrisierung unter Verwendung eines Lokationsparameters (β_i) ergibt sich als: $\beta_i = .75$, $\tau_1 = -2.25$, $\tau_2 = .60$, $\tau_3 = 1.65$ (s. horizontaler Balken)

- Die Schwellenwerte in Andrichs Modell entsprechen den Abständen zwischen Andrichs Schwierigkeitsparametern und den Parametern nach Masters: $\tau_k = \beta_{ik} - \beta_i$.

Unterschied PCM und RSM

Anmerkung: Abschließend sei noch angemerkt, dass sich dann ein Unterschied zwischen dem PCM und RSM ergibt, wenn mehrere Items betrachtet werden. Während im PCM die Itemparameter und insbesondere die Abstände der Antwortkategorien über die Items variieren können (d. h., diese separat geschätzt werden), wird im RSM davon ausgegangen, dass die Schwellenparameter für alle Items gleich sind; nur der Schwierigkeitsparameter (β_i) unterscheidet sich.

18.1.3 Graded-Response-Modell (GRM) nach Samejima

Das GRM nach Samejima (1995) beschreibt eine ganze Modellfamilie für Items mit polytomous, geordneten Itemantworten. Konkret wird davon ausgegangen, dass eine Itemantwort die Werte $Y_i = 0, 1, \dots, K_i$ annehmen kann. Die Kategoriengewahrscheinlichkeitsfunktion (Score Category Response Function, SCRF) wird ganz allgemein bezeichnet als:

$$P_{i,k}(\eta) := P(Y_i = y_i | \eta) \quad (18.16)$$

Score Category Response Function (SCRF)

Antwortmusterwahrscheinlichkeit Bei m Items ergibt sich unter der Annahme der lokalen stochastischen Unabhängigkeit ein Antwortmuster $y = (y_1, y_2, \dots, y_m)'$, sodass die Antwortmusterwahrscheinlichkeit für eine Person mit Fähigkeit η resultiert:

$$P_v(\eta) = P(Y = y | \eta) = \prod_{i=1}^m P_{i,k}(\eta) \quad (18.17)$$

■■ Grundgedanke des Ansatzes

Man stelle sich nun vor, im Rahmen eines (*kognitiven*) Prozesses werde eine Aufgabe gelöst. Dazu könne eine endliche Zahl von Stufen genommen werden. Im Rahmen dieses Prozesses wird einer Person ein Wert $y_i = k$ zugewiesen, wenn sie an der $k + 1$ -ten Stufe scheitert, aber k Stufen gemeistert hat.

Daher wird eine sog. „Processing Function“ $M_{i,k}(\eta)$ angenommen. Diese Funktion beschreibt die Wahrscheinlichkeit, Stufe k bei gegebener Fähigkeit η und bei bereits gemeisterten $k - 1$ Stufen zu erreichen. Sie ist also eine bedingte Wahrscheinlichkeit, wobei auf das Erreichen der vorherigen Stufe bedingt wird.

Da jede Person einen Itemwert von 0 erreichen und niemand einen Itemwert von $K_i + 1$ haben kann, wird Folgendes angenommen:

$$M_{i,k}(\eta) = \begin{cases} 1 & \text{wenn } k = 0 \\ 0 & \text{wenn } k = K_i + 1 \end{cases}$$

Dabei gilt dies für jede Merkmalsausprägung η . Außerdem wird angenommen, dass $P_{y_i}(\eta)$ monoton wachsend ist. Damit wird sichergestellt, dass das „Mehr-Lösen“ jedes Items mit einer steigenden Fähigkeit einhergeht.

Die Funktion der Kategorienantwortwahrscheinlichkeit (*Category Response Function*, CRF) eines geordnet kategorialen Itemwertes k ergibt sich als:

$$P_{i,k}(\eta) = \left(\prod_{j \leq k} M_{i,j}(\eta) \right) (1 - M_{i,k+1}(\eta)) \quad (18.18)$$

Wenn man Gl. (18.18) etwas umformt, erhält man alternativ:

$$P_{i,k}^*(\eta) = \prod_{j \leq k} M_{i,j}(\eta) - \prod_{j \leq k+1} M_{i,j}(\eta)$$

Und wenn man die beiden Produkte der Differenz definiert, ergibt sich folgende Funktion:

$$P_{i,k}(\eta) = P_{i,k}^*(\eta) - P_{i,k+1}^*(\eta) \quad (18.19)$$

Die Funktion $P_{i,k}^*(\eta)$ ist die sog. „Cumulative Score Category Response Function“ (CSCRF; Samejima 1995). Sie beschreibt die Wahrscheinlichkeit, dass jemand Stufe k und darüber hinaus erreicht. Daher ist es naheliegend, von zwei benachbarten Stufen die (CSCRF-)Differenz zu bilden, um die Kategorienwahrscheinlichkeit wie in Gl. (18.19) zu erhalten.

Das bisher beschriebene GRM ist allgemein gehalten. Es subsumiert eine ganze Reihe von Modellen. Um deren Zweckmäßigkeit mit Blick auf spezifische psychologische Fragestellungen zu beurteilen, ist es nötig, potenzielle Modelle auf wünschenswerte Eigenschaften hin zu prüfen. Dazu zählen folgende Eigenschaften (vgl. Samejima 2017):

1. Überprüfung der Plausibilität der gemachten Annahmen
2. „Unique Maximum Condition“, die für die spätere Schätzung sicherstellt, dass für jedes Antwortmuster die Likelihood-Funktion ein eindeutiges Maximum hat
3. Übereinstimmung der geordneten Modalwerte der CRF mit der Reihung der Itemwerte
4. Additivität der SCRF
5. Generalisierbarkeit des Modells zum sog. „Continuous-Response-Modell“

Kognitiver Prozess

Processing Function $M_{i,k}(\eta)$

Kategorienantwortwahrscheinlichkeit $P_{i,k}(\eta)$

Cumulative Score Category Response Function (CSCRF) $P_{i,k}^*(\eta)$

Zweckmäßigkeit des Modells

Logistisches Modell und Normal-Ogiven Modell

■■ Der homogene Fall

Der sog. „homogene Fall“ des GRM steht für eine besondere Unterklasse von Modellen. Diese Modelle weisen Kategorienwahrscheinlichkeiten (CSCRF) $P_{i,k}^*(\eta)$ auf, die hinsichtlich der Form und der Position auf der Fähigkeitsskala korrespondierend mit den Itemwerten $0, 1, \dots, k, \dots, K_i$ angeordnet sind. Als wünschenswerte Eigenschaft ist die eben erwähnte Additivität erfüllt. Ferner sind die Modalwerte der Kategorienwahrscheinlichkeiten (CSCRF) streng geordnet.

Samejima (1969a, 1969b) zufolge fallen das logistische Modell und das Normal-Ogiven Modell in diese Familie von homogenen Funktionen. Die entsprechenden CRF ergeben sich als:

$$P_{i,k}(\eta) = \frac{\exp[-\lambda_i(\eta - \beta_{i(k+1)})] - \exp[-\lambda_i(\eta - \beta_{ik})]}{[1 + \exp[-\lambda_i(\eta - \beta_{ik})][1 + \exp[-\lambda_i(\eta - \beta_{i(k+1)})]]]} \quad (18.20)$$

bzw.

$$P_{i,k}(\eta) = \frac{1}{(2\pi)^{1/2}} \int_{\lambda_i(\eta - \beta_{i(k+1)})}^{\lambda_i(\eta - \beta_{ik})} \exp\left(\frac{-t^2}{2}\right) dt \quad (18.21)$$

Eine alternative Darstellung des 2PL-Modells lautet:

$$P_{i,k}(\eta) = \frac{\exp[\lambda_i(\eta - \beta_{ik})]}{1 + \exp[\lambda_i(\eta - \beta_{ik})]} - \frac{\exp[\lambda_i(\eta - \beta_{i(k+1)})]}{1 + \exp[\lambda_i(\eta - \beta_{i(k+1)})]} \quad (18.22)$$

Für das zweiparametrische Normal-Ogiven Modell gilt:

$$\begin{aligned} P_{i,k}(\eta) &= \frac{1}{\sqrt{2\pi}} \int_{\lambda_i(\eta - \beta_{i(k+1)})}^{\lambda_i(\eta - \beta_{ik})} \exp\left[\frac{-t^2}{2}\right] dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda_i(\eta - \beta_{ik})} \exp\left[\frac{-t^2}{2}\right] dt - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda_i(\eta - \beta_{i(k+1)})} \exp\left[\frac{-t^2}{2}\right] dt \\ &= \Phi(\lambda_i(\eta - \beta_{ik})) - \Phi(\lambda_i(\eta - \beta_{i(k+1)})) \end{aligned} \quad (18.23)$$

Aufsteigende Modalwerte der Kategorienwahrscheinlichkeiten

Nominal-Response-Modell (NRM)

Die Abb. 18.4 veranschaulicht den logistischen Fall einer SCRF. Wie man Abb. 18.4 entnehmen kann, sind wie oben erwähnt die Modalwerte der Kategorienwahrscheinlichkeiten aufsteigend geordnet.

■■ Der heterogene Fall

Der heterogene Fall deckt jene GRM ab, für die die CSCRF $P_{i,k}^*(\eta)$ hinsichtlich ihrer Form nicht einheitlich sind. Unter bestimmten Bedingungen gehört z. B. das sog. „Nominal-Response-Modell“ (NRM) von Bock (1972) dazu:

$$P_{i,k}(\eta) = \frac{\exp(\lambda_{ik}\eta + d_{ik})}{\sum_{j \in R_i} \exp(\lambda_j\eta + d_j)} \quad (18.24)$$

Dabei entstammt k nun nicht aus einer geordneten Menge von ganzen Zahlen, sondern steht nur symbolisch für diskrete nominale Antworten auf ein Item. R_i beschreibt die Menge der nominalen Antworten eines Items i und $\lambda_j > 0$. Die Bedingungen, unter denen das NRM zu einem Untermodell des GRM wird, sind dann gegeben, wenn sich die Itemantworten tatsächlich dem Rang nach anordnen

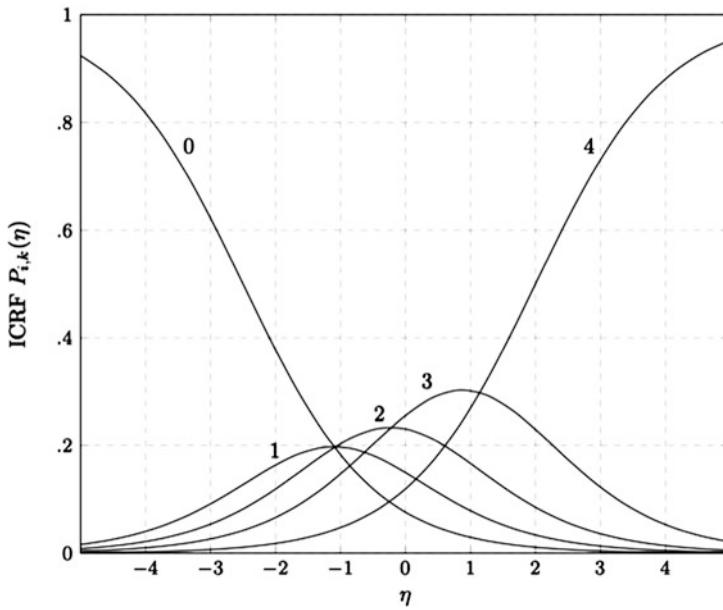


Abb. 18.4 Beispiel für ein GRM für ein Item mit fünf Kategorien. Dargestellt sind die Verläufe der Antwortkategorienwahrscheinlichkeiten (Werte der ICRF). Die Itemparameter betragen: $D_{\lambda_i} = 1$, $\beta_{i,0} = -2.5$, $\beta_{i,1} = -1.5$, $\beta_{i,2} = -0.7$, $\beta_{i,3} = 1.5$, $\beta_{i,4} = 2$

lassen und wenn sich die Diskriminationsparameter $\lambda_{i0} \leq \lambda_{i1} \leq \lambda_{i2} \dots \leq \lambda_{iM_i}$ anordnen lassen. Weitere Beispiele sind das PCM von Masters (1982) sowie das GPCM von Muraki (1992).

Ein weiteres Modell ist das sog. „Acceleration Model“ von Samejima (1995). In diesem Modell wird angenommen, dass sich die zur Bewältigung einer Aufgabe nötigen Subprozesse sequentiell entfalten, bevor das Item gänzlich beantwortet wurde. Die oben erwähnte Processing Function für jede Itemantwort $1, 2, k, \dots, m_i$ ist dann gegeben als:

$$M_{i,k}(\eta) = [\Psi_{i,k}(\eta)]^{\xi_{ik}}, \quad (18.25)$$

wobei $\xi_{ik} > 0$ der sog. „Step Acceleration Parameter“ ist. $\Psi_{i,k}(\eta)$ entstammt einer Menge von Funktionen, die streng monoton steigend, fünf Mal differenzierbar in η und mit den Asymptoten 0 und 1 versehen sind. Eine mögliche Wahl für $\Psi_{i,k}(\eta)$ ist:

$$\Psi_{i,k}(\eta) = \frac{1}{1 + \exp(-\lambda_{ik}(\eta - \beta_{ik}))} \quad (18.26)$$

mit dem Diskriminationsparameter $\lambda_{ik} > 0$ und dem Itemparameter β_{ik} . Die CSCRF $P_{i,k}^*(\eta)$ ergibt sich schließlich als

$$P_{i,k}^*(\eta) = \prod_j^k [\Psi_{i,j}(\eta)]^{\xi_{ij}} \quad (18.27)$$

und die SCRF $P_{i,k}(\eta)$ als

$$P_{i,k}(\eta) = \prod_j^k [\Psi_{i,j}(\eta)]^{\xi_{ij}} \left[1 - \prod_j^k [\Psi_{i,k+1}(\eta)]^{\xi_{i(k+1)}} \right]. \quad (18.28)$$

Acceleration Model

Step Acceleration Parameter $\Psi_{i,k}(\eta)$

Kumulative Kategorienantwort-wahrscheinlichkeit

Kategorienantwortwahrscheinlichkeit

Abschließende Anmerkungen: Vergleicht man den homogenen und den heterogenen Fall, so lässt sich festhalten, dass der homogene Fall jene genannten wünschenswerten Eigenschaften erfüllt, während der heterogene Fall sehr viel mehr Freiheiten bei der Modellierung der Kategorienwahrscheinlichkeiten zulässt. Daher ist, auch wenn in diesem Kapitel nicht auf den Modelfit eingegangen werden soll, die Modellpassung im heterogenen Fall besser. Dennoch kommt der homogene Fall bei der Einstellungsmessung (z. B. bei Likert-Ratingskalen) zum Einsatz (vgl. auch ► Kap. 5 und 16).

18.1.4 Sequential Models für geordnet kategoriale Variablen

Abhängigkeitsstrukturen innerhalb der Items und kognitive Prozesse

Während man in zahlreichen IRT-Modellen kategoriale Antworten als „gelöst“ und „nicht gelöst“ zu erklären versucht, nutzt man im sog. „Sequential Model“ konkret die Informationen aus gestuften Antworten. Insbesondere dann, wenn innerhalb der Items gewisse Schritte zur Lösung notwendig sind, um ein Item in Gänze zu lösen, sprich direkte Abhängigkeitsstrukturen innerhalb der Items angenommen werden können, eignen sich Sequential Models für die Beschreibung dieser gestuften (kognitiven) Prozesse. Wenn also eine schrittweise Lösung plausibel ist, dann erscheint die Anwendung eines Sequential Models plausibel.

Erste Überlegungen zu dieser Modellklasse gehen etwa auf Wright und Masters (1982) zurück. In ihren Überlegungen verwenden sie beispielsweise das Item $\sqrt{9/0.3 - 5} = ?$. Dann lassen sich die Stufen wie folgt unterscheiden: Auf einer untersten Stufe (Stufe 0) wird kein Teilproblem gelöst. Auf Stufe 1 wird $9 / 0.3 = 30$ als Teilproblem gelöst. Auf Stufe 2 wird $30 - 5 = 25$ gelöst. Auf Stufe 3 wird $\sqrt{25} = 5$ gelöst. Wie man der Darstellung entnehmen kann, ist das Item in Gänze nur dann lösbar, wenn die Teilaufgaben schrittweise gelöst wurden.

Weitere derartige Überlegungen zur schrittweisen Modellierung wurden auch von Tutz (1990, 2011) sowie Verhelst et al. (1997) angestellt.

■ ■ Schritte bzw. Sequenz

Für ein Item i sei die Antwortvariable $Y_i \in 0, \dots, K_i$ derartig abgestuft, dass die Antwortkategorien $0, \dots, K_i$ für ansteigende Leistungen stehen.

Die grundlegende Annahme beinhaltet, dass das Antwortverhalten Y_{ik} den Schritt von einer Stufe $k - 1$ zur nächsten Stufe k bezeichnet. Dabei beschreibt $Y_{ik} = 1$ eine erfolgreiche Bewältigung des Schrittes und $Y_{ik} = 0$ eine erfolglose Bewältigung des Schrittes.

In *Schritt 1* wird prozessual von Stufe 0 ausgegangen. Scheitert die betreffende Person und erreicht Stufe 1 nicht, dann sind $Y_{i1} = 0$ und $Y_i = 0$. Der Prozess stoppt dann. Scheitert sie nicht, weil $Y_{i1} = 1$, dann ist auch $Y_i \geq 1$.

$$Y_i = 0, \quad \text{falls } Y_{i1} = 0 \tag{18.29}$$

Variable Y_{ik}

Bedingter Prozess, Stoppbedingung

In *Schritt 2* wird prozessual von der erfolgreichen Bewältigung von Stufe 1 ausgegangen. Scheitert sie bei diesem Schritt, dann sind $Y_{i2} = 0$ und $Y_i = 1$. Erfolgt eine erfolgreiche Bewältigung des zweiten Schritts, dann ist $Y_i \geq 2$.

$$\{Y_i = 1 \text{ gegeben } Y_i \geq 1\}, \quad \text{falls } Y_{i2} = 0 \tag{18.30}$$

Entscheidend ist, dass der Schritt 2 nur dann von Relevanz ist, wenn Schritt 1 erfolgreich absolviert wurde. Daher gilt die Bedingung $Y_i \geq 1$.

Für den $(k + 1)$ -ten Schritt wird angenommen, dass alle vorhergehenden Schritte erfolgreich waren ($Y_{i1} = Y_{i2} = \dots = Y_{ik} = 1$), nämlich $Y_i \geq k$. Die abschließende Leistung ist analog zu vorher k , wenn der Schritt nach $k + 1$ nicht erfolgreich gemacht wird:

$$\{Y_i = k \text{ gegeben } Y_i \geq k\}, \quad \text{falls } Y_{i(k+1)} = 0 \tag{18.31}$$

■■ Modellierung

Wie man bisher gesehen hat, sind die Schritte binär. Man kann entweder erfolgreich die nächste Stufe nehmen oder nicht. Diese Dichotomie wird bei der Modellierung genutzt, indem man den bedingten Übergang von Stufe k nach $k + 1$ mit einem binären Wahrscheinlichkeitsmodell beschreibt:

$$P(Y_{i(k+1)} = 1|\eta) = F(\eta - \beta_{i(k+1)}), \quad \text{mit } k = 0, \dots, K_i - 1 \quad (18.32)$$

Dabei ist $\beta_{i(k+1)}$ eine Itemschwierigkeit für den Schritt $k + 1$. F ist eine Wahrscheinlichkeitsfunktion. Zum Beispiel kann dies die logistische Funktion sein, so dass ein dichotomes Rasch-Modell zur Anwendung kommt, das dann *sequentielles Rasch-Modell* genannt wird:

$$F(\eta - \beta_{i(k+1)}) = \frac{\exp(\eta - \beta_{i(k+1)})}{1 + \exp(\eta - \beta_{i(k+1)})} \quad (18.33)$$

Insgesamt ergeben sich ein Fähigkeitsparameter η und $K_i - 1$ Itemschwierigkeitsparameter ($\beta_{i1}, \dots, \beta_{i(K_i-1)}$) für das Item i .

Die *Kategorienantwortwahrscheinlichkeit* für Kategorie k ist dann gegeben als:

$$P(Y_i = k|\eta) = \begin{cases} 1 - F(\eta - \beta_{i(1)}) & \text{für } k = 0 \\ \left(\prod_{s=0}^{k-1} F(\eta - \beta_{i(s+1)}) \right) (1 - F(\eta - \beta_{i(k+1)})) & \text{für } k = 1, \dots, K_i - 1 \\ \prod_{s=0}^{K_i-1} F(\eta - \beta_{i(s+1)}) & \text{für } k = K_i \end{cases} \quad (18.34)$$

Das Antwortverhalten einer Person entspricht Kategorie k , wenn alle Schritte $Y_{i1} = Y_{i2} = \dots = Y_{ik} = 1$ erfolgreich waren, aber $Y_{i(k+1)} = 0$ nicht. Für Kategorie K_i muss auch der K_i -te Schritt ($Y_{i(K_i)} = 1$) erfolgreich sein.

■■ Veranschaulichung

Die Abb. 18.5 veranschaulicht die Kategorienantwortwahrscheinlichkeiten für vier Kategorien ($0, \dots, 3$), wobei zunächst als Schwierigkeitsparameter $\beta_{i1} = \beta_{i2} = \beta_{i3} = .3$ angenommen werden (vgl. Abb. 18.5a). Eine Person mit einem Fähigkeitsparameter von $\eta = .3$ hat dann für jeden Schritt (gegeben ist der vorhergehende erfolgreiche Schritt) eine Übergangswahrscheinlichkeit von .5. Als unbedingte Kategorienwahrscheinlichkeiten ergeben sich im Unterschied dazu $P(Y_i = 0|\eta = .3) = .5$ und $P(Y_i = 2|\eta = .3) = P(Y_i = 3|\eta = .3) = .125$.

Belässt man es für diese Person bei einem Fähigkeitsparameter von $\eta = .3$ und ändert die Itemparameter so ab, dass diese nun $\beta_{i1} = -1$, $\beta_{i2} = .3$ und $\beta_{i3} = 2$ betragen (vgl. Abb. 18.5b), dann ist der erste Schritt leichter zu absolvieren. Der zweite Schritt bleibt (unter der Bedingung des erfolgreichen ersten Schritts) gleich schwer mit einer Wahrscheinlichkeit von .5. Der dritte Schritt ist sehr schwer. Die unbedingten Kategorienwahrscheinlichkeiten für die Kategorien fallen nun wie folgt aus (gerundet): $P(Y_i = 0|\eta = .3) = .214$, $P(Y_i = 1|\eta = .3) = .393$, $P(Y_i = 2|\eta = .3) = .332$ und $P(Y_i = 3|\eta = .3) = .061$.

Die Interpretation der Itemparameter ist also analog zum Rasch-Modell (vgl. ► Kap. 17), nur dass sich jeder Parameter auf das Absolvieren eines Schrittes von k nach $k + 1$ in Bezug zum Fähigkeitsparameter η bezieht.

Binäre Modellierung

Sequentielles Rasch-Modell

Anzahl der Parameter

Kategorienantwortwahrscheinlichkeit

Interpretation von Abb. 18.5a mit $\beta_{i1} = \beta_{i2} = \beta_{i3} = .3$

Interpretation der Abb. 18.5b mit $\beta_{i1} = -1, \beta_{i2} = .3$ und $\beta_{i3} = 2$

Bezug zum Rasch-Modell

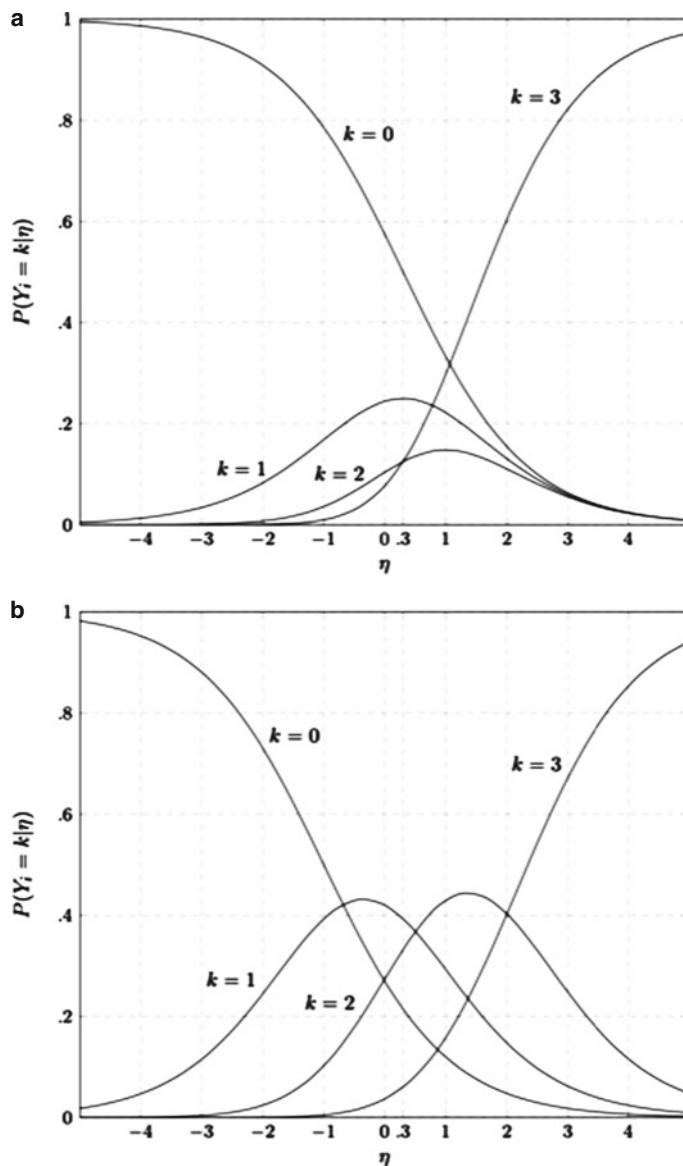


Abb. 18.5 Zwei Items, die einem *Sequential Model* folgen. **a** Das erste Item hat die Parameter $\beta_{i1} = \beta_{i2} = \beta_{i3} = .3$. **b** Das zweite Item hat die Parameter $\beta_{i1} = -1, \beta_{i2} = .3, \beta_{i3} = 2$. Erklärung siehe Text

18.2 Modelle mit mehrdimensionalen latenten Merkmalen

18.2.1 Einführung in die multidimensionalen Modelle der IRT

Grundgedanke von mIRT-Modellen

Multidimensionale IRT-Modelle (mIRT-Modelle) stellen eine bedeutsame multivariate Erweiterung der bisher dargestellten Modelle dar. Der Grundgedanke von mIRT ist, dass es mehrere Merkmale gibt, die in Zusammenhang zur gezeigten Antwort (Leistung) stehen. Während in eindimensionalen IRT-Modellen die Wahrscheinlichkeit einer Antwort anhand eines Merkmals modelliert wird, ist es das Ziel der mIRT-Modelle, den Einfluss von mindestens zwei Merkmalen auf die Antwortwahrscheinlichkeit zu beschreiben (vgl. auch Chalmers 2012; Reckase 2009).

Wir beginnen mit einem Beispiel eines zweidimensionalen 2PL-Modells, das eine dichotome Itemantwort beschreibt:

$$P_i(Y_i = 1|\eta_1, \eta_2) = \frac{\exp(\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + d_i)}{1 + \exp(\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + d_i)} \quad (18.35)$$

Wie man erkennen kann, gibt es zwei Diskriminationsparameter $\lambda_{i1}, \lambda_{i2}$ sowie zwei Merkmale η_1, η_2 . Die gewichtete Summe der Merkmale repräsentiert eine Form von heterogenem latenter Trait (z.B. lässt sich in gewissen Anwendungen diese gewichtete Summe als Zusammenspiel verschiedener Kompetenzen auffassen). Es sei an dieser Stelle angemerkt, dass d_i nicht dem Schwierigkeitsparameter β_i entspricht (► Abschn. 18.1). Für zwei latente Merkmale lässt sich ein resultierender Diskriminationsindex berechnen (sog. „Multidimensional Discrimination“, MDISC). Dieser ist dann:

$$A_i = \sqrt{\lambda_{i1}^2 + \lambda_{i2}^2} \quad (18.36)$$

Ferner lässt sich die multidimensionale Schwierigkeit (sog. „Multidimensional Difficulty“, MDIFF) wie folgt berechnen:

$$\beta_i = -\frac{d_i}{\sqrt{\lambda_{i1}^2 + \lambda_{i2}^2}} \quad (18.37)$$

Diese Kennwerte lassen sich leicht auf mehr als zwei Merkmale generalisieren (vgl. Reckase 2017).

Wie man der □ Abb. 18.6a entnehmen kann, ist die Antwortwahrscheinlichkeit nun eine Funktion zweier Merkmale (η_1 und η_2). Dabei beschreibt die Funktion nicht mehr nur eine Kurve (oder mehrere im Kontext polytomer Items), sondern eine gekrümmte Fläche. Sobald eines der Merkmale eine hohe Ausprägung aufweist, ist die Antwortwahrscheinlichkeit hoch. Dieses Verhalten wird auch *kompensatorisch* genannt.

Generalisiert man dieses Modell hinsichtlich eines Vektors von Merkmalen (bzw. Traits) $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots)'$ bei gegebenen Diskriminationsparametern $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \dots)'$, so lässt sich das Modell allgemein darstellen als:

$$P(Y_i = 1|\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\lambda}_i \cdot \boldsymbol{\eta} + d_i)}{1 + \exp(\boldsymbol{\lambda}_i \cdot \boldsymbol{\eta} + d_i)} \quad (18.38)$$

Im Falle der Aufnahme eines Rateparameters g_i und einer oberen Asymptote u_i erhält man:

$$P(Y_i = 1|\boldsymbol{\eta}) = g_i + (u_i - g_i) \frac{\exp(\boldsymbol{\lambda}_i \cdot \boldsymbol{\eta} + d_i)}{1 + \exp(\boldsymbol{\lambda}_i \cdot \boldsymbol{\eta} + d_i)} \quad (18.39)$$

Das Modell wird „kompensatorisch“ genannt, da es aufgrund des Vorhandenseins mehrerer latenter Merkmale auch dann zu einer hohen Lösungswahrscheinlichkeit führen kann, wenn eines der Merkmale niedrig ausgeprägt ist, solange andere Merkmale eine hohe Ausprägung aufweisen (vgl. auch □ Abb. 18.6).

Eine weitere Gruppe von mIRT-Modellen stellen die sog. „nicht kompensatorischen Modelle“ mit n latenten Merkmalen/Traits und mit $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \dots)$ dar:

$$P(Y_i = 1|\boldsymbol{\eta}) = g_i + (u_i - g_i) \prod_{l=1}^n \frac{\exp(\lambda_{il}\eta_l + d_{il})}{1 + \exp(\lambda_{il}\eta_l + d_{il})} \quad (18.40)$$

Wie man ihrer formalen Darstellung entnehmen kann, werden die Wahrscheinlichkeiten über ein Produkt miteinander verrechnet. Das bedeutet konkret, dass für die Bewältigung eines Items alle Merkmale η_1, \dots, η_n möglichst hoch ausgeprägt sein müssen. Wenn eines der notwendigen Merkmale niedrig ausgeprägt ist,

Zweidimensionales 2PL-Modell

Multidimensionale Diskrimination

Multidimensionale Schwierigkeit

Kompensatorisches mIRT-Modell

Nicht kompensatorisches mIRT-Modell

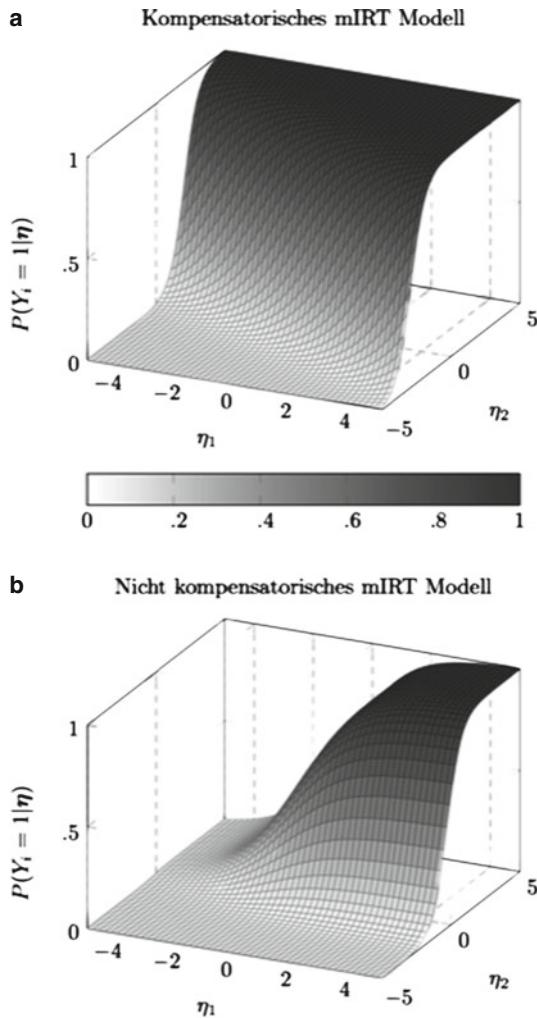


Abb. 18.6 a Beispiel für ein kompensatorisches mIRT. b Beispiel für ein nicht kompensatorisches mIRT; $\lambda_{i1} = 1, \lambda_{i2} = 2, d_{i1} = .8, d_{i2} = .6$. In beiden Fällen sind die Lösungswahrscheinlichkeiten abgebildet. Erläuterung siehe Text

mindert dies das Gesamtprodukt bzw. die Gesamtproduktwahrscheinlichkeit, so dass die Lösungswahrscheinlichkeit des Items substantiell abnimmt. Dies wurde in **Abb. 18.6b** anhand zweier Merkmale veranschaulicht. Eine Kompensation ist also nicht möglich.

18.2.2 Multidimensionales Generalized Partial-Credit-Modell (mGPCM)

Die Verallgemeinerung des GPCM (► Abschn. 18.1.1) für den zweidimensionalen Fall ist wie folgt gegeben:

$$P(Y_i = k | \eta_1, \eta_2) = \frac{\exp \left[\sum_{v=0}^k (\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + d_{iv}) \right]}{1 + \sum_{c=1}^{K_i} \exp \left[\sum_{v=0}^c (\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + d_{iv}) \right]} \quad (18.41)$$

18.2 · Modelle mit mehrdimensionalen latenten Merkmalen

Die Verallgemeinerung des zweidimensionalen 2PL-Modells in Bezug auf mehrere Merkmale, d. h. einen Vektor latenter Traits $\eta = (\eta_1, \eta_2, \dots)'$, mit $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \dots)'$, ergibt sich als (vgl. Muraki 1992):

$$P(Y_i = k|\eta) = \frac{\exp \sum_{v=0}^k (\lambda_i \cdot \eta + d_{iv})}{1 + \sum_{c=1}^{K_i} \exp \sum_{v=0}^c (\lambda_i \cdot \eta + d_{iv})} \quad (18.42)$$

Alternativ findet man auch folgende mIRT-Darstellung des mPCM:

$$P(Y_i = k|\eta) = \frac{\exp(a_{ik}\lambda_i \cdot \eta + d_{ik})}{1 + \sum_{v=1}^{K_i} \exp(a_{iv}\lambda_i \cdot \eta + d_{iv})} \quad (18.43)$$

Dabei kann $a_{ik} = \sum_{v=1}^k 1 = k$ Werte $a_{ik} \in 0, \dots, K_i$ annehmen und $d_{ik} = \sum_{v=0}^k d_{iv}$. Ebenso wird im PCM aus Normierungsgründen die Annahme gemacht, dass die Bedingungen $a_{i0} = 0$ und $d_{i0} = 0$ gelten.

Betrachtet man diese letzte Parametrisierung, so sind die a_{ik} zunächst als fixe Werte aufzufassen. Dies stellt den häufigeren Anwendungsfall dar. Dennoch lassen sich die a_{ik} auch als Parameter und nicht als fixe Werte auffassen. Die a_{ik} Parameter würden in diesem Fall geschätzt. Auf diese Weise lässt sich die empirische Ordnung der Kategorien prüfen.

18.2.3 Multidimensionales Graded-Response-Modell (mGRM)

Die Verallgemeinerung des GRM zum multidimensionalen Graded Response-Modell (mGRM) gestaltet sich wie folgt: Zunächst gilt wie auch für das GRM:

$$P(Y_i = k|\eta) = P^*(Y_i \geq k|\eta) - P^*(Y_i \geq k+1|\eta) \quad (18.44)$$

Die Kategorienantwortwahrscheinlichkeiten $P(Y_i = k|\eta)$ lassen sich erneut über die Differenzen der kumulativen Kategorienantwortwahrscheinlichkeiten $P^*(Y_i \geq k|\eta)$ (CSCRF) ermitteln (► Abschn. 18.1.3).

Für die $K_i + 1$ Kategoriegrenzen werden wie im unidimensionalen Fall für die Kategorie 0 und die größte Kategorie die Werte 1 respektive 0 angenommen. Man erhält für alle Kategorien die folgende Parametrisierung (der kumulativen Wahrscheinlichkeiten):

$$\begin{cases} P^*(Y_i \geq 0|\eta) = 1 \\ P^*(Y_i \geq 1|\eta) = \frac{\exp(\lambda_i \cdot \eta + d_{i1})}{1 + \exp(\lambda_i \cdot \eta + d_{i1})} \\ P^*(Y_i \geq 2|\eta) = \frac{\exp(\lambda_i \cdot \eta + d_{i2})}{1 + \exp(\lambda_i \cdot \eta + d_{i2})} \\ \vdots \\ P^*(Y_i \geq K_i|\eta) = 0 \end{cases} \quad (18.45)$$

Wenn man Gl. (18.45) auf Gl. (18.44) anwendet, erhält man z. B. für die Kategorie 1 die folgende Antwortwahrscheinlichkeit:

$$\begin{aligned} P(Y_i = 1|\eta) &= P^*(Y_i \geq 1|\eta) - P^*(Y_i \geq 2|\eta) \\ &= \frac{\exp(\lambda_i \cdot \eta + d_{i1})}{1 + \exp(\lambda_i \cdot \eta + d_{i1})} - \frac{\exp(\lambda_i \cdot \eta + d_{i2})}{1 + \exp(\lambda_i \cdot \eta + d_{i2})} \end{aligned} \quad (18.46)$$

Kumulative Wahrscheinlichkeiten im mGRM

Antwortwahrscheinlichkeit im mGRM

Diese ist bis auf den multidimensionalen Vektor der latenten Variablen analog zur Vorgehensweise in Gl. (18.22).

18.2.4 Multidimensionales Graded-Rating-Scale-Modell (mGRSM)

Kumulative Kategorienantwortwahrscheinlichkeit im mGRSM

Die Verallgemeinerung des RSM für den multidimensionalen Fall erhält man, indem man das vorangegangene mGRM modifiziert. Das mGRSM ist das resultierende Modell. Dazu muss man die kumulative Kategorienantwortwahrscheinlichkeit $P^*(Y_i \geq k|\eta)$ (CSCRF) für das obige mGRM wie folgt definieren:

$$P^*(Y_i \geq k|\eta) = \frac{\exp(a_k \lambda_i \cdot \eta + d_k + d_i)}{1 + \exp(a_k \lambda_i \cdot \eta + d_k + d_i)}, \quad (18.47)$$

wobei d_i wieder als Lokationsparameter dient und damit die Position des Items auf der Joint Scale bestimmt. Die Abstände zwischen den Kategorien sind wie auch im eindimensionalen RSM über die Items hinweg konstant, was durch den über Items invarianten Parameter d_k erreicht wird.

Anmerkung: Der gleiche Mechanismus wurde für das RSM und für das PCM verwendet.

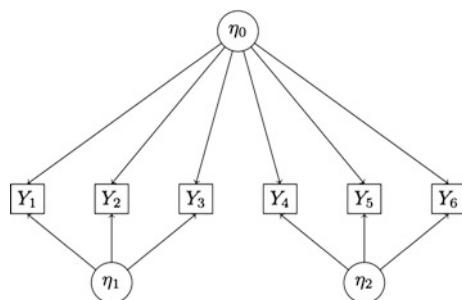
18.2.5 Bifaktormodelle

Beschreibung einer varianzstarken Domäne

Generalfaktor

Eine in den vergangenen Jahren wieder prominent gewordene Modellklasse ist die der Bifaktormodelle (Holzinger und Swineford 1937; Reise 2012, vgl. auch ▶ Kap. 15, ▶ Abschn. 15.3.1). Diese kommen dann zur Anwendung, wenn man davon ausgehen kann, dass alle Items eine übergeordnete (varianzstarke) Domäne messen und gleichzeitig für Teilmengen der Items spezifische Faktoren einen Einfluss haben. Die □ Abb. 18.7 soll beispielhaft als Pfaddiagramm veranschaulichen, wie diese Modelle aufgebaut sind.

Wie man □ Abb. 18.7 entnehmen kann, gibt es ein Merkmal η_0 , das auf alle Items einen Einfluss ausübt. Damit wird z. B. einem *Generalfaktor*, wie er in Intelligenzmodellen oder bei der Kompetenzmessung angenommen werden kann, Rechnung getragen. Weitere „kleinere“ Faktoren sind η_1 und η_2 , die jeweils nur auf spezifische Items einen Effekt haben (hier Items 1 bis 3 bzw. 4 bis 6). In diesem Beispiel wird angenommen, dass η_1 und η_2 unkorreliert sind. Diese Annahme lässt sich lockern.



□ Abb. 18.7 Beispiel für ein Bifaktormodell: η_0 ist ein Generalfaktor, der auf alle Items einen Einfluss hat. η_1 und η_2 haben nur auf spezifische Items einen Effekt. Es wird angenommen, dass η_1 und η_2 unkorreliert sind

18.3 · Ausblick auf weitere Modelle

Möchte man dieses Modell in Gleichungen ausdrücken, so ergeben sich die für die jeweiligen Items geltenden Antwortwahrscheinlichkeiten:

$$P(Y_i = 1|\eta) = \begin{cases} \frac{\exp(\lambda_0\eta_0 + \lambda_1\eta_1 + d_i)}{1 + \exp(\lambda_0\eta_0 + \lambda_1\eta_1 + d_i)} & \text{für } i = 1, 2, 3 \\ \frac{\exp(\lambda_0\eta_0 + \lambda_2\eta_2 + d_i)}{1 + \exp(\lambda_0\eta_0 + \lambda_2\eta_2 + d_i)} & \text{für } i = 4, 5, 6 \end{cases} \quad (18.48)$$

Wie man den Gleichungen entnehmen kann, unterscheidet sich die Binnenstruktur für unterschiedliche Itemgruppen. Unter Rückgriff auf die zahlreichen in diesem Kapitel vorgestellten Modelle wird deutlich, dass man vielfältige weitere Bifaktormodelle spezifizieren kann, wenn man für die Itemantworten weitere strukturelle Annahmen trifft (z. B. eine RSM- oder PCM-Struktur).

Antwortwahrscheinlichkeiten

Vielfältige weitere Bifaktormodelle

18.3 Ausblick auf weitere Modelle

An dieser Stelle soll noch der Ausblick auf ein paar exemplarische weitere Modellarten gegeben werden. Ein erschöpfender Überblick ist nicht möglich.

■■ Modelle für Antwortzeiten und kontinuierliche Antwortkategorien

Die *Modellierung von Antwortzeiten* hat in jüngerer Vergangenheit das Interesse von Forschenden geweckt (vgl. für einen Überblick Jansen 2017; Tuerlinckx et al. 2017; van der Linden 2017a). Es gibt verschiedene Ansätze und Traditionen zur Modellierung von Reaktionszeiten (auch „reaction times“, RT). Grob gesprochen können diese in zwei Formen eingeteilt werden:

1. Die erste entstammt der sog. „Mathematischen Psychologie“ und wird in den Kognitionswissenschaften zur Entscheidungsmodellierung verwendet. Es handelt sich dabei um Ratcliffs *Diffusion Model* (s. z. B. Ratcliff 1978; Wagenmakers 2009). In diesem wird Evidenz bis zu dem Zeitpunkt akkumuliert, an dem ein bestimmtes Niveau erreicht wird (s. auch sog. „Poisson Race Model“ für eine komplexere Anwendung).
2. Die zweite entstammt der Psychometrie. Dabei werden in zwei Untermodellen die Bezüge zwischen der Reaktionszeit und dem eigentlichen Antwortverhalten untersucht. In der ersten Unterart werden Parameter, die in der Verbindung zu Reaktionszeiten stehen, als erklärende Variablen in das IRT-Modell aufgenommen (z. B. um die Itemschwierigkeit in einem Rasch-Modell zu erklären). In der zweiten Unterart werden Personen und Itemparameter in Modelle aufgenommen, die spezifische Verteilungsannahmen für die Reaktionszeiten treffen.

Modellierung von Antwortzeiten

Interessanterweise bildet das sog. „Q-Diffusion Model“ (vgl., Tuerlinckx et al. 2017) eine Brücke zwischen den verschiedenen Zugängen der Mathematischen Psychologie und der Psychometrie, indem mit diesem aus einem Diffusion Modell ein IRT-Modell abgeleitet wird.

Ein Überblick über *kontinuierliche Antwortvariablen* kann man bei Mellenbergh (2017) finden (für ein frühes kontinuierliches Rasch-Modell s. auch Müller 1987). Unter diese Gruppe lassen sich Modelle verschiedener Traditionen zusammenfassen, etwa der Faktorenanalyse, der latenten Profilanalyse, der Generalisierten Linearen Modelle sowie kontinuierliche Varianten polytomer IRT-Modelle (wie das GRM in ► Abschn. 18.1.3).

Kontinuierliche Antwortvariablen

Grundsätzlich werden bei diesen die Antworten Y_i auf ein Item i als kontinuierlich verteilt angenommen. Beispielsweise können sie normalverteilt sein, $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, mit dem Erwartungswert μ_i und der Varianz σ_i^2 . Wie in multidimensionalen IRT-Modellen wird der Erwartungswert μ_i als Resultat einer Line-

arkombination der Teile des Vektors η mit Traits verstanden:

$$\mu_i = E(Y_i|\eta) = d_i + \lambda_i \cdot \eta \quad (18.49)$$

Verschiedene Modelle und Traditionen zur Schätzung können angewendet werden (z. B. aus der Faktorenanalyse). Es sei darauf hingewiesen, dass im Unterschied zu kategorialen Antwortvariablen die Varianz $\sigma_i^2 = \text{Var}(Y_i|\eta)$ mit der Itemschwierigkeit d_i und den Ladungen λ_i geschätzt werden muss. Modelle für kategoriale Variablen können dann als Spezialfälle der vorhergehenden Gleichung aufgefasst werden. Zum Beispiel nimmt das Rasch-Modell für dichotome Antwortvariablen Bernoulli-verteilte Antworten an. Diese haben eine Wahrscheinlichkeit $P(Y_i|\eta)$ mit einem Erwartungswert, der durch die Logit-Link-Funktion $\mu_i = \text{logit}P(Y_i|\eta)$ gegeben ist.

■■ Mischverteilungsmodelle

Latente diskrete Variable

In Mischverteilungsmodellen geht man davon aus, dass eine latente Variable existiert, die diskreter Natur ist (s. auch ► Kap. 22). Diese diskrete latente Variable beschreibt eine unbeobachtete Klassenzugehörigkeit jeder Person in der Stichprobe. Es wird davon ausgegangen, dass die Stichprobe heterogen ist und sich aus Individuen aus mindestens zwei Subpopulationen zusammensetzt.

Mixture-IRT-Modelle

Mit Mixture-IRT-Modellen versucht man nun zu beschreiben, wie die Messmodelle für jede dieser Subpopulationen aussehen, indem für jede Klasse/Subpopulation (Item-)Parameterschätzungen durchgeführt werden (von Davier und Rost 2017). Somit ergeben sich (wenn man nicht bewusst zusätzliche Restriktionen einführt) mindestens zwei Sätze von Parameterschätzungen für ein formales Modell (eines für jede latente Klasse). Die Verwendung solcher Modelle erlaubt es, eine vorab nicht gemessene Heterogenität in den Daten aufzufinden und z. B. Personen zu identifizieren, deren Beantwortung der Fragen abweicht.

■■ Netzwerkmodelle

Alternative Beschreibung von Phänomenen

In jüngerer Zeit wurden auch IRT-Modelle mit dem in der Physik bekannten *Ising-Modell* in Verbindung gebracht (z. B. Marsman et al. 2018). Der Grundgedanke ist es, psychologische Konstrukte in einer alternativen Weise anhand von Netzwerkmodellen zu beschreiben. Dabei wird der Zusammenhang beobachteter Variablen (z. B. Symptome der Depression) als Muster einer kausalen Interaktion der (beobachtbaren) Variablen beschrieben. Das Ising-Modell erlaubt dabei die Beschreibung von Zuständen verbundener Partikel und hat Bezüge zum Messmodell der IRT, das die Antwortwahrscheinlichkeit anhand latenter Variablen beschreibt.

18.4 Weiterführende Literatur

Einen sehr guten Überblick über die hier vorgestellten Modelle findet man bei van der Linden (2017b) oder bei Embretson und Reise (2013b). Im Falle der multidimensionalen IRT-Modelle gibt Reckase (2009) eine gute Einführung.

18.5 EDV-Hinweise

Bitte beachten Sie die EDV-Hinweise in ► Kap. 16.

18.6 Kontrollfragen

?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was ist der Ausgangspunkt des GPCM?
2. Sind im GPCM die Itemkategorienparameter geordnet?
3. Welche Idee liegt dem RSM nach Andrich zugrunde?
4. Was ist der Grundgedanke des GRM nach Smejima?
5. Wozu braucht man multidimensionale IRT-Modelle?

Literatur

- Adams, R. J., Wu, M. & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, 72, 547–573.
- Andersen, E. B. (1997). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–78.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Andrich, D. (2005). Georg Rasch: Mathematician and statistician. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 3, pp. 299–306). Amsterdam, The Netherlands: Academic Press.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, 1–29.
- Embretson, S. E. & Reise, S. P. (2013a). *Item response theories for psychologists*. Psychological Press.
- Embretson, S. E. & Reise, S. P. (2013b). *Item response theory*. Psychology Press.
- Holzinger, K. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41–54.
- Jansen, M. G. H. (2017). Poisson and gamma models for reading speed and error. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 245–259). Chapman & Hall/CRC.
- Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., Bork, R., Waldorp, L. J. et al (2018). An introduction to network psychometrics: relating Ising network models to item response theory models. *Multivariate Behavioral Research*, 53, 15–35.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Mellenbergh, G. J. (2017). Chapter 10. Models for continuous responses. In W. J. Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 153–163). Chapman & Hall/CRC.
- Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, 52, 165–181.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muraki, E. & Muraki, M. (2017). Chapter 8. Generalized Partial Credit Model. In W. J. Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 127–137). Chapman & Hall/CRC.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability IV* (pp. 321–334). Berkeley, CA: University of California Press.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, Springer.
- Reckase, M. D. (2017). Logistic multidimensional models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 189–210). Chapman & Hall/CRC.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47, 667–696.
- Samejima, F. (1969a). *Estimation of ability using a response pattern of graded scores*. Psychometrika Monograph No. 17. Richmond, VA: Psychometric Corporation.
- Samejima, F. (1969b). *A general model for free-response data*. Psychometrika Monograph No. 18. Richmond, VA: Psychometric Corporation.
- Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, 60, 549–572.
- Samejima, F. (2017). Graded response model. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 95–108). Chapman & Hall/CRC.
- Tuerlinckx, F., Molenaar, D. & van der Maas, H. L. J. (2017). Diffusion-based response-time models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 283–300). Chapman & Hall/CRC.

- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39–55.
- Tutz, G. (2011). *Regression for categorical data* (Bd. 34). Cambridge University Press.
- van der Linden, W. J. (2017a). Lognormal response-time model. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 261–282). Chapman & Hall/CRC.
- van der Linden, W. J. (2017b). *Handbook of item response theory. Volume 1: Models*. CRC Press.
- van der Linden, W. J. & Hambleton, R. K. H. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.
- Verhelst, N. D., Glas, C. A. & De Vries, H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). Springer.
- von Davier, M. & Rost, J. (2017). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume 1: Models* (pp. 421–434). Chapman and Hall/CRC.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21, 641–671.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT)

Norman Rose

Inhaltsverzeichnis

- 19.1 Verfahren der Parameterschätzung in der IRT: Überblick – 449**
- 19.2 Maximum-Likelihood-Schätzung (ML-Schätzung) – 450**
 - 19.2.1 Prinzip der ML-Schätzung – 450
 - 19.2.2 Bestimmung des ML-Schätzers – 452
 - 19.2.3 Genauigkeit der ML-Schätzung – 452
 - 19.2.4 Joint-ML (JML)-Schätzung – 453
 - 19.2.5 Conditional-ML (CML)-Schätzung – 456
 - 19.2.5.1 Suffiziente Statistiken – 456
 - 19.2.5.2 S_v -bedingte Antwortmusterwahrscheinlichkeiten im Rasch Modell – 457
 - 19.2.5.3 Bedingte ML-Funktion zur Parameterschätzung im Rasch-Modell – 459
 - 19.2.6 Marginal-ML (MML)-Schätzung – 461
 - 19.2.6.1 Gemeinsame Verteilung manifester und latenter Variablen – 461
 - 19.2.6.2 Anwendung des Bayes-Theorems in der MML-Schätzung – 462
 - 19.2.6.3 MML-Schätzung unter Verwendung des Expectation-Maximization-Algorithmus (EM-Algorithmus) – 464
 - 19.2.6.4 Anmerkungen zur MML-Schätzung – 465
- 19.3 Bayes'sche Schätzverfahren – 466**
 - 19.3.1 Grundlegendes – 466
 - 19.3.2 Bayes-Inferenz auf Basis der A-posteriori-Verteilung – 467
 - 19.3.3 Spezifikation der A-priori-Verteilung – 468
 - 19.3.4 Parameterschätzung in der Bayes-Statistik – 470
 - 19.3.4.1 Nicht simulationsbasierte Bayes'sche Schätzverfahren in der IRT – 470
 - 19.3.4.2 Simulationsbasierte Bayes'sche Schätzverfahren in der IRT (MCMC-Verfahren) – 474
- 19.4 Weitere Schätzverfahren – 483**
- 19.5 Personenparameterschätzung in der IRT – 484**

19.5.1	ML-Scoring – 485
19.5.2	Gewichtete ML-Schätzung – 486
19.5.3	Bayes'sche Personenparameterschätzer – 487
19.5.3.1	MAP-Schätzung – 487
19.5.3.2	EAP-Schätzung – 488
19.5.3.3	Plausible Values (PV)-Verfahren – 489
19.6	Reliabilitätsbeurteilung in der IRT – 490
19.6.1	Zur Erinnerung: Messgenauigkeit in der Klassischen Testtheorie (KTT) – 491
19.6.2	Testinformation und Standardfehler – 491
19.6.2.1	Grundlegendes – 491
19.6.2.2	Iteminformationsfunktion – 492
19.6.2.3	Testinformationsfunktion – 493
19.6.2.4	Beziehung von Standardfehler- und Testinformationsfunktion – 493
19.6.3	Marginale Reliabilitäten – 494
19.6.3.1	Marginale Reliabilität bei ML-Personenparameterschätzern – 494
19.6.3.2	Marginale Reliabilität bei EAP- und MAP-Personenparameterschätzern – 495
19.7	Zusammenfassung – 496
19.8	EDV-Hinweise – 497
19.9	Kontrollfragen – 498
	Literatur – 499

i Große Datenmatrizen, die oft nur die Werte null und eins beinhalten, bilden die Datengrundlage für die Modelle der IRT. Doch wie kann man aus dieser spärlich anmutenden Information Itemparameter wie die Itemsdiskrimination, Itemschwierigkeit, Schwellenparameter oder die individuellen Werte der Personen auf den latenten Variablen schätzen? Die Antwort ist so einfach wie herausfordernd: mit Mathematik!

19.1 Verfahren der Parameterschätzung in der IRT: Überblick

Die in ► Kap. 16 und 18 behandelten Modelle der IRT beschreiben die Wahrscheinlichkeit $P(y_{vi}|\theta_v)$, dass eine Person v in der Antwortkategorie y von Item i antwortet, durch eine parametrische Funktion $f(\tau_i, \theta_v)$ von zwei distinkten theoretischen Größen, den *Itemparametern* τ_i und den *Personenparametern* θ_v . In Abhängigkeit des verwendeten IRT-Modells besteht der Vektor τ_i aus den Itemschwierigkeiten β_i , den Schwellenparametern κ_{ic} , den Itemdiskriminationen α_i oder anderen Parametern, die die Eigenschaften des Items i beschreiben. Charakteristika und Eigenschaften der Personen werden in einem validen Test durch ihre individuellen Werte θ_v auf den latenten Variablen θ im Messmodell repräsentiert. θ kann dabei eine ein- oder mehrdimensionale latente Variable sein.

Bei Anwendung der IRT besteht das Schätzproblem darin, aus einer Datenmatrix, bestehend aus den individuellen Antwortmustern $\mathbf{y}_v = y_{v1}, \dots, y_{vi}, \dots, y_{vk}$ (mit $v = 1, \dots, n$), die unbekannten Itemparameter sowie die individuellen Personenparameter oder zumindest Verteilungsparameter der latenten Variablen θ zu schätzen. Hierzu bedient man sich verschiedener Maximum-Likelihood-Schätzmethoden (ML-Schätzmethoden). Die gleichzeitige Schätzung von Item- und Personenparametern wird als Joint-Maximum-Likelihood-Schätzung (JML-Schätzung) bezeichnet. Die Zahl der zu schätzenden Parameter ist bei der JML-Schätzung also abhängig von der Stichprobengröße, was grundsätzlich zu inkonsistenten Parameterschätzungen führen kann. Daher wurden Methoden entwickelt, in denen zunächst auf die simultane Schätzung der Personenparameter verzichtet wird. Im Falle der Rasch-Modelle für dichotome und ordinale Items wurde die Conditional-Maximum-Likelihood-Schätzung (CML-Schätzung) entwickelt, die die Eigenschaft der individuellen Summenscores $s_v = \sum_{i=1}^k y_{vi}$ als sog. „suffiziente Statistik“ für den Personenparameter θ_v ausnutzt (Näheres s. ► Abschn. 19.2.5.1). Dies ermöglicht die Schätzung von Itemschwierigkeiten bzw. Schwellenparametern ohne Schätzung der Personenparameter θ_v .

Für zwei-, drei- und mehrparametrische Modelle ist die CML-Methode nicht anwendbar, hier bietet die Marginal-Maximum-Likelihood-Schätzung (MML-Schätzung) eine Alternative. Bei diesem Verfahren werden die individuellen Personenwerte θ_v als Realisationen einer latenten Zufallsvariablen θ mit einer bestimmten Verteilung $f(\theta)$ aufgefasst. Anstatt alle individuellen Werte θ_v zu schätzen, wird lediglich die Verteilung der latenten Variablen mitmodelliert. Eine Möglichkeit ist die Annahme einer parametrischen Verteilung der latenten Variablen wie die Normalverteilung. Da die Normalverteilung nur durch zwei Parameter, nämlich den Erwartungswert $E(\theta)$ und die Varianz $Var(\theta)$, hinreichend beschrieben ist, müssen neben den Itemparametern lediglich diese beiden Verteilungsparameter geschätzt werden. Die MML-Schätzung hat zudem den Vorteil, dass die Zahl der zu schätzenden Parameter nicht von der Stichprobengröße abhängt. Die MML-Schätzung ist für alle parametrischen IRT-Modelle anwendbar und die am häufigsten verwendete Schätzmethode in der IRT.

Alle ML-Schätzverfahren (JML, CML und MML) basieren auf dem allgemeinen Prinzip der ML-Methode, das zu Beginn dieses Kapitels eingeführt wird (► Abschn. 19.2.1). Die JML-Schätzung ist das historisch älteste Verfahren und bildet zudem einen geeigneten theoretischen Ausgangspunkt zur Darstellung der

Item- vs. Personenparameter

JML- und CML-Schätzung

MML-Schätzung

Bayes'sche Schätzverfahren und MCMC-Verfahren

Personenparameterschätzung

Individualdiagnostik

Alternative

Personenparameterschätzer mit unterschiedlichen Eigenschaften

Parameterschätzung in der IRT (► Abschn. 19.2.4). Danach werden die CML- (► Abschn. 19.2.5) und die MML-Methode besprochen (► Abschn. 19.2.6).

Bayes'sche Schätzverfahren stellen einen alternativen Ansatz zur Bestimmung der Item- und Personenparameter dar (► Abschn. 19.3). Da die Bayes-Statistik konzeptuelle Unterschiede zur klassischen bzw. frequentistischen Statistik aufweist, werden zunächst die wichtigsten Grundbegriffe der Bayes-Statistik eingeführt, bevor die Bayes'sche Parameterschätzung in der IRT erklärt wird. Mit der Entwicklung leistungsfähiger Computer haben insbesondere simulationsbasierte *Markov-Chain-Monte-Carlo-Verfahren* (MCMC-Verfahren), d. h. Sampling-Algorithmen, immer weitere Verbreitung erfahren. Aus diesem Grund werden mit dem *Metropolis-Hastings-Algorithmus* (MH-Algorithmus), dem *Gibbs-Sampler* und deren Kombination mehrere Sampling-Algorithmen erläutert (► Abschn. 19.4).

Der Personenparameterschätzung wird in diesem Kapitel ein eigener Abschnitt (► Abschn. 19.5) gewidmet. Dafür gibt es mehrere Gründe:

1. Die CML- und MML-Schätzung erlauben zunächst nur die Itemparameterschätzung. Die Personenparameter müssen hier zwangsläufig in einem separaten zweiten Schritt geschätzt werden.
2. Bei etablierten Testverfahren werden die Itemparameter im Rahmen der Testentwicklung bestimmt. Die entsprechenden Werte lassen sich dem Testmanual entnehmen und können bei der Testanwendung zur Personenparameterschätzung verwendet werden. Dadurch werden große Stichproben zur Itemparameterschätzung überflüssig.

Die hier dargestellten Personenparameterschätzungen erfolgen separat für jede Person in Abhängigkeit von ihrem Antwortmuster und können somit auch in der Individualdiagnostik verwendet werden. Eine Anwendung ist beispielsweise das computerisierte adaptive Testen (► Kap. 20). Dabei werden die Parameter der Items des sog. „Itempools“ (Gesamtheit aller verfügbaren Items des jeweiligen Tests) in der Entwicklungsphase anhand hinreichend großer Stichproben geschätzt und in späteren Anwendungen als bekannte Größen zur Personenparameterschätzung verwendet.

Fünf verschiedene Personenparameterschätzer sind in der IRT gebräuchlich, und zwar die folgenden:

1. ML-Schätzer
2. Gewichteter ML-Schätzer nach Warm
3. Maximum-a-posteriori-Schätzer (MAP-Schätzer)
4. Expected-a-posteriori-Schätzer (EAP-Schätzer)
5. Plausible Values (PVs)

Die verschiedenen Personenparameterschätzer haben auch statistisch ganz unterschiedliche Eigenschaften, die bei der Reliabilitätsbestimmung berücksichtigt werden müssen. Daher soll die Reliabilitätsbestimmung in der IRT den Abschluss dieses Kapitels bilden, zu der u. a. sowohl die Item- wie auch die Testinformati onsfunction vorgestellt werden (► Abschn. 19.6).

19.2 Maximum-Likelihood-Schätzung (ML-Schätzung)

19.2.1 Prinzip der ML-Schätzung

Zunächst soll hier allgemein das Prinzip der ML-Schätzung erklärt werden, bevor es auf das Schätzproblem in der IRT angewendet wird. Hierfür sei Y irgendeine Zufallsvariable mit den Realisationen $Y = y$, die in der empirischen Forschung den Daten entsprechen. Die ML-Methode kann verwendet werden, um den Parametervektor $\varphi = \varphi_1, \dots, \varphi_m$ eines statistischen Modells zu schätzen, das den

19.2 · Maximum-Likelihood-Schätzung (ML-Schätzung)

beobachteten Daten zugrunde gelegt wird. Auch eine Verteilungsannahme kann als Modell aufgefasst werden. Wird z. B. angenommen, dass Y normalverteilt ist, mit $Y \sim N(\mu, \sigma)$, so sind der Erwartungswert μ und die Streuung σ die zu schätzenden Parameter, die den Vektor φ bilden, und die anhand einer Stichprobe vom Umfang N geschätzt werden sollen. In der Statistik wird dabei zunächst von einer mathematischen Stichprobe ausgegangen, die eine Folge Y_1, \dots, Y_n von stochastisch unabhängigen und gleichartig verteilten Variablen ist. Die beobachtbaren Daten y_1, \dots, y_n entsprechen den Realisationen dieser Zufallsvariablen. Aufgrund der stochastischen Unabhängigkeit kann die gemeinsame Wahrscheinlichkeitsdichte der n Variablen geschrieben werden als:

$$f(Y_1 = y_1, \dots, Y_n = y_n; \varphi) = \prod_{i=1}^n f(Y_i = y_i; \varphi) \quad (19.1)$$

Im Falle einer Stichprobe von kategorialen Variablen Y_1, \dots, Y_n lässt sich die Verbundwahrscheinlichkeit analog zu Gl. (19.1) als Produkt schreiben:

$$P(Y_1 = y_1, \dots, Y_n = y_n; \varphi) = \prod_{i=1}^n P(Y_i = y_i; \varphi) \quad (19.2)$$

Diese Gleichung beschreibt die Wahrscheinlichkeit, mit der die beobachteten Daten bei einer erneuten Stichprobenziehung unter Annahme des Modells auftreten. Diese Wahrscheinlichkeit ist eine Funktion der unbekannten zu schätzenden Parameter $\varphi = \varphi_1, \dots, \varphi_m$ und wird Likelihood-Funktion $L(\varphi)$ genannt.

Mathematische Stichprobe

Verbundwahrscheinlichkeit

Likelihood-Funktion

Definition

$Y = Y_1, \dots, Y_n$ ist eine Stichprobe mit den Realisationen $y = y_1, \dots, y_n$ und $\varphi \in \Omega_\varphi$ ein Parametervektor im Parameterraum Ω_φ eines statistischen Modells. Die **Likelihood-Funktion** $L(\varphi)$ ist die Wahrscheinlichkeits- bzw. Dichtefunktion für das Ereignis $Y = y$ in Abhängigkeit von φ .

Bei der ML-Schätzung werden die Parameter $\varphi = \varphi_1, \dots, \varphi_m$ nun so geschätzt, dass die Auftretenswahrscheinlichkeit der beobachteten Daten maximal ist. Entsprechend lässt sich auch der ML-Schätzer $\hat{\varphi}_{ML}$ definieren.

Definition

Der **ML-Schätzer** $\hat{\varphi}_{ML}$ ist definiert als der Wert von φ im Parameterraum Ω_φ , der die Likelihood-Funktion $L(\varphi)$ für das Ereignis $Y = y$ maximiert:

$$\hat{\varphi}_{ML} = \arg \max_{\varphi \in \Omega_\varphi} L(\varphi) \quad (19.3)$$

Es zu beachten, dass Wahrscheinlichkeitsaussagen nur für Ereignisse sinnvoll sind, die noch nicht eingetreten sind (sog. „Prä-facto-Perspektive“ der Wahrscheinlichkeit). Daher werden die Parameter bei der ML-Schätzung streng genommen nicht so bestimmt, dass die Wahrscheinlichkeit der Daten y maximiert wird. Die vorliegenden empirischen Daten haben weder eine Verteilung noch eine Wahrscheinlichkeit. Es lässt sich aber untersuchen, wie wahrscheinlich das zufällige Zustandekommen von Daten ist, die gleich den beobachteten Daten der vorliegenden Stichprobe sind, unter der Annahme der Gültigkeit des Modells, dessen Parameter geschätzt werden sollen. Auf die bedingte Auftretenswahrscheinlichkeit für dieses Ereignis $Y = y$ wird in der Definition des ML-Schätzers Bezug genommen.

Prä-facto-Perspektive der Wahrscheinlichkeit

19.2.2 Bestimmung des ML-Schätzers

Log-Likelihood- und Score-Funktion

Die Bestimmung des ML-Schätzers lässt sich mathematisch als ein Maximierungsproblem auffassen. Das Maximum einer Funktion lässt sich durch Nullsetzen der ersten Ableitung der Funktion nach den gesuchten Größen bestimmen. Dabei wird bei Anwendung der ML-Schätzung üblicherweise die logarithmierte Likelihood-Funktion (kurz Log-Likelihood-Funktion) $l(\boldsymbol{\varphi})$ verwendet. Das ist möglich, da der Logarithmus eine monotone Transformation ist. Somit haben $L(\boldsymbol{\varphi})$ und $l(\boldsymbol{\varphi})$ ihr Maximum an derselben Stelle $\boldsymbol{\varphi} \in \Omega_{\boldsymbol{\varphi}}$. Die Maximierung von $l(\boldsymbol{\varphi})$ führt daher auch zum gleichen ML-Schätzer $\hat{\boldsymbol{\varphi}}_{ML}$. Die erste Ableitung $l'(\boldsymbol{\varphi})$ der Log-Likelihood-Funktion nach dem Parametervektor wird auch *Score-Funktion* genannt. Bei einem Parametervektor $\boldsymbol{\varphi} = \varphi_1, \dots, \varphi_m$ ist die Score-Funktion ebenfalls ein Vektor der Länge m , der die partiellen Ableitungen nach den einzelnen Parametern beinhaltet:

$$l'(\boldsymbol{\varphi}) = \frac{\partial l(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\varphi})}{\partial \varphi_m} \end{pmatrix} \quad (19.4)$$

Ableitung der Log-Likelihood-Funktion

Die Ableitung der Log-Likelihood-Funktion nach $\boldsymbol{\varphi}$ ist mathematisch wesentlich unkomplizierter als die Ableitung der Likelihood-Funktion $L(\boldsymbol{\varphi})$, da Letztere ein Produkt ist (Gl. 19.1) und Ableitungen von Summen mathematisch wesentlich einfacher sind. Allgemein gilt, dass der Logarithmus eines Produkts gleich der Summe der logarithmierten Faktoren des Produkts ist. Somit lässt sich bei der Ableitung der Log-Likelihood-Funktion nach $\boldsymbol{\varphi}$ die Summenregel aus der Differentialrechnung angewenden. Die resultierende Score-Funktion ist dann ebenfalls eine Summe, die allgemein wie folgt geschrieben werden kann:

$$\begin{aligned} l'(\boldsymbol{\varphi}) &= \frac{\partial}{\partial \boldsymbol{\varphi}} \sum_{i=1}^n \ln[f(Y_i = y_i; \boldsymbol{\varphi})] \\ &= \sum_{i=1}^n \frac{\partial \ln[f(Y_i = y_i; \boldsymbol{\varphi})]}{\partial \boldsymbol{\varphi}} \end{aligned} \quad (19.5)$$

Numerische Optimierungsverfahren

Bei vielen Schätzproblemen der Statistik, so auch in der IRT, lässt sich die Score-Funktion oft nicht analytisch lösen. Deshalb finden numerische Optimierungsverfahren Anwendung, beispielsweise der *Newton-Raphson-Algorithmus* oder *Quasi-Newton-Verfahren* (Wright und Nocedal 1999). Auf diese Algorithmen wird bei den einzelnen ML-Schätzverfahren in der IRT genauer eingegangen.

Standardfehler als Maß der Genauigkeit des skalaren Schätzers

19.2.3 Genauigkeit der ML-Schätzung

Neben der Bestimmung des ML-Schätzers ist auch die Genauigkeit dieser Schätzung von zentraler Bedeutung. Ein Maß der Genauigkeit eines skalaren Schätzers $\hat{\boldsymbol{\varphi}}$ ist der Standardfehler. Im Falle eines m -dimensionalen Parametervektors gibt die $(m \times m)$ -dimensionale Varianz-Kovarianz-Matrix $\mathbf{V}(\hat{\boldsymbol{\varphi}})$ Auskunft über die Genauigkeit der Schätzer $\hat{\boldsymbol{\varphi}} = \hat{\varphi}_1, \dots, \hat{\varphi}_m$. Die Quadratwurzeln der Diagonalelemente von $\mathbf{V}(\hat{\boldsymbol{\varphi}})$ sind die m Standardfehler $SE(\hat{\varphi}_1), \dots, SE(\hat{\varphi}_m)$.

Zur Schätzung von $\mathbf{V}(\hat{\boldsymbol{\varphi}})$ wird auf die zweite Ableitung $l''(\boldsymbol{\varphi})$ der Log-Likelihood-Funktion zurückgegriffen, die auch als Hesse-Matrix bezeichnet wird:

$$l''(\boldsymbol{\varphi}) = \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^2} = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1^2} & \dots & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_1 \partial \varphi_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_m \partial \varphi_1} & \dots & \frac{\partial^2 l(\boldsymbol{\varphi})}{\partial \varphi_m^2} \end{pmatrix} \quad (19.6)$$

Wertet man die Hesse-Matrix nach Vorzeichenumkehr an der Stelle des ML-Schätzers $\hat{\boldsymbol{\varphi}}$ aus, so erhält man die beobachtete („observed“) Fisher-Informationsmatrix $F_o(\hat{\boldsymbol{\varphi}}_{ML}) = -l''(\hat{\boldsymbol{\varphi}}_{ML})$, bezeichnet nach dem herausragenden Statistiker Ronald A. Fisher, der die ML-Methode wesentlich mit ausformuliert hat. Die inverse Matrix der beobachteten Fisher-Informationsmatrix ist ein Schätzer der Varianz-Kovarianz-Matrix von $\hat{\boldsymbol{\varphi}}$:

$$\mathbf{V}(\hat{\boldsymbol{\varphi}}_{ML}) = F_o(\hat{\boldsymbol{\varphi}}_{ML})^{-1} \quad (19.7)$$

Es ist zu beachten, dass die Likelihood-Funktion und ihre Ableitungen ebenfalls Zufallsvariablen mit einer Verteilung sind. Dies bedeutet, dass die Maxima $\hat{\boldsymbol{\varphi}}_{ML}$ als auch die Werte der ML-Funktion am Punkt $\boldsymbol{\varphi} = \hat{\boldsymbol{\varphi}}_{ML}$ über die Stichproben hinweg variieren. Eine Eigenschaft der Score-Funktion ist, dass ihr Erwartungswert am Punkt des wahren Parametervektors $E[l'(\boldsymbol{\varphi})]$ gleich null ist. Andernfalls wäre der ML-Schätzer nicht erwartungstreu und somit verzerrt. Die Varianz-Kovarianz-Matrix von $l'(\boldsymbol{\varphi})$ ist gerade die Fisher-Informationsmatrix, die sich aufgrund der Eigenschaft $E[l'(\boldsymbol{\varphi})]$ gleich null als Erwartungswert des äußeren Produkts der Score-Funktion schreiben lässt:

$$\begin{aligned} \mathbf{V}[l'(\boldsymbol{\varphi})] &= E[(l'(\boldsymbol{\varphi}) - E[l'(\boldsymbol{\varphi})])(l'(\boldsymbol{\varphi}) - E[l'(\boldsymbol{\varphi})])^T] \\ &= E[l'(\boldsymbol{\varphi})l'^T(\boldsymbol{\varphi})] \\ &= F_e(\boldsymbol{\varphi}) \end{aligned} \quad (19.8)$$

Dabei ist $F_e(\boldsymbol{\varphi})$ die erwartete („expected“) Fisher-Informationsmatrix und $l'(\boldsymbol{\varphi})^T$ die transponierte Score-Funktion. Aus Gl. (19.8) folgt, dass die Fisher-Information auch auf der Basis der ersten Ableitungen der Log-Likelihood-Funktion anhand von Stichprobendaten geschätzt werden kann, durch:

$$\hat{\mathbf{V}}[l'(\hat{\boldsymbol{\varphi}}_{ML})] = \frac{1}{n} \sum_{i=1}^n \frac{\partial l_i(\hat{\boldsymbol{\varphi}}_{ML})}{\partial \hat{\boldsymbol{\varphi}}_{ML}} \left(\frac{\partial l_i(\hat{\boldsymbol{\varphi}}_{ML})}{\partial \hat{\boldsymbol{\varphi}}_{ML}} \right)^T \quad (19.9)$$

Die Terme $l_i(\hat{\boldsymbol{\varphi}}_{ML})$ sind dabei die individuellen logarithmierten Likelihoods der einzelnen Realisationen oder Beobachtungen $i = 1, \dots, n$, die sich zur gesamten Log-Likelihood-Funktion $l(\boldsymbol{\varphi})$ aufsummieren. Auch Gl. (19.9) kann zur Schätzung der Varianz-Kovarianz-Matrix von $\hat{\boldsymbol{\varphi}}_{ML}$ alternativ zu Gl. (19.7) verwendet werden. Im Folgenden soll nun das hier dargestellte Prinzip der ML-Schätzung auf die Item- und Personenparameterschätzung in der IRT angewendet werden.

19.2.4 Joint-ML (JML)-Schätzung

Die nachfolgend dargestellte gemeinsame Schätzung von Item- und Personenparametern, die JML-Schätzung, wurde bereits von Birnbaum (1968) dargestellt und von Baker und Kim (2004) ausführlich beschrieben. Die zugrunde liegende Logik soll ausgehend vom allgemeinen Prinzip der ML-Methode dargelegt werden. Zunächst wird wieder von der stochastischen Unabhängigkeitsannahme der Zufallsvariablen $\mathbf{Y}_1, \dots, \mathbf{Y}_v, \dots, \mathbf{Y}_n$ ausgegangen, deren Realisationen in der IRT

Hesse-Matrix

Beobachtete Fisher-Informationsmatrix

Erwartete Fisher-Informationsmatrix

Gemeinsame Schätzung von Item- und Personenparametern

die beobachteten Antwortmuster $y_1, \dots, y_v, \dots, y_n$ der Testpersonen sind. Die Likelihood-Funktion kann gemäß Gl. (19.2) geschrieben werden als das Produkt:

$$L(\boldsymbol{\varphi}) = P(Y_1 = y_1, \dots, Y_n = y_n; \boldsymbol{\varphi}) = \prod_{v=1}^n P(Y_v = y_v; \boldsymbol{\varphi}) \quad (19.10)$$

In der JML-Schätzung umfasst der Parametervektor $\boldsymbol{\varphi} = (\tau_1, \dots, \tau_k, \theta_1, \dots, \theta_n)$ alle Item- und Personenparameter. Die Modelle der IRT beschreiben bedingte Wahrscheinlichkeiten von Itemantworten bei gegebenen Werten der latenten Variablen θ_v der Personen. Entsprechend lässt sich Formel Gl. (19.10) schreiben als:

$$L(\boldsymbol{\varphi}) = \prod_{v=1}^n P(Y_v = y_v | \theta_v; \boldsymbol{\tau}) \quad (19.11)$$

Dabei ist $P(Y_v = y_v | \theta_v; \boldsymbol{\tau})$ die Antwortmusterwahrscheinlichkeit von Person v unter Annahme der Gültigkeit des Modells. Eine wichtige Annahme in vielen Modellen der IRT ist die Annahme der lokalen stochastischen Unabhängigkeit. Sie besagt, dass die Itemantworten hinsichtlich zweier Items i und j für einen festen Wert der latenten Variablen stochastisch unabhängig sind. Aufgrund dieser Unabhängigkeitsannahme kann die Antwortmusterwahrscheinlichkeit für jede Person v als Produkt der bedingten Wahrscheinlichkeiten für die Antwortkategorien $Y_{vi} = y_{vi}$ aller k Items geschrieben werden:

$$P(Y_v = y_v | \theta_v; \boldsymbol{\tau}) = \prod_{i=1}^k P(Y_{vi} = y_{vi} | \theta_v; \tau_i) \quad (19.12)$$

Lokale stochastische Unabhängigkeit

Allgemeine Likelihood-Funktion

Durch Einsetzen von Gl. (19.12) in Gl. (19.11) resultieren die allgemeine Likelihood-Funktion für die Itemantworten aller Testpersonen:

$$L(\boldsymbol{\varphi}) = \prod_{v=1}^n \prod_{i=1}^k P(Y_{vi} = y_{vi} | \theta_v; \tau_i) \quad (19.13)$$

und die entsprechende Log-Likelihood-Funktion:

$$l(\boldsymbol{\varphi}) = \sum_{v=1}^n \sum_{i=1}^k \ln P(Y_{vi} = y_{vi} | \theta_v; \tau_i) \quad (19.14)$$

Log-Likelihood-Funktion des Birnbaum-Modells

In dieser Form sind die Likelihood- und die Log-Likelihood-Funktion unter der Annahme der lokalen stochastischen Unabhängigkeit noch allgemeingültig für alle parametrischen IRT-Modelle. In Abhängigkeit von der Wahl eines konkreten IRT-Modells werden die bedingten Kategorienwahrscheinlichkeiten $P(Y_{vi} = y_{vi} | \theta_v; \tau_i)$ nun durch die jeweiligen Modellgleichungen ersetzt. Nachfolgend soll exemplarisch das zweiparametrische logistische Modell nach Birnbaum (2PL-Modell, ▶ Kap. 16) mit einer eindimensionalen latenten Variablen θ betrachtet werden, bei dem sich die folgende Log-Likelihood-Funktion ergibt:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{v=1}^n \sum_{i=1}^k \ln [P(Y_{vi} = 1 | \theta_v; \tau_i)^{y_{vi}} P(Y_{vi} = 0 | \theta_v; \tau_i)^{(1-y_{vi})}] \\ &= \sum_{v=1}^n \sum_{i=1}^k [y_{vi} \ln P(Y_{vi} = 1 | \theta_v; \tau_i) + (1 - y_{vi}) \ln P(Y_{vi} = 0 | \theta_v; \tau_i)] \\ &= \sum_{v=1}^n \sum_{i=1}^k \left[y_{vi} \ln \left(\frac{\exp[\alpha_i (\theta_v - \beta_i)]}{1 + \exp[\alpha_i (\theta_v - \beta_i)]} \right) \right. \\ &\quad \left. + (1 - y_{vi}) \ln \left(\frac{1}{1 + \exp[\alpha_i (\theta_v - \beta_i)]} \right) \right] \end{aligned} \quad (19.15)$$

Anmerkung: Es ist zu beachten, dass die Log-Likelihood-Funktion des Rasch-Modells einen Spezialfall der Gl. (19.15) darstellt, bei dem die Werte der Itemdiskriminationen für alle Items auf den Wert $\alpha_i = 1$ gesetzt werden.

Zur Bestimmung der Item- und Personenparameter wird nun die Score-Funktion, bestehend aus den ersten partiellen Ableitung $l'(\boldsymbol{\varphi})$ nach den Item- und Personenparametern, benötigt:

$$l'(\boldsymbol{\varphi}) = \frac{\partial l(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \left(\frac{\partial l(\boldsymbol{\varphi})}{\partial \alpha_1}, \dots, \frac{\partial l(\boldsymbol{\varphi})}{\partial \alpha_k}, \frac{\partial l(\boldsymbol{\varphi})}{\partial \beta_1}, \dots, \frac{\partial l(\boldsymbol{\varphi})}{\partial \beta_n} \right)^T \quad (19.16)$$

In Abhängigkeit von der Stichprobengröße n wird die Zahl der zu schätzenden Parameter schnell sehr groß. Bereits bei einem eindimensionalen Birnbaum-Modell enthält die Score-Funktion $(2k + n)$ Elemente. Die ersten Ableitungen nach den Itemparametern für jedes Item i sind

$$\frac{\partial l(\boldsymbol{\varphi})}{\partial \alpha_i} = \left(\frac{\partial l(\boldsymbol{\varphi})}{\partial \alpha_i}, \frac{\partial l(\boldsymbol{\varphi})}{\partial \beta_i} \right)^T, \quad (19.17)$$

mit

$$\begin{aligned} \frac{\partial l(\boldsymbol{\varphi})}{\partial \alpha_i} &= \sum_{v=1}^n [\theta_v - \beta_i] [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] \\ \frac{\partial l(\boldsymbol{\varphi})}{\partial \beta_i} &= -\alpha_i \sum_{v=1}^n [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)]. \end{aligned} \quad (19.18)$$

Die erste Ableitung nach den Personenparametern lautet entsprechend:

$$\frac{\partial l(\boldsymbol{\varphi})}{\partial \theta_v} = \sum_{i=1}^k \alpha_i [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] \quad (19.19)$$

Durch Nullsetzen der Score-Funktion und Lösen der Gleichung hinsichtlich der unbekannten Item- und Personenparameter werden das Maximum der Log-Likelihood-Funktion und somit die ML-Schätzer berechnet. Leider existiert keine analytische Lösung für die Score-Funktion des Rasch- oder Birnbaum-Modells, weshalb numerische Optimierungsverfahren verwendet werden müssen. Beispielsweise sei hier der Newton-Raphson-Algorithmus vorgestellt, bei dem $\hat{\boldsymbol{\varphi}}_{ML}$ iterativ bestimmt wird. Das heißt, nach Einsetzen von Startwerten $(\hat{\boldsymbol{\varphi}}^{(0)})$ in $l'(\boldsymbol{\varphi})$ wird in einem ersten Iterationsschritt $t = 1$ ein neuer Vektor $\hat{\boldsymbol{\varphi}}^{(1)}$ geschätzt, der näher am Maximum der ML-Funktion liegt. Unter Verwendung von $\hat{\boldsymbol{\varphi}}^{(1)}$ wird erneut ein Vektor $\hat{\boldsymbol{\varphi}}^{(2)}$ geschätzt, der die Likelihood wiederum erhöht usw. Allgemein werden die Parameterschätzer $\hat{\boldsymbol{\varphi}}^{(t+1)}$ im Iterationsschritt t , basierend auf den ersten und zweiten Ableitungen der Log-Likelihood-Funktion unter Verwendung von $\hat{\boldsymbol{\varphi}}^{(t)}$, wie folgt berechnet:

$$\hat{\boldsymbol{\varphi}}^{(t+1)} = \hat{\boldsymbol{\varphi}}^{(t)} - l''(\hat{\boldsymbol{\varphi}}^{(t)})^{-1} l'(\hat{\boldsymbol{\varphi}}^{(t)}) \quad (19.20)$$

Die Iterationsschritte werden so lange wiederholt, bis ein vorab festgelegtes Konvergenzkriterium erreicht ist. Der Parametervektor im letzten Iterationsschritt wird als ML-Schätzer $\hat{\boldsymbol{\varphi}}_{ML}$ akzeptiert.

In Gl. (19.20) wird die Hesse-Matrix mit den zweiten partiellen Ableitungen benötigt, die aufgrund der großen Zahl der zu schätzenden Item- und Personenparameter oft recht groß ist. Das Schätzproblem lässt sich jedoch vereinfachen. Die Annahme der Unabhängigkeit der Personen bzw. ihrer Antwortmuster wurde

Rasch-Modell als Spezialfall

Score-Funktion

Numerische Optimierungsverfahren

Konvergenzkriterium

Hesse-Matrix

bereits in der Herleitung der Likelihood-Funktion gemacht. Unter der zusätzlichen Annahme, dass sowohl die Items bzw. deren Parameterschätzungen paarweise unabhängig als auch die Item- und Personenparameterschätzer unabhängig voneinander sind, reduziert sich die Hesse-Matrix zu einer blockdiagonalen Matrix, die sich mathematisch und numerisch erheblich leichter handhaben lässt. Die Parameterschätzer $\hat{\tau}^{(t+1)}$ und $\hat{\theta}^{(t+1)}$ lassen sich so in jedem Iterationsschritt auch item- und personenweise bestimmen.

Identifikation des Modells

Eine notwendige Voraussetzung für die Parameterschätzung ist die Identifikation des Modells. Dazu können entweder die Parameter mindestens eines Items i fixiert (beim Birnbaum-Modell z. B. $\alpha_i = 1, \beta_i = 0$) werden, oder die Personenparameterschätzer werden normiert (beim Birnbaum-Modell z. B. $s(\hat{\theta}) = 1, \bar{\theta} = 0$).

■ Abschließende Anmerkungen zur JML-Schätzung

Vor- und Nachteile der JML-Schätzung

Die JML-Schätzung für Rasch-Modelle ist im Programm WINSTEPS (Linacre 2009) und der frei verfügbaren Variante BIGSTEPS (Linacre und Wright 1993) implementiert. Ein Vorteil der JML-Schätzung ist, dass keinerlei Verteilungsannahmen in Bezug auf die latenten Variablen gemacht werden müssen. Die JML-Schätzung lässt sich auch für mehrdimensionale Modelle erweitern (Rost und Carstensen 2002). Der Nachteil ist, dass die Zahl der zu schätzenden Parameter von der Stichprobengröße abhängt, da die Zahl der Personenparameter steigt. Letztere werden daher auch inzidentelle Parameter genannt. Im Unterschied dazu werden die Itemparameter als strukturelle Parameter bezeichnet, da ihre Zahl unter Verwendung eines bestimmten Modells konstant ist und nicht vom Stichprobenumfang abhängt. Aus der Statistik ist bekannt, dass inkonsistente Schätzungen von strukturellen Parametern resultieren können, wenn inzidentelle Parameter simultan geschätzt werden (Neyman und Scott 1948). Die Inkonsistenz der Parameterschätzung hat nicht zuletzt zu alternativen Schätzverfahren wie der MML-Schätzung (► Abschn. 19.2.6) geführt. Zunächst soll aber die CML-Methode näher vorgestellt werden, die für die einparametrischen Modelle (dichotomes und polytomous Rasch-Modell) entwickelt wurde und mit der das Problem der inzidentellen Personenparameter unter Verwendung suffizienter Statistiken gelöst wird.

19.2.5 Conditional-ML (CML)-Schätzung

19.2.5.1 Suffiziente Statistiken

Summenscore als suffiziente Statistik im Rasch-Modell

Sowohl für das eindimensionale Rasch-Modell als auch für dessen Erweiterungen für ordinale Items wie das Rating-Scale-Modell (RSM) nach Andrich (1978) und das Partial-Credit-Modell (PCM) nach Masters (1982) kann gezeigt werden, dass der Summenscore S_y eine suffiziente Statistik (► Exkurs 19.1) hinsichtlich des Personenparameters θ_v ist. Diese Eigenschaft kann bei der Itemparameterschätzung genutzt werden. Zur Erinnerung: Bei der JML-Schätzung sind die bedingten Antwortmusterwahrscheinlichkeiten $P(Y_{vi} = y_{vi} | \theta_v; \tau_i)$ Teil der Schätzgleichung, die ihrerseits von den unbekannten individuellen Personenparametern θ_v abhängen. Bei der CML-Schätzung (Andersen 1972) werden die bedingten Antwortmusterwahrscheinlichkeiten in Abhängigkeit der Summenscores anstelle der Personenparameter θ_v betrachtet. Aufgrund der Suffizienz des Summenscores können so die Itemparameter konsistent geschätzt werden. Im Weiteren soll die Herleitung der CML-Schätzfunktion für das eindimensionale Rasch-Modell schrittweise erfolgen. Ausgangspunkt bilden die bedingten Antwortmusterwahrscheinlichkeiten bei gegebenen Summenscores.

Exkurs 19.1**Suffizienz**

In der mathematischen Statistik ist der Begriff der Suffizienz von grundlegender Bedeutung für die Inferenz in parametrischen Modellen. Ausgangspunkt der Betrachtung bildet die (mathematische) Stichprobe $\mathbf{Y} = Y_1, \dots, Y_n$. Für all Y_i , mit $i = 1, \dots, n$ gilt, dass sie einer parametrischen Verteilung $f(Y; \boldsymbol{\varphi})$ folgen. Der Parameter $\boldsymbol{\varphi}$ ist dabei die unbekannte, d. h. zu schätzende Größe. Eine Statistik $T(\mathbf{Y})$ ist definiert als eine Stichprobenfunktion. Das heißt, dass die Werte von $T(\mathbf{Y})$ ausschließlich von \mathbf{Y} abhängen. Beispiele sind der Mittelwert, die Varianz, der Anteil richtig gelöster Testaufgaben usw., die anhand der Stichprobe berechnet werden und Schätzer der Parameter $\boldsymbol{\varphi}$ oder Funktionen von $\boldsymbol{\varphi}$ darstellen. Eine beliebige Statistik $T(\mathbf{Y})$ heißt *sufficient hinsichtlich $\boldsymbol{\varphi}$* , wenn gilt:

$$f(\mathbf{Y}|T(\mathbf{Y}); \boldsymbol{\varphi}) = f(\mathbf{Y}|T(\mathbf{Y})) \quad (19.21)$$

Das heißt, die bedingte Verteilung von \mathbf{Y} bei gegebenem $T(\mathbf{Y})$ ist unabhängig vom Parametervektor $\boldsymbol{\varphi}$. Wichtig für die Parameterschätzung sind die Implikationen, die aus der Suffizienz von $T(\mathbf{Y})$ folgen. Eine Statistik stellt üblicherweise eine Zusammenfassung von Informationen aus einer Stichprobe dar. So ist der Summenscore ein einzelner Kennwert mit Bezug zu einem ganzen Antwortmuster, der lediglich angibt, wie viele Items gelöst wurden. Die Reduktion auf einen einzelnen Kennwert bedeutet jedoch auch eine Informationsreduktion, da der Summenscore keine Information mehr darüber enthält, welche Items gelöst wurden. Entscheidend für die Parameterschätzung ist jedoch, dass keine Information hinsichtlich der zu schätzenden Parameter $\boldsymbol{\varphi}$ verloren geht. Die Suffizienz einer Statistik $T(\mathbf{Y})$ gemäß Gl. (19.21) impliziert, dass alle Informationen aus \mathbf{Y} mit Bezug zu $\boldsymbol{\varphi}$ in $T(\mathbf{Y})$ erhalten bleiben.

19.2.5.2 S_v -bedingte Antwortmusterwahrscheinlichkeiten im Rasch Modell

Die allgemeine ML-Funktion (Gl. 19.12) beinhaltet die Antwortmusterwahrscheinlichkeiten $P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\iota})$ für jede Person v bezüglich ihres Antwortmusters \mathbf{y}_v . Nun kann man auch nach der bedingten Wahrscheinlichkeit $P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v, S_v; \boldsymbol{\iota})$ für das Auftreten des Antwortmusters \mathbf{y}_v bei gegebenem Personenparameter θ_v und dem Summenscore S_v fragen. Berücksichtigt man, dass im Rasch-Modell der Itemparametervektor gleich dem Vektor der Itemschwierigkeiten ist ($\boldsymbol{\iota} = \boldsymbol{\beta}$), so lassen sich die S_v -bedingten Antwortmusterwahrscheinlichkeit wie folgt schreiben:

$$P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v, S_v; \boldsymbol{\beta}) = \frac{P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\beta})}{P(S_v = s_v | \theta_v; \boldsymbol{\beta})}. \quad (19.22)$$

Im Nenner steht nun die bedingte Wahrscheinlichkeit des Auftretens des Summenscores S_v , bei gegebener Personenvariable und den Itemschwierigkeiten, die nach dem Satz der totalen Wahrscheinlichkeit gleich der Summe der bedingten Wahrscheinlichkeiten $P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\beta})$ für alle möglichen Antwortmuster \mathbf{y}_v ist, die einen konkreten Summenscore $S_v = s_v$ ergeben:

$$P(S_v = s_v | \theta_v; \boldsymbol{\beta}) = \sum_{\mathbf{y}_v | s_v} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\beta}) \quad (19.23)$$

Satz der totalen Wahrscheinlichkeit

Aufgrund der lokalen stochastischen Unabhängigkeit ergibt sich für die Antwortmusterwahrscheinlichkeit gemäß der Produktregel:

$$\begin{aligned} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\beta}) &= \prod_{i=1}^k P(Y_{vi} = y_{vi} | \theta_v; \beta_1) \\ &= \prod_{i=1}^k \frac{\exp[y_{vi} (\theta_v - \beta_1)]}{1 + \exp(\theta_v - \beta_1)} \end{aligned} \quad (19.24)$$

Dieser Ausdruck lässt sich nun unter Verwendung von Rechenregeln für Exponentialfunktionen wie folgt weiterentwickeln:

$$\begin{aligned} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\beta}) &= \prod_{i=1}^k \frac{\exp[y_{vi} (\theta_v - \beta_1)]}{1 + \exp(\theta_v - \beta_1)} \\ &= \frac{\exp\left(\sum_{i=1}^k y_{vi} \theta_v - \sum_{i=1}^k y_{vi} \beta_i\right)}{\prod_{i=1}^k [1 + \exp(\theta_v - \beta_1)]} \\ &= \frac{\exp\left(\theta_v s_{vy} - \sum_{i=1}^k y_{vi} \beta_i\right)}{\prod_{i=1}^k [1 + \exp(\theta_v - \beta_1)]} \\ &= \frac{\exp(\theta_v s_{vy}) \exp\left(-\sum_{i=1}^k y_{vi} \beta_i\right)}{\prod_{i=1}^k [1 + \exp(\theta_v - \beta_1)]} \end{aligned} \quad (19.25)$$

Setzt man diesen Term zunächst in Gl. (19.23) ein, ergibt sich der folgende Ausdruck für den Nenner von Gl. (19.22):

$$\begin{aligned} P(S_v = s_v | \theta_v; \boldsymbol{\beta}) &= \sum_{y_v|s_v} \frac{\exp(\theta_v s_v) \exp\left(-\sum_{i=1}^k y_{vi} \beta_i\right)}{\prod_{i=1}^k [1 + \exp(\theta_v - \beta_i)]} \\ &= \frac{\exp(\theta_v s_v) \sum_{y_v|s_v} \exp(-\sum_{i=1}^k y_{vi} \beta_i)}{\prod_{i=1}^k [1 + \exp(\theta_v - \beta_i)]} \end{aligned} \quad (19.26)$$

Die letzte Zeile von Gl. (19.25) ist der Zähler von Gl. (19.22). Nach Einsetzen und Vereinfachen ergibt sich letztlich die S_v -bedingte Antwortmusterwahrscheinlichkeit:

$$P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v, S_v; \boldsymbol{\beta}) = \frac{\exp\left(-\sum_{i=1}^k y_{vi} \beta_i\right)}{\sum_{y_v|s_v} \exp\left(-\sum_{i=1}^k y_{vi} \beta_i\right)} \quad (19.27)$$

19.2 · Maximum-Likelihood-Schätzung (ML-Schätzung)

Entscheidend ist nun, dass der Ausdruck auf der rechten Seite der Gleichung keinen Personenparameter θ_v mehr enthält, sodass unmittelbar und ohne weitere Annahmen folgt:

$$P(Y_v = y_v | \theta_v, S_v; \beta) = P(Y_v = y_v | S_v; \beta) \quad (19.28)$$

Das heißt, die Antwortmuster Y_v sind bei gegebenem Summenscore S_v bedingt stochastisch unabhängig von θ_v . Damit ist auch formal gezeigt, dass der Summenscore bezüglich des Personenparameters eine suffiziente Statistik darstellt.

Der Ausdruck im Nenner von Gl. (19.27) wird auch *symmetrische Grundfunktion* oder elementarsymmetrische Funktion $\gamma(S_v; \beta)$ genannt. Die Zahl der Summanden in $\gamma(S_v; \beta)$ ergibt sich aus der Zahl möglicher Antwortmuster, die einen konkreten Summenscore S_v ergeben. Setzt man $S_v = 1$, ergeben sich so viele mögliche Antwortmuster, die zu einem Summenscore von eins führen, wie es Items im Test gibt. Also lautet die symmetrische Grundfunktion:

$$\gamma(1; \beta) = \sum_{i=1}^k \exp(-\beta_i) \quad (19.29)$$

Allgemein gibt es bei einem Summenscore $S_v = s_v$ in einem Test mit k dichotomen Items gemäß dem Binomialkoeffizienten

$$\binom{k}{s_v} = \frac{k!}{(k - s_v)(s_v!)} \quad (19.30)$$

verschiedene Antwortmuster. Entsprechend ist die symmetrische Grundfunktion für $S_v = 2$:

$$\begin{aligned} \gamma(2; \beta) &= \sum_{i=1}^k \sum_{j \neq i} \exp(-\beta_i) \exp(-\beta_j) \\ &= \exp(-\beta_1) \exp(-\beta_2) + \exp(-\beta_1) \exp(-\beta_3) + \dots \\ &\quad + \exp(-\beta_k) \exp(-\beta_{k-1}) \end{aligned} \quad (19.31)$$

und für $S_v = 3$:

$$\begin{aligned} \gamma(3; \beta) &= \sum_{i=1}^k \sum_{j \neq i} \sum_{h \notin \{i, j\}} \exp(-\beta_i) \exp(-\beta_j) \exp(-\beta_h) \\ &= \exp(-\beta_1) \exp(-\beta_2) \exp(-\beta_3) + \dots \\ &\quad + \exp(-\beta_k) \exp(-\beta_{k-1}) \exp(-\beta_{k-2}) \end{aligned} \quad (19.32)$$

usw.

19.2.5.3 Bedingte ML-Funktion zur Parameterschätzung im Rasch-Modell

Bislang wurde nur die S_v -bedingte Antwortmusterwahrscheinlichkeit für eine Person v betrachtet. Zur Parameterschätzung für eine Stichprobe mit n Personen müssen entsprechend alle n Antwortmuster berücksichtigt werden. Bei der CML-Schätzung wird dazu die bedingte Likelihood-Funktion $cL(\varphi)$ bei gegebenen Summenscores aufgestellt. Unter Verwendung von Gl. (19.27) ergibt sich folgende vollständige bedingte Likelihood-Funktion:

$$cL(\beta) = \prod_{v=1}^n \frac{\exp\left(-\sum_{i=1}^k y_{vi} \beta_i\right)}{\gamma(S_v; \beta)} = \frac{\exp\left(-\sum_{v=1}^n \sum_{i=1}^k y_{vi} \beta_i\right)}{\prod_{v=1}^n \gamma(S_v; \beta)} \quad (19.33)$$

Symmetrische Grundfunktion

Bedingte Likelihood-Funktion

Die zugehörige bedingte Log-Likelihood-Funktion lautet:

$$cl(\boldsymbol{\beta}) = - \sum_{v=1}^n \sum_{i=1}^k y_{vi} \beta_i - \sum_{v=1}^n \log[\gamma(S_v; \boldsymbol{\beta})] \quad (19.34)$$

Diese Funktion lässt sich noch etwas weiter vereinfachen. Da bei k dichotomen Items nur $k+1$ Summenscores auftreten können, können die Summen in Gl. (19.34) anstatt über die Personen $v = 1, \dots, n$ auch über Summenscores $S = 1, \dots, k-1$ laufen, sodass gilt:

$$cl(\boldsymbol{\beta}) = - \sum_{i=1}^k n(Y_i = 1) \beta_i - \sum_{S=1}^{k-1} n(S) \log[\gamma(S; \boldsymbol{\beta})] \quad (19.35)$$

Dabei ist $n(Y_i = 1)$ die Zahl der Personen, die bei Item i in Kategorie $Y_i = 1$ geantwortet haben, und $n(S)$ die Zahl der Personen in der Stichprobe, die den Summenscore S erreicht haben. Es ist zu beachten, dass Personen mit den Extremscores $S = 0$ oder $S = k$ bei der CML-Schätzung unberücksichtigt bleiben, da sie keine Information zum gesuchten Modellparameter beitragen.

Gemäß dem allgemeinen Prinzip der ML-Schätzung können die Itemschwierigkeiten geschätzt werden, indem Gl. (19.35) nach den Itemschwierigkeiten abgeleitet wird. Die resultierende Score-Funktion ist dann ein Vektor mit k partiellen ersten Ableitungen nach den Itemschwierigkeiten β_1, \dots, β_k :

$$\frac{\partial cl(\boldsymbol{\beta})}{\partial \beta_i} = -n(Y_i = 1) + \sum_{S=1}^{k-1} n(S) \frac{\exp(-\beta_i) \gamma(S-1; \boldsymbol{\beta}^{(-i)})}{\gamma(S; \boldsymbol{\beta})} \quad (19.36)$$

Dabei ist $\gamma(S-1; \boldsymbol{\beta}^{(-i)})$ die symmetrische Grundfunktion für den Summenscore $S-1$, wobei Item i ausgelassen wird. Entsprechend ist $\boldsymbol{\beta}^{(-i)} = \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k$ der Vektor der Itemschwierigkeiten ohne β_i . Durch Nullsetzen der Score-Funktion können die Schätzer für die Itemschwierigkeiten bestimmt werden. Allerdings sind, wie im Fall der JML-Methode, numerische iterative Verfahren erforderlich. Unter Verwendung des Newton-Raphson-Algorithmus (Gl. 19.20) wird auch die zweite Ableitung von $cl(\boldsymbol{\beta})$ nach den Itemparametern benötigt. Die Diagonalelemente dieser $(k \times k)$ -Matrix können nach Gl. (19.37) berechnet werden und die Elemente außerhalb der Diagonalen nach Gl. (19.38):

$$\begin{aligned} \frac{\partial^2 cl(\boldsymbol{\beta})}{\partial \beta_i^2} &= \sum_{S=1}^{k-1} n(S) \left(\frac{\exp(-\beta_i) \gamma(S-1; \boldsymbol{\beta}^{(-i)})}{\gamma(S; \boldsymbol{\beta})} \right. \\ &\quad \left. - \left[\frac{\exp(-\beta_i) \gamma(S-1; \boldsymbol{\beta}^{(-i)})}{\gamma(S; \boldsymbol{\beta})} \right]^2 \right) \end{aligned} \quad (19.37)$$

$$\frac{\partial^2 cl(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} = \sum_{S=1}^{k-1} n(S) \left(\frac{\exp[-\beta_i \cdot (-\beta_j)] \gamma(S-2; \boldsymbol{\beta}^{(-i,-j)})}{\gamma(S; \boldsymbol{\beta})} - O_{ij} \right) \quad (19.38)$$

Dabei gilt:

$$O_{ij} = \frac{\exp(-\beta_i) \gamma(S-1; \boldsymbol{\beta}^{(-i)})}{\gamma(S; \boldsymbol{\beta})} \cdot \frac{\exp(-\beta_j) \gamma(S-1; \boldsymbol{\beta}^{(-j)})}{\gamma(S; \boldsymbol{\beta})} \quad (19.39)$$

In Gl. (19.38) wird auch die symmetrische Grundfunktion $\gamma(S-2; \boldsymbol{\beta}^{(-i,-j)})$ für den Summenscore $S-2$, benötigt, die unter Auslassung der Items i und j resultiert. Der Vektor $\boldsymbol{\beta}^{(-i,-j)}$ enthält daher nur die Itemschwierigkeiten β_h , mit $h \neq i$ und $h \neq j$.

Score-Funktion

Newton-Raphson-Algorithmus

■ Abschließende Bemerkungen zur CML-Schätzung

In diesem Abschnitt wurde die CML-Methode für den Fall (0, 1)-kodierter dichotomer Items im Rasch-Modell dargestellt. Eine Reihe von Autoren haben dieses Verfahren jedoch auch auf erweiterte Rasch-Modelle für ordinale Items angewendet (Andersen 1972), z. B. auf das RSM (Andrich 1978), das Dispersion-Modell (Andrich 1982), das PCM (Masters und Wright 1997; Wright und Masters 1982) oder das linear-logistische Testmodell (Fischer 1973). Die CML-Methode ist u. a. im Programm WINMIRA (von Davier 2001a) und im R-Paket „eRm“ (Mair und Hatzinger 2007) implementiert.

Da die Personenparameter bei der CML-Methode nicht mitgeschätzt werden, müssen sie bei Bedarf in einem zweiten Schritt unter Verwendung der geschätzten Itemparameter bestimmt werden. Verschiedene Methoden der Personenparameterschätzung, die zur Verfügung stehen, werden im ▶ Abschn. 19.5 beschrieben.

Der Vorteil der CML-Methode ist die konsistente Itemparameterschätzung. Da die latente Variable θ bei der Parameterschätzung unberücksichtigt bleibt, sind auch keine Annahmen hinsichtlich der Verteilung von θ erforderlich. Ein Nachteil des CML-Verfahrens ist, dass ihre Anwendbarkeit auf eindimensionale IRT-Modelle aus der Familie der Rasch-Modelle beschränkt bleibt. Auch der Umgang mit fehlenden Werten ist schwierig. Darüber hinaus wird das Verfahren bei steigender Itemzahl extrem rechenintensiv, da bei der Berechnung der elementaren Grundfunktionen alle möglichen Antwortmuster berücksichtigt werden müssen, die dem Zustandekommen der Summenscores zugrunde liegen können. Bei großen Itemzahlen ist die CML-Schätzung daher nicht mehr anwendbar. Eine Alternative für uni- und multidimensionale ein- und mehrparametrische IRT-Modelle ist die MML-Schätzung, die nachfolgend dargestellt wird.

CML-Schätzung für Rasch-Modelle mit ordinalen Items

Vor- und Nachteile der CML-Schätzung

19.2.6 Marginal-ML (MML)-Schätzung

Die MML-Schätzung erlaubt ebenfalls die konsistente Itemparameterschätzung von ein- und mehrparametrischen IRT-Modellen und wurde von Bock und Lieberman (1970) sowie Bock und Aitkin (1981) beschrieben. Das Problem der inzidentellen Personenparameter wird dadurch gelöst, dass die Verteilung $f(\theta)$ der latenten Variablen θ modelliert wird. Unter Annahme einer parametrischen Verteilung für $f(\theta)$ werden lediglich die Verteilungsparameter anstatt der n individuellen Personenparameter geschätzt.

Modellierung der Verteilung der latenten Variable

19.2.6.1 Gemeinsame Verteilung manifester und latenter Variablen

Ausgangspunkt der Herleitung der MML-Schätzung bildet die gemeinsame Verteilung $f(Y, \theta; \iota)$ der manifesten und latenten Variablen im Messmodell unter Annahme eines bestimmten IRT-Modells, das durch den Parametervektor ι repräsentiert ist. Eine solche gemeinsame Verteilung lässt sich in folgender Weise faktorisieren:

$$f(Y, \theta; \iota) = P(Y|\theta; \iota) f(\theta) \quad (19.40)$$

Für eine bestimmte Datenmatrix $Y = y$ mit den Zeilen y_v , mit $v = 1, \dots, n$, folgt wiederum unter der Annahme, dass die Antwortmuster der Personen voneinander unabhängig sind:

$$f(Y, \theta; \iota) = \prod_{v=1}^n P(Y_v = y_v | \theta_v; \iota) f(\theta_v) \quad (19.41)$$

Um eine Schätzfunktion ohne die individuellen Werte θ_v ableiten zu können, wird nun über die Verteilung der latenten Personenvariable integriert, woraus die margi-

Marginale ML-Funktion

nale ML-Funktion $mL(\tau)$ resultiert:

$$P(\mathbf{Y}; \tau) = \prod_{v=1}^n \int_{\boldsymbol{\theta}} P(\mathbf{Y}_v = \mathbf{y}_v | \boldsymbol{\theta}; \tau) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \prod_{v=1}^n P(\mathbf{Y}_v = \mathbf{y}_v; \tau) = mL(\tau) \quad (19.42)$$

In dieser Gleichung werden nur noch die unbedingten Antwortmusterwahrscheinlichkeiten unter einem IRT-Modell als Funktion der unbekannten Parameter τ betrachtet. Die marginale Log-Likelihood-Funktion

$$ml(\tau) = \sum_{v=1}^n \ln P(\mathbf{Y}_v = \mathbf{y}_v; \tau) \quad (19.43)$$

Score-Funktion

kann wiederum nach den gesuchten Parametern τ abgeleitet werden. Die resultierende Score-Funktion lautet in allgemeiner Form wie folgt:

$$\frac{\partial ml(\tau)}{\partial \tau} = \sum_{v=1}^n \frac{\partial \ln P(\mathbf{Y}_v = \mathbf{y}_v; \tau)}{\partial \tau} \quad (19.44)$$

Fisher-Identität und individuelle A-posteriori-Verteilungen

Aufgrund der sog. „Fisher-Identität“ (Fisher 1925) lässt sich Gl. (19.44) als Summe der individuellen bedingten Erwartungswerte der ersten Ableitungen der Antwortmusterwahrscheinlichkeiten für $\mathbf{Y}_v = \mathbf{y}_v$ schreiben. Dabei wird über die individuellen A-posteriori-Verteilungen $f(\boldsymbol{\theta} | \mathbf{Y}_v = \mathbf{y}_v; \tau)$ der latenten Variablen bei gegebenem Antwortmuster der Person v integriert:

$$\frac{\partial ml(\tau)}{\partial \tau} = \sum_{v=1}^n \int_{\boldsymbol{\theta}} \frac{\partial \ln P(\mathbf{Y}_v = \mathbf{y}_v | \boldsymbol{\theta}; \tau)}{\partial \tau} f(\boldsymbol{\theta} | \mathbf{Y}_v = \mathbf{y}_v; \tau) d\boldsymbol{\theta} \quad (19.45)$$

Unter der Annahme der lokalen stochastischen Unabhängigkeit kann die partielle Ableitung der Antwortmusterwahrscheinlichkeit durch die Summe der partiellen Ableitungen der Modellgleichungen der einzelnen Items wie folgt geschrieben werden:

$$\frac{\partial ml(\tau)}{\partial \tau} = \sum_{v=1}^n \int_{\boldsymbol{\theta}} \sum_{i=1}^k \frac{\partial}{\partial \tau_i} \ln P(Y_{vi} = y_{vi} | \boldsymbol{\theta}; \tau_i) f(\boldsymbol{\theta} | \mathbf{Y}_v = \mathbf{y}_v; \tau) d\boldsymbol{\theta} \quad (19.46)$$

19.2.6.2 Anwendung des Bayes-Theorems in der MML-Schätzung

Anwendung des Bayes-Theorems

Bei Gln. (19.45) und (19.46) wird vom Bayes-Theorem Gebrauch gemacht, auch wenn die MML-Schätzung *kein* Bayes'sches Schätzverfahren ist! In der Bayes-Statistik werden auch Parameter als Zufallsvariablen aufgefasst, die eine Verteilung haben (► Abschn. 19.3). Die Verteilung eines Parameters spiegelt die Unsicherheit in Bezug auf die zu schätzende unbekannte Größe wider. Die individuellen A-posteriori-Verteilungen $f(\boldsymbol{\theta} | \mathbf{Y}_v = \mathbf{y}_v; \tau)$ sind die bedingte Wahrscheinlichkeitsverteilung des Personenparameters der Personen $v = 1, \dots, n$ unter der Bedingung ihrer jeweils beobachteten Antwortmuster \mathbf{y}_v . Weniger formal ausgedrückt bedeutet das, dass bestimmte Ausprägungen auf der latenten Variablen bei einem konkreten Antwortmuster wahrscheinlicher sind als andere. Gemäß des Bayes-Theorems gilt:

$$f(\boldsymbol{\theta} | \mathbf{Y}_v = \mathbf{y}_v; \tau) = \frac{P(\mathbf{Y}_v = \mathbf{y}_v | \boldsymbol{\theta}; \tau) f(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Y}_v = \mathbf{y}_v; \tau) f(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad (19.47)$$

Der Zähler ist das Produkt aus der bedingten Antwortmusterwahrscheinlichkeit bzw. der individuellen Antwortmuster-Likelihood $P(Y_v = y_v | \theta; \iota)$ und der sog. „A-priori-Verteilung“ $f(\theta)$ der latenten Variablen θ , die für alle Personen der Stichprobe gleich ist. Inhaltlich bedeutet dies, dass alle Personen der Stichprobe zufällig aus einer homogenen Population mit der Verteilung $f(\theta)$ gezogen wurden. Bei der Anwendung wird oft eine parametrische Verteilungsannahme getroffen, z. B. wird angenommen, dass θ multivariat normal verteilt ist mit einem Erwartungswertvektor $E(\theta)$ und einer Varianz-Kovarianz-Matrix Σ_θ . Die Verteilungsparameter in $E(\theta)$ und Σ_θ können zusätzlich zu den Itemparametern zu schätzende Größen in der MML-Schätzung sein. Oft werden sie aber auch zur Modellidentifikation oder aus theoretischen Überlegungen auf bestimmte Werte fixiert.

Bock und Lieberman (1970) konnten schließlich zeigen, dass die Score-Funktion der MML-Schätzung bezüglich der Itemparameter vergleichbar mit der Score-Funktion der JML-Schätzung ist (vgl. Gl. 19.18). Allerdings beinhalten die ersten partiellen Ableitungen von $ml(\iota)$ nach den Itemparametern die Integrale der n individuellen A-posteriori-Verteilungen. Betrachtet man wiederum beispielhaft das Birnbaum-Modell mit einer eindimensionalen latenten Variablen θ ergeben sich die folgenden ersten Ableitungen für die Itemparameter von Item i :

$$\begin{aligned} \frac{\partial ml(\varphi)}{\partial \alpha_i} &= \sum_{v=1}^n \int_{\theta} [\theta_v - \beta_i] [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] f(\theta | Y_v = y_v; \iota) d\theta \\ \frac{\partial ml(\varphi)}{\partial \beta_i} &= -\alpha_i \sum_{v=1}^n \int_{\theta} [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] f(\theta | Y_v = y_v; \iota) d\theta \end{aligned} \quad (19.48)$$

Die Integrale in Gln. (19.45) bis (19.48) sind jedoch selbst unter Annahme einer bekannten parametrischen Verteilung der latenten Variablen θ (z. B. Normalverteilung) zumeist nicht analytisch zu berechnen, da keine elementare Stammfunktion hergeleitet werden kann. Numerische Integrationsverfahren erlauben jedoch eine beliebig genaue Approximation. Eine Möglichkeit ist die Verwendung von sog. „Quadraturformeln“. Dabei wird θ durch eine diskrete Variable X_θ ersetzt, deren Wahrscheinlichkeitsfunktion annähernd der Verteilungsfunktion von θ entspricht. Aufgrund der diskreten Natur von X_θ wird aus einem Integral eine Summe über die Q verschiedenen Werte $X_\theta^{(q)}$ (mit $q = 1, \dots, Q$). Unter Verwendung einer Quadraturverteilung kann Gl. (19.48) in der folgenden Form geschrieben werden:

$$\begin{aligned} \frac{\partial ml(\varphi)}{\partial \alpha_i} &= \sum_{v=1}^n \sum_{q=1}^Q \left[X_\theta^{(q)} - \beta_i \right] \left[y_{vi} - P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \\ &\quad P(X_\theta^{(q)} | Y_v = y_v; \iota) \\ \frac{\partial ml(\varphi)}{\partial \beta_i} &= -\alpha_i \sum_{v=1}^n \sum_{q=1}^Q \left[y_{vi} - P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \\ &\quad P(X_\theta^{(q)} | Y_v = y_v; \iota) \end{aligned} \quad (19.49)$$

In Abb. 19.1 ist beispielhaft eine Quadraturverteilung mit $Q = 9$ Quadraturpunkten (sog. Stützstellen) dargestellt, die eine Standardnormalverteilung approximiert. (Anmerkung: Es werden üblicherweise 15 oder mehr Quadraturpunkte verwendet, um eine gute Approximation des Integrals zu erreichen.) Die grau dargestellten Flächenanteile die den einzelnen Quadraturpunkten $X_\theta^{(1)}, \dots, X_\theta^{(Q)}$ zugeordnet sind, entsprechen ihren A-priori-Wahrscheinlichkeiten $P(X_\theta^{(q)})$, die als Gewichte bei der Integration dienen (Gl. 19.50).

Parametrische Verteilungsannahme

Numerische Integration und Quadraturformeln

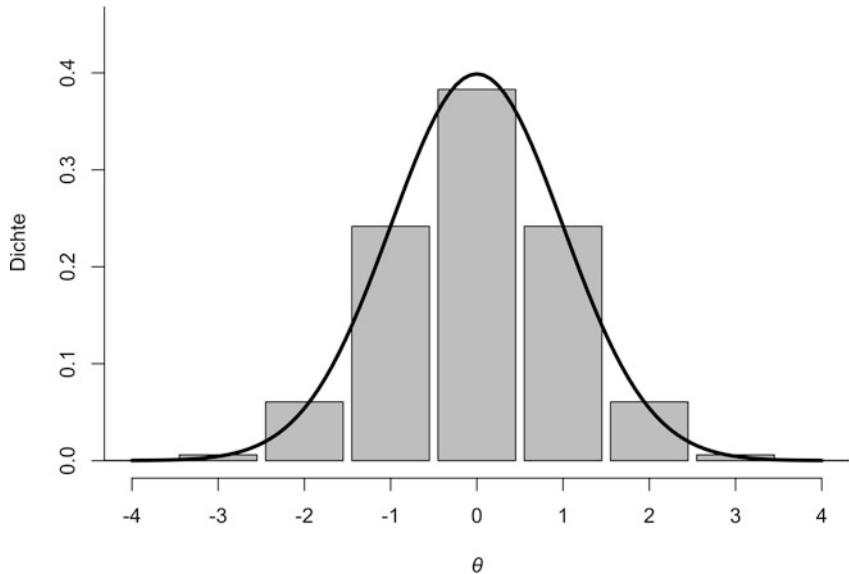


Abb. 19.1 Diskrete Quadraturverteilung zur Approximation der Dichtefunktion einer Standardnormalverteilung. Die grauen Flächenanteile entsprechen den Gewichten der einzelnen Quadraturpunkte bei der näherungsweisen Berechnung von Integralen über die Verteilung von θ

Die Berechnungen erfordern somit die A-posteriori-Wahrscheinlichkeiten $P(X_\theta^{(q)}|Y_v = \mathbf{y}_v; \boldsymbol{\iota})$, mit der die einzelnen Personen v mit ihren jeweils beobachteten Antwortmustern \mathbf{y}_v die Ausprägung $X_\theta^{(q)}$ aufweisen. Diese Wahrscheinlichkeit muss für jede Person und jeden Quadraturpunkt $X_\theta^{(q)}$, mit $q = 1, \dots, Q$, gemäß der folgenden Formel berechnet werden:

$$P\left(X_\theta^{(q)}|Y_v = \mathbf{y}_v; \boldsymbol{\iota}\right) = \frac{P\left(Y_v = \mathbf{y}_v|X_\theta^{(q)}; \boldsymbol{\iota}\right) P\left(X_\theta^{(q)}\right)}{\sum_{q=1}^Q P\left(Y_v = \mathbf{y}_v|X_\theta^{(q)}; \boldsymbol{\iota}\right) P\left(X_\theta^{(q)}\right)} \quad (19.50)$$

Diese Formel ist nichts anderes als die Anwendung des Bayes-Theorems in Gl. (19.47) auf die diskrete Quadraturverteilung. Die MML-Schätzung für die Itemparameter nach Gl. (19.49) ist jedoch numerisch anspruchsvoll und bleibt daher auf geringe Itemzahlen beschränkt.

19.2.6.3 MML-Schätzung unter Verwendung des Expectation-Maximization-Algorithmus (EM-Algorithmus)

Um die MML-Schätzung auch für Tests mit einer großen Zahl an Items anwendbar zu machen, wurde die ursprünglich von Bock und Lieberman (1970) vorgeschlagene Schätzprozedur von Bock und Aitkin (1981) weiterentwickelt. Der entwickelte Expectation-Maximation (EM)-Algorithmus liegt – mit kleineren Adaptationen – den meisten derzeitigen Implementierungen des MML-Verfahrens zugrunde. Zunächst zeigten die Autoren, dass sich durch Vertauschung der Summenzeichen und Umstellen der Gl. (19.49) folgende Ausdrücke für die ersten partiellen Ableitungen

nach α_i und β_i ergeben:

$$\begin{aligned} \frac{\partial m l(\varphi)}{\partial \alpha_i} &= \sum_{q=1}^Q \left[X_\theta^{(q)} - \beta_i \right] \left[n(Y_i = 1 | X_\theta^{(q)}) \right. \\ &\quad \left. - n(X_\theta^{(q)}) P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \\ \frac{\partial m l(\varphi)}{\partial \beta_i} &= \sum_{q=1}^Q \left[n(Y_i = 1 | X_\theta^{(q)}) - n(X_\theta^{(q)}) P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \end{aligned} \quad (19.51)$$

Dabei ist $n(X_\theta^{(q)})$ die erwartete Zahl der Personen der Stichprobe, die den Wert $X_\theta^{(q)}$ der Quadraturverteilung von θ aufweisen, und $n(Y_i = 1 | X_\theta^{(q)})$ die erwartete Zahl der Personen mit dem Wert $X_\theta^{(q)}$, die bei Item i in Kategorie $Y_i = 1$ antworten. Beide Häufigkeiten lassen sich einfach anhand der A-posteriori-Wahrscheinlichkeiten (Gl. 19.50) berechnen:

$$\begin{aligned} n(X_\theta^{(q)}) &= \sum_{v=1}^N P(X_\theta^{(q)} | Y_v = y_v; \iota) \\ \left(Y_i = 1 | X_\theta^{(q)} \right) &= \sum_{v=1}^N y_{vi} P(X_\theta^{(q)} | Y_v = y_v; \iota) \end{aligned} \quad (19.52)$$

Der EM-Algorithmus ist ebenfalls ein iteratives Verfahren zur Bestimmung von ML-Schätzern unter Berücksichtigung unbeobachteter Variablen. Jeder Iterationsschritt des EM-Algorithmus besteht wiederum aus zwei Teilschritten, die namengebend für den Algorithmus waren: erstens dem E-Schritt, bei dem der Erwartungswert der marginalen Log-Likelihood berechnet wird, und zweitens dem M-Schritt, bei dem der Erwartungswert der marginalen Log-Likelihood als Zielfunktion hinsichtlich der gesuchten Parameter maximiert wird.

Angewendet auf die MML-Schätzung in der IRT beinhaltet der *E-Schritt* in Iteration t folgende Berechnungen:

- Berechnung der individuellen A-posteriori-Wahrscheinlichkeiten $P(X_\theta^{(q)} | Y_v = y_v; \iota^{(t)})$ nach Gl. (19.50) für alle Personen $v = 1, \dots, n$ und alle Quadraturpunkte $q = 1, \dots, Q$ unter Verwendung der vorläufigen Itemparameterschätzer $\iota^{(t)}$
- Berechnung der erwarteten Häufigkeiten $n(X_\theta^{(q)})$ und $n(Y_i = 1 | X_\theta^{(q)})$ für alle Quadraturpunkte $q = 1, \dots, Q$ und alle Items $i = 1, \dots, k$ nach Gl. (19.52)

Iteratives Verfahren

E-Schritt des EM Algorithmus

Im *M-Schritt* werden die ersten partiellen Ableitungen nach den gesuchten Itemparametern (Gl. 19.51) gleich null gesetzt. Die resultierende Gleichung ist wiederum nicht analytisch lösbar, sodass auch im M-Schritt numerische Verfahren wie das Newton-Raphson-Verfahren (Gl. 19.20) verwendet werden. Dabei werden häufig nur eine oder zwei Iterationen des Newton-Raphson-Algorithmus innerhalb eines M-Schritts zugelassen, um $\iota^{(t+1)}$ zu erhalten. Dann erfolgt zunächst wieder ein erneuter E-Schritt usw. Auch der EM-Algorithmus stoppt bei Erreichen eines vorab festgelegten Konvergenzkriteriums. Oft wird die relative Differenz zwischen den Werten der marginalen Log-Likelihood zweier aufeinanderfolgender Iterationsschritte gewählt (z. B. $|ml(\iota^{(t+1)}) - ml(\iota^{(t)})| / |ml(\iota^{(t+1)})| < 10^{-5}$). Die Modellparameterschätzer des letzten Iterationsschritts werden als Punktschätzer $\hat{\iota}_{ML}$ verwendet.

M-Schritt des EM-Algorithmus

19.2.6.4 Anmerkungen zur MML-Schätzung

IRT-Modelle können auf verschiedene Weise identifiziert werden. Für zwei- und mehrparametrische Modelle können entweder Verteilungsparameter der

Mehrgruppen- und Mischverteilungs-IRT-Modelle

Erweiterungen und Vorteile der MML-Schätzung

Nachteil des MML-Verfahrens: rechenintensive numerische Integration

19

Subjektive vs. objektive/frequentistische Wahrscheinlichkeit

latenten Personenvariablen fixiert werden (z. B. $E(\theta_1), \dots, E(\theta_h) = 0$ und $Var(\theta_1), \dots, Var(\theta_h) = 1$), sodass alle Itemparameter frei geschätzt werden, oder es werden die Itemparameter mindestens eines Items je latenter Variable fixiert (z. B. $\alpha_1 = 1$ und $\beta_1 = 0$). In dem Fall werden die Erwartungswerte und Varianzen der latenten Variablen frei geschätzt. Die marginale Log-Likelihood-Funktion muss dann auch hinsichtlich dieser unbekannten Größen maximiert werden (Mislevy 1984). Dabei muss nicht zwingend von einer homogenen Verteilung der latenten Variablen ausgegangen werden. Die MML-Schätzung ist auch für Mehrgruppen-IRT-Modelle anwendbar, die Gruppenunterschiede latenter Variablen berücksichtigen und schätzen können (Baker und Kim 2004; Bock und Zimowski 1997). Die Gruppierungsvariable kann dabei selbst latent sein, sodass die Zugehörigkeit einer Person zu einer sog. „latenten Klasse“ ebenfalls unbekannt ist. In einem solchen Fall handelt es sich um Mischverteilungs-IRT-Modelle (von Davier und Rost 2007). Sowohl bei Mehrgruppen- als auch Mischverteilungs-IRT-Modellen werden im MML-Verfahren gruppen- bzw. klassenspezifische Verteilungsparameter geschätzt.

Oftmals sind bei der Anwendung nicht nur die Parameter des Messmodells und die Verteilungsparameter von Interesse, sondern auch die Zusammenhänge der latenten Variablen mit relevanten Kovariaten $Z = Z_1, \dots, Z_W$ (sozioökonomischer Status, Motivation usw.). Das MML-Verfahren erlaubt die Erweiterung des Messmodells um latente Regressionsmodelle $E(\theta|Z)$, deren Parameter ebenfalls simultan mit den Itemparametern geschätzt werden können. Auch die Annahme einer Normalverteilung für die latenten Variablen ist zwar üblich, aber nicht zwingend erforderlich. Alternative parametrische und selbst nicht parametrische Verteilungen können bei der MML-Schätzung mitmodelliert werden. Ein Beispiel sind die Spline-basierten Dichteschätzer, die in die MML-Schätzung integriert werden können und ganz ohne parametrische Verteilungsannahme auskommen (Woods 2006).

Ein Nachteil des MML-Verfahrens bzw. des EM-Algorithmus ist, dass die Standardfehlerberechnung schwieriger ist als bei der JML- oder der CML-Schätzung. Hierzu werden verschiedene Verfahren in der Literatur beschrieben (z. B. Yuan et al. 2014). Ein weiterer Nachteil der MML-Schätzung ist die begrenzte Zahl latenter Personenvariablen die gleichzeitig in mehrdimensionalen IRT-Modellen berücksichtigt werden kann. Aufgrund der rechenintensiven numerischen Integration stellen bereits fünf latente Variablen in einem mIRT-Modell eine erhebliche rechen-technische Herausforderung dar. Weiterentwickelte MML-Schätzer, die Monte-Carlo-Techniken einbeziehen (Cai 2010) und einige alternative Bayes'sche Schätzverfahren stellen hier Alternativen dar.

19.3 Bayes'sche Schätzverfahren

19.3.1 Grundlegendes

Die Bayes-Statistik ist ein alternativer Ansatz zur Bestimmung der statistischen Inferenz unbekannter zu schätzender Größen anhand beobachteter Daten, der sich grundlegend von der klassischen Statistik unterscheidet. Alle bislang betrachteten Varianten der ML-Schätzungen gehören zur klassischen Statistik, bei der die Item- und Personenparameter zwar unbekannte, aber fixe Größen sind. In der Bayes-Statistik werden die unbekannten Parameter hingegen als Zufallsvariablen betrachtet, die ihrerseits eine Verteilung haben. Das bedeutet, dass es Wertebereiche eines Parameters gibt, die wahrscheinlicher oder unwahrscheinlicher sind als andere. Unbedingt zu beachten ist dabei, dass eine solche Aussage nur Sinn ergibt, so lange man den gesuchten Parameter nicht kennt! Es handelt sich somit um einen *subjektiven Wahrscheinlichkeitsbegriff*, der in der Bayes-Statistik Anwendung findet und der sich grundlegend vom frequentistischen Wahrscheinlichkeitskonzept

19.3 · Bayes'sche Schätzverfahren

der klassischen Statistik unterscheidet. Wahrscheinlichkeiten quantifizieren in der Bayes-Statistik den Grad an Unsicherheit bzw. Ungewissheit eines Wissenschaftlers in Bezug auf einen unbekannten Sachverhalt oder eine unbekannte Größe. In der klassischen Statistik wird davon ausgegangen, dass einem jeweils betrachteten Zufallsexperiment (z. B. Münzwurf oder Würfelwurf) eine bestimmte, wenn auch oft unbekannte Wahrscheinlichkeit zugrunde liegt. Die Wahrscheinlichkeit ist in dieser Betrachtung insofern objektiv, als sie unabhängig von irgendeiner subjektiven Ungewissheit oder Sichtweise ist. Würde man nun das Zufallsexperiment unendlich oft in identischer Weise wiederholen, so wäre die relative Häufigkeit eines bestimmten Ereignisses (z. B. die Zahl Sechs beim Würfelwurf) gleich dessen Wahrscheinlichkeit – daher auch frequentistischer Wahrscheinlichkeitsbegriff.

In diesem Abschnitt soll jedoch nicht der Unterschied zwischen klassischer und Bayes'scher Statistik erörtert werden und erst recht nicht die Frage, welche Statistik die vermeintlich bessere ist. Diese Kontroverse (*Bayesian-Frequentist Debate*) wird seit über 100 Jahren nicht ganz emotionslos und leider oft kontraproduktiv geführt. Dem interessierten Leser steht jedoch eine reichhaltige Literatur zu dieser Thematik zur Verfügung (z. B. Tschirk 2014). Hier sollen Bayes'sche Ansätze vielmehr als eine wertvolle Erweiterung und Ergänzung von Verfahren zur Parameterschätzung in der IRT dargestellt werden.

Bayesian-Frequentist Debate

19.3.2 Bayes-Inferenz auf Basis der A-posteriori-Verteilung

Bei der ML-Schätzung bildet die bedingte Wahrscheinlichkeitsfunktion (ML-Funktion $L(\varphi)$) des Zustandekommens der beobachteten Daten $\mathbf{Y} = \mathbf{y}$ unter Annahme eines Modells mit dem Parametervektor φ die Grundlage der statistischen Inferenz. Konträr dazu basiert in der Bayes-Statistik die statistischen Inferenz auf der bedingten Wahrscheinlichkeitsverteilung $f(\varphi|\mathbf{Y} = \mathbf{y})$ des Parametervektors φ bei gegebenen Daten $\mathbf{Y} = \mathbf{y}$. Diese sog. „A-posteriori-Verteilung“ lässt sich für eine Datenmatrix $\mathbf{Y} = \mathbf{y}$ mit kategorialen Itemantworten gemäß dem namensgebenden Bayes-Theorem wie folgt schreiben:

$$f(\varphi|\mathbf{Y} = \mathbf{y}) = \frac{f(\mathbf{Y} = \mathbf{y}, \varphi)}{P(\mathbf{Y} = \mathbf{y})} = \frac{P(\mathbf{Y} = \mathbf{y}|\varphi) f(\varphi)}{\int_{\Omega_\varphi} P(\mathbf{Y} = \mathbf{y}|\varphi) f(\varphi) d\varphi} \quad (19.53)$$

Dabei steht $f(\mathbf{Y} = \mathbf{y}, \varphi)$ für die gemeinsame Dichtefunktion des Parametervektors φ und der manifesten Variablen \mathbf{Y} an der Stelle $\mathbf{Y} = \mathbf{y}$. Im rechten Term der Gl. (19.53) wurde diese gemeinsame Verteilung gemäß der Rechenregeln für bedingte Wahrscheinlichkeiten faktorisiert. Der erste Faktor $P(\mathbf{Y} = \mathbf{y}|\varphi)$ ist die Wahrscheinlichkeit des Zustandekommens der beobachteten Daten unter Annahme des betrachteten Modells. In der IRT ist das die Datenmatrix $\mathbf{Y} = \mathbf{y}$ mit den n Antwortmustern bei gegebenen Item- und Personenparametern in $\varphi = (\iota_1, \dots, \iota_k, \theta_1, \dots, \theta_n)$. Dieser Faktor ist also gleich der ML-Funktion nach Gl. (19.10). Der zweite Faktor $f(\varphi)$ wird als A-priori-Verteilung bezeichnet. In der Bayes-Statistik repräsentiert die A-priori-Verteilung das *Vorwissen* hinsichtlich φ , während die Likelihood-Funktion die Information aus den empirischen Daten hinsichtlich φ enthält. Der Nenner in Gl. (19.53) ist lediglich eine Normierungskonstante, die notwendig ist, damit die A-posteriori-Verteilung auch die Kriterien einer Wahrscheinlichkeitsverteilung erfüllt (z. B. das Integral der A-posteriori-Verteilung ist gleich eins).

A-posteriori-Verteilung

Bayes-Theorem

A-priori-Verteilung der Modellparameter

Definition

Die **A-posteriori-Verteilung** ist das normierte Produkt aus der Likelihood-Funktion $L(\varphi)$ und der A-priori-Verteilung $f(\varphi)$, wobei die unbedingte Wahrscheinlichkeit des Zustandekommens der beobachteten Daten $P(Y = y)$ die Normierungskonstante ist.

A-posteriori-Verteilung als Mischung aus Vorwissen und empirischer Information

Statistiken und Intervallschätzer der A-posteriori-Verteilung

Gewichtung der Vorinformation durch die A-priori-Verteilung

Die A-posteriori-Verteilung stellt also eine Mischung aus Vorwissen und empirisch gewonnener Information bezüglich der zu schätzenden Parameter in φ dar.

Obwohl die uni- oder multivariate A-posteriori-Verteilung in ihrer Gesamtheit das zentrale Ergebnis einer Bayes'schen statistischen Analyse ist, werden häufig nur Kennwerte der A-posteriori-Verteilung zur Ergebnisdarstellung verwendet. So sind der Erwartungswert (*expected a posteriori*, EAP), der Modus (*maximum a posteriori*, MAP) oder der Median der A-posteriori-Verteilung als Punktschätzer gebräuchlich. Anstelle von Standardfehlern wird in der Bayes-Statistik die Standardabweichung der A-posteriori-Verteilung eines Parameters verwendet. Diese quantifiziert die Unsicherheit in Bezug auf die Parameterschätzung aufgrund der empirischen Information aus der Stichprobe unter Berücksichtigung des Vorwissens in Form der A-priori-Verteilung. Für Intervallschätzungen werden sog. „Glaubwürdigkeits- oder Kredibilitätsintervalle“ anstelle von Konfidenzintervallen verwendet, die den Bereich angeben, in dem ein Parameter mit einer Wahrscheinlichkeit von $(1 - \alpha)$ liegt. Wählt man für die Wahrscheinlichkeit $\alpha = 0.05$ resultiert entsprechend das 95 %-Kredibilitätsintervall.

19.3.3 Spezifikation der A-priori-Verteilung

Die A-priori-Verteilung repräsentiert Annahmen oder das Vorwissen aufgrund von empirischen Voruntersuchungen, Expertenurteilen oder einfachen Plausibilitätsüberlegungen, die in eine Bayes-statistische Analyse eingehen. Grundsätzlich kann zwischen informativen und nicht informativen bzw. vagen A-priori-Verteilungen unterschieden werden. Bei *informativen A-priori-Verteilungen* wird das Vorwissen bei der Analyse stark gewichtet. Aus Bayes'scher Sicht ist das beispielsweise bei Replikationsstudien sinnvoll, um Informationen aus vorangegangenen Studien zu berücksichtigen. Mit nicht informativen oder *vagen A-priori-Verteilungen* wird das Vorwissen nicht oder nur sehr gering gewichtet, sodass die A-posteriori-Verteilung und damit die statistische Inferenz weitestgehend von der empirischen Information aus den Stichprobendaten abhängt. Das ist vor allem dann sinnvoll, wenn statt Vorwissen lediglich Annahmen unter Unsicherheit gemacht werden können. Wie informativ eine A-priori-Verteilung ist, d. h. mit welchem Gewicht die Vorinformation in die Parameterschätzung eingeht, wird durch die Wahl der Verteilungsparameter der A-priori-Verteilung gesteuert.

Leider gibt es keine eindeutigen und exakten Regeln dafür, wie man bestehende Vorwissen in eine konkrete A-priori-Verteilung übersetzt. Daraus resultiert eine gewisse Willkür bei der Datenanalyse, die von Kritikern des Bayes'schen Ansatzes als mangelnde Durchführungsobjektivität kritisiert wird. Dazu kommt, dass – unabhängig von der Entscheidung ob informativ oder vage – meist mehrere alternative Arten parametrischer Verteilungen als A-priori-Verteilung eines Parameters infrage kommen. So können für Itemdiskriminationen die Normalverteilung, die logarithmierte Normalverteilung (Log-Normalverteilung), aber auch die Gamma- oder eine Gleichverteilung in einem bestimmten Intervall gewählt werden. Es ist in konkreten Anwendungen oft nicht eindeutig, welche A-priori-Verteilung angemessen ist bzw. welche das Vorwissen am besten repräsentiert. Wohl aber gibt es Erfahrungswissen und daraus abgeleitete Empfehlungen für die Wahl von A-priori-Verteilungen für verschiedene Modellklassen und deren Parameter. Beispiele für gebräuchliche A-priori-Verteilungen in der IRT sind im ► Exkurs 19.2 dargestellt.

Exkurs 19.2

A-priori-Verteilungen am Beispiel des 2PL-Modells nach Birnbaum

Im eindimensionalen 2PL-Modell nach Birnbaum (s. ▶ Kap. 16) werden die Itemdiskriminationen α_i und die Itemschwierigkeiten β_i , sowie der Personenparameter θ_v geschätzt. Unter Verwendung Bayes'scher Schätzverfahren müssen für alle drei Größen A-priori-Verteilungen bestimmt werden. Für den Itemdiskriminationsparameter wurden verschiedene A-priori-Verteilungen vorgeschlagen, die von positiven Itemdiskriminationen $\alpha_i > 0$ ausgehen. Das setzt voraus, dass alle Items die gleiche Kodierung (z. B. 0 = „nicht gelöst“, 1 = „gelöst“) bzw. die gleiche Polung aufweisen müssen. Um negative Parameterschätzer zu vermeiden, kann eine A-priori-Verteilung gewählt werden, deren Dichte für negative Werte gleich null ist. Gemäß der sog. „Cromwell'schen Regel“ folgt dann unmittelbar, dass die A-posteriori-Dichte für Werte $\alpha_i < 0$ ebenfalls null ist. Ein Beispiel ist die *trunkierte (gestutzte), Normalverteilung*:

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha) I(\alpha_i \geq 0) \quad (19.54)$$

Dabei ist $I(\alpha_i \geq 0)$ eine Indikatorfunktion mit $I(\alpha_i \geq 0) = 1$, falls $\alpha_i \geq 0$, und $I(\alpha_i \geq 0) = 0$, falls $\alpha_i < 0$. Die Standardabweichung σ_α der A-priori-Verteilung bestimmt die Gewichtung der Vorinformation. Je kleiner die Standardabweichung σ_α gewählt wird, desto informativer ist die A-priori-Verteilung. Die Werte α_i , die im Bereich nahe des Erwartungswertes μ_α der A-priori-Verteilung liegen, haben a priori die höchste Wahrscheinlichkeit, d. h., durch die Wahl von μ_α spezifiziert man vorab den theoretisch angenommenen Wert eines gesuchten Parameter.

Die *Log-Normalverteilung* wird ebenfalls als A-priori-Verteilung für α_i verwendet, da sie ebenfalls nur positive Werte erlaubt:

$$\alpha_i \sim \ln N\left(\exp(\mu_{ln} + 0.5\sigma_{ln}^2), \sqrt{\exp(2\mu_{ln} + \sigma_{ln}^2)[\exp(\sigma_{ln}^2) - 1]} \right) \quad (19.55)$$

Die Verwendung der Log-Normalverteilung als A-priori-Verteilung für α_i ist gleichbedeutend mit der Annahme einer A-priori-Normalverteilung der logarithmierten Itemdiskriminationen $\ln(\alpha_i)$ mit einem Mittelwert von μ_{ln} und einer Standardabweichung σ_{ln} . Für Anwendungen ist die Festlegung der Verteilungsparameter (Erwartungswert und die Varianz) der Log-Normalverteilung als A-priori-Verteilung für α_i nicht

ganz einfach, da beide Größen abhängig voneinander sind. Ähnliches gilt für die *Gamma-Verteilung* $\alpha_i \sim \Gamma(a_\alpha, b_\alpha)$, die ebenfalls als A-priori-Verteilung für den Diskriminationsparameter Verwendung findet, da sie ebenfalls nur positive Werte von α_i erlaubt. Der Erwartungswert $\mu = a_\alpha/b_\alpha$ und die Varianz $\sigma^2 = a_\alpha/b_\alpha^2$ der Gamma-Verteilung werden durch den Formparameter a_α und den inversen Skalierungsparameter b_α gesteuert. Die Wahl einer *Gleichverteilung* als A-priori-Verteilung über einem Intervall $[a, b]$, mit $a \geq 0$, bedeutet, dass a priori α_i als im Bereich zwischen a und b liegend angenommen wird, wobei alle Werte in diesem Intervall als gleich plausibel angesehen werden.

Die Itemschwierigkeiten β_i können positive und negative Werte annehmen, sodass die Normalverteilung häufig als A-priori-Verteilung gewählt wird:

$$\beta_i \sim N(\mu_\beta, \sigma_\beta) \quad (19.56)$$

Je kleiner die Standardabweichung σ_β gewählt wird, desto informativer ist die A-priori-Verteilung, denn desto kleiner ist a priori auch der Bereich um den erwarteten Wert μ_β , in dem β_i vermutet wird. Bei einer nicht informativen A-priori-Verteilung werden hingegen große Werte von σ_β gewählt, sodass die plausiblen Werte von β_i a priori in einem weiten Bereich um μ_β streuen.

Auch die individuellen Personenparameter θ_v sind nicht auf einen bestimmten Wertebereich beschränkt. Wie bei der MML-Schätzung wird häufig eine A-priori-Normalverteilung verwendet, sodass für den individuellen Wert θ_v jeder Person $v = 1, \dots, n$ angenommen wird:

$$\theta_v \sim N(\mu_\theta, \sigma_\theta) \quad (19.57)$$

Nun ist es für die Anwendung wichtig, wie das Modell identifiziert ist. Werden alle Itemparameter frei geschätzt, werden zur Modellidentifikation für μ_θ und σ_θ feste Werte gewählt (z. B. $\mu_\theta = 0$ und $\sigma_\theta = 1$). Alternativ kann das Modell durch Fixierung der Itemparameter mindestens eines Items (z. B. $\beta_i = 0$ und $\alpha_i = 1$) identifiziert werden. Dann sind jedoch μ_θ und σ_θ der individuellen A-priori-Verteilungen von θ_v in Gl. (19.57) selbst unbekannte zu schätzende Größen mit jeweils eigener A-priori- und A-posteriori-Verteilung. In diesem Fall werden die A-priori-Verteilungen von μ_θ und σ_θ auch als *Hyperprior-Verteilungen* bezeichnet. Für den Erwartungswert μ_θ wird oft eine Normalverteilung und für die Varianz σ_θ^2 häufig eine inverse Gamma-Verteilung als Hyperprior-Verteilung spezifiziert.

19.3.4 Parameterschätzung in der Bayes-Statistik

Die A-posteriori-Verteilung wurde bereits in ► Abschn. 19.3.2 als essentiell für die Bayes'sche Inferenz eingeführt. Sie enthält nach der Parameterschätzung alle Informationen bezüglich der zu schätzenden Größen. In den Messmodellen der IRT sind das zunächst die Item- und Personenparameter, die den Parametervektor $\varphi = (\iota_1, \dots, \iota_k, \theta_1, \dots, \theta_n)$ bilden. Die A-posteriori-Verteilung ist entsprechend mehrdimensional. Generell folgt die A-posteriori-Verteilung, abgesehen von sehr einfachen Modellen, zumeist keiner bekannten parametrischen Verteilungsform und ist daher analytisch kaum zugänglich. Verschiedene Algorithmen sind entwickelt worden, um die A-posteriori-Verteilung oder zumindest Kennwerte der A-posteriori-Verteilung zu schätzen. Dabei können nicht simulationsbasierte und simulationsbasierte Verfahren unterschieden werden.

19.3.4.1 Nicht simulationsbasierte Bayes'sche Schätzverfahren in der IRT

MBM-Schätzung

Ein Beispiel für nicht simulationsbasierte Verfahren ist die von Mislevy (1986) entwickelte *Marginale Bayes-Modal-Schätzung* (MBM-Schätzung). Es handelt sich dabei um eine Erweiterung des bereits beschriebenen EM-Algorithmus für die MML-Schätzung (► Abschn. 19.2.6.3), bei dem die marginale Likelihood-Funktion $mL(\iota)$ (Gl. 19.42) mit der A-priori-Verteilung $f(\iota)$ der Itemparameter ι gewichtet wird. Wie bei der MML-Schätzung wird auch hier über die Verteilung der latenten Personenvariablen θ integriert, da die individuellen Personenparameter wiederum als inzidentelle Parameter aufgefasst werden. Lediglich Verteilungsparameter wie die Varianz und der Erwartungswert von θ sind relevante Parameter. Sollen sie geschätzt werden, muss auch für diese Größen eine Hyperprior-Verteilung $f(\mu_\theta, \sigma_\theta^2)$ spezifiziert werden.

Beispielhaft soll die MBM-Schätzung am Fall des eindimensionalen 2PL-Modells nach Birnbaum (1968) dargestellt werden. Die A-posteriori-Verteilung der zu schätzenden Parameter $f(\iota, \mu_\theta, \sigma_\theta^2 | Y = y)$ ist:

$$f(\iota, \mu_\theta, \sigma_\theta^2 | Y = y) = c^{-1} \underbrace{\prod_{v=1}^n \int_{\theta} P(Y_v = y_v | \theta; \iota) f(\theta | \mu_\theta, \sigma_\theta^2) f(\iota, \mu_\theta, \sigma_\theta^2) d\theta}_{mL(\iota)} \quad (19.58)$$

Dabei entspricht der erste Faktor $mL(\iota)$ der MML-Funktion von Gl. (19.42). Die Normierungskonstante ist als c bezeichnet. Nun kann die A-posteriori-Verteilung einfach als Funktion der unbekannten Parameter und die Parameterschätzung als mathematisches Maximierungsproblem aufgefasst werden, genauso wie im Fall der ML-Schätzung. Das heißt, es werden diejenigen Werte von ι , μ_θ und σ_θ^2 gesucht, bei denen die A-posteriori-Dichte maximal ist. Die Normierungskonstante kann bei der Maximierung unberücksichtigt bleiben. Wie bei der ML-Schätzung wird aufgrund der mathematisch besseren Eigenschaften der natürliche Logarithmus von Gl. (19.58) als zu maximierende Funktion $g_{MBM}(\iota, \mu_\theta, \sigma_\theta^2)$ verwendet, die dann wie folgt geschrieben werden kann:

$$\begin{aligned} g_{MBM}(\iota, \mu_\theta, \sigma_\theta^2) &= \ln \left(\prod_{v=1}^n \int_{\theta} P(Y_v = y_v | \theta; \iota) f(\theta | \mu_\theta, \sigma_\theta^2) d\theta \right) \\ &\quad + \ln [f(\iota, \mu_\theta, \sigma_\theta^2)] \end{aligned} \quad (19.59)$$

Die Zielfunktion ist somit eine Summe aus der marginalen Log-Likelihood-Funktion und der logarithmierten A-priori-Verteilung. Das Maximum der A-posteriori-Verteilung wird wiederum durch Nullsetzen der ersten partiellen Ableitungen von

Gl. (19.59) hinsichtlich der gesuchten Modellparameter ermittelt. Gemäß der Summenregel in der Differentialrechnung ist die Ableitung einer Summe gleich der Summe der Ableitungen der einzelnen Summanden. Entsprechend muss auch die logarithmierte A-priori-Verteilung aus Gl. (19.59) partiell nach den gesuchten Modellparametern abgeleitet werden. Betrachtet man die Items und somit deren Parameter sowie die Verteilungsparameter μ_θ und σ_θ^2 a priori als unabhängige Größen, ist die gemeinsame A-priori-Verteilung gleich dem Produkt der A-priori-Verteilungen der einzelnen Parameter und damit der Logarithmus entsprechend eine Summe:

$$\begin{aligned} f(\mathbf{l}, \mu_\theta, \sigma_\theta^2) &= f(\mu_\theta) f(\sigma_\theta^2) \prod_{j=1}^k f(\alpha_i) f(\beta_i) \\ \ln [f(\mathbf{l}, \mu_\theta, \sigma_\theta^2)] &= \ln [f(\mu_\theta)] + \ln [f(\sigma_\theta^2)] + \sum_{j=1}^k \ln [f(\alpha_i)] + \sum_{j=1}^k \ln [f(\beta_i)] \end{aligned} \quad (19.60)$$

Hier sollen beispielhaft folgende A-priori-Verteilungen betrachtet werden:

$$\begin{aligned} f(\alpha_i) &= \log N(\mu_\alpha, \sigma_\alpha) \\ f(\beta_i) &= N(\mu_\beta, \sigma_\beta) \\ f(\mu_\theta) &= N(\mu_{E(\theta)}, \sigma_{E(\theta)}) \\ f(\sigma_\theta^2) &= IG(u_1, u_2) \end{aligned} \quad (19.61)$$

Für den Erwartungswert der latenten Variablen θ und die Itemschwierigkeiten werden also Normalverteilungen und für die Itemdiskriminationen eine Log-Normalverteilung spezifiziert. Für die Varianz σ_θ^2 wird eine inverse Gamma-Verteilung (IG) mit dem Formparameter u_1 und dem Skalierungsparameter u_2 als A-priori-Verteilungen spezifiziert. Die ersten Ableitungen der logarithmierten A-priori-Verteilungen nach den jeweiligen Parametern sind dann:

$$\begin{aligned} \frac{d \ln [f(\mathbf{l}, \mu_\theta, \sigma_\theta^2)]}{d \alpha_i} &= \left(-\frac{\ln(\alpha_i) - \mu_\alpha}{\sigma_\alpha^2} - 1 \right) \frac{1}{\alpha_i} \\ \frac{d \ln [f(\mathbf{l}, \mu_\theta, \sigma_\theta^2)]}{d \beta_i} &= \frac{\beta_i - \mu_\beta}{\sigma_\beta^2} \\ \frac{d \ln [f(\mathbf{l}, \mu_\theta, \sigma_\theta^2)]}{d \mu_\theta} &= \frac{\mu_\theta - \mu_{E(\theta)}}{\sigma_{E(\theta)}^2} \\ \frac{d \ln [f(\mathbf{l}, \mu_\theta, \sigma_\theta^2)]}{d \sigma_\theta^2} &= (-u_1 - 1) \frac{1}{\sigma_\theta^2} + \frac{u_2}{(\sigma_\theta^2)^2} \end{aligned} \quad (19.62)$$

Die MBM-Schätzung erfolgt ebenfalls iterativ. Da es sich um eine Adaptation des bereits vorgestellten EM-Algorithmus (s. ▶ Abschn. 19.2.6.3) handelt, werden auch bei der MBM-Schätzung abwechselnd E- und M-Schritte durchlaufen, bis ein vorab festgelegtes Konvergenzkriterium erreicht ist. Die numerische Integration über die Verteilung der latenten Variablen θ kann analog zur MML-Schätzung anhand einer Quadraturverteilung erfolgen. Dabei werden im E-Schritt, basierend auf den Startwerten bzw. den vorläufigen Item- und Verteilungsparameterschätzern, zunächst die A-posteriori-Wahrscheinlichkeiten $P(X_\theta^{(q)} | Y_v = y_v; \mathbf{l}, \mu_\theta, \sigma_\theta^2)$ für jeden der Quadraturpunkte X_1, \dots, X_Q nach Gl. (19.50) sowie die a posteriori erwarteten Häufigkeiten $n(X_\theta^{(q)})$ und $n(Y_i = 1 | X_\theta^{(q)})$ berechnet (Gl. 19.52). Mit diesen Größen und unter Verwendung der ersten Ableitungen der logarithmierten A-priori-Verteilungen (Gl. 19.62) können die ersten partiellen Ableitungen von

Gemeinsame A-priori-Verteilung der Modellparameter

Ableitungen der logarithmierten A-priori-Verteilungen

Anwendung des EM-Algorithmus

$g_{MBM}(\mathbf{l}, \mu_\theta, \sigma_\theta^2)$ nach den Itemparametern α_i und β_i wie folgt geschrieben werden:

$$\begin{aligned} & \frac{\partial g_{MBM}(\mathbf{l}, \mu_\theta, \sigma_\theta^2)}{\partial \alpha_i} \\ &= \sum_{q=1}^Q \left[X_\theta^{(q)} - \beta_i \right] \left[n(Y_i = 1 | X_\theta^{(q)}) - n(X_\theta^{(q)}) P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \\ &+ \left(-\frac{\ln(\alpha_i) - \mu_\alpha}{\sigma_\alpha^2} - 1 \right) \frac{1}{\alpha_i} \\ & \frac{\partial g_{MBM}(\mathbf{l}, \mu_\theta, \sigma_\theta^2)}{\partial \beta_i} \\ &= -\alpha_i \sum_{v=1}^n \left[n(Y_i = 1 | X_\theta^{(q)}) - n(X_\theta^{(q)}) P(Y_{vi} = 1 | X_\theta^{(q)}; \alpha_i, \beta_i) \right] \\ &- \frac{\beta_i - \mu_\beta}{\sigma_\beta^2} \end{aligned} \quad (19.63)$$

Für die Schätzung der Verteilungsparameter ergeben sich folgende Schätzgleichungen:

$$\begin{aligned} \frac{\partial g_{MBM}(\mathbf{l}, \mu_\theta, \sigma_\theta^2)}{\partial \mu_\theta} &= \sum_{q=1}^Q X_\theta^{(q)} \frac{n(X_\theta^{(q)})}{N} - \frac{\mu_\theta - \mu_{E(\theta)}}{\sigma_{E(\theta)}^2} \\ \frac{\partial g_{MBM}(\mathbf{l}, \mu_\theta, \sigma_\theta^2)}{\partial \sigma_\theta^2} &= \sum_{q=1}^Q (X_\theta^{(q)} - \mu_\theta)(X_\theta^{(q)} - \mu_\theta) \frac{n(X_\theta^{(q)})}{N} \\ &+ (-u_1 - 1) \frac{1}{\sigma_\theta^2} + \frac{u_2}{(\sigma_\theta^2)^2} \end{aligned} \quad (19.64)$$

Einfluss der A-priori-Verteilung auf die Parameterschätzung – Shrinkage-Effekt

Der Einfluss der A-priori-Verteilung auf die Parameterschätzung lässt sich anhand des Vergleichs der MML-Schätzgleichung und der MBM-Schätzgleichung für die Itemschwierigkeiten β_i (Gln. 19.51 und 19.63) gut verdeutlichen. Die Gleichungen unterscheiden sich hinsichtlich der gewichteten Differenz $\beta_i - \mu_\beta$, die beim MBM-Verfahren subtrahiert wird. Das Gewicht ist dabei die inverse Varianz $1/\sigma_\beta^2$ der A-priori-Verteilung. Das heißt, wenn der MML-Schätzer unter dem a priori erwarteten Wert μ_β liegt, dann ist die Differenz $\beta_i - \mu_\beta$ negativ. Aufgrund der Subtraktion wird somit der Betrag $|(\beta_i - \mu_\beta)/\sigma_\beta^2|$ zum MML-Schätzer addiert. Wenn der MML-Schätzer jedoch über dem a priori erwarteten Wert μ_β liegt, dann ist der Wert der Differenz $(\beta_i - \mu_\beta)/\sigma_\beta^2$ positiv und wird vom MML-Schätzer abgezogen. Als Folge werden bei der MBM-Schätzung die ML-Schätzer in Richtung des A-priori-Erwartungswertes μ_β gezogen. Dieser sog. „Shrinkage-Effekt“ ist typisch für die Bayes'schen Schätzverfahren und Ausdruck der Mischung aus empirischer Information (Likelihood) und Vorwissen bzw. Vorannahmen (A-priori-Verteilung). Der Shrinkage-Effekt wird in dem ▶ Beispiel 19.1 anhand eines fiktiven Datenbeispiels illustriert.

Beispiel 19.1: Shrinkage-Effekt

Der Shrinkage-Effekt ist umso stärker ausgeprägt, je informativer bzw. je gewichtiger die A-priori-Verteilung ist. Bei der Verwendung der Normalverteilung bestimmt die Varianz, wie informativ die A-priori-Verteilung hinsichtlich der Parameterschätzung ist. Das fiktive Datenbeispiel aus □ Tab. 19.1 soll dies verdeutlichen:

■ **Tabelle 19.1** Ausprägungen auf der latenten Variablen θ_v und die mit den Werten null und eins kodierten Itemantworten y_{vi} von zehn Personen

Person	1	2	3	4	5	6	7	8	9	10
θ_v	-1.6	-1.2	-1.1	-0.7	-0.6	-0.3	-0.2	0.5	0.5	1.7
y_v	0	0	0	0	0	1	0	1	1	1

Dabei soll es sich um die mit „richtig“ ($y_{vi} = 1$) und „falsch“ ($y_{vi} = 0$) kodierten Antworten von zehn Personen hinsichtlich eines Items Y_i aus einem Mathematiktest handeln. Die individuellen Werte θ_v der Personen seien ebenfalls bekannt. In realen Anwendungen ist das üblicherweise nicht der Fall, aber man kann sich vorstellen, dass bereits sehr reliable Schätzungen für die Werte θ_v aufgrund anderer Mathematikitems des Tests vorliegen. Nun soll unter Verwendung des Rasch-Modells noch die Itemschwierigkeit β_i eines weiteren Items i geschätzt werden.

Das Item i sei schon vorab in verschiedenen Mathematiktestungen verwendet worden und habe sich dabei wiederholt als sehr schwierig erwiesen, sodass nur sehr wenige Personen das Item lösen konnten. Diese Information könnte man bei der Parameterschätzung berücksichtigen, muss sie jedoch angemessen in einer A-priori-Verteilung repräsentieren. Aufgrund der verwendeten Normierung des Tests sei der Erwartungswert der latenten Personenvariable $E(\theta) = 0$ und die Varianz $Var(\theta) = 1$. Da die Itemschwierigkeit und die Personenvariable dieselbe Metrik haben, lässt sich unmittelbar sagen, dass bei einem Wert von $\beta_i = 2$ nur 2.27 % der Personen eine Lösungswahrscheinlichkeit von mindestens 50 % haben, bei $\beta_i = 3$ sinkt dieser Anteil auf 0.13 %. Da das Item bereits als sehr schwer aufgefallen war, wird für β_i eine Normalverteilung $N(\mu_\beta, \sigma_\beta)$ als A-priori-Verteilung mit dem Erwartungswert $\mu_\beta = 3$ gewählt. Je größer man nun die Streuung σ_β wählt, umso weniger informativ ist die A-priori-Verteilung, denn umso weiter ist a priori auch der Bereich, in dem β_i erwartet wird. Je kleiner die Streuung ist, desto informativer ist die A-priori-Verteilung, da der Bereich um μ_β , in dem man β_i vermutet, immer kleiner wird. Die Streuung der A-priori-Verteilung ist somit gleichbedeutend mit der *Unsicherheit* in Bezug auf eine zu schätzende Größe.

Für das Datenbeispiel sollen exemplarisch zwei A-priori-Verteilungen gewählt werden $N(3, 1)$ und $N(3, 0.5)$, wobei beide als informativ zu bezeichnen sind. Zum Vergleich soll die Likelihood dienen, die keinerlei Vorwissen berücksichtigt.

■ Abb. 19.2 zeigt die Log-Likelihood-Funktion sowie die beiden logarithmierten A-posteriori-Verteilungen in Abhängigkeit des zu schätzenden Parameters β_i (Abszisse). Die Maxima mit den ML- bzw. MAP-Schätzern $\hat{\beta}_i$ sind mit den zugehörigen Werten durch Pfeile gekennzeichnet. Es zeigt sich, dass die ML-Schätzung mit $\hat{\beta}_i = 0.2$ am weitesten von μ_β abweicht. Je kleiner die Streuung σ_β ist, umso mehr tendieren die Bayes-Modal-Schätzer in Richtung des a priori erwarteten Wertes μ_β . Die Stärke des Shrinkage-Effekts und somit die Unterschiede zwischen ML-Schätzern und Bayes'schen Punktschätzern hängt dabei nicht nur von der Stärke der Gewichtung der A-priori-Verteilung ab, sondern auch vom Ausmaß der Diskrepanz zwischen der empirischen Information und dem Vorwissen bzw. den Vorannahmen. Im dargestellten Beispiel haben immerhin 40 % der Personen ein als besonders schwer erwartetes Item gelöst, wobei ihre Fähigkeitsausprägungen keine außergewöhnlich hohen Werte aufweisen. Das ist auch der Grund, warum die ML-Schätzung deutlich niedriger ausfällt als der a priori erwartete Wert $\mu_\beta = 3$.

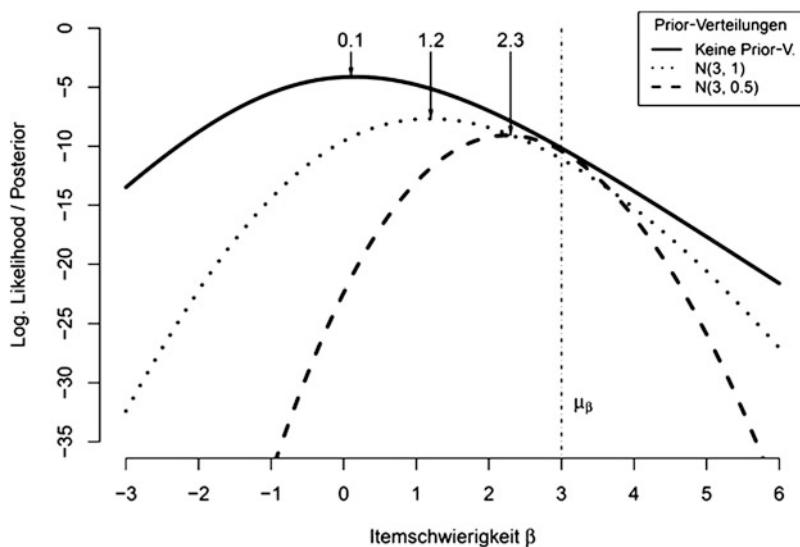


Abb. 19.2 Log-Likelihood-Funktion und logarithmierte A-posteriori-Verteilungen mit den jeweiligen Maxima (Pfeile) und dem Erwartungswert der A-priori-Verteilung (vertikale gestrichelte Linie)

19.3.4.2 Simulationsbasierte Bayes'sche Schätzverfahren in der IRT (MCMC-Verfahren)

Die dargestellte MBM-Schätzung von Mislevy (1986) ist wie die MML-Schätzung nicht für hochdimensionale IRT-Modelle geeignet, da auch sie nicht ohne eine numerische Integration auskommt. Simulationsbasierte Verfahren können bei komplexeren Modellen eine wertvolle Alternative darstellen. Zur Erinnerung: In der Bayes-Statistik werden die unbekannten Parameter als Zufallsvariablen aufgefasst. Gemäß dem sog. „Monte-Carlo-Prinzip“ kann man alles über eine Zufallsvariable in Erfahrung bringen, wenn man nur hinreichend viele Zufallsvariablen aus der Verteilung ziehen kann. Das heißt, sämtliche Verteilungsparameter wie der Erwartungswert, der Modus, das Maximum, der Median, die Varianz, die Schiefe, der Exzess usw. können anhand einer hinreichend großen Zahl von Realisierungen der Zufallsvariablen geschätzt werden. Aber auch Bereichs- und Intervallschätzungen oder die Visualisierung der gesamten Verteilung mittels Histogramm oder Kerndichteschätzern sind möglich. Je größer die Zahl der Realisationen der Zufallsvariable ist, desto präziser kann die Verteilung charakterisiert werden. Das gilt natürlich auch für die A-posteriori-Verteilung in der Bayes-Statistik. Bayes'sche Simulationsverfahren unter Verwendung von MCMC-Verfahren sind für genau diesen Zweck entwickelt worden. Sie erlauben selbst bei unbekanntem Verteilungstyp die Zufallsziehungen aus komplexen uni- oder multidimensionalen A-posteriori-Verteilungen. Genau wie das Glücksspiel, worauf der Begriff „Monte Carlo“ in MCMC anspielt, sind MCMC-Verfahren eine Anwendung der Wahrscheinlichkeitstheorie. Das zweite MC steht für Markov-Chains bzw. -Ketten; die in der Bayes-Statistik verwendeten Algorithmen zur Generierung von Zufallszahlen aus A-posteriori-Verteilungen. Der Gibbs-Sampler oder der Metropolis Hastings-(MH)-Algorithmus sind Anwendungsbeispiele von Markov-Ketten.

**Monte-Carlo-Prinzip,
MCMC-Verfahren und
Markov-Ketten**

Markov-Ketten

Markov-Ketten sind stochastische Prozesse. Sei $(X^{(t)})_{t \in \mathbb{N}}$ eine Folge von Zufallsvariablen $X^{(1)}, \dots, X^{(t)}, \dots, X^{(T)}$, wobei der Index $t = 0, 1, \dots, T$ für die Zeitpunkte steht, so handelt es sich bei dieser Folge um eine Markov-Kette, wenn für die beding-

te Verteilung von $\mathbf{X}^{(t)}$ in Bezug auf die zeitlich vorgeordneten Variablen X_0, \dots, X_t gilt:

$$f(\mathbf{X}_{t+1}|\mathbf{X}_0, \dots, \mathbf{X}_t) = f(\mathbf{X}_{t+1}|\mathbf{X}_t) \quad (19.65)$$

Diese sog. „Markov-Eigenschaft“ ist eine bedingte stochastische Unabhängigkeit. Fasst man die Zufallsvariablen $\mathbf{X}^{(t)}$ als Zustände zu den jeweiligen Zeitpunkten t auf, so ist der Zustand zu einem Zeitpunkt $t + 1$, unter der Bedingung des Zustands zum unmittelbar vorgeordneten Zeitpunkt t , stochastisch unabhängig von allen früheren Zuständen. Der Wertebereich Ω_X der Variablen $\mathbf{X}^{(t)}$ (für alle $t = 0, \dots, T$) beinhaltet alle möglichen Zustände und wird auch Zustandsraum („state space“) genannt. $\mathbf{X}^{(0)}$ ist dabei der Ausgangszustand.

Doch warum sind Markov-Ketten in der Bayes-Statistik von so großer Bedeutung? Grundsätzlich können Computer nicht wirklich Zufallszahlen generieren, sondern lediglich Pseudozufallszahlen. Außerdem zeigen aufeinanderfolgende Zufallszahlen gängiger MCMC-basierter Bayes-Algorithmen aufgrund ihrer Konstruktion oft nicht unerhebliche Autokorrelationen, d. h. Korrelation zwischen aufeinander folgenden Zufallszahlen. Der namensgebende russische Mathematiker Andrei A. Markov (* 1856, † 1922) zeigte, dass das Gesetz der großen Zahlen auch für *abhängige Zufallsvariablen* gilt. Es ist also möglich, mit einer hinreichend großen Zahl von Realisationen abhängiger Zufallsvariablen die zugrunde liegende *stationäre Verteilung* mit jeglicher gewünschter Präzision näherungsweise zu bestimmen. Stationäre Verteilung bedeutet, dass es genau eine zeitlich invariante Verteilung $f(\mathbf{X})$ für alle $\mathbf{X}^{(t)}$ gibt, sodass gilt: $f(\mathbf{X}) = f(\mathbf{X}^{(t)})$, für alle $t = 0, \dots, T$. Mit zunehmender Abhängigkeit bzw. Autokorrelation zwischen $\mathbf{X}^{(t)}$ und $\mathbf{X}^{(t+1)}$ verringert sich lediglich die Effizienz, mit der eine Information über $f(\mathbf{X})$ anhand von realisierten Werten $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}$ gewonnen werden kann.

Damit eine Markov-Kette jedoch überhaupt eine stationäre Verteilung aufweist, müssen bestimmte Bedingungen erfüllt sein. So muss die Markov-Kette *irreduzibel* sein, d. h., sie muss ausgehend von einem beliebigen Zustand $\mathbf{x}^{(t)} \in \Omega_X$ in einer endlichen Zeit jeden anderen Zustand in Ω_X erreichen können. Auch muss jeder beliebige Zustand wiederholt auftreten können (nicht müssen!), sodass $\mathbf{x}^{(t)} = \mathbf{x}^{(t+r)}$ ergibt, wobei r eine natürliche Zahl größer null ist. Die notwendige Bedingung der *positiven Rekurrenz* der Markov-Kette ist erfüllt, wenn die erwartete Wiederkehrzeit endlich ist. Zudem muss die Markov-Kette *aperiodisch* sein, sodass die Auftretenswahrscheinlichkeiten der Zustände unabhängig von t sind, also keine zeitliche Systematik aufweisen. Sind die Eigenschaften Irreduzibilität, Aperiodizität und positive Rekurrenz gegeben, so bezeichnet man die Markov-Kette auch als *ergodisch*. Lässt sich für eine ergodische Markov-Kette ein sog. *Markov-Kern* $K(\mathbf{X}^{(t)}, \mathbf{X}^{(t+1)})$ konstruieren, der mathematisch beschreibt, wie – ausgehend von einem Wert $\mathbf{x}^{(t)} \in \Omega_X$ – Realisationen $\mathbf{x}^{(t+1)}$ der Zufallsvariablen $\mathbf{X}^{(t+1)}$ generiert werden können, für die gilt

$$P\left(\mathbf{X}^{(t+1)} \in A | \mathbf{X}^{(t)} = \mathbf{x}^{(t)}\right) = \int_A K(\mathbf{x}^{(t)}, \mathbf{X} \in A), \forall A \subset \Omega_X, \quad (19.66)$$

so konvergiert die Markov-Kette mit großem T gegen die zugrunde liegende Verteilung $f(\mathbf{X})$. Zur Erläuterung: Die Teilmenge A auf Ω_X in Gl. (19.66) wurde lediglich eingeführt, um auch für den Fall stetiger Zufallsvariablen $\mathbf{X}^{(t)}$ sinnvoll den Begriff der Wahrscheinlichkeit verwenden zu können. Denn die Wahrscheinlichkeit $P(\mathbf{X}^{(t+1)} = \mathbf{x}^{(t+1)} | \mathbf{X}^{(t)} = \mathbf{x}^{(t)})$ für einen konkreten Wert $\mathbf{x}^{(t+1)}$ einer stetigen Zufallsvariable $\mathbf{X}^{(t+1)}$ ist immer null. Ist A beispielsweise ein Intervall $[a, b]$ auf $\Omega_X = \mathbb{R}^m$, so ist $P(\mathbf{X}^{(t+1)} \in A | \mathbf{X}^{(t)} = \mathbf{x}^{(t)})$ die Wahrscheinlichkeit, dass der Wert $\mathbf{x}^{(t+1)}$ in diesem Intervall des Wertebereichs von \mathbf{X} liegt. Mit größer werdendem

T gilt dann, dass die relativen Häufigkeiten $h(\mathbf{x}^{(t)} \in A)$ für das Ereignis, dass die realisierten Werte $\mathbf{x}^{(t)}$ aus der Markov-Kette in A liegen, gegen die wahren Wahrscheinlichkeiten konvergieren, sodass gilt:

$$\lim_{T \rightarrow \infty} \underbrace{\frac{1}{T} \sum_{t=0}^T I(\mathbf{x}^{(t)} \in A)}_{h(\mathbf{x}^{(t)} \in A)} = P(X \in A), \forall A \subset \Omega_X \quad (19.67)$$

Dabei ist $I(\mathbf{x}^{(t)} \in A)$ eine Indikatorvariable, mit $I(\mathbf{x}^{(t)} \in A) = 1$, falls \mathbf{x}_t in A liegt; und $I(\mathbf{x}^{(t)} \in A) = 0$, falls dies nicht der Fall ist, sodass der Mittelwert von $I(\mathbf{x}^{(t)} \in A)$ gleich der relativen Häufigkeit $h(\mathbf{x}^{(t)} \in A)$ ist.

Sampling-Algorithmus

Wendet man die Begrifflichkeit der Markov-Ketten auf die simulationsbasierte Bayes-Schätzung an, so ist die stationäre Verteilung, die man anhand der Realisationen einer Markov-Kette approximieren will, die multidimensionale A-posteriori-Verteilung $f(\boldsymbol{\varphi}|\mathbf{Y} = \mathbf{y})$ des gesamten Parametervektors $\boldsymbol{\varphi}$ mit allen Item- und Personenparametern und ggf. weiteren Verteilungsparametern der latenten Personenvariablen $\boldsymbol{\theta}$ (z. B. $E(\boldsymbol{\theta})$ und $\Sigma(\boldsymbol{\theta})$). Der Zustandsraum entspricht dem Parameterraum $\Omega_{\boldsymbol{\varphi}}$. Der sog. „Sampling-Algorithmus“ generiert nun T Realisierungen $\boldsymbol{\varphi}^{(1)}, \dots, \boldsymbol{\varphi}^{(T)}$ des Parametervektors, wobei die Herausforderung in der Wahl bzw. Konstruktion des Markov-Kerns $K(\boldsymbol{\varphi}^{(t)}, \boldsymbol{\varphi}^{(t+1)})$ liegt. Dies soll hier an den zwei prominentesten Sampling-Algorithmen, dem MH-Algorithmus und dem Gibbs-Sampler, erläutert werden. Beide teilen die Markov-Eigenschaft, d. h., die Markov-Kerne $K(\boldsymbol{\varphi}^{(t)}, \boldsymbol{\varphi}^{(t+1)})$ in beiden Algorithmen generieren $\boldsymbol{\varphi}^{(t+1)}$ bei gegebenen Daten $\mathbf{Y} = \mathbf{y}$ und dem Parametervektor $\boldsymbol{\varphi}^{(t)}$ des vorangegangenen Iterationsschritts aus der bedingten Verteilung $f(\boldsymbol{\varphi}^{(t+1)}|\boldsymbol{\varphi}^{(t)}, \mathbf{Y} = \mathbf{y})$.

Metropolis Hastings-(MH)Algorithmus

Beim MH-Algorithmus erfolgt die Generierung von $\boldsymbol{\varphi}^{(t+1)}$ ausgehend von $\boldsymbol{\varphi}^{(t)}$ schrittweise, beginnend mit dem Vektor $\boldsymbol{\varphi}^{(0)}$ mit den Startwerten. Die Schritte des MH-Algorithmus lassen sich wie folgt beschreiben:

■ Schritt 1

Vorschlagsdichte

Zunächst wird aus einer sog. „Vorschlagsdichte“ (*Proposal Density*) $J(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)})$ bei gegebenem $\boldsymbol{\varphi}^{(t)}$ zufällig ein Wert $\boldsymbol{\varphi}^*$ als „Kandidat“, d. h. als potentieller Wert für $\boldsymbol{\varphi}^{(t+1)}$ gezogen.

■ Schritt 2

Akzeptanzverhältnis

Als Nächstes wird das folgende Akzeptanzverhältnis $r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)})$ aus der Vorschlagsdichte und der A-posteriori-Dichte von $\boldsymbol{\varphi}^{(t)}$ und $\boldsymbol{\varphi}^*$ berechnet:

$$r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)}) = \frac{f(\boldsymbol{\varphi}^*|\mathbf{Y} = \mathbf{y}) / J(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)})}{f(\boldsymbol{\varphi}^{(t)}|\mathbf{Y} = \mathbf{y}) / J(\boldsymbol{\varphi}^{(t)}|\boldsymbol{\varphi}^*)} \quad (19.68)$$

■ Schritt 3

Akzeptanzwahrscheinlichkeit

Die Akzeptanzwahrscheinlichkeit mit der $\boldsymbol{\varphi}^*$ als Wert für $\boldsymbol{\varphi}^{(t+1)}$ angenommen wird, ist $P(\boldsymbol{\varphi}^* = \boldsymbol{\varphi}^{(t+1)}) = \min(r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)}), 1)$. Das heißt, ist $r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)}) \geq 1$, so folgt $\boldsymbol{\varphi}^{(t+1)} = \boldsymbol{\varphi}^*$. Ist jedoch $r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)}) < 1$, so wird der Kandidat $\boldsymbol{\varphi}^*$ nur mit einer Wahrscheinlichkeit von $r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)})$ als Wert für $\boldsymbol{\varphi}^{(t+1)}$ akzeptiert. Mit der Gegenwahrscheinlichkeit $1 - r(\boldsymbol{\varphi}^*|\boldsymbol{\varphi}^{(t)})$ wird der Wert $\boldsymbol{\varphi}^{(t)}$ jedoch beibehalten, sodass $\boldsymbol{\varphi}^{(t+1)} = \boldsymbol{\varphi}^{(t)}$ ist.

■ Schritt 4

Die Schritte 1 bis 3 werden so lange wiederholt bis die Maximalzahl T oder ein anderes vorab spezifiziertes Abbruchkriterium erfüllt ist.

Die Vorschlagsdichte kann eine beliebige Verteilung sein, aus der möglichst einfach Zufallszahlen generiert werden können. Das Akzeptanzverhältnis quantifiziert die Plausibilität des Kandidaten φ^* hinsichtlich der unbekannten A-posteriori-Verteilung. Kandidaten φ^* mit höherer A-posteriori-Dichte werden entsprechend häufiger akzeptiert als Kandidaten mit einer niedrigen A-posteriori-Dichte, und zwar in der Weise, dass der Algorithmus gegen die A-posteriori-Verteilung konvergiert. Die Stärke des MH-Algorithmus liegt in der Flexibilität, eine beliebige Vorschlagsdichte zum Generieren der Zufallszahlen φ^* verwenden zu können. Das heißt jedoch nicht, dass alle Vorschlagsdichten gleich gut geeignet sind. Vorschlagsdichten, die viele unplausible Kandidaten φ^* generieren, führen zu niedrigen Akzeptanzraten, hoher Autokorrelation und somit ineffizienten Schätzalgorithmen. Insbesondere bei Modellen mit sehr vielen Parametern kann es durchaus schwierig sein, geeignete Vorschlagsdichten für die Kandidaten φ^* zu spezifizieren.

Gibbs-Sampler

Für die Parameterschätzung von komplexen Modellen mit vielen Parametern, wie es oft auch bei IRT-Modellen der Fall ist, kann der Gibbs-Sampler verwendet werden. Dieser basiert auf dem Faktorisierungssatz der Wahrscheinlichkeit, der besagt, dass die gemeinsame Verteilung von Zufallsvariablen vollständig durch die bedingten Verteilungen der einzelnen Variablen unter der Bedingung der übrigen Variablen beschrieben werden kann. Anhand des Faktorisierungssatzes lässt sich die A-posteriori-Verteilung von φ wie folgt schreiben:

$$\begin{aligned} f(\varphi | \mathbf{Y} = \mathbf{y}) &= f(\varphi_1, \varphi_2, \dots, \varphi_M | \mathbf{Y} = \mathbf{y}) \\ &= f(\varphi_1 | \varphi_2, \dots, \varphi_M, \mathbf{Y} = \mathbf{y}) \\ &\quad f(\varphi_2 | \varphi_3, \dots, \varphi_M, \mathbf{Y} = \mathbf{y}) \dots f(\varphi_M | \mathbf{Y} = \mathbf{y}) \end{aligned} \quad (19.69)$$

Die gemeinsame A-posteriori-Verteilung aller Modellparameter $\varphi_1, \dots, \varphi_M$ ist somit gleich dem Produkt der bedingten A-posteriori-Verteilungen der einzelnen Parameter φ_m unter der Bedingung der übrigen Parameter und der Daten. In allgemeiner Form kann der Algorithmus des Gibbs-Samplers wie folgt geschrieben werden:

■ Schritt 1

Ziehe $\varphi_1^{(t+1)}$ zufällig aus der A-posteriori-Verteilung $f(\varphi_1^{(t+1)} | \varphi_2^{(t)}, \dots, \varphi_M^{(t)}, \mathbf{Y} = \mathbf{y})$ von φ_1 , bei gegebenen Daten und aller übrigen in Iteration t generierten Werte $\varphi_2^{(t)}, \dots, \varphi_M^{(t)}$.

■ Schritt 2

Ziehe $\varphi_2^{(t+1)}$ zufällig aus der A-posteriori-Verteilung $f(\varphi_2^{(t+1)} | \varphi_1^{(t+1)}, \varphi_3^{(t)}, \dots, \varphi_M^{(t)}, \mathbf{Y} = \mathbf{y})$ von φ_2 , bei gegebenen Daten und $\varphi_1^{(t+1)}$ sowie aller übrigen in Iteration t generierten Werte $\varphi_3^{(t)}, \dots, \varphi_M^{(t)}$.

...

■ Schritt M

Ziehe $\varphi_M^{(t+1)}$ zufällig aus der A-posteriori-Verteilung $f(\varphi_M^{(t+1)} | \varphi_1^{(t+1)}, \dots, \varphi_{M-1}^{(t+1)}, \mathbf{Y} = \mathbf{y})$ von φ_M , bei gegebenen Daten und aller übrigen in Iteration $t+1$ generierten Werte $\varphi_1^{(t+1)}, \dots, \varphi_{M-1}^{(t+1)}$.

Der Algorithmus wird so lange durchlaufen, bis für alle Parameter ein vorab festgelegte Zahl T von Zufallszahlen aus den bedingten A-posteriori-Verteilungen

Faktorisierungssatz der Wahrscheinlichkeit

Algorithmus des Gibbs-Samplers

generiert wurde oder ein anderes Abbruchkriterium erfüllt ist. Jede bedingte A-posteriori-Verteilung eines Parameters φ_m aus φ ist dabei:

$$f\left(\varphi_m^{(t+1)} | \varphi_1^{(t+1)}, \varphi_{m-1}^{(t)}, \varphi_{m+1}^{(t)}, \dots, \varphi_M^{(t)}, Y = y\right) = \frac{L\left(\varphi_m^{(t+1)}\right) f(\varphi_m)}{\int_{\varphi_m} L\left(\varphi_m^{(t+1)}\right) f(\varphi_m) d\varphi_m} \quad (19.70)$$

Gibbs-Sampler als Spezialfall des MH-Algorithmus

Gibbs-Sampler für die Parameterschätzung in der IRT

Dabei werden in den ML-Funktionen $L(\varphi_m^{(t+1)})$ alle übrigen Parameter $\varphi_1^{(t+1)}, \varphi_{m-1}^{(t)}, \varphi_{m+1}^{(t)}, \dots, \varphi_M^{(t)}$ auf den jeweiligen Werten konstant gehalten. $f(\varphi_m)$ ist die A-priori-Verteilung von φ_m . Streng genommen ist der Gibbs-Sampler ein Spezialfall des MH-Algorithmus bezüglich der einzelnen Parameter φ_m , wobei die Vorschlagsdichte gerade der A-posteriori-Verteilung (Gl. 19.70) entspricht und das Akzeptanzverhältnis $r(\varphi_m^* | \varphi_m^{(t)})$ und somit die Akzeptanzwahrscheinlichkeit immer gleich eins sind.

Wendet man den Gibbs-Sampler für die Parameterschätzung in der IRT an, so umfasst der Parametervektor alle Itemparameter, alle Personenparameter und ggf. die Verteilungsparameter der latenten Personenvariablen (z. B. Erwartungswerte, Varianzen und Kovarianzen). Für alle diese Größen müssen für die Anwendung die jeweils angemessenen A-priori-Verteilungen $f(\varphi_m)$ spezifiziert werden. Im ▶ Beispiel 19.2 wird die Anwendung des Gibbs-Samplers an einem Minimalbeispiel illustriert.

Beispiel 19.2: Gibbs-Sampler zur Schätzung der Itemparameter im 2PL-Modell

Hier soll noch einmal auf das Datenbeispiel aus □ Tab. 19.1 zurückgegriffen werden. Zur Erinnerung: Es liegen von $N = 10$ Personen die richtigen ($y_{vi} = 1$) oder falschen ($y_{vi} = 0$) Antworten zu einem Mathematikitem Y_i sowie die individuellen Ausprägungen θ_v , wobei $\theta \sim N(0, 1)$, vor. Es wird das 2PL-Modell nach Birnbaum als psychometrisches Modell gewählt. Nun sollen unter Verwendung eines Gibbs-Samplers die Itemschwierigkeit β_i und die Itemdiskrimination α_i des Items geschätzt werden. Bei der Spezifikation der A-priori-Verteilung soll wiederum berücksichtigt werden, dass sich das Item in früheren Erhebungen als sehr schwer erwiesen hat, sodass wiederum eine Normalverteilung $N(3, 1)$ als informative A-priori-Verteilung von β_i spezifiziert wird. Für den Diskriminationsparameter wird eine Log-Normalverteilung $\ln N(0, 1)$ als informative A-priori-Verteilung gewählt, sodass negative Werte für α_i a priori ausgeschlossen werden. Die Zahl der Iterationen wird auf $T = 4000$ festgelegt. Zur Parameterschätzung wurde das Programm WinBUGS (Lunn et al. 2000) verwendet. In □ Tab. 19.2 sind die wichtigsten Kennwerte (Mittelwert, Median, Modus, Standardabweichung, 95 %-Kredibilitätsintervall) der A-posteriori-Verteilungen beider Itemparameter und die Rubin-Gelman-Statistik (auch Potential-Scale-Reduction-Faktor, kurz PSR-Faktor, genannt) zusammengefasst. Letztere dient der Konvergenzdiagnostik (▶ Abschn. 19.3.4.2), wobei Werte des PSR-Faktors < 1.1 als Indikator für die Konvergenz des Gibbs-Samplers interpretiert werden.

Es fällt auf, dass die Kennwerte der zentralen Tendenz (Mittelwert, Median und Modus) der A-posteriori-Verteilung von β_i sehr ähnlich sind, was auf eine symmetrische Verteilung hinweist. Dies gilt jedoch nicht für die Itemdiskrimination! Der Median und der Modus der A-posteriori-Verteilung sind im Vergleich zum Mittelwert deutlich linksverschoben, was auf eine asymmetrische, linkssteile bzw. rechtsschiefe Verteilung verweist. Das bestätigt sich auch grafisch in der gemeinsamen A-posteriori-Verteilung von α_i und β_i , die in □ Abb. 19.3 dargestellt ist. Die beiden Itemparameter sind a posteriori deutlich negativ korreliert

($r = .60$). Der Zusammenhang ist jedoch keineswegs linear. Generell entspricht die gemeinsame A-posteriori-Verteilung keiner bekannten parametrischen Verteilung, was für den Gibbs-Sampler jedoch unproblematisch ist. Insgesamt weisen die recht großen Standardabweichungen und weiten 95 %-Kredibilitätsintervalle beider Itemparameterschätzer auf eine recht unpräzise Schätzung hin, was auf die sehr kleine Stichprobe von nur $N = 10$ Itemantworten zurückzuführen ist. Mit steigendem Stichprobenumfang nimmt die Schätzgenauigkeit wie bei der ML-Schätzung zu, wodurch die Standardabweichung der A-posteriori-Verteilung sowie die Weite der 95 %-Kredibilitätsintervalle abnehmen.

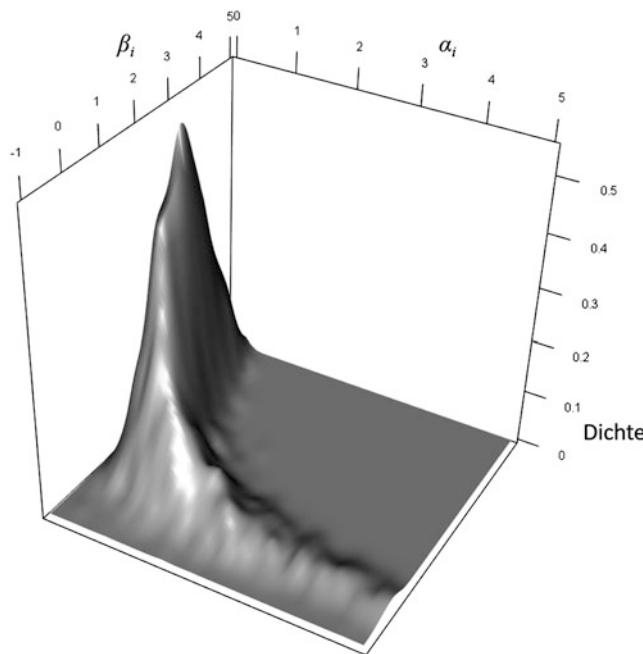
Ein Vorteil des Gibbs-Samplers ist, dass – im Gegensatz zur ML-Schätzung – nicht zwischen inzidentellen und strukturellen Parametern unterschieden werden muss. So ist es theoretisch möglich, simultan auch die A-posteriori-Verteilungen und die daraus abgeleitete Punktschätzer für die individuellen Personenparameter zu bestimmen.

Vorteil des Gibbs-Samplers

■ **Tabelle 19.2** Kennwerte der A-posteriori-Verteilung der Itemschwierigkeit und Itemdiskrimination

	Mittelwert	Median	Modus	SD	95 %-Kredibilitätsintervall	\hat{R}
β_i	1.895	1.881	1.828	1.189	[-.145, 4.246]	1.001
α_i	0.888	.525	.377	1.098	[.117, 4.197]	1.001

SD: Standardabweichung. \hat{R} : PSR-Faktor



■ **Abb. 19.3** Gemeinsame A-posteriori-Verteilung von Itemschwierigkeit β_i und Itemdiskrimination α_i

Metropolis-Hastings-within-Gibbs-Sampler

Kombination von MH-Algorithmus und Gibbs-Sampler

Der Gibbs-Sampler und der MH-Algorithmus können auch kombiniert werden. So kann innerhalb des Gibbs-Algorithmus für die Zufallsziehung der Einzelkomponenten $\varphi_M^{(t+1)}$ von φ aus den jeweiligen bedingten A-posteriori-Verteilungen (Gl. 19.70) der MH-Algorithmus (*Single Component MH-Algorithmus*) verwendet werden. Die Kombination beider Algorithmen (auch als sog. „Metropolis-Hastings-within-Gibbs-Sampler“ bezeichnet) wurde von Levy und Mislevy (2016, S. 261 ff.) für die Parameterschätzung in IRT-Modellen beschrieben. Der Algorithmus soll hier kurz am Beispiel des 2PL-Modells nach Birnbaum skizziert werden:

■ Schritt 1

Zufallsziehungen der Personenparameter $\theta_1, \dots, \theta_N$: Ziehe unabhängig für jede Person $v = 1, \dots, N$ einen Wert aus einer Vorschlagsdichte $J(\theta_v^* | \theta_v^{(t)})$. Berechne dann für jede Person $v = 1, \dots, N$ das Akzeptanzverhältnis $r(\theta_v^* | \theta_v^{(t)})$:

$$r(\theta_v^* | \theta_v^{(t)}) = \frac{P(Y_v = y_v | \theta_v^*; \mathbf{t}^{(t)}) f(\theta_v^*; \mu_\theta, \sigma_\theta) / J(\theta_v^{(t)} | \theta_v^*)}{P(Y_v = y_v | \theta_v^{(t)}; \mathbf{t}^{(t)}) f(\theta_v^{(t)}; \mu_\theta, \sigma_\theta) / J(\theta_v^* | \theta_v^{(t)})} \quad (19.71)$$

Dabei ist $P(Y_v = y_v | \theta_v; \mathbf{t}^{(t)})$ die bedingte Antwortmusterwahrscheinlichkeit bei gegebenen Itemparametern $\mathbf{t}^{(t)}$ aus der Iteration t und $f(\theta_v; \mu_\theta, \sigma_\theta)$ die A-priori-Verteilung der latenten Personenvariablen. Setze nachfolgend $\theta_v^{(t+1)} = \theta_v^*$, mit der Wahrscheinlichkeit $P(\theta_v^{(t+1)} = \theta_v^*) = \min(r(\theta_v^* | \theta_v^{(t)}), 1)$, bzw. $\theta_v^{(t+1)} = \theta_v^{(t)}$, mit der Gegenwahrscheinlichkeit $1 - P(\theta_v^{(t+1)} = \theta_v^*)$.

■ Schritt 2

Zufallsziehungen der Itemparameter $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k$: Ziehe unabhängig für jedes Item $i = 1, \dots, k$ aus einer Vorschlagsdichte $J(\beta_i^* | \beta_i^{(t)})$ für die Itemschwierigkeiten β_i . Berechne dann für jeden Parameter β_i mit $j = 1, \dots, k$, das Akzeptanzverhältnis $r(\beta_i^* | \beta_i^{(t)})$:

$$r(\beta_i^* | \beta_i^{(t)}) = \frac{P(\underline{\mathbf{Y}}_i = \underline{\mathbf{y}}_i | \beta_i^*; \boldsymbol{\theta}^{(t+1)}, \alpha_i^{(t)}) f(\beta_i^*; \mu_{\beta_i}, \sigma_{\beta_i}) / J(\beta_i^{(t)} | \beta_i^*)}{P(\underline{\mathbf{Y}}_i = \underline{\mathbf{y}}_i | \beta_i^{(t)}; \boldsymbol{\theta}^{(t+1)}, \alpha_i^{(t)}) f(\beta_i^{(t)}; \mu_{\beta_i}, \sigma_{\beta_i}) / J(\beta_i^* | \beta_i^{(t)})} \quad (19.72)$$

Dabei ist $\underline{\mathbf{y}}_i$ der beobachtete Itemvektor, d.h. die Spalte i aus der Datenmatrix $\mathbf{Y} = \mathbf{y}$ mit den (0, 1)-kodierten Itemantworten von Item i . Setze $\beta_i^{(t+1)} = \beta_i^*$, mit der Wahrscheinlichkeit $P(\beta_i^{(t+1)} = \beta_i^*) = \min(r(\beta_i^* | \beta_i^{(t)}), 1)$, bzw. $\beta_i^{(t+1)} = \beta_i^{(t)}$, mit der Gegenwahrscheinlichkeit $1 - P(\beta_i^{(t+1)} = \beta_i^*)$. Ziehe unabhängig für jedes Item $i = 1, \dots, k$ aus einer Vorschlagsdichte $J(\alpha_i^* | \alpha_i^{(t)})$ für die Itemdiskrimination α_i und berechne für jeden Parameter α_i mit $i = 1, \dots, k$ das Akzeptanzverhältnis $r(\alpha_i^* | \alpha_i^{(t)})$:

$$r(\alpha_i^* | \alpha_i^{(t)}) = \frac{P(\underline{\mathbf{Y}}_i = \underline{\mathbf{y}}_i | \alpha_i^*; \theta^{(t+1)}, \beta_i^{(t+1)}) f(\alpha_i^*; \mu_{\alpha_i}, \sigma_{\alpha_i}) / J(\alpha_i^{(t)} | \alpha_i^*)}{P(\underline{\mathbf{Y}}_i = \underline{\mathbf{y}}_i | \alpha_i^{(t)}; \theta^{(t+1)}, \beta_i^{(t+1)}) f(\alpha_i^{(t)}; \mu_{\alpha_i}, \sigma_{\alpha_i}) / J(\alpha_i^* | \alpha_i^{(t)})} \quad (19.73)$$

Setze $\alpha_i^{(t+1)} = \alpha_i^*$, mit der Wahrscheinlichkeit $P(\alpha_i^{(t+1)} = \alpha_i^*) = \min(r(\alpha_i^* | \alpha_i^{(t)}), 1)$, bzw. $\alpha_i^{(t+1)} = \alpha_i^{(t)}$, mit der Gegenwahrscheinlichkeit $1 - P(\alpha_i^{(t+1)} = \alpha_i^*)$.

■ Schritt 3

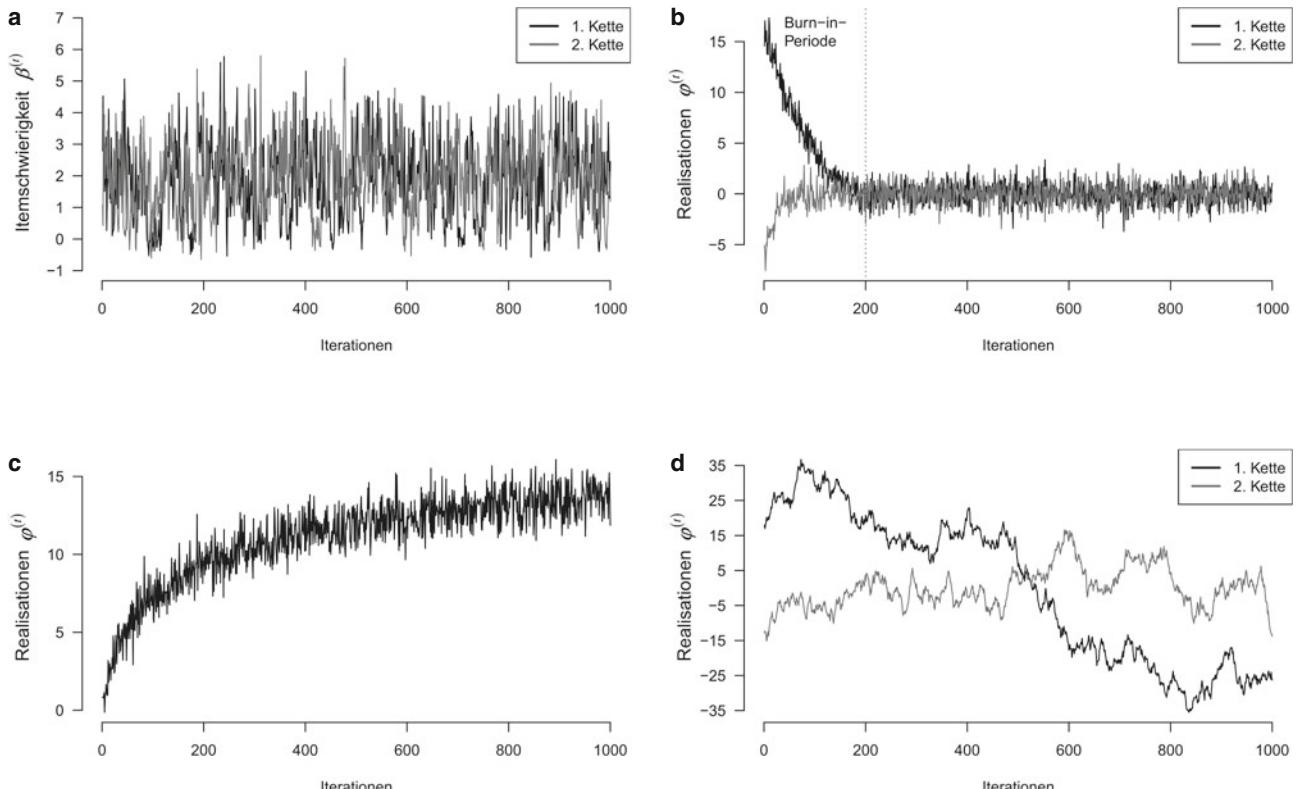
Wiederhole die Schritte 1 und 2 so lange, bis eine vorher festgelegte Iterationszahl T oder ein anderes Abbruchkriterium erreicht ist.

Sollten die Verteilungsparameter μ_θ und σ_θ selbst zu schätzende Größen sein, so kann der Algorithmus um einen entsprechenden Schritt erweitert werden, sodass auch für diese beiden Parameter aus einer Vorschlagsdichte Kandidaten generiert, die zugehörigen Akzeptanzverhältnisse und die Akzeptanzwahrscheinlichkeiten berechnet und die Werte $\mu_\theta^{(t+1)}$ und $\sigma_\theta^{(t+1)}$ bei gegebenen $\mu_\theta^{(t)}$ und $\sigma_\theta^{(t)}$ gezogen werden können.

Konvergenzdiagnostik, Burn-in-Periode und Anzahl der Iterationen

Bei der praktischen Anwendung sind in Bezug auf MCMC-Verfahren einige Dinge zu beachten. Gemäß der Theorie der Markov-Ketten ist eine hinreichend große Zahl T von Iterationen, d. h. Zufallsziehungen aus den A-posteriori-Verteilungen, für eine valide Inferenz bezüglich der gesuchten Parameter nötig. Dabei gibt es keine klaren Richtlinien, was als hinreichend anzusehen ist. Üblicherweise wird T nicht einfach festgelegt. Vielmehr wird eine sog. „Konvergenzdiagnostik“ empfohlen, mit der anhand grafischer Methoden und statistischer Kennwerte untersucht wird, ob und ab wann der Sampling-Algorithmus Realisierungen $\varphi^{(t)}$ aus einer stationären A-posteriori-Verteilung generiert. Zu den am häufigsten verwendeten grafischen Methoden gehören die sog. „Traceplots“. Das sind Liniendiagramme, mit den Iterationen $t = 1, \dots, T$ auf der Abszisse und den Parameterwerten $\varphi^{(t)}$ auf der Ordinate. Dabei sollten die Parameter unsystematisch, d. h. unabhängig von t um den Mittelwert der A-posteriori-Verteilung streuen. In ▶ Abb. 19.4 sind vier verschiedene Traceplots dargestellt. ▶ Abb. 19.4a zeigt den Traceplot mit den ersten 1000 realisierten Werten $\beta^{(t)}$ der Itemschwierigkeit aus dem oben eingeführten Datenbeispiel (▶ Beispiel 19.2). Um die Konvergenz von MCMC-Algorithmen zu prüfen, werden oft mehrere unabhängige Markov-Ketten gleichzeitig zur Parameterschätzung verwendet. Diesem Vorgehen liegt die Idee zugrunde, dass

Konvergenzdiagnostik und Traceplots



■ Abb. 19.4 Darstellung verschiedener Traceplots. **a** Konvergenz zweier unabhängiger Ketten eines Gibbs-Samplers; **b** Effekt unterschiedlicher Startwerte in der Burn-in-Periode; **c** und **d** zeigen die systematische Drift und die Oszillation bei fehlender Konvergenz

bei identischen Daten unterschiedliche Ketten trotz unterschiedlicher Startwerte zu der gleichen A-posteriori-Verteilung konvergieren müssen. Ist dem so, sollten die Werte beider Ketten unsystematisch um den gleichen mittleren Wert der A-posteriori-Verteilung des jeweiligen Parameters streuen (Abb. 19.4a).

In Abhängigkeit von den Startwerten und anderen simultan geschätzten Modellparametern braucht es in manchen Anwendungen eine gewisse Zeit, bis die Realisationen $\varphi^{(t)}$ aus einem MCMC-Algorithmus repräsentativ für die A-posteriori-Verteilung sind. Aus diesem Grund werden häufig die initialen Werte $\varphi^{(t)}$ der sog. „Burn-in-Periode“ verworfen. Abb. 19.4b zeigt zwei unabhängige Ketten mit unterschiedlichen Startwerten. Nach ungefähr 200 Iterationen streuen die realisierten Werte $\varphi^{(t)}$ beider Ketten gleich stark um einen gemeinsamen Wert, dem Erwartungswert der A-posteriori-Verteilung. Beide Ketten sind daher gegen die gleiche stationäre A-posteriori-Verteilung konvergiert. Die Realisationen $\varphi^{(t)}$ aus beiden Ketten sind somit erst nach den ersten 200 Iterationen der Burn-in-Periode gleich repräsentativ bezüglich der A-posteriori-Verteilung. Die Werte $\varphi^{(t)}$ der Burn-in-Periode werden verworfen, da sie wie Extrem- oder Ausreißerwerte zu verzerrten Schlüssen in Bezug auf φ führen können. Fehlende Konvergenz und damit nicht vertrauenswürdige Parameterschätzungen fallen durch eine systematische Parameter-Drift (Abb. 19.4c) oder oszillierende Graphen in Traceplots (Abb. 19.4d) auf. In beiden Fällen zeigt sich zu keinem Zeitpunkt, dass die Werte $\varphi^{(t)}$ aus einer eindeutig bestimmten Verteilung mit fixen Parametern (z. B. Erwartungswert und Streuung) stammen. Wenn zwei oder mehr Ketten verwendet werden, zeigt sich die Parameter-Drift oft auch in unterschiedliche Richtungen. Das heißt, die Werte $\varphi^{(t)}$ aus einer ersten Kette tendieren zu immer größeren Werten, während die Realisationen aus einer zweiten Kette tendentiell immer kleiner werden. Ändern diese Trends scheinbar abrupt immer wieder ihre Richtung und das für unterschiedliche Ketten in ganz unterschiedlicher Weise, resultieren die oszillierenden Graphen wie sie in Abb. 19.4d dargestellt sind. Nicht selten sind hierfür Modellfehlspezifikationen oder eine fehlende Identifikation des Modells verantwortlich.

Zur Konvergenzdiagnostik wurde auch eine ganze Reihe statistischer Kennwerte vorgeschlagen. Ein gute Übersicht ist bei Cowles und Carlin (1996) zu finden. Hier soll mit dem *Potential Scale Reduction (PSR)-Faktor* von Gelman und Rubin (1992) nur einer der gebräuchlichsten Kennwerte erläutert werden. Zur Bestimmung des PSR-Faktors müssen mindestens $P \geq 2$ unabhängige Markov-Ketten (z. B. Gibbs-Sampler) simultan iterieren. Für einen einzelnen Parameter φ_m aus einem Parametervektor $\varphi = \varphi_1, \dots, \varphi_M$ werden nun in jeder Kette $p = 1, \dots, P$ aus den M univariaten A-posteriori-Verteilungen zufällig $t = 1, \dots, T$ Werte $\varphi_m^{(pt)}$ generiert. Nun können die Varianzen von $\varphi_m^{(pt)}$ sowohl innerhalb der Ketten als auch zwischen den Ketten verglichen werden. Die Varianz innerhalb der Ketten gibt an, wie stark die Werte $\varphi_m^{(pt)}$ um die jeweiligen Mittelwerte $\bar{\varphi}_m^{(p)}$ innerhalb der einzelnen Ketten variieren. Sie wird wie folgt berechnet:

$$W = \frac{1}{P(T-1)} \sum_{p=1}^P \sum_{t=1}^T (\varphi_m^{(pt)} - \bar{\varphi}_m^{(p)})^2 \quad (19.74)$$

Die Varianz zwischen den Ketten gibt an, wie stark die P Mittelwerte $\bar{\varphi}_m^{(1)}, \dots, \bar{\varphi}_m^{(P)}$ der Ketten um den Gesamtmittelwert $\bar{\bar{\varphi}}$ aller $P \cdot T$ Werte $\varphi_m^{(pt)}$ variieren:

$$B = \frac{1}{P-1} \sum_{p=1}^P (\bar{\varphi}_m^{(p)} - \bar{\bar{\varphi}})^2 \quad (19.75)$$

Burn-in-Periode

Parameter-Drift

PSR-Faktor

19.4 · Weitere Schätzverfahren

Letztlich kann unter Verwendung von W und B auch die Varianz der A-posteriori-Verteilung von φ_m bei gegebenen Daten geschätzt werden:

$$\widehat{\text{Var}}(\varphi_m | \mathbf{Y} = \mathbf{y}) = \frac{T-1}{T} W + \frac{M+1}{TM} B \quad (19.76)$$

Wenn nun alle M Ketten gegen die gleiche stationäre A-posteriori-Verteilung konvergieren, muss die Varianz zwischen den Ketten mit steigender Iterationszahl T immer kleiner werden, sodass sich das Verhältnis von $\widehat{\text{Var}}(\varphi_m | \mathbf{Y} = \mathbf{y})$ und W bei großer Zahl T dem Wert eins annähern sollte. Der PSR-Faktor ist letztlich definiert als die Quadratwurzel dieses Verhältnisses:

$$PSR = \sqrt{\frac{\widehat{\text{Var}}(\varphi_m | \mathbf{Y} = \mathbf{y})}{W}} \quad (19.77)$$

Gemäß Konvention werden Werte von $PSR \leq 1.1$ oder $PSR \leq 1.05$ als Konvergenzkriterium verwendet. Große Werte des PSR-Faktors bedeuten, dass entweder die Varianz zwischen den Ketten zu hoch ist, um eine Konvergenz annehmen zu können, oder die Varianz innerhalb der Ketten noch zu gering ist. In letzterem Fall sind weitere Iterationen nötig, um die gesamte A-posteriori-Verteilung von φ_m hinreichend zu approximieren. Der PSR-Faktor wird üblicherweise für jeden Parameter eines Modells separat berechnet, die Konvergenz gilt dann als erreicht, wenn für jeden Modellparameter $PSR \leq 1.1$ oder $PSR \leq 1.05$ gilt. In □ Tab. 19.2 des ► Beispiels 19.2 sind die PSR-Faktoren für die Itemdiskrimination und die Itemschwierigkeit für $T = 4000$ angegeben. Für dieses Anwendungsbeispiel kann Konvergenz angenommen werden, da beide Parameter bei 1.001 liegen und damit deutlich kleiner als 1.05 sind.

Konvergenzkriterium: $PSR \leq 1.1$ oder $PSR \leq 1.05$

19.4 Weitere Schätzverfahren

Mit den verschiedenen Verfahren der ML-Schätzung (JML, CML und MML) sowie den simulationsbasierten und nicht simulationsbasierten Bayes'schen Ansätzen wurden längst nicht alle Methoden zur Parameterschätzung in IRT-Modellen beschrieben. Eine umfassende Darstellung würde den Rahmen dieses Kapitels allerdings bei Weitem übersteigen. Es soll jedoch zumindest erwähnt werden, dass auch anhand von Strukturgleichungsmodellen für dichotome und ordinale Variablen die Parameter von 1PL- und 2PL-Modellen geschätzt werden können. In diesen speziellen Strukturgleichungsmodellen wird angenommen, dass den manifesten Variablen Y_i latente, aber messfehlerbehaftete kontinuierliche, normalverteilte Antwortvariablen Y_i^* („latent response variables“) zugrunde liegen. Das Messmodell ist zweiteilig. Mithilfe von Schwellenparametern δ_{ic} für jedes Item i und den Kategorien $c = 1, \dots, C$ werden die Beziehungen zwischen den manifesten Variablen Y_i und den latenten Antwortvariablen Y_i^* beschrieben. Dabei gilt:

$$\delta_{ic} < Y_i^* \leq \delta_{i(c+1)} \Leftrightarrow Y_i = c \quad (19.78)$$

Strukturgleichungsmodelle für dichotome und ordinale Variablen

Bei dichotomen Items gibt es nur einen Schwellenparameter δ_{i1} , sodass gilt: Wenn der Wert der latenten Antwortvariablen größer ist als der Schwellenparameter ($\delta_{i1} < Y_i^*$), so ist $Y_i = 1$. Ist der Wert von Y_i^* jedoch kleiner oder gleich δ_{i1} , so folgt $Y_i = 0$. Unter Annahme der Standardnormalverteilung $Y_i^* \sim N(0, 1)$ lassen sich die Schwellen leicht anhand der relativen Kategorienhäufigkeiten und der Quantilfunktion der Standardnormalverteilung schätzen. Der zweite Teil des Messmodells beschreibt den linearen regressiven Zusammenhang zwischen Y_i^* und der messfehlerfreien latenten Personenvariable θ . In einem eindimensionalen

Schwellenparameter

Messmodell lautet die Modellgleichung für jedes Item i wie folgt:

$$Y_i^* = \lambda_{1i}\theta + \varepsilon_i^* \quad (19.79)$$

Dabei ist λ_{1i} die Faktorladung und ε_i^* das Regressionsresiduum bzw. der Messfehler von Y_i^* . Die Parameter der 2PL-Modelle lassen sich leicht aus den Modellparametern der Strukturgleichungsmodelle errechnen. Werden alle Faktorladungen und Schwellenparameter frei geschätzt, so müssen die Verteilungsparameter der latenten Personenvariablen θ fixiert werden. Wird $E(\theta) = 0$ und $Var(\theta) = 1$ gewählt, lassen sich die Itemdiskrimination und die Itemschwierigkeit des 2PL-Modells nach Birnbaum anhand folgender Formeln schätzen:

$$\alpha_i \approx \frac{1.7\lambda_i}{\sqrt{1 - \lambda_i^2}}; \quad \beta_i \approx \frac{\delta_{i1}}{\lambda_{11}} \quad (19.80)$$

Kleinste-Quadrat- und paarweise ML-Schätzung

Mehrebenen-Regressionsmodelle für dichotome oder ordinale Variablen

Die Parameterschätzung der Strukturgleichungsmodelle für kategoriale Variablen kann anhand verschiedener *Kleinste-Quadrat-Schätzer* (Forero und Maydeu-Olivares 2009), basierend auf den tetrachorischen (bei dichotomen Variablen Y_i) oder polychorischen Korrelationen (bei mehrkategorialen ordinalen Variablen Y_i) erfolgen. Die tetra- und polychorischen Korrelationen entsprechen dabei den Produkt-Moment-Korrelationen $Kor(Y_j^*, Y_h^*)$ zwischen den latenten Antwortvariablen, die anhand der bivariaten Häufigkeitsverteilungen der manifesten Items Y_i und Y_h , für alle Paare $i \neq h$, geschätzt werden können. Alternativ kann auch die paarweise ML-Schätzung nach Katsikatsou et al. (2012) verwendet werden, die im R-Paket lavaan (Rosseel 2012) implementiert ist.

Auch *Mehrebenen-Regressionsmodelle* für dichotome oder ordinale Variablen (z. B. logistische Mehrebenen-Regressionsmodelle) finden bei der Parameterschätzung von IRT-Modellen Anwendung. Die Grundidee ist, dass die Itemantworten als genestet in Personen und in Items betrachtet werden (sog. „kreuzklassifizierte Nestungsstruktur“; De Boeck 2008; Van den Noortgate et al. 2003). Die verwendeten Schätzmethoden sind folglich nicht spezifisch für die IRT und können in der einschlägigen Literatur zu generalisierten Mehrebenen-Regressionsmodellen nachgelesen werden (z. B. Demidenko 2013).

19.5 Personenparameterschätzung in der IRT

In den vorangegangenen Abschnitten zur Parameterschätzung in IRT-Modellen lag der Fokus auf der Bestimmung der Itemparameter. Wie bei der JML-Methode erläutert, werden die Personenparameter als inzidentelle Parameter aufgefasst und bei den CML- und MML-Schätzungen zunächst nicht mitgeschätzt, sondern erst in einem zweiten Schritt auf Basis der initial geschätzten Itemparameter bestimmt.

Auch bei der Personenparameterschätzung können ML- oder Bayes'sche Schätzer verwendet werden. Da die verschiedenen Schätzer unterschiedliche Eigenschaften haben, muss in Abhängigkeit vom jeweiligen Verwendungszweck entschieden werden, welchem der Vorzug zu geben ist. In diesem Abschnitt sollen die bekanntesten Personenparameterschätzer, und zwar der ML-Schätzer, der gewichtete ML-Schätzer nach Warm (1989), der MAP-Schätzer, der EAP-Schätzer sowie das PV-Verfahren vorgestellt werden (vgl. ▶ Abschn. 19.1). Neben der Bestimmung dieser Schätzer soll kurz auf ihre wichtigsten anwendungsrelevanten Eigenschaften eingegangen werden.

Verschiedene Personenparameterschätzer mit unterschiedlichen Eigenschaften

19.5.1 ML-Scoring

Sind die Itemparameter bekannt oder wurden sie in einem ersten Schritt anhand der CML- oder MML-Schätzung bestimmt, so können die individuellen Ausprägungen auf der latenten PersonenvARIABLEN auf Basis der Antwortmuster $\mathbf{Y}_v = \mathbf{y}_v$ separat für alle Personen $v = 1, \dots, N$ geschätzt werden. Wird dazu die ML-Schätzung verwendet, wird das Verfahren auch als „ML-Scoring“ bezeichnet. Dabei ist der ML-Schätzer $\hat{\theta}_v$ einer Person v der Wert der latenten Variablen für den die Antwortmusterwahrscheinlichkeit $P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{t})$ (Gl. 19.12) maximal ist. Die individuellen zu maximierenden Likelihood-Funktionen sind nun also gerade die Antwortmusterwahrscheinlichkeiten, die unter Verwendung des jeweiligen Modells als Funktion der latenten Variablen beschrieben werden. Die Itemparameter werden beim ML-Scoring als fixe Größen und nicht als unbekannte Parameter betrachtet. Für das eindimensionale 2PL-Modell nach Birnbaum lautet die individuelle Antwortmusterwahrscheinlichkeit für das Antwortmuster $\mathbf{y}_v = y_{v1}, \dots, y_{vk}$ von Person v wie folgt:

$$\begin{aligned} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{t}) &= \prod_{i=1}^k P(Y_{vi} = 1 | \theta_v; t_i)^{y_{vi}} P(Y_{vi} = 0 | \theta_v; t_i)^{1-y_{vi}} \\ &= \prod_{i=1}^k \left(\frac{\exp[\alpha_i(\theta_v - \beta_i)]}{1 + \exp[\alpha_i(\theta_v - \beta_i)]} \right)^{y_{vi}} \\ &\quad \cdot \left(\frac{1}{1 + \exp[\alpha_i(\theta_v - \beta_i)]} \right)^{1-y_{vi}} \end{aligned} \quad (19.81)$$

Genau wie bei der ML-Schätzung der Itemparameter kann das Maximum dieser Funktion ermittelt werden, indem die Log-Likelihood-Funktion $l(\theta_v)$ (der natürliche Logarithmus aus Gl. 19.81) nach θ_v abgeleitet und gleich null gesetzt wird. Es ergibt sich die folgende Schätzgleichung:

$$\frac{dl'(\theta_v)}{d\theta_v} = \sum_{i=1}^k \alpha_i [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] \quad (19.82)$$

Diese ist identisch mit der partiellen Ableitungen der Log-Likelihood $l(\boldsymbol{\varphi})$ der Gesamtdaten, wie sie bereits bei der JML-Schätzung eingeführt wurde (vgl. Gl. 19.19). Gl. 19.82 lässt sich wiederum nicht analytisch lösen, sodass iterative Verfahren, z. B. der Newton-Raphson-Algorithmus, zur Lösung verwendet werden. Dabei wird durch Iteration t der Wert $\hat{\theta}_v^{(t+1)}$ nach folgender Gleichung bestimmt:

$$\hat{\theta}_v^{(t+1)} = \hat{\theta}_v^{(t)} - \frac{l'(\hat{\theta}_v^{(t)})}{l''(\hat{\theta}_v^{(t)})} \quad (19.83)$$

Es wird also die zweite Ableitung $l''(\hat{\theta}_v^{(t)})$ der Log-Likelihood benötigt, die in allgemeiner Form durch folgende Gleichung gegeben ist:

$$\frac{d^2 l(\theta_v)}{d\theta_v^2} = - \sum_{i=1}^k \alpha_i^2 P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i) P(Y_{vi} = 0 | \theta_v; \alpha_i, \beta_i) \quad (19.84)$$

Der Algorithmus iteriert bis zum Erreichen eines vorab festgelegten Abbruchkriteriums (z. B. dass die Werte der Log-Likelihood zwischen zwei aufeinanderfolgenden Iterationsschritten eine bestimmte Grenze unterschreiten). Der Wert $\hat{\theta}_v^{(T)}$ aus

Testinformationsfunktion

der letzten Iteration wird dann als ML-Schätzer verwendet. Gemäß der allgemeinen Schätztheorie lässt sich der Standardfehler eines ML-Schätzers anhand der negativen zweiten Ableitung der Log-Likelihood-Funktion bestimmen, die auch als *beobachtete Fisher-Information* bezeichnet wird. Im Fall der Personenparameterschätzung in der IRT wird diese auch Testinformationsfunktion $T(\theta)$ genannt. Die inverse beobachtete Fisher-Information ist gleich der Varianz des ML-Schätzers (► Abschn. 19.2.3). Der Standardfehler ist die Quadratwurzel aus der Varianz des Schätzers. Folglich ergibt sich der Standardfehler für den ML-Personenparameterschätzer unter Verwendung von Gl. (19.84) wie folgt:

$$\begin{aligned} SE_{ML}(\hat{\theta}_v) &= \sqrt{\frac{1}{T(\hat{\theta}_v)}} \\ &= \frac{1}{\sqrt{\sum_{i=1}^k \alpha_i^2 P(Y_{vi}=1|\hat{\theta}_v; \alpha_i, \beta_i) P(Y_{vi}=0|\hat{\theta}_v; \alpha_i, \beta_i)}} \end{aligned} \quad (19.85)$$

**Standardfehler für den
ML-Personenparameterschätzer**

Die Standardfehler für die geschätzten individuellen Ausprägungen einer latenten Variablen sind zugleich ein inverses Reliabilitätsmaß. Je kleiner der Standardfehler von $\hat{\theta}_v$ ist, desto reliabler sind die Personenparameterschätzer, die aus dem Test resultieren. Auf diesen Punkt wird in ► Abschn. 19.6 noch vertiefend eingegangen.

Gewichtete ML-Funktion**19.5.2 Gewichtete ML-Schätzung**

Warm (1989) stellte eine gewichtete ML-Schätzung der Personenfähigkeit in IRT-Modellen vor, die einen geringeren Schätzfehler im Vergleich zur klassischen ML-Schätzung aufweist und aus diesem Grund häufig bevorzugt wird. Außerdem erlaubt die gewichtete ML-Schätzung auch beim Vorliegen von Extremwerten (wenn alle Items oder gar kein Item gelöst/bejaht wurden) eine Punktschätzung für die Ausprägung der latenten Personenvariablen. Anstelle der ML-Funktion für ein Antwortmuster (Gl. 19.81) wird die gewichtete (weighted) ML-Funktion $wL(\theta_v)$ bzw. deren natürlicher Logarithmus $wl(\theta_v)$ maximiert:

$$wl(\theta_v) = \ln [P(Y_v = y_v | \theta_v; \mathbf{t})] + \ln [w(\theta_v)] \quad (19.86)$$

Der zweite Summand ist dabei ein recht komplexer Korrekturterm, der die Verzerung des ML-Schätzers ausgleicht:

$$\ln [w(\theta_v)] = \frac{J(\theta)}{2T(\theta)} \quad (19.87)$$

wobei:

$$J(\theta) = \sum_{i=1}^k \frac{\frac{\partial}{\partial \theta} P(Y_{vi}=1|\theta; \alpha_i, \beta)}{P(Y_{vi}=1|\theta; \alpha_i, \beta)} \frac{\frac{\partial^2}{\partial \theta^2} P(Y_{vi}=1|\theta; \alpha_i, \beta)}{P(Y_{vi}=0|\theta; \alpha_i, \beta)} \quad (19.88)$$

$T(\theta)$ ist wiederum die Testinformationsfunktion. Die Terme $\frac{\partial}{\partial \theta} P(Y_{vi}=1|\theta; \alpha_i, \beta)$ und $\frac{\partial^2}{\partial \theta^2} P(Y_{vi}=1|\theta; \alpha_i, \beta)$ stellen die erste und zweite partielle Ableitung der Modellgleichung des jeweiligen IRT-Modells (z. B. Rasch- oder Birnbaum-Modell) nach der latenten Variablen θ dar. Der gewichtete ML-Schätzer nach Warm ist jener Wert von θ , für den die Gl. (19.86) den Wert null annimmt. Die

Lösung muss durch numerische Verfahren iterativ bestimmt werden (z. B. Newton-Raphson-Algorithmus). Aufgrund des Korrekturterms ist auch der Standardfehler für den gewichteten ML-Schätzer etwas komplizierter:

$$SE_{ML}(\hat{\theta}_v) = \sqrt{\frac{1}{T(\hat{\theta}_v) + \frac{\frac{\partial}{\partial \theta} T(\hat{\theta}_v) J(\hat{\theta}_v) + \frac{\partial}{\partial \theta} J(\hat{\theta}_v) T(\hat{\theta}_v)}{2T(\hat{\theta}_v)^2}}}} \quad (19.89)$$

Im Vergleich zur Formel, mit der der Standardfehler des ML-Schätzers berechnet wird (Gl. 19.85), kommt in Abhängigkeit vom Korrekturterm immer noch ein mehr oder minder großer Betrag im Nenner dazu. Daher hat der gewichtete ML-Schätzer generell etwas kleinere Standardfehler und somit eine geringere Variabilität als der herkömmliche ML-Schätzer.

Standardfehler des gewichteten ML-Schätzers

19.5.3 Bayes'sche Personenparameterschätzer

Alle im Folgenden vorgestellten Bayes'schen Personenparameterschätzungen beruhen auf den individuellen A-posteriori-Verteilungen $f(\theta_v | \mathbf{Y}_v = \mathbf{y}_v)$ der latenten Variablen bei gegebenen individuellen Antwortmustern. Wie bereits bei der Itemparameterschätzung erläutert, können verschiedene Kennwerte einer A-posteriori-Verteilung als Punktschätzer verwendet werden. Der Modus bzw. das Maximum der A-posteriori-Verteilung ist der *MAP-Schätzer*. Der Erwartungswert der A-posteriori-Verteilung ist der *EAP-Schätzer*. Simulationsbasierte Bayes'sche Algorithmen wie der *Gibbs-Sampler* (► Abschn. 19.3.4.2) oder der *Metropolis Hastings (MH)-Algorithmus* (► Abschn. 19.3.4.2) erlauben es außerdem, beliebig oft zufällig Werte aus den individuellen A-posteriori-Verteilung zu ziehen. Diese Realisierung wird als *PV-Verfahren* bezeichnet, wobei die PVs in der weiterführenden statistischen Analyse von Testdaten von großer Bedeutung sind, obwohl sie sich für individualdiagnostische Zwecke nicht eignen. Im Folgenden soll kurz die Berechnung der verschiedenen Kennwerte beschrieben werden.

19.5.3.1 MAP-Schätzung

In ► Abschn. 19.3.2 wurde die A-posteriori-Verteilung bereits in allgemeiner Form als normiertes Produkt von Likelihood und A-priori-Verteilung eingeführt (Gl. 19.53). Für das eindimensionale 2PL-Modell nach Birnbaum lautet die A-posteriori-Verteilung der latenten Variablen für eine Person v bei gegebenem Antwortmuster $\mathbf{Y}_v = \mathbf{y}_v$:

$$f(\theta_v | \mathbf{Y}_v = \mathbf{y}_v) = \frac{P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\iota}) f(\theta_v; \mu_\theta, \sigma_\theta)}{\int_{\theta} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \boldsymbol{\iota}) f(\theta_v; \mu_\theta, \sigma_\theta) d\theta} \quad (19.90)$$

Als A-priori-Verteilung $f(\theta_v; \mu_\theta, \sigma_\theta)$ wird bei der Personenparameterschätzung üblicherweise die Normalverteilung $N(\mu_\theta, \sigma_\theta)$ gewählt. Die Verteilungsparameter μ_θ und σ_θ sind hier die Populationsparameter, d. h. die Kennwerte der Verteilung der latenten Personenvariablen in der Population, aus der die betreffende Person stammt. Diese sind entweder bekannt oder wurden zusammen mit den Itemparametern in einem ersten Schritt geschätzt. Bei der MAP-Schätzung wird das Maximum der A-posteriori-Verteilung gesucht. Gl. (19.90) lässt sich dabei einfach als Funktion der latenten Variablen auffassen, deren Maximum der gesuchte MAP-Schätzer ist. Es handelt sich also genau wie bei der ML-Schätzung um ein Maximierungsproblem. Dabei reicht es, den Zähler von Gl. (19.90) als zu maximierende Funktion zu betrachten, da der Nenner der A-posteriori-Verteilung lediglich

Maximum der A-posteriori-Verteilung

eine Normierungskonstante ist. Wie bei der ML-Schätzung und auch der Bayes-Modal-Schätzung der Itemparameter (► Abschn. 19.3.4.1) wird aufgrund besserer mathematischer Eigenschaften auch hier der natürliche Logarithmus der Funktion verwendet. Die zu maximierende Funktion ist dann

$$\begin{aligned} & \ln [P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{l}) f(\theta_v; \mu_\theta, \sigma_\theta)] \\ &= \ln [P(Y_v = y_v | \theta_v; \mathbf{l})] + \ln [f(\theta_v; \mu_\theta, \sigma_\theta)] \\ &= l(\theta_v) + \ln [f(\theta_v; \mu_\theta, \sigma_\theta)] \end{aligned} \quad (19.91)$$

Es handelt sich also um die Summe aus der Log-Likelihood-Funktion und der logarithmierten A-priori-Verteilung. Leitet man Gl. (19.91) nach den latenten Variablen ab, ergibt sich die Schätzgleichung für den MAP-Schätzer:

$$\begin{aligned} & \frac{d}{d\theta_v} \ln [P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{l}) f(\theta_v; \mu_\theta, \sigma_\theta)] \\ &= \frac{d}{d\theta_v} l(\theta_v) + \frac{d}{d\theta_v} \ln [f(\theta_v; \mu_\theta, \sigma_\theta)] \\ &= \sum_{i=1}^k \alpha_i [y_{vi} - P(Y_{vi} = 1 | \theta_v; \alpha_i, \beta_i)] - \frac{\theta_v - \mu_\theta}{\sigma_\theta^2} \end{aligned} \quad (19.92)$$

Im Vergleich zum ML-Schätzer (Gl. 19.82) wird bei der MAP-Schätzung die mit der inversen Varianz σ_θ^2 gewichtete Abweichung $\theta_v - \mu_\theta$ subtrahiert. Dadurch ergibt sich wiederum der bereits beschriebene Shrinkage-Effekt, der für Bayes'sche Schätzer charakteristisch ist. Das heißt, sowohl unter- als auch überdurchschnittliche potentielle Werte θ_v werden in Richtung des Erwartungswertes der A-priori-Verteilung gezogen.

Auch die Bestimmung der Variabilität, d. h. der Unreliabilität des MAP-Schätzers, kann analog zur ML-Schätzung erfolgen. Bildet man die negative zweite Ableitung von Gl. (19.91), so erhält man die A-posteriori-Information. Die Quadratwurzel aus dem Kehrwert der A-posteriori-Information ist dann eine Schätzung der Streuung der A-posteriori-Verteilung. Für den MAP-Schätzer der Personenparameters ergibt sich folgender Standardfehler:

$$SE_{MAP}(\hat{\theta}_v) = \sqrt{\frac{1}{\sum_{i=1}^k \alpha_i^2 P(Y_{vi} = 1 | \hat{\theta}_v; \alpha_i, \beta_i) P(Y_{vi} = 0 | \hat{\theta}_v; \alpha_i, \beta_i) + \frac{1}{\sigma_\theta^2}}} \quad (19.93)$$

19.5.3.2 EAP-Schätzung

Im Unterschied zur MAP-Schätzung bedarf es bei der EAP-Schätzung anstelle einer Maximierung der Integration über die individuelle A-posteriori-Verteilung. Definiert ist der EAP-Schätzer für eine Person v bei gegebenem Antwortmuster $\mathbf{Y}_v = \mathbf{y}_v$ als bedingter Erwartungswert:

$$\begin{aligned} E(\theta_v | \mathbf{Y}_v = \mathbf{y}_v) &= \int_{\theta_v} \theta_v f(\theta_v | \mathbf{Y}_v = \mathbf{y}_v) d\theta_v \\ &= \int_{\theta_v} \theta_v P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{l}) f(\theta_v; \mu_\theta, \sigma_\theta) d\theta_v \\ &= \frac{\int_{\theta_v} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{l}) f(\theta_v; \mu_\theta, \sigma_\theta) d\theta_v}{\int_{\theta_v} P(\mathbf{Y}_v = \mathbf{y}_v | \theta_v; \mathbf{l}) f(\theta_v; \mu_\theta, \sigma_\theta) d\theta_v} \end{aligned} \quad (19.94)$$

19.5 · Personenparameterschätzung in der IRT

Beide Integrale in Gl. (19.94) sind analytisch nicht berechenbar (d.h., sind als geschlossener Ausdruck nicht darstellbar). Wie bei der MML-Schätzung der Itemparameter kommen auch hier zur Berechnung der EAP-Schätzung numerische Integrationsverfahren unter Verwendung von Quadraturformeln zur Anwendung. (► Abschn. 19.2.6). Die Integrale in Gl. (19.94) sind dann gewichtete Summen über die Q verschiedenen Quadraturpunkte $X_{\theta}^{(q)}$ (mit $q = 1, \dots, Q$), sodass der EAP-Schätzer über die folgende Gleichung näherungsweise bestimmt werden kann:

$$E(\theta_v | Y_v = y_v) \approx \frac{\sum_{q=1}^Q X_{\theta}^{(q)} P(Y_v = y_v | X_{\theta}^{(q)}; \mathbf{t}) P(X_{\theta}^{(q)})}{\sum_{q=1}^Q P(Y_v = y_v | X_{\theta}^{(q)}; \mathbf{t}) P(X_{\theta}^{(q)})} \quad (19.95)$$

Die konkreten Werte der Gewichte $P(X_{\theta}^{(q)})$ hängen von der Wahl der jeweiligen A-priori-Verteilung $f(\theta_v; \mu_{\theta}, \sigma_{\theta})$ und des spezifischen Integrationsverfahrens ab. Auch die Standardabweichung der A-posteriori-Verteilung (*Posterior Standard Deviation, PSD*) kann anhand der numerischen Integration berechnet werden. Theoretischer Ausgangspunkt ist die Definition der Varianz einer beliebigen Zufallsvariable Z als Erwartungswert der quadrierten Abweichungen $(Z - \mu_z)^2$, wobei μ_z der Erwartungswert von Z ist. Angewendet auf die Varianz der A-posteriori-Verteilung ergibt sich die folgende Gleichung:

$$E[(\theta_v - \mu_{\theta})^2 | Y_v = y_v] = \frac{\int (\theta_v - \mu_{\theta})^2 P(Y_v = y_v | \theta_v; \mathbf{t}) f(\theta_v; \mu_{\theta}, \sigma_{\theta}) d\theta_v}{\int P(Y_v = y_v | \theta_v; \mathbf{t}) f(\theta_v; \mu_{\theta}, \sigma_{\theta}) d\theta_v} \quad (19.96)$$

Genau wie in Gl. (19.94) zur Bestimmung des EAP ist hier das Integral über die A-posteriori-Verteilung erforderlich. Die Standardabweichung der A-posteriori-Verteilung ist schließlich die Quadratwurzel von Gl. (19.96). Sie kann durch Quadraturformeln wie folgt bestimmt werden:

$$PSD(\hat{\theta}_v) \approx \sqrt{\frac{\sum_{q=1}^Q (X_{\theta}^{(q)} - \mu_{\theta})^2 P(Y_v = y_v | X_{\theta}^{(q)}; \mathbf{t}) P(X_{\theta}^{(q)})}{\sum_{q=1}^Q P(Y_v = y_v | X_{\theta}^{(q)}; \mathbf{t}) P(X_{\theta}^{(q)})}} \quad (19.97)$$

19.5.3.3 Plausible Values (PV)-Verfahren

Plausible Values (PVs) stellen keine geeigneten Punktschätzer für die individuelle Ausprägung latenter Variablen einer Person dar. Vielmehr sind sie Zufallsziehungen aus den individuellen A-posteriori-Verteilungen der latenten Variablen einer Person mit ihrem individuellen Antwortmuster. Entsprechend werden für jede Person separat mehrere PVs gezogen, wobei simulationsbasierte Bayes-Algorithmen wie der Gibbs-Sampler oder der MH-Algorithmus verwendet werden können (► Abschn. 19.3.4.2). Anhand der PVs kann nun die gesamte A-posteriori-Verteilung einer Person mit der gewünschten Präzision beschrieben werden, wobei die Präzision mit der Zahl der generierten PVs steigt; PVs müssen dabei nicht zwingend in einem vollständigen Bayes-IRT-Modell generiert werden. Oft werden

Quadraturformeln

Standardabweichung der A-posteriori-Verteilung

PVs sind Zufallsziehungen aus den individuellen A-posteriori-Verteilungen

initial die Itemparameter unter Verwendung einer der beschriebenen ML-Verfahren (► Abschn. 19.2.4 bis 19.2.6) geschätzt und die PVs in einem zweiten Schritt generiert, wobei die Itemparameterschätzer als fixe Größen angenommen werden.

Multiple Imputationen

PVs lassen sich ausgehend von der statistischen Theorie fehlender Werte nach Rubin (1976) als Multiple Imputationen (MI) der nicht beobachtbaren und somit inhärent „fehlenden“ latenten Variablen begreifen. Die MI ist ein etabliertes Verfahren, um mit fehlenden Werten in statistischen Analysen umzugehen. Bei der Anwendung der MI auf latente Variablen in der Psychometrie werden die latenten Variablen nicht anders behandelt als manifeste Variablen, die fehlende Werte aufweisen. Um die Unsicherheit bzw. den Messfehler von imputierten Werten hinsichtlich der latenten Variablen in nachfolgenden statistischen Analysen zu berücksichtigen, werden zufällig mehrere Werte aus der A-posteriori-Verteilung gezogen, die für das beobachtete Antwortmusters $Y_v = y_v$ und das jeweils gewählte IRT-Modell plausibel sind.

PVs sind für weiterführende statistische Analysen geeignet, nicht aber zur Individualdiagnostik

Zwar sind PVs für individualdiagnostische Zwecke ungeeignet, jedoch in der sozialwissenschaftlichen Forschung von größter Bedeutung (von Davier et al. 2009). Sie werden in nahezu allen nationalen und internationalen Bildungsvergleichsstudien wie für das *Programme for International Student Assessment* (PISA) und die *Trends in International Mathematics and Science Study* (TIMSS) als Personenparameterschätzer der Wahl für die weiterführenden statistischen Analysen herangezogen. Der Grund ist, dass sie – genauso wie manifeste Variablen – für weitere statistische Analysen verwendet werden können, jedoch ohne die Gefahr von Verzerrungen oder Verfälschungen aufgrund des Messfehlers. Das unterscheidet die PVs von allen anderen hier vorgestellten Personenparameterschätzern! Der einzige Nachteil besteht darin, dass jede statistische Analyse genau so oft wiederholt werden muss, wie PVs generiert wurden. Hinterher müssen die Ergebnisse aus den einzelnen Analysen nach bestimmten Regeln, den sog. „Rubin’s rules“, aggregiert werden, um die finalen Ergebnisse und Statistiken zu erhalten.

Berechnung von Verteilungsparametern der individuellen A-posteriori-Verteilungen anhand von PVs

Da anhand von PVs die gesamte individuelle A-posteriori-Verteilung beschrieben werden kann, lassen sich mit ihnen auch sehr einfach Verteilungsparameter dieser Verteilungen berechnen. So ist das arithmetische Mittel einer hinreichend großen Zahl von PVs ein erwartungstreuer Schätzer für den Erwartungswert der A-posteriori-Verteilung, was nichts anderes ist als der EAP-Schätzer (► Abschn. 19.5.3.2). Auch die Standardabweichungen der individuellen A-posteriori-Verteilung können ganz einfach als Quadratwurzel der Varianz der generierten PVs für jede Person berechnet werden. Wichtig ist dabei, dass diese Standardabweichung nicht als Maß der Schätzgenauigkeit eines einzelnen PVs interpretiert werden sollte. Die Bedeutung der A-posteriori-Standardabweichung ist unabhängig davon, ob sie anhand von Gl. (19.93) bzw. Gl. (19.97) oder als Standardabweichung der PVs berechnet wird. In jedem Fall quantifiziert sie im Bayes'schen Sinne die Unsicherheit hinsichtlich des unbekannten zu schätzenden Personenparameters, bei gegebenen Daten und der A-priori-Information, und hat daher eine andere Bedeutung als der Standardfehler des ML-Schätzers.

19.6 Reliabilitätsbeurteilung in der IRT

In der Testtheorie ist die Genauigkeit bzw. die Reliabilität, mit der eine Personenvariable erfasst wird, ein zentrales Gütekriterium eines Tests. Wird der Test unter Verwendung von IRT-Modellen ausgewertet, stellt sich folglich die Frage, wie reliabel die verschiedenen modellbasierten Personenparameterschätzer sind. Dieser Frage wird in im Folgenden nachgegangen.

19.6.1 Zur Erinnerung: Messgenauigkeit in der Klassischen Testtheorie (KTT)

Die Reliabilität wurde bereits als ein wesentliches Gütekriterium von psychometrischen Tests eingeführt (► Kap. 2). Die Reliabilität ist ein standardisiertes Maß der Messgenauigkeit eines Tests. In der KTT wird eine manifeste Testwertvariable Y als Summe von True-Score τ und Messfehler ε konzeptualisiert ($Y = \tau + \varepsilon$). Da die True-Score-Variable τ und der Messfehler ε unkorreliert sind, ist die Varianz der manifesten Testwertvariable gleich der Summe $Var(Y) = Var(\tau) + Var(\varepsilon)$. Die Reliabilität von Y ist letztlich definiert als Varianzverhältnis $Rel(Y) = Var(\tau)/Var(Y)$. Neben der Reliabilität ist auch der Standardmessfehler $SD(\varepsilon) = \sqrt{Var(Y)[1 - Rel(Y)]}$ als inverses Maß der Messgenauigkeit gebräuchlich (► Kap. 13).

Die Reliabilität ist als Maß der Messgenauigkeit auch in der IRT gebräuchlich, wird jedoch auf andere Weise bestimmt als in der KTT. Ausgangspunkt in der IRT bildet nicht die Zerlegung der manifesten Testwertvariable (z. B. des Summenscores eines Tests), sondern die Zerlegung des Personenparameterschätzers $\hat{\theta}$:

$$\hat{\theta} = \theta + \varepsilon_{\hat{\theta}} \quad (19.98)$$

Dabei ist θ die latente Personvariable mit den wahren zu schätzenden Werten der Personen. Die Differenz zwischen geschätztem Wert und wahren Wert ist der Schätzfehler $\varepsilon_{\hat{\theta}} = \hat{\theta} - \theta$, der dem Messfehler in der KTT entspricht. Die Streuung des Schätzfehlers $\varepsilon_{\hat{\theta}}$ ist der Standardfehler $SE(\hat{\theta})$ des Personenparameterschätzers $\hat{\theta}$. Da es in der IRT verschiedene Personenparameterschätzer mit unterschiedlichen Eigenschaften gibt, ist auch die Bestimmung der Standardfehler und der Reliabilität je nach Schätzer verschieden. Im Folgenden wird zunächst auf die Bestimmung der Standardfehler der Personenparameterschätzer eingegangen, darauf aufbauend wird anschließend gezeigt, wie sich in der IRT Kennwerte für die Reliabilität eines Tests bestimmen lassen und wie sie zu interpretieren sind.

**Unterschiedliche
Personenparameterschätzer mit
spezifischen Standardfehlern und
Reliabilitätsmaßen**

19.6.2 Testinformation und Standardfehler

19.6.2.1 Grundlegendes

In den vorangegangenen Abschnitten sind fünf verschiedene Personenparameterschätzer vorgestellt worden. Die Berechnung der Standardfehler ist für die einzelnen Schätzer ebenfalls unterschiedlich. Die entsprechenden Formeln wurden dabei bereits dargestellt (Gln. 19.85, 19.89, 19.93 und 19.97). In diesem Abschnitt soll daher nicht die konkrete Berechnung der einzelnen Standardfehler im Vordergrund stehen, sondern der Zusammenhang dieser Maße zu den Konzepten der Item- und Testinformation.

Auffallend ist bei genauer Betrachtung der Standardfehler der Personenparameterschätzer in der IRT, dass sie in Abhängigkeit der latenten Variablen variieren. Im Gegensatz zur KTT gibt es nicht nur einen Standardmessfehler, sondern jede Merkmalsausprägung kann einen anderen Standardmessfehler bzw. Standardfehler aufweisen. Somit ist es in der IRT kein Widerspruch, wenn ein Test bestimmte Merkmalsausprägungen sehr genau erfasst (niedriger Standardfehler) und zugleich die Messung in anderen Bereichen des gleichen Merkmals sehr ungenau ausfällt (großer Standardfehler). Die Größe der Standardfehler hängt dabei von der Testinformation ab. Je höher die Testinformation ausfällt, desto geringer ist der Standardfehler. Die Testinformation ist ihrerseits eine Funktion der latenten zu schätzenden Variablen, weswegen sie auch als *Testinformationsfunktion* $T(\theta)$ bezeichnet wird.

Variable Standardfehler

Der Begriff der Information wurde bereits bei der ML-Schätzung (► Abschn. 19.2) in Form der Informationsmatrix eingeführt. Es wurde auch gezeigt, wie diese Informationsmatrix zur Berechnung der Standardfehler von ML-Schätzern verwendet wird. Bei der Testinformation handelt es sich um das gleiche Konzept. Auch die Testinformationsfunktion ist die negative zweite Ableitung der logarithmierten ML-Funktion, hier der logarithmierten bedingten Antwortwahrscheinlichkeit in Abhängigkeit von der latenten Personenvariablen. Bei mehrdimensionalen latenten Personenvariablen θ ist die inverse Testinformation $T(\theta)^{-1}$ gleich der Varianz-Kovarianz-Matrix des ML-Schätzers von θ . In eindimensionalen IRT-Modellen entspricht die inverse Testinformation $T(\theta)^{-1}$ der Messfehlervarianz, deren Quadratwurzel gleich dem Standardfehler des ML-Schätzers gemäß Gl. (19.85) ist. Diese Erklärung ist jedoch recht technisch. Zum besseren Verständnis wird im Folgenden die Testinformationsfunktion ausgehend von der Varianz der Itemantworten in Abhängigkeit der latenten Personenvariablen schrittweise entwickelt.

19.6.2.2 Iteminformationsfunktion

Bedingte Varianzfunktion

Zunächst soll wiederum ein eindimensionales IRT-Modell für dichotome Items Y_i , mit $i = 1, \dots, k$, betrachtet werden. Die bedingte Varianzfunktion $Var(Y_i|\theta)$ eines Items i wurde bereits in ► Kap. 16 eingeführt. Sie gibt die Variabilität hinsichtlich der Antwortkategorien von Item i in Abhängigkeit der latenten Variablen an, bei dichotomen Items ist sie gleich dem Produkt der beiden bedingten Kategorienwahrscheinlichkeiten:

$$Var(Y_i|\theta) = P(Y_i = 1|\theta; \tau_i) \cdot P(Y_i = 0|\theta; \tau_i) \quad (19.99)$$

Daraus folgt, dass die bedingte Varianzfunktion maximal den Wert $Var(Y_i|\theta) = 0.25$ erreichen kann, und zwar dann, wenn $P(Y_i = 1|\theta; \tau_i) = P(Y_i = 0|\theta; \tau_i) = 0.5$. Beim Rasch- und beim Birnbaum-Modell (1PL- und 2PL-Modell) ist das der Fall, wenn der Wert der latenten Personenvariablen gleich der Itemschwierigkeit ist. Interessanterweise ist die bedingte Varianzfunktion $Var(Y_i|\theta)$ gleich der bedingten Varianz des Residuums $Var(\varepsilon_i|\theta)$, wobei $\varepsilon_i = Y_i - P(Y_i = 1|\theta; \tau_i)$. Inhaltlich bedeutet das: Je geringer die Differenz zwischen θ und der Itemschwierigkeit β_i und je höher somit die Varianz der beobachteten Itemantworten ist, desto größer ist die Unsicherheit bei der Vorhersage der manifesten Itemantwort und somit die Residualvarianz des Items. Trotzdem ist gerade dann die Information aus den beobachteten Itemantworten, d. h. den Werten der manifesten Variablen Y_i , hinsichtlich der zu schätzenden latenten Variable θ maximal! Das ist intuitiv einsichtig, wenn man das Beispiel eines Fähigkeitstests betrachtet. Sind sämtliche Items des Tests viel zu schwer, kann man zwar (fast sicher) vorhersagen, dass die Items nicht gelöst werden, aus den resultierenden Itemantworten resultiert aber auch nahezu keine Information über die Ausprägung der Fähigkeit der Person(en). Das gilt analog auch für Items, die viel zu leicht sind, sodass sie (fast sicher) alle gelöst werden. Erst aus der Varianz der Itemantworten einer Person hinsichtlich der verschiedenen Items eines Tests lässt sich eine verwertbare Information über die zu schätzende Fähigkeit gewinnen.

Formal ist die Iteminformationsfunktion $I_i(\theta)$ für ein Item i im Birnbaum-Modell gleich der mit der quadrierten Itemdiskrimination gewichteten bedingten Varianzfunktion:

$$\begin{aligned} I_i(\theta) &= \alpha_i^2 P(Y_i = 1|\theta; \tau_i) P(Y_i = 0|\theta; \tau_i) \\ &= \alpha_i^2 Var(Y_i|\theta) \end{aligned} \quad (19.100)$$

Ist die Itemdiskrimination wie beim Rasch-Modell gleich eins, sind die Iteminformationsfunktion und die bedingte Varianzfunktion identisch.

19.6.2.3 Testinformationsfunktion

Nun lassen sich die Werte der latenten Variablen anhand eines IRT-Modells unter Verwendung nur eines Items kaum sinnvoll schätzen. Vielmehr wird dazu ein Test bestehend aus $k > 1$ Items verwendet. Die Information aus den Antworten aller Items ist dann als Testinformationsfunktion formalisiert und lässt sich als Summe der Iteminformationsfunktionen berechnen, sodass gilt:

$$T(\theta) = \sum_{i=1}^k I_i(\theta) \quad (19.101)$$

Setzt man für das Birnbaum-Modell Gl. (19.100) für die Iteminformationsfunktion gemäß Gl. (19.101) ein und vergleicht man das Ergebnis mit Gl. (19.84), so bestätigt sich, dass die Testinformationsfunktion gleich der negativen zweiten Ableitung der logarithmierten Antwortmusterwahrscheinlichkeit in Abhängigkeit von der latenten Variablen θ ist. Hohe Werte der Testinformationsfunktion ergeben sich also dann, wenn die Werte der aufsummierten Iteminformationsfunktionen hoch sind, d. h. möglichst viele Itemschwierigkeiten nahe dem zu schätzenden Wert der latenten Variablen liegen, und die Itemdiskriminationen der Items hoch sind.

19.6.2.4 Beziehung von Standardfehler- und Testinformationsfunktion

Für den ML-Personenparameterschätzer (► Abschn. 19.5.1) besteht eine funktionale Beziehung zwischen dessen Standardfehler bzw. dessen Standardfehlerfunktion und der Testinformationsfunktion. Je höher die Information aus dem beobachteten Antwortverhalten ausfällt, desto höher ist die Testinformationsfunktion und desto geringer ist der Standardfehler des ML-Schätzers der latenten Personenvariablen. Für den gewichteten ML-Schätzer nach Warm gilt das in ganz ähnlicher Weise.

Bei den Bayes'schen Personenparameterschätzern ist der Zusammenhang zwischen Testinformation und Standardfehler jedoch etwas komplizierter. Beim MAP-Schätzer wird der Logarithmus mit der anhand der A-priori-Verteilung gewichteten Antwortmusterwahrscheinlichkeit maximiert. Als negative zweite Ableitung dieser Schätzfunktion erhält man die sog. „Präzision“, die in der Bayes-Statistik als Kehrwert der Varianz definiert ist. Sie ist beim MAP-Schätzer gleich der folgenden Summe:

$$T(\theta) + 1/\sigma_\theta^2 \quad (19.102)$$

Das entspricht dem Quadrat des Nenners von Gl. (19.93). Dabei ist $T(\theta)$ wiederum die Testinformationsfunktion und $1/\sigma_\theta^2$ der Kehrwert der Varianz der A-priori-Verteilung. Der Standardfehler des MAP-Schätzers (Gl. 19.93) hängt somit von zwei Faktoren ab. Wie beim ML-Schätzer gilt auch hier, dass der Standardfehler umso kleiner ist, je höher die Testinformationsfunktion ausfällt. Zugleich gilt, dass der Standardfehler abnimmt, je kleiner die Varianz der A-priori-Verteilung gewählt wird (d. h. je höher das Vorwissen über den zu schätzenden Personenparameter gewichtet wird).

Für EAP-Personenparameterschätzer wird die individuelle A-posteriori-Standardabweichung $PSD(\hat{\theta}_v)$ (Gl. 19.97) als Maß der Ungenauigkeit der Schätzung verwendet. Leider ist der Zusammenhang zwischen der Testinformationsfunktion und der Schätzgleichung der A-posteriori-Standardabweichung gemäß Gl. (19.97) nicht leicht zu illustrieren. Aber auch für die A-posteriori-Standardabweichung gilt, dass sie umso kleiner ist, je höher die Testinformationsfunktion und je kleiner die Varianz der A-priori-Verteilung ausfallen.

Präzision

**Individuelle
A-posteriori-Standardabweichung
als Maß der Ungenauigkeit der
Personenparameterschätzung**

19.6.3 Marginale Reliabilitäten

Messgenauigkeit der Personenparameterschätzungen ist abhängig von der latenten Personenvariablen

In der KTT ist es üblich, für einen Test einen Kennwert für die Reliabilität anzugeben. Unter der Verwendung psychometrischer Messmodelle der IRT erscheint es zunächst schwierig, einen solchen Kennwert zu berechnen, da die Standardfehler der Personenparameterschätzungen in Abhängigkeit der latenten Personenvariablen variieren. Es gibt also nicht nur eine Reliabilität, sondern der Test ist bezüglich der Schätzung verschiedener Ausprägungen der latenten Variablen unterschiedlich reliabel bzw. genau. Um dennoch eine allgemeine Aussage über die Reliabilität eines Tests in Form eines Kennwertes zu erlauben, wurden in der IRT sog. „marginale Reliabilitätsmaße“ entwickelt. Sie werden interpretiert als die durchschnittliche Reliabilität, gemittelt über die Verteilung der latenten Variablen. Auch die marginalen Reliabilitäten werden für verschiedene Personenparameterschätzer in unterschiedlicher Weise berechnet.

19.6.3.1 Marginale Reliabilität bei ML-Personenparameterschätzern

Andrich's reliability

Die marginale Reliabilität für den ML-Personenparameterschätzer wird auch als „separation reliability“ oder „Andrich's reliability“ bezeichnet, da sie auf Arbeiten von David Andrich (1988) zurückgehen. Ausgangspunkt bildet die bereits eingeführte Zerlegung eines Schätzers $\hat{\theta}$ in den wahren Wert θ und den Messfehler $\varepsilon_{\hat{\theta}}$ (Gl. 19.98). Unter der Annahme, dass die zu schätzende latente Variable und der Messfehler unkorreliert sind, ist auch die Varianz des Personenparameterschätzers gleich der folgenden Summe:

$$\text{Var}(\hat{\theta}) = \text{Var}(\theta) + \text{Var}(\varepsilon_{\hat{\theta}}) \quad (19.103)$$

In der KTT ist die Reliabilität definiert als das Verhältnis der Varianzen der wahren Werte (True-Score-Variablen) und Varianz der beobachteten Testwerte. Angewendet auf die Personenparameterschätzung in der IRT entspricht das dem Verhältnis:

$$\text{Rel}(\hat{\theta}) = \frac{\text{Var}(\theta)}{\text{Var}(\hat{\theta})} \quad (19.104)$$

Die Reliabilität ist daher gleich dem Anteil der Varianz der Personenparameterschätzungen, die auf die Varianz der latenten Variablen, d. h. auf tatsächliche interindividuelle Unterschiede der Ausprägungen von θ , zurückzuführen ist. Dieser Anteil lässt sich alternativ anhand der Fehlervarianz berechnen:

$$\text{Rel}(\hat{\theta}) = 1 - \frac{\text{Var}(\varepsilon_{\hat{\theta}})}{\text{Var}(\hat{\theta})} \quad (19.105)$$

Marginale Fehlervarianz

Da in der IRT die Standardfehler in Abhängigkeit der latenten Variablen variieren, ist entsprechend die Fehlervarianz $\text{Var}(\varepsilon_{\hat{\theta}})$ eine marginale, d. h. durchschnittliche Fehlervarianz, gemittelt über die Verteilung der latenten Variablen:

$$\text{Var}(\varepsilon_{\hat{\theta}}) = \int_{\theta} \text{Var}(\varepsilon_{\hat{\theta}}|\theta) f(\theta) d\theta \quad (19.106)$$

Bedingte Fehlervarianzfunktion

Dabei ist die bedingte Fehlervarianzfunktion $\text{Var}(\varepsilon_{\hat{\theta}}|\theta)$ gleich der quadrierten Standardfehlerfunktion. Glücklicherweise kann man bei der Anwendung auf die direkte Berechnung des Integrals in Gl. (19.106) verzichten. Die marginale Reliabilität nach Gl. (19.105) kann hinreichend gut approximiert werden, indem

19.6 · Reliabilitätsbeurteilung in der IRT

die marginale Fehlervarianz durch den Mittelwert der quadrierten Standardfehler $SE(\hat{\theta}_v)$ geschätzt und die Stichprobenvarianz der ML-Personenparameterschätzer verwendet wird, sodass sich Gl. (19.107) ergibt:

$$Rel(\hat{\theta}) \approx 1 - \frac{\overline{SE}(\hat{\theta}_v)^2}{s_{\hat{\theta}_v}^2} = 1 - \frac{\frac{1}{n} \sum_{v=1}^n SE(\hat{\theta}_v)^2}{\frac{1}{n-1} \sum_{v=1}^n (\hat{\theta}_v - \bar{\hat{\theta}}_v)^2} \quad (19.107)$$

Diese Form der Berechnung der marginalen Reliabilität lässt sich für den ML-Schätzer und den gewichteten ML-Schätzer anwenden, nicht jedoch für MAP- und EAP-Schätzer! Für diese Personenparameterschätzer sind alternative Berechnungen notwendig.

19.6.3.2 Marginale Reliabilität bei EAP- und MAP-Personenparameterschätzern

Zur Bestimmung der marginalen Reliabilität unter Verwendung von EAP- oder MAP-Personenparameterschätzern sind die Gl. (19.105) bzw. Gl. (19.106) nicht geeignet, da der EAP- und der MAP-Schätzer als Bayes'sche Schätzer dem Shrinkage-Effekt unterliegen (► Beispiel 19.1). Das heißt, unterdurchschnittliche Ausprägungen von θ werden überschätzt, überdurchschnittliche Ausprägungen hingegen unterschätzt. Daraus folgt eine negative Kovarianz $Cov(\theta, \varepsilon_{\hat{\theta}}) < 0$ zwischen den wahren Werten und dem Messfehler, sodass die Varianzzerlegung nach Gl. (19.103) nicht mehr gültig ist. Als Folge ist die Varianz $Var(\hat{\theta})$ der EAP-/MAP-Schätzer immer kleiner oder maximal gleich der $Var(\theta)$. Reliabilitätsberechnungen nach Gl. (19.105) ergäben daher nicht definierte Werte ≥ 1 . Allerdings ist der Shrinkage-Effekt selbst ein Ausdruck der Unreliabilität der Personenparameterschätzung. Bei einer perfekten Reliabilität von 1 gibt es keinen Shrinkage-Effekt und die Varianz der EAP-/MAP-Schätzer ist gleich der wahren Varianz der latenten Variablen. Je geringer jedoch die Reliabilität ist, desto stärker ist der Shrinkage-Effekt und desto mehr unterschätzt die Varianz der EAP-/MAP-Schätzer die wahre Varianz. Eingangs wurde bereits dargelegt, dass die A-posteriori-Verteilung eines Parameters das Produkt aus Likelihood und A-priori-Verteilung des Parameters ist. Entsprechend ist der Erwartungswert der A-posteriori-Verteilung, der EAP, gleich dem gewichteten Mittel aus dem ML-Personenparameterschätzer $\hat{\theta}_v^{(ML)}$ und dem Erwartungswert μ_{θ} der A-priori-Verteilung der latenten Variablen:

$$E(\theta_v | Y_v = y_v) = \rho \hat{\theta}_v^{(ML)} + (1 - \rho) \mu_{\theta} \quad (19.108)$$

Dabei sind die Reliabilität ρ und die Unreliabilität $(1 - \rho)$ die Gewichte. Verwendet man die rechte Seite von Gl. (19.108) für die Varianz des EAP-Schätzers, folgt:

$$\begin{aligned} Var[E(\theta_v | Y_v = y_v)] &= Var[\rho \hat{\theta}_v^{(ML)} + (1 - \rho) \mu_{\theta}] \\ &= Var[\rho \hat{\theta}_v^{(ML)}] \\ &= \rho^2 Var[\hat{\theta}_v^{(ML)}] \end{aligned} \quad (19.109)$$

Nun kann man die rechte Seite der Definitionsgleichung der Reliabilität (Gl. 19.104) für ρ einsetzen und erhält schließlich:

$$Var[E(\theta_v | Y_v = y_v)] = \left(\frac{Var(\theta)}{Var[\hat{\theta}_v^{(ML)}]} \right)^2 Var[\hat{\theta}_v^{(ML)}] = \rho Var(\theta) \quad (19.110)$$

Shrinkage-Effekt und EAP-Schätzer als gewichtetes Mittel

Berechnung der EAP-Reliabilität

Das heißt, die Reliabilität ist der Faktor, um den sich die Varianz der EAP-Schätzer aufgrund der Schätzgenauigkeit vermindert. Somit kann die Reliabilität als Varianzverhältnis zwischen EAP-Schätzer und wahrer Varianz geschätzt werden:

$$\rho = \frac{\text{Var}[E(\theta_v|Y_v = y_v)]}{\text{Var}(\theta)} = \text{Rel}(\hat{\theta}) \quad (19.111)$$

Bei der Anwendung kann für den Zähler einfach die Stichprobenvarianz $s_{\hat{\theta}}^2$ der zuvor bestimmten EAP-Schätzer verwendet werden. Für die wahre Varianz $\text{Var}(\theta)$ im Nenner von Gl. (19.111) wird die vom Modell geschätzte Varianz $s_{\hat{\theta}}^2$ eingesetzt. Wurde das Modell jedoch durch die Fixierung der Varianz der latenten Variablen auf einen bestimmten Wert identifiziert, so wird dieser Wert im Nenner von Gl. (19.111) eingesetzt. Wurde das Messmodell z. B. durch Fixierung der Skala der latenten Variablen ($\text{Var}(\theta) = 1$) identifiziert, ist die geschätzte EAP-Reliabilität gleich der Varianz $s_{\hat{\theta}}^2$ der EAP-Schätzer. Auch die Varianz von PVs kann als unverfälschter Schätzer der wahren Varianz $\text{Var}(\theta)$ im Nenner von Gl. (19.111) verwendet werden. Der resultierende Kennwert wird dann oft auch als EAP/PV-Reliabilität bezeichnet.

Berechnung der marginalen MAP-Reliabilität

Zur Herleitung der marginalen Reliabilität wurden zwar die EAP-Schätzer verwendet, aber Gl. (19.111) lässt sich auch auf die MAP-Schätzer anwenden. Dazu wird im Zähler die Stichprobenvarianz $s_{\hat{\theta}}^2$ der MAP-Schätzer eingesetzt. Theoretisch lässt sich dieses Vorgehen dadurch begründen, dass die A-posteriori-Verteilungen asymptotisch normal verteilt sind und sowohl die EAP- als auch die MAP-Schätzer damit identisch sind. Insbesondere bei kleinen Itemzahlen ist die Normalverteilungsannahme häufig verletzt. Erfahrungsgemäß ergeben die Schätzungen der marginalen Reliabilität anhand der MAP-Schätzer unter Verwendung von Gl. (19.111) aber dennoch sehr ähnliche Werte im Vergleich zur Verwendung der EAP-Schätzer.

19.7 Zusammenfassung

In der IRT existieren verschiedene Verfahren der Item- und Personenparameterschätzung, wobei sich grundsätzlich ML- und Bayes'sche Schätzverfahren unterscheiden lassen. Innerhalb beider Verfahrensklassen gibt es wiederum verschiedene Schätzalgorithmen mit unterschiedlichen Eigenschaften. Die wichtigsten wurden in diesem Kapitel am Beispiel ein- und zweiparametrischer IRT-Modelle dargestellt. Von den ML-Verfahren wurden die JML-Schätzung, die CML-Schätzung und die MML-Schätzung erläutert. Bevor detailliert auf verschiedene Bayes'sche Schätzverfahren eingegangen wurde, sind zunächst einführend die Grundlagen der statistischen Inferenz in der Bayes-Statistik dargestellt worden, wobei näher auf die zentrale Bedeutung der A-priori- und der A-posteriori-Verteilung bei der Parameterschätzung eingegangen wurde. Nachfolgend wurden nicht simulationsbasierte und simulationsbasierte Bayes-Schätzer erläutert. Der MH-Algorithmus und der Gibbs-Sampler wurden als Beispiele der zunehmend populären simulationsbasierten MCMC-Verfahren dargestellt. Item- und Personenparameter werden oft (aus gutem Grund) nicht simultan, sondern separat geschätzt. Daher wurde der Personenparameterschätzung in der IRT ein eigener Abschnitt gewidmet, in dem der ML-Schätzer, der gewichtete ML-Schätzer, der EAP-Schätzer, der MAP-Schätzer und die PVs als gebräuchliche Schätzer für die latente Personenvariable erläutert wurden. Da die Reliabilität in der IRT in Abhängigkeit der zu schätzenden Personenparameter variiert, gibt es streng genommen nicht nur einen Wert der Reliabilität für einen Test. Ausgehend von der Item-, der Testinformations- und Standardfehlerfunktion wurde gezeigt, wie die Genauigkeit der Personenparameterschätzung für konkrete Werte von der latenten Variable abhängt. Abschließend wurden marginale, d. h. durchschnittliche Reliabilitätskoeffizienten als ein-

fach zu kommunizierende, aber populationsabhängige Gütemaße der Messgenauigkeit eines Tests vorgestellt und ihre exakte Interpretation und Berechnung erläutert.

19.8 EDV-Hinweise

Zur Anwendung von IRT-Modellen stehen verschiedene Computerprogramme zur Verfügung, die sich bezüglich der schätzbaren Modelle und der implementierten Schätzverfahren unterscheiden. Zur Parameterschätzung von dichotomen und polytomen Rasch-Modellen mit dem JML-Verfahren stehen beispielsweise die Programme BIGSTEPS (Linacre und Wright 1993) und WINSTEPS (Linacre 2009) zur Verfügung. Auch das Programm ConQuest 4 (Adams et al. 2015), das insbesondere für Large-Scale-Assessments in der empirischen Bildungsforschung angewendet wird, erlaubt neben der MML-Schätzung auch die JML-Schätzung von ein- und mehrdimensionalen dichotomen und polytomen Rasch-Modellen. Mehrdimensionale Rasch-Modelle für dichotome und ordinale Items können auch mit dem Programm MULTIRIA (Carstensen und Rost 2000) berechnet werden, das ebenfalls die JML-Schätzung verwendet. Die CML-Schätzung für 1PL-Modelle der Rasch-Familie ist im Programm WINMIRA (von Davier 2001b) implementiert. Das Programm Bilog-MG 3 (Zimowski et al. 1996) erlaubt die MML-Schätzung und die marginale Bayes-Modal-Schätzung von eindimensionalen 1PL-, 2PL- und 3PL-Modellen für dichotome Items. Beide Schätzverfahren sind auch in der Software PARSCALE 4 (Muraki und Bock 2003) verfügbar, die im Gegensatz zu BILOG-MG 3 auch bei polytomen Items angewendet werden kann. Eine Vielzahl von ein- und mehrdimensionalen IRT-Modellen kann mit dem Programm IRTPRO 4 (Cai et al. 2011) geschätzt werden, in dem neben der MML-Schätzung nach Bock und Aitkin (1981) auch der Metropolis-Hastings-Robbins-Monro-Algorithmus (Cai 2010) implementiert ist. Eine umfassende Software für latente Variablen Modelle ist *Mplus*, mit dem ab der 8. Version (Muthén und Muthén 2017) eine breite Palette von 1PL- bis 4PL-Modellen für zwei- und mehrkategoriale Items geschätzt werden kann. Dazu kann entweder auf die MML-Schätzung oder den Gibbs-Sampler als Bayes'schen Schätzer zurückgegriffen werden. Auch Strukturgleichungsmodelle für dichotome und ordinale Variablen, basierend auf tetra- und polychorischen Korrelationen, können mit *Mplus* berechnet werden.

In der freien Software R (R Core Team 2018) sind eine ganze Reihe von Paketen zur Parameterschätzung von IRT-Modellen entwickelt worden. Hier sollen nur ein paar Beispiele genannt werden: Im Paket „eRm“ (Mair und Hatzinger 2007) ist die CML-Schätzung für eindimensionale Rasch-Modelle implementiert, wobei dichotome und polytome Items verwendet werden können. 1PL-, 2PL- und 3PL-Modelle für dichotome und polytome Items können unter Verwendung der MML-Schätzung mit dem Paket „ltm“ (Rizopoulos 2006) berechnet werden. Multidimensionale IRT-Modelle können mit dem MIRT-Paket (Chalmers 2012) oder dem TAM-Paket (Robitzsch et al. 2017) geschätzt werden, die ebenfalls beide die MML-Schätzung verwenden. Strukturgleichungsmodelle für dichotome und ordinale Variablen können anhand von gewichteten Kleinste-Quadrat-Schätzern auf Basis tera- und polychorischer Korrelationen mit dem Paket „lavaan“ (Rosseel 2012) geschätzt werden. Auch die Pairwise-ML-Schätzung von Katsikatsou et al. (2012) ist im lavaan-Paket verfügbar. MCMC-Verfahren zur Bayes'schen Schätzung von Parametern für eine Reihe IRT-Modellen stehen im Paket „MCMCpack“ (Martin et al. 2011) zur Verfügung. Darüber hinaus gibt es eine ganze Reihe von R-Paketen zur IRT mit ganz unterschiedlicher Funktionalität. Einen guten Überblick über relevante Pakete im Bereich der Psychometrie und speziell der IRT ist auf der Internetseite ► <https://cran.r-project.org/web/views/Psychometrics.html> zu finden.

In der Software STATA (StataCorp LLC 2017b) ist mit dem irt-Paket (StataCorp 2017a) und dem gllamm-Paket (Rabe-Hesketh et al. 2004) ebenfalls eine große Zahl von ein- und mehrdimensionalen IRT-Modellen mit unterschiedlichen Schätzverfahren verfügbar. Einen Überblick über die schätzbaren IRT-Modelle des PROC-IRT-Moduls der Software SAS und die verwendeten Schätzalgorithmen bietet der Artikel von Choi (2016). Die Liste der Programme, die auf die hier dargestellten Schätzverfahren zurückgreifen, ist keinesfalls vollständig. Die Auswahl enthält weit verbreitete Programme, für die auf gute Erfahrungswerte zurückgegriffen werden kann und zu denen umfangreiche Dokumentationen und reichlich Hintergrundliteratur zur Verfügung stehen.

Die in diesem Kapitel erläuterten mathematischen Grundlagen können zudem als Ausgangspunkt für das selbstständige Programmieren von Schätzalgorithmen dienen. MATLAB und R eignen sich für den Einstieg hervorragend, da sie bereits eine umfangreiche Funktionalität für die numerische Berechnung und Optimierung mitbringen. Außerdem gibt es für beide Sprachen hinreichend Literatur und Ressourcen im Internet.

19.9 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion). Unter dem angegebenen Link in den Zusatzmaterialien sind zudem zwei weitere Rechenbeispiele zur Parameterschätzung und Messgenauigkeit in der IRT zu finden.

1. Begründen Sie unter Betrachtung des allgemeinen Maximum-Likelihood-Prinzips (ML-Prinzips) und der Definition der A-posteriori-Verteilung, warum ML-Schätzer und Bayes'sche Schätzer nicht identisch sind.
2. Warum sind die Parameterschätzer unter Verwendung der Joint-Maximum-Likelihood-Schätzung (JML-Schätzung) inkonsistent, und wie wird dieses Problem bei der bedingten (CML-) und der marginalen ML-Schätzung (MML-Schätzung) gelöst?
3. Welche Kennwerte werden bei der ML-Schätzung und bei Bayes'schen Schätzverfahren als Maße der (Un-)Genauigkeit der Parameterschätzung verwendet, und wie sind diese Maße jeweils genau zu interpretieren?
4. Wann bezeichnet man eine A-priori-Verteilung als informativ, und wie wirkt sich eine informative A-priori-Verteilung im Vergleich zu einer nicht informativen A-priori-Verteilung auf die Parameterschätzung aus?
5. Es sei ein eindimensionaler Mathematikkompetenztest mit ausschließlich schweren Items für zwei unabhängige Stichproben A und B eingesetzt worden. Während Stichprobe A aus Personen mit durchschnittlicher Mathematikfähigkeit besteht, ist Stichprobe B aus Personen mit überdurchschnittlicher Mathematikkompetenz zusammengesetzt. Die Daten beider Stichproben werden mit dem Birnbaum-Modell ausgewertet. Die Itemparameterschätzer aus beiden Stichproben unterscheiden sich bis auf kleine zufällige Schwankungen nicht. Welche Befunde erwarten Sie für die Testinformationsfunktion, die Standardfehlerfunktion und die marginale Reliabilität in den beiden Gruppen?
6. Vergleichen Sie die Maße der Messgenauigkeit bzw. -unge nauigkeit der Personenparameterschätzung, die in der KTT und der IRT verwendet werden, und erläutern Sie Gemeinsamkeiten und Unterschiede.
7. Vergleichen Sie die fünf hier vorgestellten Personenparameterschätzer (ML-, gewichteter ML-, EAP-, MAP-Schätzer und PVs) hinsichtlich ihrer Eignung für die Individualdiagnostik.

Literatur

- Adams, R. J., Wu, M. L. & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modeling software (Version 4)*. Camberwell, Victoria: Australian Council for Educational Research.
- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 42–54.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47, 105–113.
- Andrich, D. (1988). *Rasch models for measurement* (Vol. 68). Newbury Park, CA: Sage.
- Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. Boca Raton, FL: CRC.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179–197.
- Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In van der Linden, W. J. & Hambleton, R. K. H. (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.
- Cai, L. (2010). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics*, 35, 307–335.
- Cai, L., Thissen, D. & du Toit, S. H. C. (2011). *IRTPRO 4 for Windows* (Computer software). Lincolnwood, IL: Scientific Software International.
- Carstensen, C. H. & Rost, J. (2000). *MULTIRA ein Programmsystem zur Analyse mehrdimensionaler Rasch-Modelle. Handbuch zum Computerprogramm MULTIRA*. Kiel: IPN Kiel – Institut für die Pädagogik der Naturwissenschaften. Verfügbar unter https://docplayer.org/61132930-Claus-h-carstensen-und-juergen-rost-ipn-kiel-multira-ein-programmsystem-zur-analyse-mehrdimensionaler-rasch-modelle.html#show_full_text [29.12.2019]
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, 1–29.
- Choi, J. (2016). A Review of PROC IRT in SAS. *Journal of Educational and Behavioral Statistics*, 42, 195–205.
- Cowles, M. K. & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883–904.
- De Boeck, P. (2008). Random Item IRT Models. *Psychometrika*, 73, 533–559.
- Demidenko, E. (2013). *Mixed models: Theory and applications with R*. Hoboken, NJ: John Wiley & Sons.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Forero, C. G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14, 275–299.
- Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.
- Katsikatsou, M., Moustaki, I., Yang-Wallentin, F. & Joreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, 56, 4243–4258.
- Levy, R. & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Linacre, J. M. (2009). *Winsteps (Version 3.68. 0)* (Computer software). Chicago, IL: winsteps.com.
- Linacre, J. M. & Wright, B. D. (1993). *A user's guide to BIGSTEPS: Rasch-model computer program*. Chicago, IL: MESA Press.
- Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.
- Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20, 1–20. <https://doi.org/10.18637/jss.v020.i09>
- Martin, A. D., Quinn, K. M. & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, 42, 1–21. <https://doi.org/10.18637/jss.v042.i09>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Masters, G. N. & Wright, B. D. (1997). The partial credit model. In W. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NJ: Springer.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359–381.
- Mislevy, R. J. (1986). Bayes Modal Estimation in Item Response Models. *Psychometrika*, 51, 177–195.

- Muraki, E. & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales*. Lincolnwood, IL: Scientific Software International.
- Muthén, B. O. & Muthén, L. K. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 16, 1–32.
- R Core Team. (2018). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.
- Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). *GLLAMM Manual*. Retrieved from <https://biostats.bepress.com/ucbbiostat/paper160> [29.12.2019]
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17, 1–25.
- Robitzsch, A., Kiefer, T. & Wu, M. (2017). TAM: Test analysis modules. R package version 2.2-49. Retrieved from <https://CRAN.R-project.org/package=TAM> [29.12.2019]
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, 48, 1–36.
- Rost, J. & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, 26, 42–56.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- StataCorp LLC (2017a). *Stata item response theory reference manual (Release 15)*. College Station, TX: Stata Press.
- StataCorp LLC (2017b). *Stata statistical software*. College Station, TX: StataCorp LLC.
- Tschirk, W. (2014). *Statistik: Klassisch oder Bayes*. Berlin, Heidelberg: Springer.
- Van den Noortgate, W., De Boeck, P. & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28, 369–386.
- von Davier, M. (2001a). WINMIRA 2001 (Software). St. Paul, MN: Assessment Systems Corp. Retrieved from <http://208.76.80.46/~svfklumu/wmira/index.html> [29.12.2019]
- von Davier, M. (2001b). *WINMIRA 2001 user's guide*. Kiel: IPN.
- von Davier, M., Gonzalez, E. & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 9–36). Hamburg, Princeton: IEA-ETS Research Institute.
- von Davier, M. & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Volume 26: Psychometrics* (pp. 643–661). Amsterdam: Elsevier.
- Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.
- Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11, 253–270.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.
- Wright, S. & Nocedal, J. (1999). *Numerical optimization*. New York, NJ: Springer.
- Yuan, K.-H., Cheng, Y. & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, 79, 232–254.
- Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). BILOG-MG 3: Item analysis and test scoring with binary logistic models (Computer software). Chicago, IL: Scientific Software International.



Computerisiertes adaptives Testen

Andreas Frey

Inhaltsverzeichnis

- 20.1 Was ist computerisiertes adaptives Testen? – 502**
- 20.2 Grundgedanke – 503**
- 20.3 Elementare Bausteine – 506**
 - 20.3.1 Itempool – 507
 - 20.3.2 Testbeginn – 509
 - 20.3.3 Personenparameterschätzung – 509
 - 20.3.4 Itemauswahl – 510
 - 20.3.4.1 Voll adaptive Itemauswahl – 510
 - 20.3.4.2 Eingeschränkt adaptive Itemauswahl – 513
 - 20.3.5 Berücksichtigung von Einschränkungen – 514
 - 20.3.6 Testende – 515
- 20.4 Auswirkungen des adaptiven Testens – 516**
 - 20.4.1 Messeffizienz – 517
 - 20.4.2 Validität – 517
 - 20.4.3 Motivation zur Testbearbeitung – 519
- 20.5 Multidimensionales adaptives Testen – 520**
- 20.6 Zusammenfassung und Anwendungsempfehlungen – 521**
- 20.7 EDV-Hinweise – 522**
- 20.8 Kontrollfragen – 522**
- Literatur – 523**

i Computerisiertes adaptives Testen ist ein spezielles Vorgehen zur computerbasierten Messung individueller Merkmalsausprägungen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am vorherigen Antwortverhalten der Testperson orientiert. Der Grundgedanke besteht darin, nur solche Items eines Tests vorzugeben, die möglichst viel diagnostische Information über die individuelle Merkmalsausprägung liefern. Dieses Anliegen wird durch die Spezifikation der sechs elementaren Bausteine Itempool, Testbeginn, Personenparameterschätzung, Itemauswahl, Berücksichtigung von Einschränkungen und Testende umgesetzt. Der Hauptvorteil des computerisierten adaptiven Testens besteht in einer Messeffizienzsteigerung. Darüber hinaus sind positive Auswirkungen auf die Validität der adaptiv erhobenen Testergebnisse zu verzeichnen. Um unerwünschte Effekte beim computerisierten adaptiven Testen zu vermeiden, sollte die Funktionsweise eines adaptiven Tests im Rahmen der Instruktion transparent erläutert werden.

20.1 Was ist computerisiertes adaptives Testen?

Bei den meisten Fragebogen- und Testverfahren wird allen getesteten Personen eine festgelegte Menge von Items in einer festen Reihenfolge vorgegeben. Beim adaptiven Testen werden abweichend davon einer Testperson nur solche Items zur Bearbeitung vorgelegt, von denen aufgrund des bislang bei der Testung gezeigten Antwortverhaltens davon auszugehen ist, dass sie besonders viel diagnostische Information über die individuelle Ausprägung des Individuums im Bezug auf das gemessene Merkmal liefern.

Definition

Computerisiertes adaptives Testen ist ein spezielles Vorgehen zur computerbasierten Messung individueller Merkmalsausprägungen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am vorherigen Antwortverhalten der Testperson orientiert.

Beispiel einer Prüfungssituation

Das Vorgehen beim computerisierten adaptiven Testen kann man sich gut am Beispiel eines typischen Verhaltens von Prüfenden bei mündlichen Prüfungen verdeutlichen. Prüfende passen den Schwierigkeitsgrad ihrer Fragen nämlich typischerweise dem Leistungsvermögen des Prüflings an. Kann ein Prüfling beispielsweise mehrere als mittelschwer eingeschätzte Fragen korrekt beantworten, werden viele Prüfende im weiteren Verlauf schwierigere Fragen stellen, um herauszufinden, wie fundiert die Kenntnisse des Prüflings sind. Für einen anderen Prüfling stellen mittelschwere Fragen unter Umständen bereits eine Überforderung dar. Es ist wahrscheinlich, dass er auch weitere Fragen mittlerer Schwierigkeit nicht korrekt beantworten können wird. In diesem Fall werden viele Prüfende im weiteren Verlauf leichtere Fragen stellen, um zu erfahren, ob die Kenntnisse zum Bestehen der Prüfung ausreichen oder nicht. Die abschließende Beurteilung der Prüfungsleistung erfolgt durch eine integrierende Betrachtung der gegebenen Antworten unter Berücksichtigung, welche Fragen gestellt wurden. Durch die geschilderte an das Antwortverhalten angepasste Auswahl von Fragen wird erreicht, dass über einen breiten Leistungsbereich differenzierende Aussagen getroffen werden können. Würden hingegen nur schwere Fragen gestellt, dann könnte nicht gut im niedrigen Leistungsbereich differenziert werden; würden nur leichte Fragen gestellt, wäre eine Differenzierung leistungsfähiger Prüflinge kaum möglich. Eine Differenzierung über einen breiten Leistungsbereich wäre natürlich auch bei zufälliger Auswahl der Fragen möglich, allerdings müssten dann deutlich mehr Fragen gestellt werden, um sicherzustellen, dass das Schwierigkeitskontinuum angemessen abgedeckt ist. Das Beispiel beschreibt ein Vorgehen, bei dem solche Fragen ausgewählt werden, die in Abhängigkeit des vorherigen Antwortverhaltens als angemessen erscheinen.

20.2 · Grundgedanke

Das Vorgehen des Prüfenden wird vermutlich meistens implizit und ohne explizite Regel ablaufen.

Die Itemauswahl beim computerisierten adaptiven Testen verläuft ganz ähnlich wie das im Beispiel geschilderte Vorgehen: Auch hier orientiert sich die Auswahl der vorzugebenden Items am zuvor gezeigten Antwortverhalten. Computerisiertes adaptives Testen ist allerdings bezüglich zweier Punkte exakter. Erstens werden ausschließlich Items eingesetzt, die die Annahmen eines psychometrischen Modells nicht verletzen. Hierdurch wird u. a. ermöglicht, dass aufgrund des beobachteten Antwortverhaltens jeweils auf das gleiche Merkmal geschlossen werden kann, auch wenn von unterschiedlichen Testpersonen unterschiedliche Items bearbeitet wurden. Zweitens orientiert sich das Vorgehen bei der Testung an einem vorab festgelegten adaptiven Algorithmus. Ein adaptiver Algorithmus ist ein Regelsystem, das zu Beginn und während des Tests die Itemauswahl trifft und den Umgang mit nicht statistischen Einschränkungen, die Schätzung der individuellen Merkmalsausprägung sowie Kriterien der Testbeendigung spezifiziert.

Die Geschichte des computerisierten adaptiven Testens – wie wir es heute verstehen – beginnt in den 1970er-Jahren. Obschon bereits vorher einzelne Versuche unternommen wurden, antwortabhängige Testverfahren zu entwickeln (vgl. Chang 2015; Weiss 2004), wurden erst in dieser Zeit tragfähige Konzepte zum computerisierten adaptiven Testen formuliert und erprobt. In den wegweisenden Büchern von Lord (1980) und Weiss (1983) wurden die wesentlichen Aspekte des adaptiven Testens erstmals zusammenhängend dargestellt. Seit Beginn der 1980er-Jahre wurden dann zahlreiche grundlegende Fragen, die für eine operationale (= praktische) Anwendung des adaptiven Testens von Relevanz sind, im Rahmen der Entwicklung der „Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery“ (CAT-ASVAB; Sands et al. 1997) untersucht. Diese Testbatterie wird beim US-amerikanischen Militär bis heute zur Personalauswahl eingesetzt. Die CAT-ASVAB ist eines der am gründlichsten untersuchten und mit rund 400.000 Testpersonen pro Jahr eines der am meisten verwendeten psychodiagnostischen Testinstrumente überhaupt. Heutzutage finden sich CAT-Anwendungen darüber hinaus in ganz verschiedenen Bereichen wie der Kompetenzdiagnostik, der Erfassung gesundheitlicher Aspekte, der Intelligenzdiagnostik, der Messung beruflicher Eignung oder dem universitären Zulassungswesen.

In den folgenden Abschnitten werden die zentralen Aspekte des computerisierten adaptiven Testens dargestellt. In ► Abschn. 20.2 wird zunächst der Grundgedanke dieser speziellen Form des Testens erläutert. Zur Umsetzung des Grundgedankens werden sechs elementare Bausteine benötigt. Diese Bausteine werden in ► Abschn. 20.3 einzeln beschrieben. Im nachfolgenden ► Abschn. 20.4 werden die Auswirkungen der Verwendung computerisierter adaptiver Tests auf psychometrische Kenngrößen und auf Reaktionen seitens der Testpersonen zusammengefasst, um nachfolgend in ► Abschn. 20.5 das multidimensionale adaptive Testen vorzustellen. Das Kapitel wird durch eine Zusammenfassung mit Anwendungsempfehlungen in ► Abschn. 20.6 sowie Hinweisen auf CAT-Software in ► Abschn. 20.7 abgeschlossen. Wenn nichts gesondert angegeben wird, beziehen sich alle Ausführungen der Einfachheit halber immer auf Leistungstests mit dichotomem Antwortmodus, wobei anzumerken ist, dass das computerisierte adaptive Testen grundsätzlich auch für die Messung nicht leistungsbezogener Merkmale wie Persönlichkeitsmerkmale oder Einstellungen sowie bei polytomem Antwortmodi verwendet werden kann.

20.2 Grundgedanke

Beim computerisierten adaptiven Testen orientiert sich die Auswahl der vorzugebenden Items am vorher gezeigten Antwortverhalten der Testperson. Dieses Vorgehen verfolgt das Ziel, nur solche Items vorzugeben, die für das getestete Indi-

Itemauswahl

Adaptiver Algorithmus

Geschichte

Übersicht

Optimierte Itemauswahl

viduum „optimal“ sind. Um das jeweils optimale Item zu identifizieren, wird für alle noch verfügbaren Items das Ausmaß an diagnostischer Information über die interessierende individuelle Merkmalsausprägung bestimmt, das aufgrund seiner Beantwortung zu erwarten ist. Das Item mit dem höchsten Wert wird dann für die Vorgabe ausgewählt. Hat beispielsweise eine Testperson im bisherigen Testverlauf von fünf vorgegebenen mittelschweren Aufgaben alle korrekt beantwortet, dann ist eine Antwort auf eine sehr leichte Aufgabe (für die man begründet annehmen kann, dass sie mit hoher Wahrscheinlichkeit ebenfalls korrekt beantwortet werden wird) deutlich weniger informativ als eine Antwort auf eine schwierige Aufgabe (für die man schwerer einschätzen kann, ob sie korrekt beantwortet werden kann oder nicht). Die Itemauswahl beim computerisierten adaptiven Testen besteht üblicherweise darin, dass Testpersonen mit hoher Merkmalsausprägung schwierigere Items vorgelegt bekommen als Testpersonen mit niedrigerer Merkmalsausprägung. Neben der Optimierung der diagnostischen Information wird hierdurch zudem vermieden, dass leistungsfähige Testpersonen wiederholt Items bearbeiten müssen, die für sie deutlich zu leicht sind und problemlos gelöst werden können bzw. dass Testpersonen mit geringer Leistungsfähigkeit wiederholt deutlich zu schwere Items bearbeiten müssen, die nicht gelöst werden können.

Die Tatsache, dass bei einer adaptiven Anpassung der Itemauswahl verschiedene Testpersonen in der Regel unterschiedliche Items vorgelegt bekommen, führt zu einem Problem bei der Testwertbildung. Wenn Testpersonen mit hoher Merkmalsausprägung schwierigere Items vorgelegt bekommen als Testpersonen mit niedrigerer Merkmalsausprägung, kann ein fairer interindividueller Vergleich nicht mit Testwerten erfolgen, die – wie im Rahmen der Klassischen Testtheorie (KTT, ▶ Kap. 13) üblich – aus der Anzahl der korrekt beantworteten Items gebildet werden: Zehn korrekt beantwortete Items hoher Schwierigkeit sind Ausdruck einer höheren Merkmalsausprägung als zehn korrekt beantwortete Items niedriger Schwierigkeit. Um anstelle von einfachen Summen korrekter Antworten besser geeignete Testwerte bestimmen zu können, setzt man bei computerisierten adaptiven Tests Modelle der Item-Response-Theorie (IRT, ▶ Kap. 16; van der Linden 2016a) als psychometrische Modelle ein. Da bei IRT-Modellen die Charakteristika von Items (als Itemparameter bezeichnet) und die individuelle Merkmalsausprägung (als Personenparameter bezeichnet) separat bestimmt werden, können Testergebnisse auch dann problemlos zwischen Personen verglichen werden, wenn von den Testpersonen unterschiedliche Items bearbeitet wurden.

Hierbei ist es von zentraler Bedeutung, dass vor Beginn der Testung die Menge aller Testitems, der sog. *Itempool*, kalibriert wurde und sichergestellt ist, dass sie die Annahmen des verwendeten IRT-Modells nicht verletzen. Die aus der Kalibrierung hervorgehenden Itemparameter ermöglichen es, auf Basis der von einer Testperson gegebenen Antworten, die individuelle Ausprägung des zu messenden Merkmals zu schätzen. Das heißt, dass eine Testperson, die von zehn schwierigen Items fünf korrekt und die anderen fünf inkorrekt beantwortet hat, einen höheren Personenparameter als Schätzung der individuellen Ausprägung im zu messenden Merkmals erhält als eine Testperson, die von zehn leichten Items fünf korrekt und fünf inkorrekt beantwortet hat.

Prinzipiell kann eine sehr breite Palette von IRT-Modellen als psychometrisches Modell beim computerisierten adaptiven Testen genutzt werden. Bei den meisten im Einsatz befindlichen computerisierten adaptiven Tests werden jedoch eindimensionale Modelle für dichotome Items mit einem Parameter (einparametrisches logistisches Modell, 1PL-Modell), zwei Parametern (2PL-Modell) und seltener auch drei Parametern (3PL-Modell) verwendet (▶ Kap. 16 und 18).

Das Grundprinzip des computerisierten adaptiven Testens lässt sich mit dem in □ Abb. 20.1 gezeigten Flussdiagramm zusammenfassen. Der typische Verlauf findet sich in allen computerisierten adaptiven Tests.

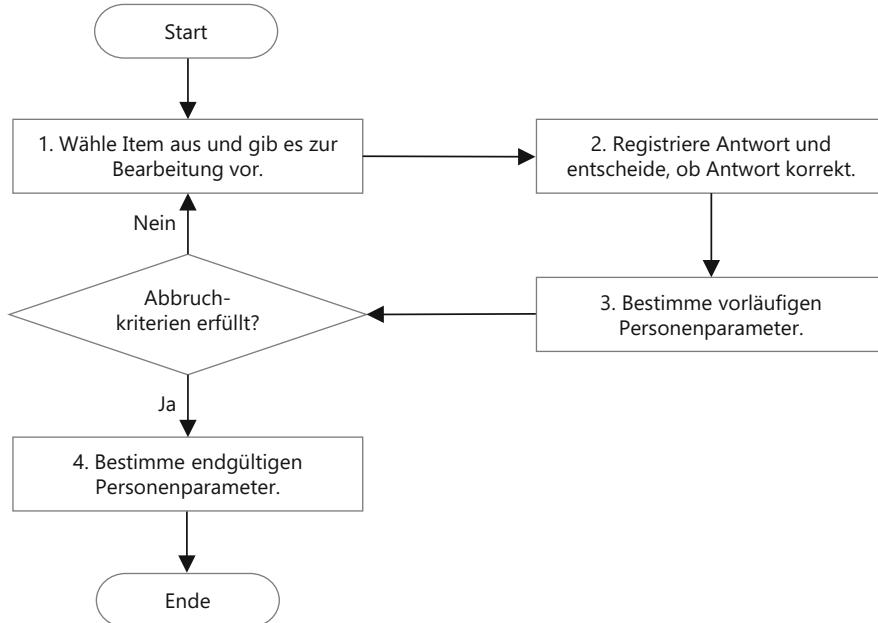
Ungeeignete und geeignete Testwerte

Itemkalibrierung

Psychometrisches Modell

Typischer Ablauf des computerisierten adaptiven Testens

Der Test startet mit der Auswahl des ersten Items aus dem Itempool (▶ Abschn. 20.3.1) und dessen Vorgabe zur Bearbeitung (▶ Abschn. 20.3.2). Die Antwort der



■ Abb. 20.1 Flussdiagramm zum typischen Ablauf computerisierter adaptiver Tests

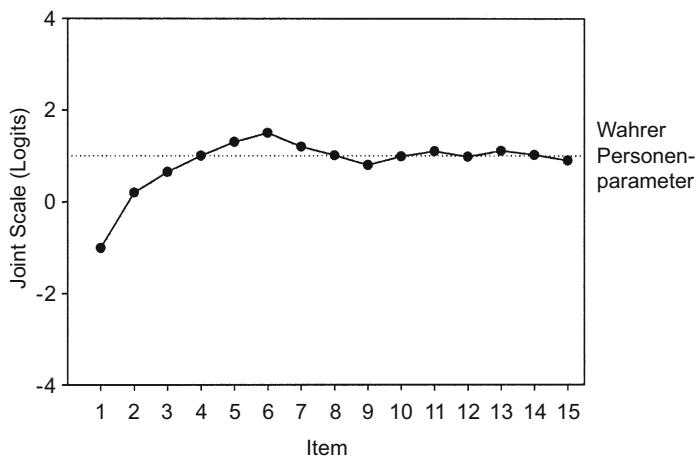
Testperson auf dieses Item wird seitens des Computers registriert und auf Korrektheit geprüft. Mit dem Resultat der Prüfung liegt die empirisch gewonnene Information vor, die es ermöglicht, die individuelle Merkmalsausprägung der Testperson etwas genauer einzuschätzen als zuvor. Diese Information wird für eine vorläufige Schätzung des Personenparameters genutzt (► Abschn. 20.3.3). Danach wird geprüft ob ein oder mehrere vorab aufgestellte Kriterien zur Beendigung des Tests erfüllt sind (► Abschn. 20.3.6). Dies wird nach dem ersten Item üblicherweise noch nicht der Fall sein. Deshalb werden die Schritte 1 bis 3 so lange durchlaufen, bis alle Abbruchkriterien erfüllt sind. Bei der Itemauswahl (► Abschn. 20.3.4) werden neben einem statistischen Optimalitätskriterium üblicherweise weitere nicht statistische Einschränkungen berücksichtigt (► Abschn. 20.3.5). Nachdem alle Abbruchkriterien erfüllt sind, wird die endgültige Schätzung der individuellen Ausprägung des zu messenden Merkmals bestimmt und die Testung beendet.

Optimalitätskriterium

Beispiel 20.1: Ablauf eines typischen computerisierten adaptiven Tests

Der konkrete Ablauf eines typischen computerisierten adaptiven Tests soll an einem Beispiel illustriert werden: Gegeben sei eine Menge von 400 dichotomen Items. Die Items verletzen die Annahmen des 1PL-Modells nicht und messen demnach alle die gleiche Merkmalsdimension, die mit θ bezeichnet werden soll. Ferner sei aus einer Kalibrierungsstudie für jedes Item i die Itemschwierigkeit b_i bekannt. Personenparameter und Itemschwierigkeiten sind unter dem 1PL-Modell auf einer gemeinsamen Skala (*Joint Scale*) lokalisiert. Der Mittelwert von 0 auf dieser Skala zeige eine mittlere Merkmalsausprägung an. Getestet werde ein Individuum v mit einer überdurchschnittlichen Ausprägung des zu messenden Merkmals von $\theta_v = 1.00$ (diese Information liegt in realen Testsituationen nicht vor, wird aber zu Illustrationszwecken hier als bekannt angenommen). Der Ablauf einer exemplarischen adaptiven Testung des Individuums ist in ■ Abb. 20.2 dargestellt.

Es ist zu erkennen, dass zu Beginn ein Item mit einer Schwierigkeit von $b_1 = -1.00$ vorgegeben wird. Solche Items mit unterdurchschnittlicher Schwierigkeit werden beim Testen öfters zu Beginn vorgegeben, um einen problemlosen



■ Abb. 20.2 Illustration eines adaptiven Testablaufs. Die *Punkte* repräsentieren Itemschwierigkeiten

Einstieg in den Test zu fördern (► Abschn. 20.3.2). Die Wahrscheinlichkeit, dass die untersuchte Testperson v eine korrekte Antwort y auf das erste Item i geben kann ($y_{vi} = 1$), ist mit $P(y_{vi} = 1|\theta_v) = .88$ relativ hoch, da die Ausprägung des zu messenden Merkmals mit $\theta_v = 1.00$ deutlich höher ist als die Schwierigkeit des präsentierten Items. In der Tat wird in dem Beispiel das erste Item auch gelöst, sodass der Testperson als zweites Item ein schwierigeres Item mit $b_2 = .20$ vorgegeben wird, das ebenfalls gelöst werden kann, weshalb nachfolgend ein noch schwierigeres Item mit $b_3 = .65$ präsentiert wird (die zur Itemauswahl genutzten Kriterien werden in ► Abschn. 20.3.4 erläutert). Nachdem dieses wiederum korrekt beantwortet wurde, entspricht die Schwierigkeit des vierten Items mit $b_4 = 1.00$ exakt der Merkmalsausprägung der Testperson. Unter dem hier genutzten 1PL-Modell beträgt für das untersuchte Individuum die Wahrscheinlichkeit, dieses Item korrekt zu beantworten $P(y_{vi} = 1|\theta_v) = .50$. Im vorliegenden Beispiel kann die Testperson das Item in der Tat richtig beantworten. Nachfolgend wird deshalb ein Item ausgewählt, dessen Schwierigkeit mit $b_5 = 1.30$ abermals höher ausfällt.

Der weitere Verlauf der Testung folgt der gleichen Logik: Wird ein Item korrekt beantwortet, dann wird als nächstes ein schwierigeres Item vorgegeben; wird ein Item nicht korrekt beantwortet, dann wird als nächstes ein leichteres Item vorgegeben. Bei der Wahl der Itemschwierigkeit wird das Antwortverhalten auf alle zuvor bearbeiteten Items beachtet. Im Beispiel auf Basis des 1PL-Modells pendelt sich die Schwierigkeit der vorgegebenen Items recht schnell um die individuelle Ausprägung der Testperson in dem zu messenden Merkmal ein. Sie ist in ■ Abb. 20.2 durch eine gestrichelte Linie gekennzeichnet.

20.3 Elementare Bausteine

Es können sechs elementare Bausteine computerisierter adaptiver Tests unterschieden werden, die in den nachfolgenden Abschnitten einzeln besprochen werden:

- Itempool (► Abschn. 20.3.1)
- Itemauswahl zu Beginn der Testung (► Abschn. 20.3.2)
- Schätzung der individuellen Ausprägung des zu messenden Merkmals (► Abschn. 20.3.3)
- Auswahl von Items während der Testung (► Abschn. 20.3.4)

- Umgang mit Einschränkungen bei der Itemauswahl (► Abschn. 20.3.5)
- Kriterien für die Beendigung der Testung (► Abschn. 20.3.6)

20.3.1 Itempool

Als Itempool wird die Menge der Items bezeichnet, auf die bei der Durchführung eines computerisierten adaptiven Tests zurückgegriffen werden kann. Während der eigentlichen adaptiven Testung wird dann jeweils aus den noch nicht vorgegebenen Items des Itempools dasjenige selegiert, von dessen Beantwortung ein Maximum an diagnostischer Information zu erwarten ist. Um das Ausmaß an diagnostischer Information bestimmen zu können, werden statistische Charakteristika für alle Items im Itempool benötigt. Diese statistischen Charakteristika werden im Rahmen einer Kalibrierungsstudie ermittelt.

Die *Kalibrierungsstudie* ist ein integraler Bestandteil der Entwicklung eines computerisierten adaptiven Tests. Sie ist dessen operationaler Anwendung vorschaltet. Das primäre Ziel der Kalibrierungsstudie liegt in der Bestimmung statistischer Charakteristika der Testitems. Diese Charakteristika werden im Rahmen der IRT als Itemparameter bezeichnet. Beim 1PL-Modell handelt es sich dabei um die Itemschwierigkeit, beim 2PL-Modell um die Itemschwierigkeit und die Itemdiskrimination und beim 3PL-Modell kommt zu den beiden genannten Itemparametern noch ein Pseudo-Rateparameter hinzu. Die geschätzten Itemparameter werden bei der nachfolgenden Anwendung des computerisierten adaptiven Tests nicht erneut geschätzt. Somit kommt der Kalibrierungsstudie eine hohe Wichtigkeit zu.

Insbesondere ist sicherzustellen, dass die Testbedingungen bei der Kalibrierungsstudie vergleichbar mit den Testbedingungen der eigentlichen Anwendung des fertigen Tests sind. Wenn die Testungen mit dem fertigen Test beispielsweise als Einzeltestungen am PC stattfinden sollen, dann sollte bei der Kalibrierungsstudie der Test in gleicher Weise vorgegeben werden. Eine Gruppentestung mit Tablets wäre indes nicht geeignet, da nicht auszuschließen ist, dass einige oder alle Items bei dieser Art der Testvorgabe anders funktionieren, sodass andere Itemparameter nötig wären.

Während hinsichtlich der Testbedingungen strenge Vergleichbarkeitsmaßstäbe zwischen Kalibrierungsstudie und eigentlicher Testanwendung anzulegen sind, fallen aufgrund der Nutzung von IRT-Modellen die Anforderungen an die *Repräsentativität der Kalibrierungsstichprobe* weniger strikt aus. Sicherzustellen ist lediglich, dass

- a. alle bei der Kalibrierungsstudie getesteten Individuen über eine Ausprägung des zu messenden Merkmals verfügen (was beispielsweise bei einer Stichprobe von Achtklässlerinnen und Achtklässlern bei der Entwicklung eines Wissenstests für KFZ-Mechatroniker nicht anzunehmen wäre) und
- b. die Merkmalsausprägungen der getesteten Personen über den gesamten Bereich des Merkmals streuen, den der fertige Test letztendlich abbilden soll.

Insofern diese beiden Punkte gewährleistet sind, ist darüber hinaus keine strikte Vergleichbarkeit von Kalibrierungsstichprobe und anvisierter Zielstichprobe erforderlich. Im Hinblick auf eine präzise Schätzung der Itemparameter von Items mit sehr niedriger oder sehr hoher Schwierigkeit ist vielmehr sogar anzustreben, mehr Testpersonen mit besonders niedriger und solche mit besonders hoher Merkmalsausprägung in die Kalibrierungsstichprobe aufzunehmen, als bei einer Zufallsstichprobe zu erwarten wären.

Neben der Frage der Merkmalsverteilung in der Kalibrierungsstichprobe ist deren Größe vorab festzulegen. Zentrale Einflussgrößen auf die benötigte *Stichprobengröße* sind die Anzahl der zu kalibrierenden Items und das genutzte IRT-Modell. Um eine Antwortabhängige Itemauswahl zu ermöglichen, umfassen Item-

Itempool

Kalibrierungsstudie

Vergleichbarkeit der Testbedingungen

Repräsentativität der Kalibrierungsstichprobe

Größe der Kalibrierungsstichprobe

Randomisierte Itemzuteilung

pools computerisierter adaptiver Tests notwendigerweise mehr Items, als während der vorgesehenen Testzeit beantwortet werden können. Aus diesem Grund ist für die Kalibrierungsstudie zu entscheiden, welche Teilmengen aus allen verfügbaren Items von den einzelnen Testpersonen zu bearbeiten sind. Die Zuteilung von Items zu Testpersonen erfolgt üblicherweise unter Zuhilfenahme eines balancierten unvollständigen Block-Designs (z. B. Frey et al. 2009). Im Resultat bekommt jede Testperson eine Teilmenge der zu kalibrierenden Items vorgelegt, wobei wichtige potentielle Störgrößen wie itemspezifische Vorgabehäufigkeiten und Itemposition während der Testungen ausbalanciert werden. Dieses Vorgehen ist im Übrigen einer randomisierten Itemzuteilung überlegen, da sich die angestrebte Ausbalancierung (z. B. der itemspezifischen Vorgabehäufigkeiten) bei randomisierter Itemzuteilung erst bei deutlich größeren Stichproben als bei der Verwendung balancierter unvollständiger Block-Designs asymptotisch einstellt.

Im Hinblick auf die benötigte Stichprobengröße bei der Kalibrierungsstudie ist für jedes Item sicherzustellen, dass alle benötigten Itemparameter zuverlässig geschätzt werden können. Daraus folgt, dass die Anzahl der Testpersonen umso größer ausfällt, je größer die Anzahl der zu kalibrierenden Items ist und je mehr Parameter pro Item zu schätzen sind. Als grobe Daumenregel empfiehlt de Ayala (2009) bei guten Voraussetzungen (z. B. gute Passung zwischen Itemschwierigkeiten und Merkmalsverteilung, viele Items je Testperson) für das 1PL-Modell ein Minimum von einigen Hundert Antworten, für das 2PL-Modell von 500 Antworten und für das 3PL-Modell 1000 Antworten je Item. Wichtig zu erwähnen ist, dass diese Richtlinien nur als grober Anhaltspunkt verwendet werden können, da die Anforderungen an die Stichprobengröße erheblich in Abhängigkeit der jeweiligen Testsituation (Itemanzahl, Items je Person, Merkmalsstreuung, Passung Merkmalsverteilung mit Itemschwierigkeitsverteilung, Anteil fehlender Werte etc.) und der verfolgten Zielsetzung variieren. Da die genannten Daumenregeln von de Ayala allgemein für Kalibrierungen unter Nutzung von IRT-Modellen formuliert wurden, stellen sie beim computerisierten adaptiven Testen aufgrund der zentralen Bedeutung der Itemparameterschätzungen für Itemauswahl und Personenparameterschätzung untere Grenzen dar.

Die Skalierung der bei der Kalibrierungsstudie erhobenen Daten erfolgt mithilfe eines IRT-Modells. Dabei werden die bei IRT-basierten Testkonstruktionen üblichen Schritte der letztendlich in den Itempool aufzunehmenden Items durchlaufen, d. h. die Prüfung der Modellpassung, die Itemparameterschätzung und die Itemselektion (► Kap. 16; vgl. Khorramdel und von Davier 2016). Vorab sind die angestrebten Charakteristika des Itempools festzulegen, die sich auf die *Anzahl der Items* (ggf. unterteilt nach weiteren Aspekten wie Inhaltsfacetten oder Antwortmodi) und die *Verteilung von Itemparametern* bezieht. Im Hinblick auf die Anzahl der im Itempool enthaltenen Items ist offensichtlich, dass eine adaptive Anpassung an das Antwortverhalten der Testpersonen dann erfolgen kann, wenn der Itempool mehr Items enthält, als man in der geplanten Testzeit vorgeben würde. Je umfangreicher der Itempool ist, desto besser kann die Itemvorgabe an das Antwortverhalten angepasst werden. Bei Tests mit vorab festgelegter Itemanzahl (► Abschn. 20.3.6) kann bestimmt werden, wie der Itempool mindestens beschaffen sein muss, um eine optimale Anpassungsfähigkeit zu gewährleisten. Wird das 1PL-Modell verwendet, ist dies ein Itempool, bei dem für alle Testpersonen zu allen möglichen Testzeitpunkten jeweils mindestens ein Item vorhanden ist, dessen Itemschwierigkeit der aktuellen provisorischen Merkmalsschätzung $\hat{\theta}_j$ entspricht. Die Größe des dafür nötigen optimalen Itempools kann mit dem von He und Reckase (2014) beschriebenen Algorithmus unter Berücksichtigung nicht statistischer Einschränkungen (► Abschn. 20.3.5) bestimmt werden. Gerade bei komplex definierten Konstrukten mit vielen zu operationalisierenden Teilespekten resultieren dabei sehr große Itemanzahlen. Dies ist deshalb der Fall, da nicht nur alle theoretisch spezifizierten Teilespekte des zu messenden Merkmals (z. B. verschiedene Facetten der Intelligenz bei der Intelligenzmessung) durch die Items im

Daumenregeln für Anzahl der Testpersonen in der Kalibrierungsstudie in Abhängigkeit von der Komplexität der Modelle

Angestrebte Charakteristika des Itempools

Größe des Itempools

Pool abgedeckt werden müssen, sondern dies auch für alle Schwierigkeitsstufen zutreffen muss, die für den zu messenden Merkmalsbereich relevant sind. Da optimale Itempools aufgrund ihrer Größe im zuweilen vierstelligen Bereich häufig nicht mit vertretbarem Aufwand konstruiert und kalibriert werden können, kann mit Simulationsstudien vor Verwendung der Tests (sog. „präoperationale Simulationsstudien“) ein für die jeweilige Testentwicklung vernünftiger und realisierbarer Kompromiss zwischen optimaler Itempoolgröße und Messgenauigkeit bestimmt werden. Die resultierenden Itempoolgrößen fallen dabei üblicherweise deutlich kleiner aus als die von optimalen Itempools. Nicht selten sind brauchbare Lösungen bereits mit Poolgrößen um die 100 Items zu realisieren (für Beispiele s. Spoden et al. 2018; Ziegler et al. 2016).

20.3.2 Testbeginn

Laut Definition orientiert sich die Itemauswahl bei einem computerisierten adaptiven Test am vorherigen Antwortverhalten der untersuchten Testperson. Dies wirft die Frage auf, welches Item ganz zu Beginn der Testung zu wählen ist, also zu einem Zeitpunkt, an dem die Testperson noch kein Antwortverhalten gezeigt hat.

Bei vielen computerisierten adaptiven Tests wird als Erstes ein Item vorgelegt, das eine mittlere Schwierigkeit aufweist. Um einen problemlosen Einstieg in den Test zu ermöglichen, werden manchmal auch sehr leichte Items mit einer Lösungswahrscheinlichkeit von $\approx .80$ vorgegeben (sog. *Eisbrecheritems*).

Liegen zusätzliche Vorinformationen über die Ausprägung der Testperson in dem zu messenden Merkmal vor, können diese für eine Auswahl des ersten Items genutzt werden. Als Vorinformationen können Testresultate aus vorherigen Testungen mit dem gleichen Test, Resultate bei Tests, die ähnliche Merkmale messen, sowie alle anderen Maße verwendet werden, von denen ein enger Zusammenhang mit dem zu messenden Merkmal angenommen wird. Aufgrund von Vorinformationen kann eine mehr oder weniger genaue A-priori-Schätzung der Merkmalsausprägung des untersuchten Individuums erfolgen, auf deren Basis es dann möglich ist, ein Item mit optimalen Eigenschaften aus dem Itempool auszuwählen und zur Bearbeitung vorzugeben (vgl. z. B. van der Linden 1999a).

Die Entscheidung, wie bei der Auswahl des ersten Items vorzugehen ist, sollte nach diagnostischer Zielsetzung, untersuchter Stichprobe und Verfügbarkeit von Vorinformationen über die untersuchten Individuen erfolgen. Sowohl bei Simulations- als auch bei empirischen Studien zeigte sich jedoch, dass bei computerisierten adaptiven Tests mit üblicher Länge die Auswahl des ersten Items zumindest auf Stichprobenebene einen vergleichsweise geringen Einfluss auf die Präzision des Testwerts am Ende der Testung hat. Diesen Ergebnissen entsprechend ordnen die meisten Autoren der Wahl des ersten Items bei einem computerisierten adaptiven Test nur eine untergeordnete Wichtigkeit zu (z. B. Hambleton et al. 1991).

Präoperationale Simulationsstudien

Eisbrecheritems

Nutzung von Vorinformationen

20.3.3 Personenparameterschätzung

Beim computerisierten adaptiven Testen werden zu verschiedenen Zeitpunkten Personenparameter als Messwert für die interessierende individuelle Merkmalsausprägung geschätzt; einerseits vorläufige Personenparameter während der Testung (mehrmales Schritt 3 im Flussdiagramm in □ Abb. 20.1) und andererseits der endgültige Personenparameter am Ende der Testung (Schritt 4 im Flussdiagramm in □ Abb. 20.1). Zur Bestimmung der Personenparameter können die im IRT-Bereich üblichen Schätzmethoden genutzt werden (► Kap. 19; van der Linden 2016b). Grundlegend können zwei Ansätze unterschieden werden. Einerseits werden Maximum-Likelihood-Schätzer (ML-Schätzer; z. B. de Ayala 2009), wozu auch der

ML- und Bayes-Schätzung

WLE („Weighted Likelihood Estimate“; Warm 1989) gezählt wird, und andererseits Bayes-Schätzer wie der EAP („Expected A Posteriori“; Bock und Mislevy 1982) oder der BME („Bayes Modal Estimate“; Mislevy 1986), der auch als MAP („Maximum A Posteriori Estimate“) bezeichnet wird, verwendet. Beide Schätzansätze eignen sich prinzipiell gut zur Bestimmung der individuellen Ausprägung latenter Merkmalsausprägungen, weisen aber kleine Unterschiede bezüglich der Schätzgüte auf. Bayes'sche Ansätze haben im Vergleich zu ML-Ansätzen den Vorteil kleinerer bedingter Standardfehler. Sie weisen aber gleichzeitig den Nachteil eines größeren Bias auf, vor allem in extremen θ -Bereichen. Die Größe dieser Unterschiede bewegt sich jedoch nur bei Tests mit geringer Itemanzahl in relevanten Größenordnungen und wird ab einer Testlänge von 20 Items vernachlässigbar klein (van der Linden 1998). Vergleiche verschiedener Methoden der Personenparameterschätzung findet man für eindimensionale adaptive Tests bei van der Linden und Pashley (2010) und Cheng und Liou (2000) sowie für multidimensionale adaptive Tests bei Diao (2010).

Da ML-Schätzer bei invariantem Antwortverhalten nicht bestimmt werden können (z. B. wenn im bisherigen Testverlauf nur richtige oder nur falsche Antworten gegeben wurden), geht man in der Praxis zuweilen so vor, dass die Schätzung der vorläufigen Personenparameter während der Testung mit Bayes'schen Methoden erfolgt und die abschließende Schätzung mit einem ML-basierten Schätzer.

20.3.4 Itemauswahl

20.3.4.1 Voll adaptive Itemauswahl

Maßgeschneiderte Tests (Tailored Testing)

In diesem Kapitel werden vornehmlich solche computerisierte adaptive Tests beschrieben, die von Weiss (2011) als voll adaptive computerisierte Tests bezeichnet werden. Diese Art des computerisierten adaptiven Testens, bei dem nach der Beantwortung eines jeden Items eine Anpassung stattfindet, wird auch als maßgeschneidertes Testen (Tailored Testing) bezeichnet. Es handelt sich dabei um die flexibelste Art des computerisierten adaptiven Testens, die gleichzeitig auch am weitesten verbreitet ist.

Bei der voll adaptiven Testdarbietung kann aufgrund der Antwort auf das erste Item eine erste grobe Schätzung der Merkmalsausprägung der Testperson erfolgen und ein weiteres passendes Item ausgewählt und zur Bearbeitung vorgelegt werden. Die Itemauswahl basiert dabei auf einem statistischen *Optimalitätskriterium* und berücksichtigt in der Regel zusätzliche nicht statistische Einschränkungen (► Abschn. 20.3.5).

Optimalitätskriterium

Maximale Iteminformation

Ein häufig zum Einsatz gebrachtes Optimalitätskriterium ist die Auswahl nach maximaler Iteminformation. Dieses Kriterium besteht darin, jeweils dasjenige Item aus den noch nicht der aktuellen Testperson vorgegebenen Items auszuwählen, das für deren vorläufige Personenparameterschätzung $\hat{\theta}$ über maximale Information verfügt. Dabei geht man so vor, dass für alle noch verfügbaren Items der itemspezifische Wert der Informationsfunktion am Punkt $\hat{\theta}$ berechnet und das Item mit dem höchsten Wert ausgewählt wird. Die Berechnung der Iteminformationsfunktion variiert in Abhängigkeit des genutzten IRT-Modells. Beim 3PL-Modell ergibt sie sich aus (z. B. de Ayala 2009):

$$I_i(\theta) = a_i^2 \left(\frac{p_i(\theta) - c_i}{1 - c_i} \right)^2 \frac{q_i(\theta)}{p_i(\theta)}, \quad (20.1)$$

Lösungs- und Gegenwahrscheinlichkeit

wobei $p_i(\theta)$ die Wahrscheinlichkeit ausdrückt, dass eine Person mit latenter Merkmalsausprägung θ das Item i lösen kann, und $q_i(\theta)$ die Gegenwahrscheinlichkeit, dass das Item i von einer Person mit latenter Merkmalsausprägung θ nicht gelöst werden kann, sodass $q_i(\theta) = 1 - p_i(\theta)$ gilt. Da die Iteminformation $I_i(\theta)$

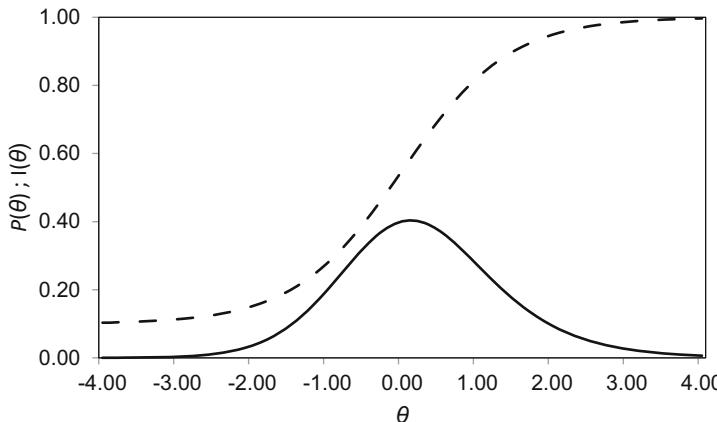


Abb. 20.3 Iteminformationsfunktion (durchgezogene Linie) mit zugehöriger IC-Funktion (gestrichelte Linie) für ein Item mit $a = 1.4$, $b = .0$ und $c = .1$

in Abhängigkeit der latenten Merkmalsausprägung θ formuliert ist, kann pro Item kein konstanter Informationswert angegeben werden. Vielmehr variiert die Iteminformation in Abhängigkeit von der latenten Merkmalsausprägung. Weiterhin ist Gl. (20.1) zu entnehmen, dass sowohl der Diskriminationsparameter a als auch der Pseudo-Rateparameter c einen direkten Einfluss auf die Höhe der Iteminformation haben. Während hohe a -Parameter mit hoher Iteminformation einhergehen, fällt die Iteminformation mit steigenden c -Parametern zunehmend geringer aus. Wird das Maximum des c -Parameters von eins erreicht, verfügt das Item über keinerlei Information. Dies ist direkt intuitiv verständlich, wenn man bedenkt, dass in diesem Extremfall das gesamte Antwortverhalten einzig auf den Zufall zurückzuführen ist und damit nichts über das zu messende Merkmal aussagt. Eine exemplarische Iteminformationsfunktion für ein Item mit $a = 1.4$, $b = .0$ und $c = .1$ ist zusammen mit der zugehörigen itemcharakteristischen Funktion (IC-Funktion) in Abb. 20.3 zu sehen. Auf der x -Achse ist das zu messende latente Merkmal abgetragen. Je höher die Werte, desto höher die Merkmalsausprägung. Das Maximum der Iteminformation des gezeigten Items liegt bei ungefähr $\theta = .1$. An dieser Stelle ist das Item am informativsten – man könnte auch sagen, dass es an dieser Stelle am besten misst. Das bedeutet, dass es Individuen, deren Merkmalsausprägung im Bereich von 0.1 liegt, besser differenzieren kann als Individuen mit niedrigeren oder höheren Merkmalsausprägungen. Bei Merkmalsausprägungen unter -3 oder über 4 liefert das Item fast keine Information mehr.

Das 2PL-Modell und das 1PL-Modell verfügen über weniger Itemparameter als das 3PL-Modell. Aus diesem Grund kann für diese beiden Modelle die Iteminformation einfacher bestimmt werden als für das 3PL-Modell. Für das 2PL-Modell berechnet sie sich durch:

$$I_i(\theta) = a_i^2 p_i(\theta) q_i(\theta) \quad (20.2)$$

Da der Itemdiskriminationsparameter beim 1PL-Modell für alle Items des Tests den gleichen Wert hat (z. B. 1), vereinfacht sich die Berechnung der Iteminformationsfunktion für das 1PL-Modell weiter zu:

$$I_i(\theta) = p_i(\theta) q_i(\theta) \quad (20.3)$$

Im Gegensatz zu den Iteminformationsfunktionen für das 3PL- und 2PL-Modell enthält dieser Ausdruck weder den Itemparameter a noch den Itemparameter c . Es besteht beim 1PL-Modell also keine direkte Abhängigkeit der Iteminformation von der Diskrimination oder der Pseudo-Ratewahrscheinlichkeit. Anders sieht

Maximum der Iteminformation

Iteminformation im 2PL-Modell und im 1PL-Modell

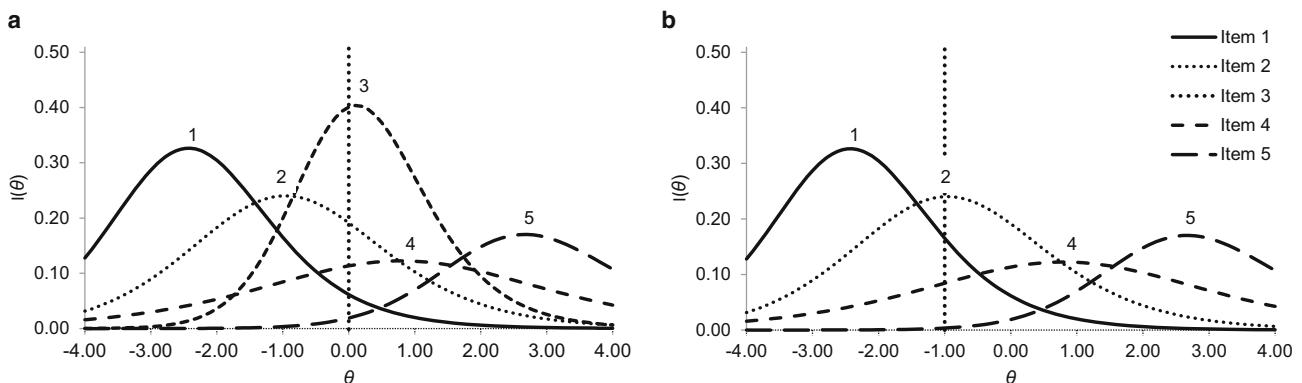
dies für den Itemschwierigkeitsparameter b aus. Das in der Gl. (20.3) für das 1PL-Modell angeführte Produkt aus der Lösungswahrscheinlichkeit und seiner Gegegenwahrscheinlichkeit nimmt den maximalen Wert von .25 dann an, wenn beide Wahrscheinlichkeiten gleich groß sind und somit bei $p_i(\theta) = q_i(\theta) = .50$. Dieser Punkt entspricht in dem 1PL-Modell dem Wendepunkt der IC-Funktion und somit per Definition der Itemschwierigkeit. Bei der Itemauswahl nach dem Optimalitätskriterium der maximalen Information wird bei Verwendung des 1PL-Modell somit jeweils dasjenige Item ausgewählt, dessen Itemschwierigkeit b_i bestmöglich mit dem vorläufigen Personenparameter $\hat{\theta}$ übereinstimmt (vgl. hierzu auch ▶ Kap. 16). Hieraus begründet sich auch die vor allem in älterer Literatur zu findende Definition, dass computerisierte adaptive Tests das Schwierigkeitsniveau der Items an die Leistungsfähigkeit der Testpersonen anpassen. Streng genommen trifft dies nur beim 1PL-Modell vollständig zu. Bei anderen IRT-Modellen wie dem 2PL-Modell und dem 3PL-Modell resultiert nicht zwangsläufig eine exakte Passung von Itemschwierigkeit und provisorischer Merkmalsschätzung, da bei diesen Modellen der Diskriminationsparameter bei der Itemselektion aufgrund seiner quadrierten Berücksichtigung bei der Berechnung der Iteminformation im Vordergrund steht. In der Regel werden jedoch auch beim 2PL- und 3PL-Modell Items ausgewählt, die für das untersuchte Individuum ungefähr mittlere Schwierigkeiten aufweisen, jedoch nicht zwangsläufig immer jenes Item, dessen Itemschwierigkeit die geringste Differenz zu $\hat{\theta}$ aufweist.

Zur Illustration des Vorgehens bei der Itemauswahl nach dem Optimalitätskriterium der maximalen Information, sind in □ Abb. 20.4a die Informationsfunktionen eines Itempools von fünf Items abgebildet (Der Itempool eines operationalen computerisierten adaptiven Tests wäre natürlich deutlich größer.).

Die Iteminformationsfunktionen unterscheiden sich schon auf den ersten Blick deutlich. Während die Iteminformationsfunktion von Item 3 einen hohen Maximalwert aufweist und an den Rändern relativ schnell abfällt, verläuft die Kurve von Item 4 insgesamt flacher, erreicht bei extremen θ -Werten aber leicht höhere Werte als die Kurve von Item 1. Der zentrale Unterschied zwischen den beiden Informationsfunktionen ist deren unterschiedliche Itemdiskrimination ($a_3 = 1.4$; $a_4 = .7$). □ Abb. 20.4a ist weiterhin zu entnehmen, dass es sich bei Item 5 um das schwierigste der fünf Items handelt, da dessen Maximum am weitesten rechts liegt. Es ist insbesondere für die Messung von Individuen mit hoher Merkmalsausprägung informativ. Item 1 eignet sich aufgrund seiner sehr niedrigen Schwierigkeit hingegen vor allem für die Differenzierung von Personen mit sehr niedriger Merkmalsausprägung. Die gepunktete vertikale Linie repräsentiert den in diesem Beispiel zu Beginn der adaptiven Testung angenommenen Personenparameter von $\hat{\theta} = 0.00$. An dieser Stelle der latenten Merkmalsdimension verfügt Item 3 über die höchste

Itemauswahl gemäß Iteminformation

Vorläufige Personenparameterschätzung



□ Abb. 20.4 Iteminformationsfunktionen für fünf Items gemäß 3PL-Modell, **a** vor und **b** nach Vorgabe des ersten Items (hier: Item 3). Die gepunktete vertikale Linie repräsentiert die im ersten bzw. nächsten Schritt angenommene Merkmalsausprägung der Testperson. Näheres siehe Text

Iteminformation und wird deshalb zur Bearbeitung vorgegeben. Die Antwort der Testperson wird registriert und auf Korrektheit geprüft. Im vorliegenden Fall sei die Antwort der Testperson falsch. Die aktualisierte vorläufige Personenparameterschätzung fällt mit $\hat{\theta} = -1.00$ deshalb nun niedriger aus als die initiale Annahme bezüglich des Personenparameters, s. □ Abb. 20.4b.

Neben der aktualisierten vorläufigen Personenparameterschätzung zeigt □ Abb. 20.4b die nach Vorgabe von Item 3 verbleibenden vier Iteminformationsfunktionen. Für den Wert von $\hat{\theta} = -1.00$ hat das Item 2 die höchste Iteminformation, sodass es ausgewählt und vorgegeben wird. Die Antwort wird wieder registriert, auf richtig/falsch geprüft und die vorläufige Personenparameterschätzung aktualisiert. Wenn die Testperson das Item richtig beantworten kann, würde nun ein höherer Wert von $\hat{\theta} = -.50$ resultieren. Im nächsten Schritt würde dann das Item 1 gewählt und vorgegeben. Der geschilderte Ablauf wiederholt sich so lange, bis alle Abbruchkriterien erfüllt sind.

Das Vorgehen führt zu zwei wesentlichen Resultaten: Erstens werden die Schritte zwischen den vorläufigen Merkmalsschätzungen im Verlauf der Testung üblicherweise immer kleiner, sodass die Personenparameterschätzung zunehmend auf einen stabilen Wert konvergiert; zweitens verringern sich die Standardfehler der θ -Schätzungen von Item zu Item und konvergieren ebenfalls. Aufgrund des Zusammenhangs $SE(\theta) = 1/\sqrt{I(\theta)}$ ist bei der Itemauswahl gemäß dem Optimalitätskriterium der maximalen Information sichergestellt, dass sich der Standardfehler bei jeder Itemselektion in größtmöglicher Weise verringert.

In den letzten Jahrzehnten wurden weitere Optimalitätskriterien vorgeschlagen, die zur Itemauswahl beim computerisierten adaptiven Testen genutzt werden können (Übersicht in van der Linden und Pashley 2010). In einigen besonderen Anwendungssituationen kann mit alternativen Optimalitätskriterien die Performance von computerisierten adaptiven Tests im Vergleich zur Auswahl nach dem Kriterium der maximalen Iteminformation leicht gesteigert werden. Das Kriterium der maximalen Iteminformation erwies sich aber auch in diesen Fällen als nur geringfügig schlechter und überzeugt insgesamt durch einen sehr breiten Einsatzbereich. Vor diesem Hintergrund ist es weiterhin als das Standardkriterium bei der Itemauswahl beim computerisierten adaptiven Testen anzusehen.

20.3.4.2 Eingeschränkt adaptive Itemauswahl

Bestimmte Rahmenbedingungen legen es manchmal nahe, die Flexibilität bei der adaptiven Itemauswahl einzuschränken. Manchmal ist es wünschenswert die antwortabhängigen Verzweigungen bereits vor der Testung festzulegen. Dies ist beispielsweise der Fall, wenn keine digitalen Endgeräte für die Testdurchführung zur Verfügung stehen, auf denen die Personenparameterschätzungen umgehend realisiert werden können. Tests, bei denen bereits vor Testbeginn feststeht, welche Items bei welchem Antwortverhalten vorgelegt werden, bezeichnet man als *fest verzweigte Tests* („fixed-branched“). Ein Beispiel für einen solchen Test ist das Adaptive Intelligenz-Diagnostikum II (AID-II; Kubinger 2009).

Eine weitere Notwendigkeit, die Flexibilität voll adaptiver computerisierter Tests einzuschränken, liegt bei Items im sog. „Testlet-Format“ vor. Bei *Testlets* handelt es sich um Itemgruppen, die sich auf einen gemeinsamen Stimulus wie eine Rahmengeschichte, ein Bild oder eine Tabelle beziehen (s. auch ▶ Kap. 5). Derartige Items werden beispielsweise bei groß angelegten Vergleichsstudien wie Programme for International Student Assessment (PISA) eingesetzt. Da es wahrscheinlich ist, dass eine wiederholte Vorgabe von Stimuli in Kombination mit jeweils anderen Einzelitems den Testablauf stört und die Übertragbarkeit der Itemparameter von der Kalibrierungsstudie auf die operationale Phase des adaptiven Tests erheblich infrage stellt, wurden Möglichkeiten zur Nutzung kompletter Testlets bei computerisierten adaptiven Tests entwickelt (Frey et al. 2016; Keng 2011).

Itemauswahl für den ersten Schritt

Itemauswahl für den nächsten Schritt

Konvergenz der Schätzungen und Verringerung des Standardfehlers

Alternativen zum Maximum der Iteminformation

Fest verzweigte Tests

Testlets

Multistage-Tests

Weiterhin erfuhren antwortabhängige Tests aus der Anfangszeit des adaptiven Testens, in der noch keine für maßgeschneidertes adaptives Testen hinreichend leistungsfähigen Computer zur Verfügung standen, jüngst eine Renaissance. Es handelt sich dabei um fest verzweigte mehrstufige Tests (sog. „Multistage-Tests“; Yan et al. 2014b) mit zwei bis fünf vorab zusammengestellten Stufen, die heutzutage ebenfalls mithilfe des Computers durchgeführt werden. Das mehrstufige Testen geriet vermutlich deshalb wieder verstärkt in den Blick, da mit dieser Art der Testadministration organisatorische Probleme bei groß angelegten Vergleichsstudien vermeintlich einfacher gelöst werden können als mit voll adaptiven computerisierten Tests (z. B. Berücksichtigung inhaltlicher Einschränkungen bei komplexen Merkmalen, ► Abschn. 20.3.5). Bei mehrstufigen Tests geht man so vor, dass die Testpersonen zu Beginn einen sog. „Routing-Test“ (vgl. Lord 1971, 1980; Yan et al. 2014a) zur Bearbeitung vorgelegt bekommen. Testpersonen, die bei dem Routing-Test gut abschneiden, bekommen im nächsten Schritt einen weiteren Subtest vorgelegt, dessen mittlere Schwierigkeit höher ist als für Testpersonen, die bei dem Routing-Test weniger gut abgeschnitten haben. Dieses Vorgehen wiederholt sich dann auf insgesamt zwei bis fünf Stufen, wobei auf jeder der den Routing-Tests nachfolgenden Stufen zwischen zwei oder mehr Subtests mit unterschiedlicher mittlerer Itemschwierigkeit verzweigt werden kann.

Mehrstufige Tests

Das Grundprinzip des mehrstufigen Testens ist somit dem des computerisierten adaptiven Testens sehr ähnlich, wobei die Anpassung jedoch auf einem größeren Niveau erfolgt. Somit stellen fest verzweigte mehrstufige Tests eine suboptimale Zwischenstufe zwischen herkömmlichem nicht adaptivem und voll adaptivem computerisiertem Testen dar.

20.3.5 Berücksichtigung von Einschränkungen

Nicht statistische Anforderungen

Der Grundgedanke des adaptiven Testens zielt auf die Maximierung diagnostischer Information je Item ab. Dies wird durch die Orientierung der Itemauswahl an einem statistischen Optimalitätskriterium wie dem der maximalen Information umgesetzt. Bei der praktischen Anwendung computerisierter adaptiver Tests sind üblicherweise aber zusätzliche nicht statistische Anforderungen zu erfüllen, die die rein statistisch getriebene Itemauswahl einschränken können. Bei der Selektion von Items ist in einem solchen Fall ein Kompromiss zwischen statistischer Optimalität und der Erfüllung nicht statistischer Anforderungen zu finden.

Constraint-Management-Methoden

Die zur Berücksichtigung nicht statistischer Anforderungen genutzten Methoden werden als Constraint-Management-Methoden bezeichnet. Sie lassen sich in zwei Gruppen differenzieren. Die erste Gruppe von Methoden strebt die generelle Kontrolle der Vorgabehäufigkeit von Items über Personen an (*Exposure Control*). Die zweite Gruppe von Methoden regelt den Umgang mit inhaltlichen Anforderungen für jede einzelne Testperson (*Content-Management*).

Problematik der Vorgabehäufigkeit

Methoden zur generellen Kontrolle von Itemvorgabehäufigkeiten werden nötig, da die Verwendung eines statistischen Optimalitätskriteriums bei der Itemauswahl dazu führen kann, dass einzelne Items oder ganze Reihen von Items sehr vielen Testpersonen, andere Items hingegen sehr wenigen oder im Extremfall keiner Testperson zur Bearbeitung vorgegeben werden. Mit der Häufigkeit der Vorgabe einzelner Items steigt allerdings auch die Wahrscheinlichkeit, dass der Iteminhalt seitens der Testpersonen im Gedächtnis behalten und weiterkommuniziert wird. Dies ist besonders bei Tests zu erwarten, von deren Resultat persönlich relevante Entscheidungen abhängen (z. B. die Zuweisung eines Studienplatzes). Werden Items in der Population potenzieller Testpersonen bekannt, ist es fraglich, ob mit diesen Items noch die intendierten Inhalte gemessen werden. Dies ist deshalb der Fall, weil Antworten auf vorab bekannte und ggf. auswendig gelernte Items nicht mehr eindeutig auf das zu messende Merkmal zurückgeführt werden können, sondern

auch auf andere Merkmale wie Gedächtnisprozesse oder auch die Einbindung in ein soziales Netz, in dem die Iteminhalte bekannt sind. Zudem dürfte erfolgreiches Auswendiglernen eines Items in der Breite zu einem Abfall der Itemschwierigkeit führen.

Unter dem Begriff „Exposure Control“ wurden zur Vermeidung unerwünschter Verteilungen der Vorgabehäufigkeiten verschiedene Strategien entwickelt. Eine besteht darin, dem statistischen Itemauswahlkriterium eine stochastische Komponente hinzuzufügen. So kann alternativ zu der Auswahl des jeweils informativsten Items für die gegenwärtige Merkmalschätzung $\hat{\theta}$ ein Item aus den 5 (8, 10, ...) informativsten Items für die gegenwärtige Merkmalschätzung $\hat{\theta}$ per Zufall ausgewählt werden. Bei einem hinreichend großen Itempool kann mit diesem Ansatz eine unerwünscht ungleichmäßige Vorgabehäufigkeit der Items vermieden werden. Oft kann die Verteilung der Vorgabehäufigkeiten durch das genannte Vorgehen aber nicht exakt genug gesteuert werden, sodass elaboriertere Algorithmen wie die „Sympson-Hetter-Methode“ (Sympson und Hetter 1985), die „Progressive Method“ (Revuelta und Ponsoda 1998) oder die „ α -Stratification“ (Cheng et al. 2009) zum Einsatz gebracht werden. Detaillierte Beschreibungen der Funktionsweise der Methoden findet sich in den genannten Originalarbeiten. Einen Vergleich der Leistungsfähigkeit von Methoden zur Kontrolle von Itemvorgabehäufigkeiten geben Leroux et al. (2013).

Neben der generellen Kontrolle von Itemvorgabehäufigkeiten über Personen ist es oft notwendig oder zumindest wünschenswert, sicherzustellen, welche Arten von Items jeder einzelnen Testperson vorgegeben werden. Methoden zur Realisierung von inhaltlichen Anforderungen auf Individualebene werden als Methoden zum *Content-Management* bezeichnet. Mit ihnen kann beispielsweise kontrolliert werden, wie groß die Anteile vorgegebener Items je Subfacette eines gemessenen Merkmals, je Antwortmodus, je Position der korrekten Antwort bei Multiple-Choice-Items, für Items mit Bild gegenüber Items ohne Bild im Aufgabenstamm etc. ausfallen. In modernen Content-Management-Methoden können mehrere solcher Anforderungen simultan berücksichtigt werden. Der aktuell vermeintlich leistungsstärkste Ansatz ist die von van der Linden und Reese (1998) vorgeschlagene *Shadow-Testing-Methode*. Mit ihr werden sehr gute Resultate auch für eine hohe Anzahl nicht statistischer Einschränkungen erzielt. Die Methode eignet sich vor allem für große Testprogramme, da ihre Umsetzung aufwendig ist, tiefergehende Kenntnisse des linearen Programmierens erfordert und spezielle Solver-Software benötigt wird. Eine Alternative zur Shadow-Testing-Methode stellen andere heuristische Verfahren wie das *Weighted Deviation Model* (Stocking und Swanson 1993), das *Weighted Penalty Model* (Shin et al. 2009) sowie der *Maximum Priority Index* (Cheng und Chang 2009) dar. Aufgrund ihrer einfacheren Implementierbarkeit bei trotzdem guter Leistungsfähigkeit stellen die heuristischen Verfahren für viele Testentwicklungen eine gute Alternative zur Shadow-Testing-Methode dar. Simulationsstudien zum Vergleich heuristischer Methoden zum Umgang inhaltlicher Anforderungen wurden von Born und Frey (2017) sowie He et al. (2014) vorgelegt.

20.3.6 Testende

Ein adaptiver Test wird so lange fortgesetzt, bis ein oder mehrere vorab definierte Abbruchkriterien erfüllt sind. Nachfolgend werden vier häufig verwendete Abbruchkriterien angeführt (vgl. Linacre 2000).

Ein adaptiver Test ist zu beenden, wenn

1. eine bestimmte Anzahl von Items vorgelegt wurde und/oder
2. der Standardfehler der Personenparameterschätzung hinreichend klein ist und/ oder

Exposure Control

Elaborierte Algorithmen

Content-Management

Abbruchkriterien

3. eine maximale Testzeit erreicht wurde oder
4. alle im Itempool verfügbaren Items vorgelegt wurden.

Wahl des Abbruchkriteriums

Segall (2005) weist darauf hin, dass die Wahl des Abbruchkriteriums stark vom jeweiligen Anwendungskontext, der Beschaffenheit des Itempools und einschränkenden Rahmenbedingungen bei der Durchführung des Tests abhängt. Sollen die individuellen Testwerte beispielsweise für interindividuelle Vergleiche (z. B. in der Persönlichkeitspsychologie) oder für individuelle Entscheidungen (z. B. zur Begründung der Zulassung zu einem Studienplatz) verwendet werden, sind Schätzungen individueller Merkmalsausprägungen mit über die Stichprobe vergleichbaren Standardfehlern wünschenswert, wie sie bei Abbruchkriterium 2 resultieren. Um sicherzustellen, dass der gewünschte Standardfehler der geschätzten Personenparameter für alle getesteten Personen erreicht werden kann, ist ein großer Itempool mit hinreichend vielen Items über das gesamte Merkmalskontinuum nötig (vgl. Babcock und Weiss 2012). Werden hingegen Gruppen (z. B. Schulklassen) gemeinsam getestet, dann wird es aufgrund von Rahmenbedingungen oft nicht möglich sein, einen computerisierten adaptiven Test mit flexibler Testlänge durchzuführen. Hier wird die Testung für alle Testpersonen in der Regel nach einer bestimmten Zeit (Kriterium 3) zu beenden sein. Die aus der Verwendung gleicher Testzeiten resultierende interindividuell variierende Präzision der Personenparameterschätzung ist bei solchen Studien aufgrund der üblichen Auswertung auf aggregierter Ebene in der Regel unproblematisch. Da nicht sichergestellt werden muss, dass über das gesamte Merkmalskontinuum genügend Items vorliegen, kann dieses Abbruchkriterium weiterhin gut mit kleinen Itempools oder Itempools mit steilgipfliger Informationsfunktion verwendet werden. In der Praxis wird häufig eine Kombination von Abbruchkriterien genutzt. Eine Simulationsstudie, bei der verschiedene Abbruchkriterien und Kombinationen von Abbruchkriterien in Abhängigkeit der Beschaffenheit des Itempools untersucht wurden, findet sich bei Babcock und Weiss (2012).

20.4 Auswirkungen des adaptiven Testens

Vor- und Nachteile der computerbasierten Testadministration

Da computerisierte adaptive Tests an die Nutzung eines Computers gebunden sind, weisen sie die Vor- und Nachteile auf, die mit einer computerbasierten Testadministration (► Kap. 3) verbunden sind. Solche Effekte gehen auf den Computer als Administrationsmedium und nicht auf den Einsatz adaptiver Testalgorithmen zurück, weshalb sie hier nur in Stichpunkten angeführt werden. Zu den möglichen *Vorteilen* einer computerbasierten Testadministration im Vergleich zu Papier- und Bleistift-Tests gehören eine hohe Testsicherheit, der standardisierte Testablauf, die individuumbestimmte Testgeschwindigkeit, eine schnelle und fehlerfreie Testwertbestimmung, eine Auswertung ohne psychometrisches Fachwissen, die schnelle Ergebnisrückmeldung sowie die Möglichkeit zur Verwendung innovativer Itemformate (z. B. interaktive Items). Die potenziellen *Nachteile* einer computerbasierten Testadministration sind weniger dem konzeptuellen, sondern vielmehr dem organisatorischen Bereich zuzurechnen. Sie bestehen in hohem Entwicklungsaufwand, in Aufwand, der durch die Bereitstellung von Computern am Testort entsteht, in teilweise höheren Kosten und in problematischer Fairness hinsichtlich computerbezogener Persönlichkeitsmerkmale (z. B. Erfahrung mit Computern, Computerängstlichkeit). Eine weiterführende Diskussion der Vor- und Nachteile computerbasierter Testadministration findet sich beispielsweise bei Frey und Hartig (2013).

Darüber hinaus liegen Kenntnisse zu Auswirkungen des computerisierten adaptiven Testens auf die Messeffizienz (► Abschn. 20.4.1), die Validität (► Abschn. 20.4.2) und die Motivation zur Testbearbeitung bei den Testpersonen (► Abschn. 20.4.3) vor.

20.4.1 Messeffizienz

Der Hauptvorteil computerisierter adaptiver Tests im Vergleich zu nicht adaptiven Tests besteht in der Möglichkeit einer beachtlichen Messeffizienzsteigerung. Die Messeffizienz eines Tests ist als Quotient von Messpräzision und Testlänge definiert (Segall 2005). Die Testlänge wird in der Regel durch die Anzahl präsentierter Items quantifiziert. Unter *Messpräzision* wird der Grad der Genauigkeit von Testwerten verstanden. Bei IRT-Modellen kann die Messpräzision als Funktion der zu messenden Merkmalsdimension (θ) variieren und ist durch die Testinformationsfunktion darstellbar (► Kap. 16). Soll für einen Test ein einzelner Wert für die Messeffizienz bestimmt werden, kann vereinfachend für einen konkreten θ -Bereich die mittlere Testinformation als Maß der Messpräzision berechnet und durch die durchschnittliche Anzahl präsentierter Items dividiert werden.

Die Itemauswahl erfolgt beim computerisierten adaptiven Testen häufig nach dem Kriterium maximaler Information (► Abschn. 20.3.4). Hierbei wird also explizit eine Maximierung der Messpräzision und nachfolgend der Messeffizienz angestrebt. In der Testpraxis zeigen sich entsprechend zwei Vorteile des computerisierten adaptiven Testens hinsichtlich der Messpräzision. Erstens ergibt sich im Vergleich zum nicht adaptiven Testen typischerweise eine beachtliche Verringerung der Anzahl vorzugebender Items bei vergleichbarer Messpräzision. Oft werden bei adaptiver Itemvorgabe nur 40–60 % der Items benötigt, um genauso präzise Messungen zu erhalten wie bei nicht adaptiver Itemvorgabe (z. B. Frey und Ehmke 2007; Segall 2005).

Zweitens werden die Standardfehler der Personenparameterschätzungen zwischen Testpersonen angeglichen; dies allerdings nur dann in bestmöglichem Umfang, wenn ein variables Abbruchkriterium, z. B. das Kriterium 2 (► Abschn. 20.3.6), verwendet wird. Inwieweit sich bei einer Anwendung des computerisierten adaptiven Testens tatsächlich für alle getesteten Individuen gleiche Standardfehler ergeben, hängt auch von der Beschaffenheit des Itempools ab. Für eine optimale Angleichung müssen für die individuelle Merkmalsausprägung einer jeden Testperson hinreichend Items vorhanden sein. Das heißt, dass die Informationsfunktion des Itempools über den gesamten Merkmalsbereich, für den Testpersonen zu erwarten sind, eine bestimmte Höhe aufweisen muss. Fällt die Information in einem Bereich unter den Wert, müssen weniger informative Items vorgegeben werden, sodass der Standardfehler dort höher ausfällt.

Messeffizienzsteigerung und Messpräzision

Verringerung der Anzahl vorzugebender Items

Vergleichbare Standardfehler

20.4.2 Validität

Unter Validität versteht man gemäß den Standards for Educational and Psychological Testing (► Kap. 11; AERA et al. 2014) „das Ausmaß, in dem empirische Befunde und theoretische Argumente die Interpretationen von Testwerten für die beabsichtigten Verwendungen von Tests unterstützen“ (► Abschn. 21.1). Bei empirischen Validitätsuntersuchungen wird neben anderen Aspekten üblicherweise erklärt, inwieweit die ermittelten Testergebnisse inhaltlich erwartete statistische Zusammenhänge

- mit auf andere Weise erhobenen Messwerten für das gleiche Merkmal oder mit theoretisch verwandten Merkmalen aufweisen (sog. „konvergente Evidenzen“) sowie
- keinen bzw. nur einen niedrigen statistischen Zusammenhang mit theoretisch abgegrenzten Merkmalen aufweisen (sog. „diskriminante Evidenzen“).

Mit Blick auf den ersten Aspekt sind bei der Anwendung des computerisierten adaptiven Testens höhere statistische Zusammenhänge mit konvergenten Variablen zu erwarten, sofern der Messeffizienzvorteil zur Erhöhung der Messpräzision genutzt

Konvergente Evidenz

wird. Die Erhöhung der Messpräzision ist gleichbedeutend mit dem Anheben des Anteils systematischer, auf das zu messende Merkmal zurückzuführender Varianz der Personenparameterschätzungen in Relation zur zufälligen, unsystematischen Fehlervarianz (vgl. Frey 2006). Der zusätzliche systematische Varianzanteil wird mit konvergenten Variablen kovariieren, während für den unsystematischen Varianzanteil eine Nullkorrelation zu erwarten ist. Wird indes der Messeffizienzvorteil des computerisierten adaptiven Testens zur Verringerung der Itemanzahl bei gleichbleibender Messpräzision genutzt, sind entsprechend keine steigernden Effekte auf konvergente Zusammenhänge zu erwarten.

Während Möglichkeiten zur Steigerung konvergenter Zusammenhänge direkt aus dem bei der adaptiven Itemauswahl genutzten Optimalitätskriterium abgeleitet werden können, sind etwaige Effekte auf diskriminante Variablen etwas schwieriger vorherzusagen. Durch eine Erhöhung des Anteils systematischer Varianz an der Gesamtvarianz der Personenparameterschätzungen sind jedenfalls keine größeren Änderungen bezüglich des Zusammenhangs mit diskriminanten Variablen zu erwarten. Dies ist deshalb der Fall, da bei gegebener Validität sowohl die auf das zu messende Merkmal zurückgehende systematische Varianz als auch die unsystematische Varianz mit diskriminanten Variablen in keinem statistischen Zusammenhang stehen. Eine Verschiebung von Varianzanteilen lässt somit keinen Effekt auf Zusammenhänge mit diskriminanten Variablen erwarten. Naheliegend ist jedoch, dass sich die beim computerisierten adaptiven Testen realisierte spezielle Art des Testens auf psychische Variablen auswirken kann, die dann ggf. mit den Testergebnissen interferieren. Beispielsweise könnte der Sachverhalt, dass beim computerisierten adaptiven Testen im Mittel nur ca. die Hälfte der Aufgaben korrekt beantwortet wird, zumindest für erfolgsgewöhnte Personen zu negativen emotionalen Reaktionen wie Ängstlichkeit, Ärger oder Absenkungen des subjektiven Kontrollempfindens führen. Diese Effekte können sich dann auf die mit dem adaptiven Test erhobenen Merkmale auswirken. Nun besteht aber das Ziel bei Leistungstests üblicherweise in der Messung von Maximalleistungen. Wären gemessene Leistungen durch potenzielle Störvariablen (z. B. Testangst) abgesenkt, würden die gemessenen Werte nicht die maximale Leistung, sondern eine Mischung der interessierenden Maximalleistung und der potenziellen Störvariablen repräsentieren. Hierdurch wäre die Interpretierbarkeit der Testergebnisse eingeschränkt oder im schlimmsten Fall unmöglich. Aus diesem Grund werden als diskriminante Evidenzen meistens Korrelationen nahe null zwischen Leistungswerten und potenziellen Störvariablen angestrebt (für eine weiterführende Diskussion s. Frey 2006).

Empirische Untersuchungen zu den Effekten des computerisierten adaptiven Testens auf diskriminante Validitätsevidenz anhand des Frankfurter Adaptiven Konzentrationsleistungs-Tests (FAKT; Moosbrugger und Goldhammer 2007; Moosbrugger und Heyden 1997) zeichnen ein für computerisiertes adaptives Testen vorteilhaftes Bild. Durch Anwendung der adaptiven FAKT-Testformen konnten im Vergleich zur nicht adaptiven Testform Verzerrungen vermieden werden. Während die untersuchten potenziellen Störvariablen bei adaptiver Testung keinen signifikanten Einfluss auf die Konzentrationsleistung hatten, zeigten sich bei nicht adaptiver Testung in fast allen Fällen signifikante verzerrende Effekte auf die Konzentrationsleistung. Dieses einheitliche Befundmuster ergab sich für die potenziellen Störvariablen Aktivierung, State-Ärger vor der Testung, Veränderung von negativem Affekt während der Testung, Trait-Prüfungsangst vor der Testung und Lärm während der Testung. Diese Befunde können aber nicht ohne Weiteres auf alle computerisierten adaptiven Tests generalisiert werden. So berichten Ortner und Caspers (2011) für ihre Studie an $N = 110$ Studierenden gegenläufige Effekte. Es zeigten sich signifikante Zusammenhänge zwischen dispositioneller Testangst und Intelligenz nur beim computerisierten adaptiven Testen, nicht aber beim nicht adaptiven Testen. Evidenz für die Möglichkeit, diskriminant valide Testwertinterpretationen ableiten zu können, liegt hier also zunächst nur für die nicht adaptive Testform vor. Etwaige verzerrende Zusammenhänge zwischen dispositioneller

Diskriminante Evidenz

Beispiel für Kontrolle von Störvariablen: FAKT

Testangst und Intelligenz konnten aber durch die transparente Beschreibung der Funktionsweise der adaptiven Testversion in der Instruktion vermieden werden. Damit passen die Ergebnisse gut zu den für den FAKT gewonnenen Ergebnissen, da bei diesem die Adaptivität ebenfalls in der Instruktion erläutert wird. Zum jetzigen Zeitpunkt ist somit davon auszugehen, dass bei transparenter Instruktion beim computerisierten adaptiven Testen Zusammenhänge mit diskriminanten Variablen vermieden oder zumindest verringert werden können.

Wichtig ist eine transparente Beschreibung der adaptiven Funktionsweise

20.4.3 Motivation zur Testbearbeitung

Viele Jahre lang galt es als gesichertes Lehrbuchwissen, dass computerisiertes adaptives Testen die Motivation zur Testbearbeitung der Testpersonen steigert. Diese Annahme lässt sich bis zu frühen Arbeiten zum computerisierten adaptiven Testen von Betz (1975) sowie Betz und Weiss (1976a, 1976b) zurückverfolgen. Auf Basis experimenteller Ergebnisse folgerten Betz und Weiss (1976b), dass computerisiertes adaptives Testen vor allem bei leistungsschwächeren Individuen eine Steigerung der Motivation zur Testbearbeitung bewirke. Der Befund wurde damit erklärt, dass die Testpersonen bei adaptiver Testung Items vorgelegt bekämen, die auf ihr individuelles Leistungsniveau abgestimmt seien. Dabei würde die Vorgabe von Items vermieden, die für ein Individuum viel zu leicht seien und damit Langeweile hätten auslösen können oder viel zu schwer seien und damit Frustration hätten auslösen können. Über diese Befunde hinaus liegen allerdings keine fundierten empirischen Ergebnisse vor, die die Annahmen einer motivationssteigernden Wirkung von computerisierten adaptiven Tests stützen.

Pro Motivationssteigerung

Arbeiten aus den 1990er- und 2000er-Jahren stellen die Annahme einer motivationssteigernden Wirkung des computerisierten adaptiven Testens allerdings nachdrücklich infrage. Sie stützen sich auf die Argumentation, dass die bei computerisierten adaptiven Tests realisierte Vorgabe von Items mit mittlerer individueller Lösungswahrscheinlichkeit nicht zu einer hohen Motivation zur Testbearbeitung führen könne (z. B. Bergstrom et al. 1992; Eggen 2004; Ponsoda et al. 1999). Gerade für leistungsfähige Testpersonen stelle ein solcher Test im Gegenteil ein ungewohnt demotivierendes Ereignis dar, da im Mittel nur ca. die Hälfte der vorgelegten Items gelöst werden kann und dies unabhängig von der eigenen Anstrengung. Bei den meisten nicht adaptiven Tests hingegen können leistungsfähige Testpersonen größere Anteile der Items lösen, weshalb eine höhere Motivation zur Testbearbeitung zu vermuten sei. Es wird angenommen, dass derartige Effekte stärker zu Buche schlagen als die oben erwähnten möglichen positiven Effekte auf leistungsschwache Testpersonen. Die Position wird durch die Ergebnisse des Experiments von Frey et al. (2009) unterstützt. Bei diesem zeigte sich, dass die Motivation zur Testbearbeitung bei Vorgabe einer adaptiven Testform des FAKT signifikant niedriger ausfällt als bei Vorgabe der nicht adaptiven Testform.

Kontra Motivationssteigerung

Das von Asseburg (2011) vorgeschlagene, Positionen an einer Stichprobe von $N = 703$ Schülerinnen und Schülern aus neunten Klassen empirisch untersuchte Erwartung-Wert-Modell der Motivation zur Testbearbeitung erlaubt eine Integration der zunächst widersprüchlichen Einzelergebnisse zu den motivationalen Auswirkungen des computerisierten adaptiven Testens. Dies wird vor allem durch die separate Betrachtung von Erwartungs- und Wertkomponenten der Motivation zur Testbearbeitung ermöglicht. Die Befunde der Studie von Asseburg sind vielfältig und münden in differenzierte Empfehlungen zur Gestaltung computerisierter adaptiver Tests für verschiedene Testsituationen und Personengruppen. Ein wichtiger und zudem einfach umzusetzender Punkt ist dabei die Art der Gestaltung der Testinstruktion. Wird die Funktionsweise des Tests in der Instruktion transparent erläutert (bei computerisierten adaptiven Tests z. B. die Nennung, dass auf korrekte Antworten schwierigere Aufgaben folgen und auf inkorrekte Antworten

Integration

leichtere Aufgaben) werden zahlreiche Wechselwirkungen zwischen Adaptivität (adaptiv vs. nicht adaptiv), Persönlichkeitsmerkmalen (z. B. Fähigkeitsselbstkonzept, Selbstwirksamkeitserwartung) und Gruppenzugehörigkeit (Schulart) auf die Motivation zur Testbearbeitung vermieden. Zwischen Personengruppen variierende motivationale Auswirkungen werden somit vermieden. Vor dem Hintergrund, dass 8–12 % der Testleistung durch die Motivation zur Testbearbeitung erklärt werden kann (Asseburg 2011), ist dies zusammen mit den oben berichteten Befunden von Ortner und Caspers (2011) eine wichtige Voraussetzung für die faire und diskriminant valide Interpretation der erhobenen Testwerte.

20.5 Multidimensionales adaptives Testen

Eine vielversprechende Generalisierung des herkömmlich eindimensional angelegten computerisierten adaptiven Testens ist im multidimensionalen adaptiven Testen zu sehen. Der Grundgedanke des multidimensionalen adaptiven Testens entspricht dem seines eindimensionalen Pendants. Während beim eindimensionalen computerisierten adaptiven Testen das Antwortverhalten jedoch auf eine einzelne latente Dimension zurückgeführt wird, werden im multidimensionalen Fall mehrere latente Dimensionen als ursächlich für das beobachtete Antwortverhalten angesehen. Mit multidimensionalen adaptiven Tests können individuelle Ausprägungen mehrerer Merkmale simultan gemessen werden. Hierdurch eröffnen sich neue Möglichkeiten für die psychologische Diagnostik, da komplexe theoretische Annahmen über zu messende mehrdimensionale Merkmalsstrukturen direkt durch das Messinstrument abgebildet werden können. Als psychometrische Modelle kommen beim multidimensionalen adaptiven Testen mehrdimensionale IRT-Modelle (► Kap. 18; z. B. Reckase 2009) zum Einsatz.

Die ersten umfassenden Beschreibungen multidimensionaler adaptiver Algorithmen wurden von Segall (1996), Luecht (1996) und van der Linden (1999b) Ende der 1990er-Jahre vorgelegt.

Auch beim multidimensionalen adaptiven Testen können zur Personenparametterschätzung sowohl ML-Schätzer als auch Bayes-Schätzer verwendet werden. Der von Segall (1996) eingeführte Bayes'sche Ansatz erfuhr in der Literatur bislang die größte Resonanz. Er berücksichtigt, dass die Antworten einer Testperson auf Items, die zur Messung einer Dimension konstruiert wurden, nicht nur Informationen über die individuelle Ausprägung des Probanden auf dieser einen Merkmalsdimension liefern, sondern auch über die Ausprägung auf weiteren Dimensionen, zu denen eine korrelative Beziehung besteht. Der Nutzen von Vorinformationen über die Interkorrelationen der gemessenen Merkmalsdimensionen kann wie folgt verdeutlicht werden: Verfügt beispielsweise ein Schüler über eine hohe mathematische Kompetenz, so kann mit einer bestimmten Wahrscheinlichkeit, wenn auch nicht sicher, darauf geschlossen werden, dass er auch eine hohe naturwissenschaftliche Kompetenz aufweist. Beim Bayes'schen Algorithmus von Segall wird im Rahmen einer maßgeschneiderten Strategie jeweils dasjenige Item als nächstes zur Bearbeitung ausgewählt, das die größte Reduktion im Volumen des Kredibilitätsellipsoids (mehrdimensionales Bayes'sches Pendant eines Konfidenzintervalls) des geschätzten Merkmalsvektors $\hat{\theta}_v = (\hat{\theta}_{v1}, \dots, \hat{\theta}_{vP})$ von Testperson v auf den P zu messenden Dimensionen bewirkt. Es wird also das Item ausgewählt, dessen Vorgabe die größte Steigerung der Messpräzision hinsichtlich aller zu messenden Dimensionen liefert.

In mehreren Simulationsstudien zeigte sich, dass die Nutzung von Vorinformationen über Zusammenhänge zwischen den zu messenden Dimensionen im Rahmen des multidimensionalen adaptiven Testens signifikante Steigerungen der Messeffizienz im Vergleich zu mehreren eindimensionalen adaptiven Tests bewirkt (zusammenfassend s. Frey und Seitz 2009 sowie Frey et al. 2017). Die

Simultane Messung mehrerer Merkmale

Multidimensionale adaptive Algorithmen

Nutzung von Vorinformationen

Steigerung der Messeffizienz

Messeffizienzsteigerung fällt dabei umso größer aus, je stärker der korrelative Zusammenhang zwischen den gemessenen Dimensionen ist (Frey und Seitz 2010). Bei der Einordnung der Ergebnisse reiner Simulationsstudien ist jedoch zu beachten, dass bei diesen in der Regel Itempools verwendet wurden, die sehr gut für adaptives Testen geeignet sind, und dass die Itemauswahl ohne Einschränkungen erfolgen konnte. In der Praxis werden jedoch nicht immer optimale Itempools zur Verfügung stehen und Einschränkungen bei der Itemauswahl zu berücksichtigen sein. Weiterführende Simulationsstudien auf Basis empirischer Daten, sog. „Echtdatensimulationen“, zeigten, dass auch bei Verwendung von Itempools, die für nicht adaptive Tests entwickelt wurden, durchaus eine sehr hohe Messeffizienz mit multidimensionalen adaptiven Tests erzielt werden kann (z. B. Frey und Seitz 2011). Die sehr hohe Messeffizienz kann dabei genutzt werden, um deutlich differenziertere Messergebnisse zu generieren als mit herkömmlichen Verfahren. Mikolajetz und Frey (2016) konnten diesbezüglich zeigen, wie durch den Einsatz von multidimensionalen adaptiven Tests elf theoretisch begründete Subdimensionen mathematischer Kompetenz mit hinreichender Präzision gemessen werden können. Die bisherige Art der Messung und Berichtlegung auf Basis konventioneller Papier-und-Bleistift-Tests musste sich auf eine eindimensionale Skala (als hauptsächliche Berichtsskala) sowie eine fünfdimensionale Auswertung beschränken. Ähnliche Befunde berichten Frey et al. (2013) für PISA. Auch hier konnte die Anzahl der Skalen, auf denen hinreichend präzise Messwerte erhoben werden können, durch multidimensionales adaptives Testen deutlich gesteigert werden.

Die grundlegende Forschung zum multidimensionalen adaptiven Testen ist als weitgehend abgeschlossen anzusehen. Für die kommenden Jahre ist mit den ersten größeren Anwendungen dieser hocheffizienten Art des Testens zu rechnen.

20.6 Zusammenfassung und Anwendungsempfehlungen

Computerisiertes adaptives Testen ist ein spezielles Vorgehen zur computerbasierten Messung individueller Merkmalsausprägungen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am vorherigen Antwortverhalten der Testperson orientiert. Der Grundgedanke besteht darin, keine starre Abfolge von Items vorzugeben, sondern nur solche Items, die möglichst viel diagnostische Information über die individuelle Ausprägung des zu messenden Merkmals liefern. Dieses Anliegen wird durch die Spezifikation von sechs elementaren Bausteinen umgesetzt. Es handelt sich dabei um den Itempool, die Art den Test zu beginnen, die Schätzung der individuellen Merkmalsausprägung, die Itemauswahl, die Berücksichtigung nicht statistischer Einschränkungen und die Art, den Test zu beenden. Für alle Bausteine liegen mehrere Optionen vor, die je nach Anforderung der Testsituation in bestmöglicher Weise miteinander kombiniert werden können. Der Hauptvorteil des computerisierten adaptiven Testens im Vergleich zum nicht adaptiven Testen besteht in einer Messeffizienzsteigerung, die in den meisten Fällen beträchtlich ausfällt. Darüber hinaus sind positive Auswirkungen auf die Validität der adaptiv erhobenen Testergebnisse zu verzeichnen. Um unerwünschte Effekte beim computerisierten adaptiven Testen zu vermeiden, sollte die Funktionsweise eines adaptiven Tests im Rahmen der Instruktion transparent erläutert werden. Die Konstruktion eines computerisierten adaptiven Tests ist aufwendig. Neben der Erstellung und Kalibrierung eines geeigneten Itempools, sind präoperationale Simulationsstudien durchzuführen, sodass ein dem Gegenstand und Einsatzbereich angemessener adaptiver Algorithmus spezifiziert werden kann.

20.7 EDV-Hinweise

Mit dem kommerziell vertriebenen Computerprogramm FastTest Pro (► <http://www.fasttestweb.com>) können Itempools verwaltet, adaptive Algorithmen spezifiziert, Tests wahlweise lokal oder über das Internet administriert und ausgewertet werden. Über diese relativ kostspielige Lösung hinaus existieren mehrere für Forschung und Lehre kostenfrei nutzbare Open-Source-Programme zum computerisierten adaptiven Testen. Ein gut dokumentiertes, webbasiertes System mit umfangreicher Funktionalität, verschiedenen Arten zur Testadministration, zahlreichen Beispielen und der Möglichkeit, Workshops zu besuchen, stellt die Plattform Concerto der University of Cambridge dar (► <https://www.psychometrics.cam.ac.uk/newconcerto>). In Chalmers (2016) wird weiterhin beschrieben, wie mit dem R-Package „mirtCAT“ (Chalmers 2017) in Kombination mit dem R-Package „shiny“ (Chang 2017) computerisierte adaptive Tests konfiguriert und HTML-basiert vorgegeben werden können. Dieser Ansatz wurde bei der R-basierten App KAT-HS (Fink et al. 2019) genutzt und um eine grafische Benutzeroberfläche und aktuelle Entwicklungen wie die Online-Kalibrierung (Fink et al. 2018) ergänzt. Nützlich für die Durchführung von CAT-Simulationsstudien sind die R-Packages „mirtCAT“, „catR“ (Magis et al. 2016) und „MAT“ (Choi und King 2014).

20.8 Kontrollfragen

- ?(?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Warum ist die Verwendung der klassischen Testtheorie (KTT) beim computerisierten adaptiven Testen nicht angezeigt?
 2. Welches sind die sechs elementaren Bausteine des computerisierten adaptiven Testens?
 3. Durch computerisiertes adaptives Testen können substanzelle Steigerungen der Messeffizienz im Vergleich zu nicht adaptiven Tests erzielt werden. Für welche beiden Verbesserungen kann diese Messeffizienzsteigerung genutzt werden?
 4. Im Bereich der klinischen Psychologie soll ein Test entwickelt werden, mit dem die Ängstlichkeit sowohl von Gesunden als auch von pathologisch Ängstlichen gemessen werden soll. Warum ist in diesem Fall computerisiertes adaptives Testen als vorteilhaft anzusehen?
 5. Wie sollte der Itempool eines computerisierten adaptiven Tests beschaffen sein, damit dieser über die gesamte Breite der zu messenden Merkmalsdimension in gleicher Weise differenzierungsfähig ist?
 6. Welches Kriterium zur Itemauswahl wird beim computerisierten adaptiven Testen am häufigsten genutzt?
 7. Durch welche Maßnahme können unerwünschte Auswirkungen des computerisierten adaptiven Testens auf die individuelle Motivation zur Testbearbeitung minimiert werden?
 8. Gegeben sei ein Konstrukt, das auf theoretischer Ebene durch sieben korrelierende Subdimensionen spezifiziert wird. Die siebendimensionale Struktur ist aufgrund bereits vorliegender empirischer Ergebnisse als gesichert anzusehen. Zur Messung des Konstruktts soll nun ein neues Testverfahren konstruiert werden. Warum bietet sich in diesem Fall multidimensionales adaptives Testen an?

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 Psychological Association
- Asseburg, R. (2011). *Leistungsbereitschaft in Testsituationen. Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests*. Marburg: Tectum.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.
- Babcock, B. & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, 1, 1–18.
- Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, 5, 137–149.
- Betz, N. E. (1975). New types of information and psychological implications. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and Prospects (Research Report 75-5)* (pp. 32–43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Betz, N. E. & Weiss, D. J. (1976a). *Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Betz, N. E. & Weiss, D. J. (1976b). *Psychological effect of immediate knowledge of results and adaptive ability testing (Research Report 76-4)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.
- Born, S. & Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educational and Psychological Measurement*, 77, 241–262.
- Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71, 1–38. <https://doi.org/10.18637/jss.v071.i05>
- Chalmers, P. (2017). *mirtCAT: Computerized adaptive testing with multidimensional item response theory. R package. Version 1.3*. Retrieved from <https://CRAN.R-project.org/package=mirtCAT> [29.12.2019]
- Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1–20.
- Chang, W. (2017). shiny: web application framework for R. *R package. Version 1.0.0*. Retrieved from <https://cran.r-project.org/web/packages/shiny/index.html> [29.12.2019]
- Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, 24, 257–265.
- Cheng, Y. & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369–383.
- Cheng, Y., Chang, H. H., Douglas, J. & Guo, F. (2009). Constraint-weighted α -stratification for computerized adaptive testing with nonstatistical constraints balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, 69, 35–49.
- Choi, S. W. & King, D. (2014). MAT: Multidimensional adaptive testing. *R package. Version 2.2*. Retrieved from <https://rdrr.io/cran/MAT/> [29.12.2019]
- Diao, Q. (2010). *Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing*. Ann Arbor, MI: UMI Research Press.
- Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Enschede: Print Partners Ipskamp.
- Fink, A., Born, S., Frey, A. & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60, 327–346.
- Fink, A., Spoden, C., Kroll, P. & Frey, A. (2019). *KAT-HS-App Benutzerhandbuch*. Frankfurt am Main: Johann-Wolfgang Goethe-Universität Frankfurt.
- Frey, A. (2006). *Validitätssteigerungen durch adaptives Testen*. Frankfurt am Main: Peter Lang.
- Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 169–184.
- Frey, A. & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen Anstelle von Papier- und Bleistift-basierten Verfahren eingesetzt werden? *Zeitschrift für Erziehungswissenschaft*, 16, 53–57.
- Frey, A. & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, 35, 89–94.
- Frey, A. & Seitz, N. N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. *Zeitschrift für Pädagogik, Beiheft 56*, 40–51.
- Frey, A. & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*, 71, 503–522.

- Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung. *Diagnostica*, 55, 20–28.
- Frey, A., Hartig, J. & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28, 39–53.
- Frey, A., Seitz, N. N. & Kroehne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.), *Research on PISA* (pp. 103–120). Dordrecht: Springer.
- Frey, A., Seitz, N. N. & Brandt, S. (2016). Testlet-based multidimensional adaptive testing. *Frontiers in Psychology*, 7, 1–14.
- Frey, A., Kröhne, U., Seitz, N. N. & Born, S. (2017). Multidimensional adaptive measurement of competences. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Eds.), *Competence assessment in education. Research, models, and instruments*. Cham: Springer.
- Hambleton, R. K., Zaal, J. N. & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing. Theory and applications* (pp. 341–366). New York, NY, US: Kluwer Academic/Plenum Publishers.
- He, W. & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*, 74, 473–494.
- He, W., Diao, Q. & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, 74, 677–696.
- Keng, L. (2011). *A Comparison of the performance of testlet-based computer adaptive tests and multistage tests*. Ann Arbor, MI: Proquest.
- Khorramdel, L. & von Davier, M. (2016). Item response theory as a framework for test construction. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction*. Göttingen: Hogrefe.
- Kubinger, K. D. (2009). *Adaptives Intelligenz Diagnostikum – Version 2.2 (AID 2) samt AID 2-Türkisch*. Göttingen: Beltz.
- Leroux, A. J., Lopez, M., Hembry, I. & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, 73, 857–874.
- Linacre, J. M. (2000). Computer-Adaptive Testing: A methodology whose time has come. In C. Sun-hee, K. Unson, J. Eunhwa and J. M. Linacre (Eds.), *Development of computerized middle school achievement test*. Seoul, South Korea: Komesa Press.
- Lord, F. M. (1971). The self-scoring flexilevel test. *Educational and Psychological Measurement*, 8, 147–151.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389–404.
- Magis, D., Raiche, G. & Barrada, J. R. (2016). *catR: Generation of IRT response patterns under computerized adaptive testing. R package. Version 3.11*. Retrieved from <https://drri.io/cran/catR/> [29.12.2019]
- Mikolajetz, A. & Frey, A. (2016). Differentiated assessment of mathematical competence with multidimensional adaptive testing. *Psychological Test and Assessment Modeling*, 58, 617–639.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195.
- Moosbrugger, H. & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test*. Bern: Huber.
- Moosbrugger, H. & Heyden, M. (1997). *Frankfurter Adaptiver Konzentrationsleistungs-Test*. Bern: Huber.
- Ortner, T. M. & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, 27, 157–163.
- Ponsoda, V., Olea, J., Rodriguez, M. S. & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education*, 12, 167–184.
- Reckase, M. D. (2009). *Multidimensional item response theory*. Dordrecht: Springer.
- Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311–327.
- Sands, W. A., Waters, B. K. & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331–354.
- Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. Amsterdam: Elsevier.
- Shin, C. D., Chien, Y., Way, W. D. & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. San Antonio, TX: Pearson.
- Spoden, C., Frey, A. & Bernhardt, R. (2018). Implementing three CATs within eighteen months. *Journal of Computerized Adaptive Testing*, 60, 38–55.

Literatur

- Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277–292.
- Sympson, J. B. & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In Navy Personnel Research and Development Center (Ed.), *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.
- van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, 63, 201–216.
- van der Linden, W. J. (1999a). A procedure for empirical initialization of the trait estimator on ability estimates. *Applied Psychological Measurement*, 23, 21–29.
- van der Linden, W. J. (1999b). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 28, 398–412.
- van der Linden, W. J. (Ed.). (2016a). *Handbook of item response theory. Volume 1: Models*. Boca Raton: Chapman & Hall/CRC.
- van der Linden, W. J. (Ed.). (2016b). *Handbook of item response theory. Volume 2: Statistical tools*. Boca Raton: Chapman & Hall/CRC.
- van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In van der Linden, W. J. & Glas, C. A. W. (Eds.), *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer.
- van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, 54, 427–450.
- Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37, 70–84.
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2, 1–27.
- Yan, D., Lewis, C. & von Davier, A. A. (2014a). Overview of computerized multistage tests. In D. Yan, A. A. von Davier & C. Lewis, C. (Eds.). *Computerized multistage testing: Theory and applications* (pp. 3–20). Boca Raton: Chapman & Hall/CRC.
- Yan, D., von Davier, A. A. & Lewis, C. (Eds.). (2014b). *Computerized multistage testing: Theory and applications*. Boca Raton: Chapman & Hall/CRC.
- Ziegler, B., Frey, A., Seeber, S., Balkenhol, A. & Bernhardt, R. (2016). Adaptive Messung allgemeiner Kompetenzen (MaK-adapt). In K. Beck, M. Landenberger & F. Oser (Hrsg.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT* (S. 33–54). Bielefeld: wbv.

Validität und Möglichkeiten ihrer Überprüfung

Inhaltsverzeichnis

- Kapitel 21 Validität von Testwertinterpretationen – 529**
Johannes Hartig, Andreas Frey und Nina Jude
- Kapitel 22 Latent-Class-Analyse (LCA) – 547**
Mario Gollwitzer
- Kapitel 23 Exploratorische Faktorenanalyse (EFA) – 575**
Holger Brandt
- Kapitel 24 Konfirmatorische Faktorenanalyse (CFA) – 615**
Jana C. Gäde, Karin Schermelleh-Engel und Holger Brandt
- Kapitel 25 Multitrait-Multimethod-Analysen
(MTMM-Analysen) – 661**
Karin Schermelleh-Engel, Christian Geiser und G. Leonard Burns
- Kapitel 26 Latent-State-Trait-Theorie (LST-Theorie) – 687**
Augustin Kelava, Karin Schermelleh-Engel und Axel Mayer
- Kapitel 27 Konvergente und diskriminante Validität über die Zeit:
Integration von Multitrait-Multimethod-Modellen
(MTMM-Modellen) und der Latent-State-Trait-Theorie
(LST-Theorie) – 713**
*Fridtjof W. Nussbeck, Michael Eid, Christian Geiser,
Delphine S. Courvoisier und David A. Cole*

Validität von Testwertinterpretationen

Johannes Hartig, Andreas Frey und Nina Jude

Inhaltsverzeichnis

- 21.1 Einleitung – 530**
- 21.2 Validität im fachgeschichtlichen Wandel – 530**
 - 21.2.1 Unterscheidung von Validitätsarten – 531
 - 21.2.2 Konstruktvalidität und nomologische Netze – 532
 - 21.2.3 Validität von Testwertinterpretationen – 534
- 21.3 Argumentationsbasierter Ansatz der Validierung – 535**
 - 21.3.1 Spezifikation der Testwertinterpretation – 535
 - 21.3.2 Formulierung von prüfbaren Grundannahmen – 536
 - 21.3.3 Sammlung empirischer Evidenz – 538
 - 21.3.4 Zusammenfassende Bewertung des Validitätsarguments – 538
- 21.4 Beispiele für Validierungsprozesse – 539**
 - 21.4.1 Beispiel 1: Interpretation von Testwerten als Indikatoren für ein theoretisches Konstrukt – 539
 - 21.4.2 Beispiel 2: Verwendung von Testwerten bei der Auswahlagnostik – 540
 - 21.4.3 Beispiel 3: Interpretation eines Testwertes als Screening in der klinisch-psychologischen Diagnostik – 541
 - 21.4.4 Beispiel 4: Interpretation eines Testwertes bezogen auf das Erreichen von Lernzielen – 542
- 21.5 Zusammenfassung – 544**
- 21.6 Kontrollfragen – 544**
- Literatur – 544**

i Das Gütekriterium der Validität bezieht sich darauf, inwieweit Interpretationen von Testwerten im Hinblick auf die beabsichtigten Verwendungen von Tests gerechtfertigt sind. In diesem Kapitel wird das aktuelle Verständnis von Validität und ihrer Überprüfung beschrieben und an mehreren Beispielen illustriert. Hierbei wird Validität nicht als eine Eigenschaft eines Testverfahrens betrachtet, sondern als Qualitätskriterium hinsichtlich der Zulässigkeit von Testwertinterpretationen. Zur Absicherung der angestrebten Testwertinterpretation ist a) die Formulierung von Grundannahmen erforderlich, die für die angestrebte Interpretation erfüllt sein müssen, b) die Sammlung empirischer Evidenz, mit der die Grundannahmen gestützt oder widerlegt werden können um c) eine abschließende integrierende Beurteilung der Angemessenheit der intendierten Testwertinterpretation vorzunehmen. Der Abschluss eines Validierungsprozesses hat immer vorläufigen Charakter, da zukünftige Befunde einzelne Grundannahmen und damit die Belastbarkeit der Testwertinterpretationen infrage stellen können.

21.1 Einleitung

Was ist Validität?

Das Gütekriterium der Validität (engl. „validity“ = Gültigkeit) wird häufig zusammengefasst als das Ausmaß, in dem „ein Test misst, was er messen soll“. Schon diese vereinfachende Zusammenfassung drückt aus, dass Validität ein umfassendes und sehr wichtiges Gütekriterium zur Beurteilung eines diagnostischen Verfahrens darstellt. Genauer genommen bezieht sich Validität darauf, inwieweit spezifische *Interpretationen von Testwerten für die beabsichtigten Verwendungen* eines Tests gerechtfertigt sind. Validität ist den Gütekriterien der Objektivität und Reliabilität übergeordnet: Wenn die aus einem Test resultierenden Werte nicht so interpretiert werden können, wie es den Zielsetzungen des Tests entspricht, sind weder die Objektivität noch die Reliabilität von Belang. Validität ist das komplexeste der drei Gütekriterien, da die Prüfung der Validität einen Forschungsprozess umfasst, dessen empirische und theoretische Einzelbefunde zu einer integrierenden Bewertung zusammenzufassen sind.

Definition

Validität ist das Ausmaß, in dem empirische Befunde und theoretische Argumente die Interpretationen von Testwerten für die beabsichtigten Verwendungen von Tests unterstützen (AERA et al. 2014, S. 11, Übersetzung der Autoren).

Überblick

In diesem Kapitel wird zunächst der Begriff der Validität im fachgeschichtlichen Wandel beschrieben (► Abschn. 21.2), da sich die genaue Definition dieses Gütekriteriums in den letzten Jahrzehnten deutlich verändert hat. Anschließend wird der sog. argumentationsbasierte Ansatz der Validierung dargestellt, der das aktuelle Verständnis von Validität maßgeblich prägt (► Abschn. 21.3). Das schrittweise Vorgehen der argumentationsbasierten Validierung wird detailliert erläutert. Anhand von vier Beispielen werden sodann Validierungsprozesse für unterschiedliche Anwendungsfelder der psychologischen Diagnostik dargestellt (► Abschn. 21.4).

21.2 Validität im fachgeschichtlichen Wandel

Veränderungen des Validitätskonzepts

Das Konzept der Validität ist aufgrund seiner Komplexität und der zentralen Bedeutung für die psychologische Diagnostik seit seiner erstmaligen Verwendung zu Beginn des 20. Jahrhunderts stetig weiterentwickelt worden (s. z. B. Frey 2006; Kane 2001). In den letzten Jahrzehnten sind zwei zentrale Entwicklungen zu verzeichnen: Die erste besteht darin, dass seit den 1950er-Jahren zunächst immer

mehr „Validitätsarten“ unterschieden wurden (► Abschn. 21.2.1). Diese Unterscheidung steht im aktuellen Verständnis von Validität nicht mehr im Vordergrund, vielmehr wird Validität als ein *zusammenfassendes Gütekriterium* betrachtet, das verschiedene Befunde bezüglich eines Tests integriert. Diese Entwicklung nahm mit den Überlegungen von Cronbach und Meehl (1955) zur Konstruktvalidität ihren Anfang (► Abschn. 21.2.2). Die zweite Entwicklung ist darin zu sehen, dass Validität früher als „Eigenschaft eines Tests“ betrachtet wurde, während sich das aktuelle Verständnis von Validität auf die Verwendung und die Belastbarkeit von *Testwertinterpretationen* (und damit auf den gesamten diagnostischen Prozess) bezieht (► Abschn. 21.2.3).

21.2.1 Unterscheidung von Validitätsarten

Die verschiedenen im Laufe der Fachgeschichte entwickelten Validitätsarten bezogen sich auf unterschiedliche Testwertinterpretationen und wurden mit unterschiedlichen empirischen Vorgehensweisen untersucht. Im Verlauf der Entwicklung des Validitätskonzepts wurden u.a. die folgenden „Validitätsarten“ unterschieden (vgl. dazu auch ► Kap. 2):

- *Kriteriumsvalidität* wurde über den korrelativen Zusammenhang des Testwertes mit einem „Außenkriterium“ geprüft, im Idealfall mit einem direkten Maß des mit dem Test zu messenden Merkmals. Die Kriteriumsvalidität ist der fachgeschichtlich älteste Aspekt der Validität. Die Korrelation eines Testwertes mit dem Außenkriterium wurde zeitweise auch als „Validitätskoeffizient“ bezeichnet (z. B. Taylor und Russell 1939), womit die Korrelation mit der Validität als solcher gleichgesetzt wurde.
- Bei zeitgleicher Erhebung von Testwert und Kriterium wurde der korrelative Zusammenhang auch als *Übereinstimmungsvalidität* bezeichnet.
- Die Vorhersage eines Kriteriums mithilfe einer zeitlich vorgeordneten Testung wurde als *prognostische Validität* bezeichnet.
- Das Ausmaß, in dem ein Test die Vorhersage eines Kriteriums verbessert, wenn er zusätzlich zu anderen Verfahren in die Vorhersage aufgenommen wird, wurde als *inkrementelle Validität* bezeichnet.
- *Inhaltsvalidität* wurde durch eine zumeist durch Experten vorgenommene Prüfung der Passung zwischen Testinhalten und der Konstruktdefinition untersucht.
- Der Begriff der *Augenscheininvalidität* („face validity“) bezog sich auf eine unmittelbar offensichtliche, z. B. auch für Laien ohne weitere Prüfung ersichtliche Gültigkeit eines Testwertes; er war in seiner genauen Definition jedoch unscharf (Mosier 1947).
- Das Konzept der *Konstruktvalidität* bezog sich darauf, dass ein Test ein zugrunde liegendes theoretisches Konstrukt messen soll.
- Mit *konvergenter Validität* wurde der korrelativ bestimmte Zusammenhang eines Testwertes mit Tests bezeichnet, die dasselbe oder ein stark verwandtes Konstrukt erfassen.
- Als *diskriminante Validität* wurden im Gegensatz dazu niedrige korrelative Zusammenhänge eines Testwertes mit anderen Tests bezeichnet, die andere Konstrukte erfassen sollen.
- *Faktorielle Validität* bezeichnete Befunde faktorenanalytischer Verfahren, mit denen die erwartete dimensionale Struktur eines Tests bestätigt wurde.

Mittlerweile wird von einer Differenzierung in verschiedene Validitätsarten abgesehen. Die Befunde, die früher unter Bezug auf die einzelnen Validitätsarten gesammelt wurden, sind weiterhin von Bedeutung. Dem aktuellen Verständnis von

Validität zufolge werden sie jedoch in einem gemeinsamen Qualitätsurteil zusammengefasst.

21.2.2 Konstruktvalidität und nomologische Netze

Integration verschiedener Validitätsansätze

Hypothetisch-deduktiver Ansatz

Nomologisches Netz

Die große Vielfalt von Validitätsarten hat in der Praxis zu einer gewissen Beliebigkeit geführt, welche der verschiedenen Validitäten herangezogen wurden. Während insbesondere Kriteriumsvalidität und Inhaltsvalidität zunächst noch als separate Alternativen behandelt wurden, stellte das Konzept der Konstruktvalidität in den 1950er- bis 1970er-Jahren einen ersten Versuch dar, Validität als ein einheitliches Gütekriterium zu betrachten und verschiedene Vorgehensweisen zur Untersuchung der Validität zusammenzuführen.

Eine zentrale Arbeit zur Konstruktvalidität wurde von Cronbach und Meehl (1955) vorgelegt. In dieser wird eine Idealvorstellung der Prüfung der Konstruktvalidität konzeptualisiert. Das Vorgehen basiert auf dem in den empirisch arbeitenden Sozialwissenschaften verwendeten *hypothetisch-deduktiven Ansatz* der Erkenntnisgewinnung und unterscheidet zwischen einem Bereich der Theorie und einem Bereich der Beobachtung. Im Bereich der Theorie werden nicht direkt beobachtbare theoretische Konstrukte und die auf theoretischer Ebene bestehenden Zusammenhänge zwischen diesen Konstrukten definiert. Idealerweise sollten die theoretischen Zusammenhänge durch sog. „Axiome“ formalisiert sein. Das Herzstück einer Theorie besteht hierbei aus einem Satz von Axiomen, die theoretische Interdependenzen zwischen Konstrukten mathematisch beschreiben. Die semantische Interpretation der theoretischen Axiome erfolgt durch Verbindungen einiger oder aller Terme der Axiome mit beobachtbaren Variablen. Die Verbindungen werden *Korrespondenzregeln* genannt. Sie verknüpfen den Bereich der Theorie mit dem Bereich der Beobachtung. Aufgrund der theoretischen Axiome über die Zusammenhänge von Konstrukten lassen sich entsprechende Vorhersagen für den Bereich der Beobachtung ableiten, die anhand von beobachtbaren Variablen empirisch überprüft werden können. Diese vorhergesagten Zusammenhänge werden als *empirische Gesetze* bezeichnet.

Eine Theorie besteht im Ansatz von Cronbach und Meehl (1955) aus einem sog. „nomologischen Netz“, das aus einem durch Korrespondenzregeln interpretierten axiomatischen System und allen daraus abgeleiteten empirischen Gesetzen besteht. Das nomologische Netz umspannt folglich Elemente des Bereichs der Theorie und des Bereichs der Beobachtung. In Abb. 21.1 werden die wesentlichen Elemente eines nomologischen Netzes schematisch dargestellt.

Das Ziel der Prüfung der Konstruktvalidität im Sinne von Cronbach und Meehl (1955) besteht darin, die Korrektheit des nomologischen Netzes schrittweise zu überprüfen. Hierbei wird untersucht, ob beobachtete Testwerte zulässige Indika-

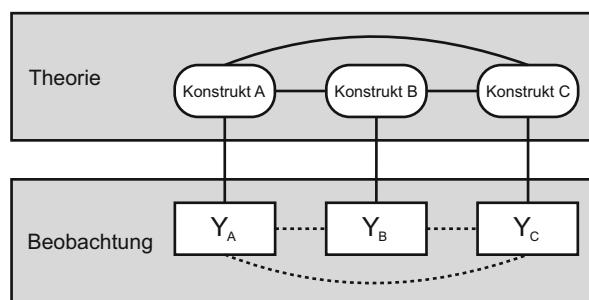


Abb. 21.1 Schematische Darstellung eines nomologischen Netzes von drei Konstrukten A, B und C mit zugehörigen manifesten Messwerten Y_A , Y_B und Y_C sowie punktierten Korrespondenzregeln

toren für die individuellen Ausprägungen von nicht direkt beobachtbaren Konstrukten darstellen. Hierzu werden die empirischen Gesetze mittels empirischer Beobachtungen geprüft. Stimmen theoretische Vorhersagen und empirische Beobachtungen überein, dann bedeutet dies eine Bestätigung sowohl der Theorie als auch der Interpretation der Testwerte als individuelle Ausprägungen in dem theoretischen Konstrukt. Stimmen die vorhergesagten Zusammenhänge empirisch jedoch nicht mit der Theorie überein, dann sind Teile des nomologischen Netzes zu verwerfen oder in modifizierter Form einer erneuten empirischen Prüfung zu unterziehen. Inkonsistenzen zwischen Theorie und Beobachtung können auf Fehler in den Axiomen, den abgeleiteten Korrespondenzregeln oder dem verwendeten Testverfahren zurückgehen. Da sich der Ansatz der Konstruktvalidität des hypothetisch-deduktiven Ansatzes der Erkenntnisgewinnung bedient, kann die Validität einer konstruktbezogenen Testwertinterpretation prinzipiell nie endgültig belegt oder gar bewiesen werden. Vielmehr ist die Annahme, dass ein Testergebnis auf ein bestimmtes Konstrukt zurückzuführen ist, genau so lange gerechtfertigt, bis sie falsifiziert wird.

In ▶ Beispiel 21.1 werden die konkreten Zusammenhänge in einem nomologischen Netz exemplarisch dargestellt.

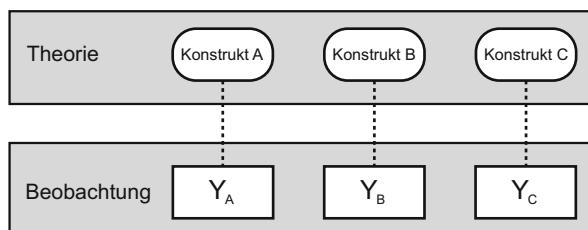
Übereinstimmung von theoretischen Vorhersagen und empirischen Beobachtungen

Beispiel 21.1: Nomologisches Netzwerk

Ein Forscher hat eine Theorie über ein Konstrukt A entworfen. Seine Theorie nimmt an, dass A mit zwei bereits etablierten Konstrukten B und C unterschiedliche Zusammenhänge aufweist. Während ein Einfluss von A auf B angenommen wird, soll C von A und B unabhängig sein. Aufgrund dieser theoretischen Annahmen lassen sich drei empirische Gesetze ableiten (Abb. 21.2):

1. B ist abhängig von A.
2. C ist unabhängig von A.
3. C ist unabhängig von B.

Während für B und C geeignete Testverfahren vorliegen, findet sich keines für A, sodass der Forscher einen neuen Test zur Messung von A entwickelt. Zur Untersuchung, ob die Interpretation der mit dem neuen Test erhobenen Testwerte als Maß von A konstrukt valide ist, sind Daten zu erheben und die empirischen Gesetze 1 und 2 zu überprüfen. Stimmen die beobachteten Zusammenhänge mit den theoretisch vorhergesagten überein, dann können mit dem Test erhobene Testwerte konstrukt valide als A interpretiert werden. Diese Interpretation gilt so lange, bis sie aufgrund von neuen empirischen Ergebnissen zu verwerfen ist. In der Forschungspraxis können auch größere nomologische Netze vorliegen, sodass bei der Konstruktvalidierung entsprechend mehr empirische Gesetze untersucht werden müssen.



■ Abb. 21.2 Schematische Darstellung eines nomologischen Netzes von drei als Ellipsen dargestellten Konstrukten A, B, und C mit zugehörigen manifesten Messwerten Y_A , Y_B und Y_C

Problem: Korrelationen ohne verbindende theoretische Argumentation sind beliebig (Konstruktvalidität als Mülleimerkategorie)

Als Problem des von Cronbach und Meehl (1955) dargelegten Vorgehens zur Prüfung der Konstruktvalidität erwies sich allerdings, dass psychologische Theorien in vielen Bereichen zu schwach entwickelt sind, um empirisch überprüfbare Hypothesen über ein Konstrukt ableiten zu können (s. Cronbach und Meehl 1955). Infolgedessen wurde die Konstruktvalidität, wie Cronbach (1980) beklagte, oft als eine „Mülleimerkategorie“ verwendet: Beliebige Korrelationen eines Testwertes mit anderen Variablen wurden als Belege für Konstruktvalidität bezeichnet, ohne dass eine verbindende theoretische Argumentation formuliert wurde. Grundsätzlich ist die von Cronbach und Meehl (1955) vorgeschlagene Einbettung der Untersuchung der Validität von Testwertinterpretationen noch heute als angemessen anzusehen.

21.2.3 Validität von Testwertinterpretationen

Die zweite zentrale Entwicklung in der Geschichte des Validitätskonzepts ist, dass Validität über lange Zeit als *Eigenschaft eines Tests* betrachtet wurde. Diese Sichtweise wurde im angelsächsischen Raum beginnend mit den 1980er-Jahren aufgegeben, da jeder Test auf unterschiedliche Weise interpretiert und verwendet werden kann und es nicht möglich ist, alle denkbaren Interpretationen und Verwendungen zu validieren (► Exkurs 21.1). Die Verwendung des undifferenzierten Begriffs der „Validität eines Tests“ wird mittlerweile ausdrücklich als falsch betrachtet.

- » It is incorrect to use the unqualified phrase ‘the validity of the test’. (AERA et al. 2014, S. 11)

Interpretationen und Verwendungen von Testwerten

Validität bezieht sich hingegen auf spezifische, explizit zu definierende *Interpretationen und Verwendungen von Testwerten*. Jede unterschiedliche Testwertinterpretation erfordert eine separate Prüfung ihrer Validität.

Exkurs 21.1

„Die“ Validität eines Tests gibt es nicht

Dass ein Test *nicht per se* „gültig“ und damit valide sein kann, lässt sich an Beispielen wie dem folgenden illustrieren: Eine Fragebogenskala soll das Persönlichkeitsmerkmal „Neurotizismus“ erfassen. Der zugrunde liegenden Persönlichkeitstheorie zufolge sollten Personen mit einer hohen Ausprägung dieses Merkmals anfälliger für Stimmungsveränderungen sein als weniger neurotische (z. B. Gray 1981). Bei einer experimentell induzierten Stimmungsveränderung kann gezeigt werden, dass die Induktion einer negativen Stimmung bei Testpersonen mit einem hohen Testwert auf der Neurotizismuskala zu einem besonders starken Anstieg der negativen Stimmung führt. Dieser Befund stützt die Annahme, dass der Testwert valide als Maß für das theoretische Persönlichkeitsmerkmal „Neurotizismus“ interpretiert werden kann. In einer psychotherapeutischen Praxis wird nun dieselbe Fragebogenskala verwendet, um zu entscheiden, welche Patienten auf einer Warteliste für Psychotherapieplätze vorrangig behandelt werden. Die Testwertinterpretation, dass die Neurotizismuskala ein geeignetes Instrument zur Einschätzung der Dringlichkeit einer Psychotherapie darstellt, wird jedoch durch die vorliegenden Befunde keineswegs gestützt. Hierfür müssten andere Befunde vorgelegt werden, z. B. dass hohe Neurotizismuswerte mit einem erhöhten Suizidrisiko einhergehen. Es kann also eine spezifische Testwertinterpretation (hier der Bezug zu einer Persönlichkeitstheorie) valide sein, eine andere (hier die Verwendung des Testwertes in der klinischen Praxis) jedoch nicht. Die Skala kann daher nicht für sich genommen valide sein, ohne eine spezifische Interpretation und Verwendung anzugeben.

21.3 Argumentationsbasierter Ansatz der Validierung

Die Differenzierung in eine Vielzahl von Validitätsarten bei gleichzeitiger Betrachtung von Validität als Eigenschaft eines Tests hat in der Vergangenheit häufig dazu geführt, dass Ansammlungen von theoretisch kaum begründeten empirischen Befunden (vorwiegend korrelative Zusammenhänge) unter „Validität“ zusammengefasst wurden, ohne dass dies notwendigerweise für die Interpretation und Verwendung des Tests von Belang war. Das aktuelle Verständnis von Validität, das sich auf Testwertinterpretationen bezieht und von einer Unterscheidung verschiedener Validitätsarten absieht, überwindet diese mit den älteren Validitätskonzepten verbundenen Schwierigkeiten. Es zielt darauf ab, Testwertinterpretationen und Testanwendungen gezielt durch hypothesenleitete Forschung zu untermauern. Die früher gebräuchlichen Begrifflichkeiten finden sich dennoch naturgemäß noch in der Fachliteratur, insbesondere zu älteren Testverfahren.

Mit *Validierung* wird der Prozess bezeichnet, unterstützende Belege (Evidenz) für die beabsichtigten Testwertinterpretationen zu sammeln und diese einer Prüfung zu unterziehen. Das Vorgehen hierbei folgt einem argumentationsbasierten Ansatz der Validierung. Das *Validitätsargument* ist die Gesamtheit der Argumentation zur Stützung der Validität einer spezifischen Testwertinterpretation. Es erlaubt eine zusammenfassende Bewertung der Evidenz für (und gegen) diese Testwertinterpretation (Cronbach 1988; Kane 2013). Das Validitätsargument beinhaltet die folgenden Elemente, die als aufeinander aufbauende Schritte im Prozess der Validierung betrachtet werden können:

1. Spezifikation der angestrebten Testwertinterpretation
2. Identifikation und Formulierung von empirisch prüfbaren Grundannahmen, auf denen die Testwertinterpretation aufbaut
3. Sammlung von Evidenz für und gegen die einzelnen Grundannahmen
4. Zusammenfassende Bewertung der Evidenz

Auf die einzelnen Schritte wird im Folgenden genauer eingegangen. Im daran anschließenden ► Abschn. 21.4 wird das Vorgehen bei der argumentationsbasierten Validierung an mehreren Beispielen veranschaulicht.

21.3.1 Spezifikation der Testwertinterpretation

Zu Beginn einer jeden Validierung muss zunächst genau spezifiziert werden, wie ein Testwert interpretiert werden soll und welche Bedeutung die Interpretation bei den beabsichtigten Verwendungen hat (AERA et al. 2014). Die genaue Spezifikation der beabsichtigten Testwertinterpretationen zu Beginn der Validierung ist ausgesprochen wichtig, da dieselben Testwerte in der Regel auf mehr als eine Weise interpretiert werden können (vgl. dazu auch ► Kap. 9). So können Ergebnisse eines Intelligenztests beispielsweise dazu verwendet werden, um

- die individuelle Intelligenzausprägung eines Individuums mit dem einer Referenzgruppe zu vergleichen,
- eine Diagnose der Dyskalkulie zu begründen (wofür eine ausgeprägte Rechenschwäche, aber gerade keine stark unterdurchschnittliche Intelligenz gegeben sein muss) oder
- Studierende zu bestimmten Studiengängen zuzulassen.

Die erfolgreiche Stützung einzelner dieser Testwertinterpretationen mit wissenschaftlichen Belegen bedeutet aber nicht, dass die anderen möglichen Testwertinterpretationen damit ebenfalls gestützt wären. Der Prozess der Validierung muss deshalb für jede Interpretation, die für einen Test relevant ist, gesondert erfolgen; unkritische Generalisierungen sind zu vermeiden.

**Hypothesengeleitete Forschung
zur Untermauerung
von Testwertinterpretationen und
Testanwendungen**
**Validierung als Prozess
der Sammlung von
Validitätsargumenten**

**Genaue Spezifikation der
beabsichtigten
Testwertinterpretationen**

**Unkritische Generalisierungen
vermeiden**

Exkurs 21.2**Arten von Testwertinterpretationen**

Mit den „Interpretationen eines Testergebnisses“ ist gemeint, welche *Schlussfolgerungen* (Inferenzen) aus dem Testergebnis gezogen werden. Kane (2001) nennt fünf Inferenzbereiche, die bei der Interpretation von Testergebnissen häufig wichtig sind:

- **Bewertung:** Ein bewertender Schluss erfolgt, wenn Individuen aufgrund ihrer Testwerte hinsichtlich ihrer Ausprägung des gemessenen Merkmals verglichen werden (z. B. Aussagen über höhere/niedrigere Intelligenz auf Basis von Intelligenztestwerten).
- **Verallgemeinerung:** Eine verallgemeinernde Schlussfolgerung beruht auf der Annahme, dass eine getestete Person mit einem hohen Testwert auch andere, inhaltlich ähnlich gelagerte Testitems in Richtung einer hohen Merkmalsausprägung beantworten wird (z. B. ähnliche Leistungstestaufgaben ebenfalls lösen wird).
- **Extrapolation:** Bei einer extrapolierenden Schlussfolgerung wird von den Testergebnissen auf Bereiche außerhalb der Testsituation geschlossen (z. B. Schluss vom Ergebnis eines Eignungstests auf den späteren beruflichen Erfolg).
- **Erklärung:** Ein erklärender Schluss besteht darin, das Testergebnis mit Bezug auf Theorien (z. B. zur allgemeinen Intelligenz) als Indikator für ein theoretisches Konstrukt zu interpretieren.
- **Entscheidungsfindung:** Schließlich könnte aus dem Testergebnis eine schlussfolgernde Entscheidung abgeleitet werden (z. B. wenn auf Basis des Testergebnisses Studienplätze vergeben werden).

Definition des Messgegenstandes

Um zu bestimmen, wie ein Testergebnis interpretiert werden soll, ist es in der Regel erforderlich, den Messgegenstand, d. h. das Konstrukt, das der Test erfassen soll, genauer zu definieren. Zweckmäßig für die Validierung ist es zudem, die Bedeutung der intendierten Testwertinterpretationen für die primär beabsichtigten Verwendungen des Tests schriftlich darzulegen. Verschiedene Verwendungen von Testergebnissen erfordern in der Regel jeweils spezifische Interpretationen eines Testwertes, wobei man verschiedene Arten von typischen Testwertinterpretationen unterscheiden kann (► Exkurs 21.2).

Verantwortung liegt bei Testentwicklern und Testanwendern

Aus dem Umstand, dass jede spezifische Testwertinterpretation gesondert validiert werden muss, folgt auch, dass nicht nur Testentwickler, sondern auch die Testanwender Verantwortung dafür tragen, dass Testwerte in einer gerechtfertigten Weise interpretiert und verwendet werden können. So müssen die Testentwickler Evidenz dafür zur Verfügung stellen, dass die primär intendierten Testwertinterpretationen eines diagnostischen Verfahrens gerechtfertigt sind, und die Anwender sind letztlich dafür verantwortlich, dass sich die tatsächliche Verwendung des Tests innerhalb des Rahmens bewegt, der durch diese Evidenz gedeckt ist (AERA et al. 2014).

21.3.2 Formulierung von prüfbaren Grundannahmen

Um eine spezifische Testwertinterpretation empirisch stützen oder widerlegen zu können, müssen im Prozess der Validierung *empirisch prüfbare Grundannahmen* identifiziert und explizit benannt werden, die der Testwertinterpretation zugrunde liegen. Soll beispielsweise ein Englischlesetest verwendet werden, um Entscheidungen für die Vergabe von Plätzen in einem Masterstudiengang der Psychologie

21.3 · Argumentationsbasierter Ansatz der Validierung

herbeizuführen, liegen der Interpretation der Englischtestwerte als Auswahlkriterium mindestens die folgenden beiden Grundannahmen zugrunde:

- Der Testwert ist ein Indikator für die Ausprägung der Lesekompetenz im Englischen.
- Lesekompetenz im Englischen ist – z. B. aufgrund des hohen Anteils englischsprachiger Fachliteratur – ein bedeutsamer Prädiktor für den Erfolg im Psychologiestudium.

Die Formulierung der Grundannahmen für eine spezifische Testwertinterpretation ist ein anspruchsvoller und oft schwieriger Schritt. Es kann durchaus strittig sein, ob eine bestimmte Grundannahme für eine Testwertinterpretation tatsächlich bedeutsam ist. Umgekehrt muss im Zuge der Validierung auch begründet werden, dass die formulierten Grundannahmen erschöpfend sind, was bedeutet, dass keine relevante Grundannahme ausgelassen wurde. Ein gutes Kriterium für die Identifikation relevanter Grundannahmen ist die Frage, ob durch ein Widerlegen der Annahme auch die Belastbarkeit der Testwertinterpretation beeinträchtigt würde. Im Beispiel wäre die Interpretation der Englischtestwerte als Auswahlkriterium für Psychologiestudienplätze tatsächlich nicht haltbar, wenn der Test als Indikator für Lesekompetenz im Englischen ungeeignet wäre oder Lesekompetenz im Englischen für den Erfolg im Psychologiestudium nur eine geringe Rolle spielen würde. Inwieweit die formulierten Grundannahmen erschöpfend sind, ist durch intensives Elaborieren und ggf. Abstimmungen mit weiteren Experten sicherzustellen.

Ein weiteres hilfreiches Kriterium bei der Identifikation von Grundannahmen ist die Frage, ob der Test das zugrunde liegende Konstrukt hinreichend gut repräsentiert. Hierbei sollten die möglichen Probleme einer Konstruktunterrepräsentation und der Messung von konstruktirrelevanter Varianz in Betracht gezogen werden:

- *Konstruktunterrepräsentation* bedeutet, dass wesentliche Aspekte der für die Interpretation relevanten Konstruktdefinition im Test nicht abgedeckt sind. Dies könnte im obigen Beispiel der Fall sein, wenn der Englischlesetest nur erzählerische Texte mit einfachem, umgangssprachlichem Vokabular umfasst und Fachtexte mit anspruchsvollerem Inhalt und Vokabular fehlen.
- *Konstruktirrelevante Varianz* bedeutet, dass außer den wesentlichen Aspekten der für die Interpretation relevanten Konstruktdefinition im Test auch noch weitere konstruktirrelevante systematische Quellen von interindividuellen Unterschieden erfasst werden, die nicht Bestandteil der Konstruktdefinition sind. Im Beispiel könnte eine solche Varianzquelle darin bestehen, dass für die Lösung der Testaufgaben spezifisches Sachwissen bezüglich der Textinhalte erforderlich ist, das von der zu erfassenden sprachlichen Fähigkeit unabhängig ist.

Abb. 21.3 veranschaulicht in schematisierter Weise die Missverhältnisse von Messgegenstand (Konstrukt) und Messungen (Items) sowohl für den Fall von Konstruktunterrepräsentation als auch von konstruktirrelevanter Varianz.

Die Grundannahmen müssen einer empirischen Prüfung zugänglich sein; sie müssen daher – ebenso wie gute wissenschaftliche Hypothesen im Allgemeinen – so formuliert werden, dass eine Widerlegung mit empirischen Befunden gut möglich ist. Generell kann die Identifikation und Formulierung der Grundannahmen im Validitätsargument dadurch erleichtert werden, dass gezielt *konkurrierende Hypothesen* in Betracht gezogen werden, die im Widerspruch zu der zu validierenden Testwertinterpretation stehen (AERA et al. 2014).

Beispiel für Grundannahmen

Anforderungen bei der Formulierung von Grundannahmen

Ist das Konstrukt angemessen repräsentiert?

Grundannahmen müssen empirisch widerlegbar sein

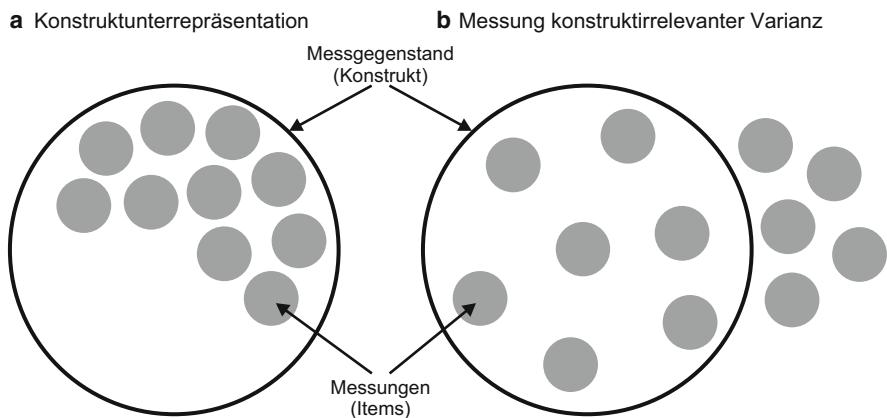


Abb. 21.3 Schematische Veranschaulichung der Problematik **a** der Konstruktunterrepräsentation, bei der wesentliche Bereiche des Messgegenstands nicht abgedeckt werden, und **b** der konstruktirrelevanten Varianz, bei der auch Inhalte erfasst werden, die außerhalb des Messgegenstands liegen

21.3.3 Sammlung empirischer Evidenz

Sind die Grundannahmen formuliert, kann geprüft werden, inwieweit sich empirische Evidenz für sie finden lässt. Dies kann durch den Rückgriff auf bereits publizierte wissenschaftliche Befunde oder durch eigens durchgeführte neue Untersuchungen erfolgen. Bei der Prüfung der Grundannahmen kann auf unterschiedliche Informationsquellen zurückgegriffen werden, mit denen unterschiedliche Aspekte der Validität betrachtet werden können. Zur Gewinnung von empirischer Evidenz zur Stützung der Grundannahmen können gemäß den *Standards for Educational and Psychological Testing* (AERA et al. 2014, s. ▶ Kap. 10 sowie 11) u. a. folgende Quellen herangezogen werden:

- Testinhalte
- Prozesse bei der Testbeantwortung
- die interne Struktur der Testdaten
- Beziehungen der Testwerte zu anderen Variablen

Betont sei, dass dieselben Grundannahmen und Testwertinterpretationen mit Evidenz aus verschiedenen Quellen gestützt werden können – daher sind die verschiedenen Quellen nicht als Basis für unterschiedliche „Validitätsarten“ im Sinne der Verwendung wie in älteren Validitätskonzepten zu verstehen (▶ Abschn. 21.2.1).

Wie bei der Formulierung der Grundannahmen ist auch bei der Auswahl der Evidenz zu ihrer Prüfung zu beachten, dass eine empirische Widerlegung tatsächlich möglich sein muss. Empirische Evidenz für eine Grundannahme ist nur dann wertvoll, wenn die empirischen Befunde bei einer anderen Datenlage auch potenziell geeignet wären, die Grundannahme zu widerlegen.

21.3.4 Zusammenfassende Bewertung des Validitätsarguments

Den Abschluss findet der Prozess der Validierung in der Integration der Evidenz für die verschiedenen Grundannahmen, die einer Testwertinterpretation zugrunde liegen. Die Testwertinterpretation gilt dann als gestützt, wenn für alle relevanten Grundannahmen belastbare Evidenz vorliegt oder – anders formuliert – alle Versuche, die relevanten Grundannahmen zu widerlegen, erfolglos geblieben sind. Die Widerlegung auch nur einer Grundannahme hat zur Folge, dass die zu validierende

21.4 · Beispiele für Validierungsprozesse

Testwertinterpretation nicht aufrechterhalten werden kann – sie muss mindestens abgeschwächt oder sogar grundsätzlich verworfen werden.

Der Prozess der Validierung gleicht theoriegeleiteter und hypothesenbasierter Forschung, mit der die Grundannahmen gestützt oder auch falsifiziert werden können. Damit hat der Abschluss eines Validierungsprozesses immer einen vorläufigen Charakter. Es besteht auch nach zusammenfassender Bewertung des Validitätsarguments weiterhin die Möglichkeit, dass zukünftige Befunde einzelne Grundannahmen und somit eine Testwertinterpretation zur Gänze infrage stellen.

Validierungsprozess ist nicht endgültig

21.4 Beispiele für Validierungsprozesse

In den folgenden Abschnitten wird der argumentationsbasierte Ansatz der Validierung anhand verschiedener, für psychologische Tests typischer Testwertinterpretationen veranschaulicht. Es werden jeweils exemplarisch zentrale Grundannahmen, die den Testwertinterpretationen zugrunde liegen, expliziert sowie mögliche Evidenz zur Stützung dieser Grundannahmen angeführt. Die Beispiele zeigen, dass eine Validierung kein immer gleiches Routineverfahren ist. Befunde, die für eine Testwertinterpretation relevant sind, können für eine andere Interpretation derselben Testwerte unbedeutend sein. Bei der Betrachtung der Beispiele ist zu beachten, dass die Auswahl der Grundannahmen aus didaktischen Gründen jeweils nicht ausführlich begründet wird und dass nicht immer eine erschöpfende Anzahl von Grundannahmen angeführt wird. Die Begründung muss für echte Validitätsargumente umfangreicher sein und aus einer zugrunde liegenden Theorie abgeleitet und auf die exakten Charakteristika des Anwendungskontexts abgestimmt werden.

Vorschau

21.4.1 Beispiel 1: Interpretation von Testwerten als Indikatoren für ein theoretisches Konstrukt

In der psychologischen Forschung werden Testwerte häufig dahingehend interpretiert, dass sie interindividuelle Unterschiede in einem theoretischen Konstrukt widerspiegeln. Demzufolge sollten Personen mit niedrigen Testwerten eine niedrige, Personen mit hohen Testwerten eine hohe Ausprägung in Bezug auf das zugrunde liegende Konstrukt haben. Eine derartige *erklärende Inferenz* lässt sich, abhängig von der zugrunde liegenden Theorie, auf vielfältige Weise widerlegen oder stützen, wie das folgende Beispiel veranschaulicht. Validiert werden soll die genannte Testwertinterpretation:

■ Testwertinterpretation

„Die mit Ravens Matrizentest (Raven 1962) erfassten Testwerte sind Indikatoren für den Intelligenzfaktor *Logisches Schlussfolgern (Reasoning)*.“

Eine Argumentation für diese *erklärende Interpretation* lässt sich auf mehreren Grundannahmen aufbauen:

1. Die Antworten im Test lassen sich auf eine einzige zugrunde liegende Fähigkeitsdimension zurückführen.
2. Die Testleistungen hängen hoch mit allgemeiner Problemlösefähigkeit zusammen.
3. Die Testleistungen sind unabhängig von anderen theoretisch abgrenzbaren kognitiven Fähigkeiten.

Ravens Matrizen-Test zur Erfassung des Intelligenzfaktors „Reasoning“

Diese drei Grundannahmen können durch Evidenz aus verschiedenen Quellen gestützt werden.

1. Ob den Testantworten wie theoretisch angenommen eine einzige Dimension interindividueller Unterschiede zugrunde liegt, lässt sich durch eine Analyse der

Evidenz auf Basis der internen Struktur

internen Struktur der Antwortdaten untersuchen. Mit einer konfirmatorischen Faktorenanalyse (CFA, ▶ Kap. 24) kann geprüft werden, ob die Daten tatsächlich eine eindimensionale Struktur aufweisen.

2. Ob die Testergebnisse tatsächlich hoch mit der allgemeinen Problemlösefähigkeit zusammenhängen, lässt sich anhand des *Zusammenhangs mit anderen Variablen* untersuchen. Die statistische Analyse erfolgt im einfachsten Fall durch die Berechnung einer Korrelation der mit dem Ravens Matrizen test ermittelten Testwerte mit Testwerten, die mit einem Instrument zur Messung der allgemeinen Problemlösefähigkeit an denselben Personen erhoben wurden. Natürlich können auch komplexere statistische Modellierungsmethoden zum Einsatz kommen. So nutzten beispielsweise Schweizer et al. (2007) mehrdimensionale Strukturgleichungsmodelle, um zu klären, ob die Dimension in Ravens Matrizen test mit einem allgemeinen, durch andere Tests definierten Problemlösefaktor auf latenter Ebene hoch korreliert ist. Ein derartiger Zusammenhang wird als *konvergente Evidenz* bezeichnet.
3. Der Beleg, dass Testwerte *nicht* mit anderen theoretisch abgrenzbaren Konstrukten zusammenhängen, wird als *diskriminante Evidenz* bezeichnet. Sie kann analog der konvergenten Evidenz durch die Untersuchung der korrelativen *Zusammenhänge mit anderen Variablen* erbracht werden. Im Beispiel könnte durch die simultane Erhebung mit geeigneten anderen Verfahren gezeigt werden, dass die Leistungen in den Raven-Matrizen nicht mit räumlichem Vorstellungsvermögen oder Konzentrationsfähigkeit zusammenhängen. Eine gemeinsame statistische Modellierung von konvergenter und diskriminanter Evidenz kann mithilfe des Multitrait-Multimethod-Ansatzes (MTMM-Ansatz, ▶ Kap. 25) erfolgen.

Konvergente Evidenz

Diskriminante Evidenz

Alle Grundannahmen müssen gestützt sein

Falsifikationsprinzip beachten

Die oben angeführte Testwertinterpretation gilt dann als vorläufig gestützt, wenn alle Grundannahmen durch die gesammelten Befunde gestützt werden. Wenn nur eine der Grundannahmen widerlegt würde – z. B. wenn sich ein hoher Zusammenhang zwischen der Leistung in den Raven-Matrizen und räumlichen Vorstellungsvermögen finde – wäre die gesamte Testwertinterpretation infrage zu stellen.

Es ist zu betonen, dass es sehr wichtig ist, bei der Validierung dem *Falsifikationsprinzip* streng zu folgen. Es ist nicht zulässig, lediglich unterstützende Evidenzen zu sammeln und ggf. sogar empirische Ergebnisse zu unterdrücken, die den aufgestellten Grundannahmen widersprechen. Die Grundannahmen sind explizit mit dem Ziel einer Falsifikation auf die Probe zu stellen; dies ist durch die Planung angemessener Studien sicherzustellen. Beispielsweise ist die Stichprobengröße auf Basis zu erwartender Effekte so zu planen, dass Hypothesen zur Testung der Grundannahmen mit angemessener Teststärke geprüft werden können. Weiterhin sind Grenzwerte für maximale oder minimale korrelative Zusammenhänge als spezifische Alternativhypotesen bei inferenzstatistischen Tests (vgl. Frey 2006; Hartig et al. 2012) zu formulieren. Aufgrund vorliegender Erkenntnisse und Theorien sind nämlich bei Untersuchung konvergenter Evidenz in der Regel zwar keine perfekten Zusammenhänge von eins, aber dennoch über einem bestimmten Grenzwert liegende Korrelationen zu erwarten. Gleichermaßen ist bei der Untersuchung diskriminanter Evidenz nicht zwangsläufig eine Nullkorrelation notwendig bzw. realistisch, um eine hinreichende Abgrenzung zu anderen Konstrukten zu gewährleisten.

21.4.2 Beispiel 2: Verwendung von Testwerten bei der Auswahl diagnostik

Viele psychologische Test- und Fragebogenverfahren werden bei der Auswahl diagnostik verwendet. Auch der im vorangegangenen Beispiel verwendete Matri-

21.4 · Beispiele für Validierungsprozesse

zentest kann zur Personalauswahl genutzt werden, z. B. um aus einer Menge von Bewerbern für einen Ausbildungsplatz als Kfz-Mechatroniker geeignete Personen auszuwählen. Die Testwertinterpretation, die hierfür zu validieren wäre, nimmt eine extrapolierende Inferenz vor:

■ Testwertinterpretation

„Personen mit einem hohen Testwert im Ravens Matrizen test sind für eine Ausbildung in Kfz-Mechatronik besser geeignet als Personen mit einem niedrigen Testwert.“

Diese *extrapolierende Interpretation* desselben Testwertes baut auf folgenden Grundannahmen auf:

1. Der Test erfasst eine Fähigkeit, die in der Ausbildung in Kfz-Mechatronik gebraucht wird.
2. Personen mit einem hohen Testwert sind in der Ausbildung erfolgreicher als Personen mit einem niedrigen Testwert.

Zur Stützung dieser Grundannahmen könnte wieder die Evidenz aus verschiedenen Quellen herangezogen werden:

1. Ob die erfassten Fähigkeiten für die infrage stehende Tätigkeit relevant sind, kann bezogen auf den *Testinhalt* untersucht werden. Die Grundannahme kann gestützt werden, wenn Arbeitsplatzanalysen oder Interviews mit Experten aus dem Berufsfeld ergeben, dass logisches Schlussfolgern für die Tätigkeit eines Kfz-Mechatronikers als bedeutsam angesehen wird.
2. Ob der Testwert tatsächlich mit größerem Erfolg in der Ausbildung einhergeht, kann durch den *Zusammenhang mit anderen Variablen* untersucht werden. Ein solcher Beleg wäre erbracht, wenn sich der Testwert zu Ausbildungsbeginn in einer Längsschnittstudie tatsächlich als bedeutsamer Prädiktor für den Ausbildungserfolg (z. B. die Abschlussnote) erweist.

Eine Stützung der Testwertinterpretation erfolgt, wenn beide Grundannahmen nicht durch empirische Daten widerlegt werden können.

Die beiden Beispiele auf Basis des Ravens Matrizen tests veranschaulichen, dass einerseits derselbe Testwert in verschiedenen Kontexten unterschiedlich verwendet werden kann und dass andererseits verschiedene Interpretationen unterschiedliche Evidenz erfordern können. Für die Interpretation des Testwertes aus den Raven-Matrizen als Indikator für allgemeine Problemlösefähigkeit (► Abschn. 21.4.1) ist es beispielsweise irrelevant, ob die Testwerte den Erfolg in einer spezifischen Berufsausbildung vorhersagen. Für die Verwendung im Rahmen der Auswahlagnostik (► Abschn. 21.4.2) hingegen spielt es keine Rolle, ob die Zusammenhänge der Testwerte mit anderen Tests zum Problemlösen oder zum räumlichen Vorstellungsvermögen theoriekonform ausfallen.

Intelligenztest zur Personalauswahl für einen Ausbildungsplatz

Verschiedene Interpretationen erfordern unterschiedliche Evidenz

21.4.3 Beispiel 3: Interpretation eines Testwertes als Screening in der klinisch-psychologischen Diagnostik

In der klinisch-psychologischen Praxis werden Fragebogen u. a. eingesetzt, um ein Screening auf häufig auftretende psychische Störungen vorzunehmen und erste Informationen über deren Schweregrad zu ermitteln (z. B. der Gesundheitsfragebogen für Patienten, PHQ-D; Löwe et al. 2002). Wenn ein Fragebogen beispielsweise neu entwickelt wird, um das Vorliegen und den Schweregrad von depressiven Symptomen zu erfassen, kann die angestrebte Testwertinterpretation wie folgt lauten:

Screening in der klinisch-psychologischen Diagnostik

Extrapolierende Testwertinterpretation

■ Testwertinterpretation

„Bei Patientinnen und Patienten mit einem hohen Testwert im Fragebogen besteht der dringende Verdacht des Vorliegens einer schweren Depression (*Major Depression*).“

Diese *extrapolierende Testwertinterpretation* basiert auf den folgenden Grundannahmen:

- Der Fragebogen erfasst die wichtigsten Symptome einer schweren Depression und keine konstruktfremden Aspekte.
- Personen mit einer Major Depression haben höhere Testwerte im Fragebogen als Personen, die die Diagnosekriterien für eine Major Depression nicht erfüllen.

Ähnlich wie in der beruflichen Auswahlagnostik kann Evidenz für diese Annahmen auf Basis des Testinhalts und des Zusammenhangs mit anderen Variablen gewonnen werden:

- Die Grundannahme, dass der Fragebogen die wichtigsten Symptome einer Major Depression erschöpfend abdeckt, bezieht sich auf den *Testinhalt*. Evidenz zur Stützung dieser Grundannahme kann erbracht werden, wenn Experten aus der klinischen Psychologie dahingehend übereinstimmen, dass sowohl die zentralen Kriterien für eine Major Depression im Sinne des Diagnostic and Statistical Manual of Mental Disorders (DSM-5; Falkai und Wittchen 2015) vollständig im Fragebogen enthalten sind als auch im Fragebogen keine diesbezüglich irrelevanten Inhalte abgefragt werden.
- Die Grundannahme, dass Personen mit einer vorliegenden Major Depression im Fragebogen höhere Testwerte aufweisen, kann über den Nachweis des *Zusammenhangs mit einer anderen Variablen* gestützt werden. Hierzu kann eine Patientengruppe, bei der das Vorliegen einer Major Depression z. B. durch ein ausführliches klinisches Interview und weitere diagnostische Verfahren gesichert ist, mit einer Gruppe verglichen werden, bei der sichergestellt ist, dass die Kriterien für eine Depressionsdiagnose nicht erfüllt sind. Zur Stützung der Grundannahme sollte sich ein großer Mittelwertunterschied (ein großer Effekt im Sinne von Cohen 1988) zwischen beiden Gruppen in der erwarteten Richtung finden. Sofern ein derartiger Unterschied nachgewiesen wird, kann für die Interpretation des Testwertes in der Praxis ein optimaler Schwellenwert zur Trennung zwischen depressiven und nicht depressiven Patienten ermittelt werden (► Kap. 9).

Alle relevanten Grundannahmen müssen gestützt sein

Auch in diesem Beispiel kann die angestrebte Testwertinterpretation nur dann als validiert betrachtet werden, wenn beide Grundannahmen gestützt werden können. Würden beispielsweise die Expertenurteile ergeben, dass ein definierendes Diagnosekriterium für eine Major Depression im Fragebogen fehlt und/oder würden sich die Testwerte depressiver und nicht depressiver Personen nicht oder nur geringfügig unterscheiden, könnte der Fragebogen nicht wie intendiert interpretiert und verwendet werden.

21.4.4 Beispiel 4: Interpretation eines Testwertes bezogen auf das Erreichen von Lernzielen

Kompetenztests zur Erfassung von Bildungsstandards

Im Bildungswesen werden Kompetenztests (Frey und Hartig 2018) als spezielle Leistungstests eingesetzt, um zu untersuchen, inwieweit angestrebte Bildungsziele erreicht wurden (z. B. Stanat et al. 2012). In Deutschland werden die Bildungsziele allgemeinbildender Schulen für zentrale Schulfächer durch bundeseinheitliche fachspezifische Bildungsstandards konkretisiert. Die Bildungsstandards formulieren Erwartungen darüber, über welche Kompetenzen Schüler bis zu einem be-

21.4 · Beispiele für Validierungsprozesse

stimmten Zeitpunkt ihres Bildungsgangs verfügen sollen. Im Rahmen der im Jahr 2003 begonnenen Einführung der Bildungsstandards erfolgte auch eine Ausrichtung der Lehrpläne und des Unterrichts auf die entsprechende Kompetenzvermittlung. Daher können die Ergebnisse der Tests zu den Bildungsstandards auch zur Evaluation des Erfolgs des Unterrichts bei der Kompetenzvermittlung genutzt werden. Wird z. B. ein Kompetenztest im Bereich der Mathematik eingesetzt, um das Erreichen der in den Bildungsstandards in Mathematik für die Grundschule (Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004) definierten Anforderungen zu untersuchen, wäre die zu validierende Testwertinterpretation folgende:

■ Testwertinterpretation

„Schülerinnen und Schüler mit hohen Testwerten im Mathematiktest haben die angestrebten Bildungsziele im Fach Mathematik zum Ende der Grundschule in höherem Umfang erreicht als Schülerinnen und Schüler mit niedrigeren Testwerten.“

Diese Testwertinterpretation enthält einen *verallgemeinernden* und einen *erklärenden Aspekt* und basiert auf den folgenden Grundannahmen:

1. Der Test enthält Aufgaben, deren Inhalte dem implementierten Lehrplan entsprechen.
2. Zum Lösen der Testaufgaben sind die im Unterricht vermittelten Kompetenzen erforderlich.

Zur Stützung dieser Grundannahmen könnte wieder die Evidenz aus verschiedenen Quellen herangezogen werden:

1. Ob die Aufgaben eines Kompetenztests mit dem Lehrplan übereinstimmen, kann bezogen auf die *Testinhalte* belegt werden. Evidenz für diese Grundannahme kann erbracht werden, wenn Inhaltsexperten (z. B. Lehrkräfte oder Verantwortliche in den Bildungsministerien) die Aufgabeninhalte als passend zu den in den Bildungsstandards definierten Zielen einschätzen und Aufgaben zu allen inhaltlichen Bereichen enthalten sind.
2. Die Grundannahme, dass für das Lösen der Testaufgaben die im Unterricht vermittelten Kompetenzen benötigt werden, bezieht sich auf *Antwortprozesse*. Diese können z. B. mit der Methode des „lauten Denkens“ analysiert werden. Hierbei werden Testpersonen angehalten, ihre Lösungsprozesse beim Bearbeiten der Testaufgaben kontinuierlich zu verbalisieren. Die Protokolle dieser Verbalisierungen werden mit einem vorab formulierten theoretischen Prozessmodell der Aufgabenbearbeitung verglichen (Ericsson und Simon 1993). Evidenz für die Grundannahme kann erbracht werden, wenn an einer Stichprobe aus der Zielgruppe (hier Grundschüler) anhand der Protokollanalyse gezeigt werden kann, dass Lösungen der Testaufgaben tatsächlich nur dann erreicht werden, wenn die im Unterricht vermittelten Kompetenzen (z. B. spezifische Rechenmethoden) angewandt werden.

Die im Beispiel angestrebte Testwertinterpretation könnte als (vorläufig) validiert betrachtet werden, wenn die bezogen auf die Testinhalte und Antwortprozesse gesammelte Evidenz die Grundannahmen nicht widerlegt. Würden befragte Experten z. B. feststellen, dass der Test zu erheblichen Teilen Aufgaben zu Inhalten enthält, die im Hinblick auf die Bildungsziele irrelevant sind (konstruktirrelevante Varianz), müsste die Testwertinterpretation verworfen werden. Ebenfalls verworfen werden müsste sie, wenn die Analyse der Antwortprozesse zeigen würde, dass viele Aufgaben auch ohne die im Unterricht vermittelten Kompetenzen gelöst werden können (z. B. durch Raten oder logisches Erschließen ohne Rückgriff auf mathematikspezifische Aspekte).

**Beispiele, wann
Testwertinterpretationen verworfen
werden müssen**

21.5 Zusammenfassung

Das Gütekriterium der Validität ist ein zentrales Qualitätskriterium, das den Gütekriterien Objektivität oder Reliabilität übergeordnet ist. Es bezieht sich darauf, inwieweit Interpretationen von Testwerten und beabsichtigte Verwendungen von Tests gerechtfertigt sind. Das Verständnis von Validität hat sich in den letzten Jahrzehnten deutlich weiterentwickelt. Während sich im vergangenen Jahrhundert zunächst eine wenig praktikable Vielzahl „verschiedener Validitäten“ herausgebildet hatte, wird Validität inzwischen als ein einheitliches Qualitätskriterium betrachtet, das Informationen aus verschiedenen Quellen integriert. Zudem wurde Validität früher als Eigenschaft eines Tests per se aufgefasst, heute bezieht sie sich auf die Interpretation von Testwerten im Hinblick auf die intendierte Nutzung. Ein Test kann demnach nicht als solcher valide sein, sondern jede unterschiedliche Testwertinterpretation erfordert eine separate Prüfung ihrer Validität.

Die Prüfung der Validität (Validierung) einer Testwertinterpretation erfolgt im Rahmen eines argumentationsbasierten Ansatzes. Als erster Schritt muss die zu *validierende Testwertinterpretation* präzise formuliert werden. Anschließend werden *prüfbare Grundannahmen* identifiziert, auf denen die Testwertinterpretation aufbaut. Im nächsten Schritt wird *empirische Evidenz* gesammelt, anhand derer die Grundannahmen widerlegt oder vorläufig gestützt werden können. Wichtige Quellen für Evidenz zur Prüfung der Grundannahmen sind die Testinhalte, die bei der Testbeantwortung ablaufenden kognitiven Prozesse, die interne Struktur der Testdaten und die Beziehungen der Testwerte zu anderen Variablen. Bei der abschließenden *zusammenfassenden Bewertung* wird eine Testwertinterpretation dann als valide betrachtet, wenn keine der zugrunde liegenden Annahmen widerlegt werden konnte. Die argumentationsbasierte Validierung ist kein immer gleiches Routineverfahren. In Analogie zur theoriegeleiteten und hypothesenbasierten Forschung muss die argumentationsbasierte Validierung für jede Testwertinterpretation spezifisch hergeleitet werden. Der Abschluss eines Validierungsprozesses hat insoweit einen vorläufigen Charakter, als zukünftige Befunde einzelne Grundannahmen und damit eine Testwertinterpretation zur Gänze entkräften können.

21.6 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Häufig wird vereinfachend von „der Validität eines Tests“ gesprochen. Warum ist diese Vereinfachung nach einem modernen Verständnis von Validität potenziell irreführend?
2. Welche beiden zentralen Bereiche werden bei der Konzeptualisierung der Konstruktvalidität von Cronbach und Meehl (1955) unterschieden?
3. Welches sind wichtige Evidenzquellen für die Validierung von Testwertinterpretationen?
4. Welche Schritte werden beim argumentationsbasierten Ansatz der Validierung üblicherweise durchlaufen?

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 Psychological Association.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Literatur

- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L. J. (1980). Selection theory for a political world. *Public Personnel Management*, 9, 37–50.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.
- Falkai, P. & Wittchen, H.-U. (Hrsg.). (2015). *Diagnostisches und statistisches Manual psychischer Störungen DSM-5*. Göttingen: Hogrefe.
- Frey, A. (2006). *Validitätssteigerungen durch adaptives Testen*. Frankfurt am Main: Peter Lang.
- Frey, A. & Hartig, J. (2018). Kompetenzdiagnostik. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Hrsg.), *Handbuch Schulpädagogik* (S. 849–858). Münster: Waxmann.
- Gray J. A. (1981). A critique of Eysenck's theory of personality. In H. J. Eysenck (Ed.), *A model for personality* (pp. 246–277). Berlin, Heidelberg: Springer.
- Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Test- und Fragebogenkonstruktion* (2. Aufl., S. 143–171). Berlin, Heidelberg: Springer.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. Köln: Luchterhand.
- Löwe, B., Spitzer, R. L., Zipfel, S. & Herzog, W. (2002). *Gesundheitsfragebogen für Patienten (PHQ-D)*. Komplettversion und Kurzform. Testmappe mit Manual, Fragebögen, Schablonen (2. Aufl.). Karlsruhe: Pfizer.
- Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, 7, 191–205.
- Raven, J. C. (1962). *Advanced progressive matrices*. London: Lewis & Co. Ltd.
- Schweizer, K., Goldhammer, F., Rauch, W. & Moosbrugger, H. (2007). On the validity of Raven's matrices test: does spatial ability contribute to performance? *Personality and Individual Differences*, 43, 1998–2010.
- Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrs.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.
- 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.



Latent-Class-Analyse (LCA)

Mario Gollwitzer

Inhaltsverzeichnis

- 22.1 Einleitung und Überblick – 549**
 - 22.1.1 Quantitative vs. qualitative Personenvariablen – 549
 - 22.1.2 Die LCA im Überblick – 550
- 22.2 Herleitung der Modellgleichung – 552**
 - 22.2.1 Bedingte Wahrscheinlichkeiten für einzelne Testitems – 552
 - 22.2.2 Bedingte Wahrscheinlichkeiten für Antwortmuster – 553
 - 22.2.3 Unbedingte Wahrscheinlichkeiten für Antwortmuster – 554
 - 22.2.4 Bedingte Klassenzuordnungswahrscheinlichkeiten – 554
- 22.3 Parameterschätzung und Überprüfung der Modellgüte – 556**
 - 22.3.1 Likelihood-Funktion – 556
 - 22.3.2 Likelihood-Ratio-Test – 558
 - 22.3.3 Klassischer χ^2 -Test – 559
 - 22.3.4 Bootstrap-Verfahren – 559
 - 22.3.5 Informationskriterien – 560
 - 22.3.6 Genauigkeit der Klassenzuordnung – 561
 - 22.3.7 Eliminierung nicht trennscharfer Items – 561
- 22.4 Exploratorische und konfirmatorische Anwendungen der LCA – 562**
 - 22.4.1 Exploratorische Anwendungen der LCA: Finden des besten Modells – 562
 - 22.4.2 Konfirmatorische Anwendungen der LCA: Testen von Modellrestriktionen – 564
 - 22.4.3 Modellvergleichstests – 566
- 22.5 Erweiterte Anwendungen der LCA – 567**
 - 22.5.1 LCA für polytome Antwortformate – 567
 - 22.5.2 Mischverteilungs-Rasch-Modelle – 569

22.6 Zusammenfassung – 571

22.7 EDV-Hinweise – 572

22.8 Kontrollfragen – 572

Literatur – 573

i Handelt es sich bei den latenten Personenvariablen, die zur Erklärung des unterschiedlichen Verhaltens von verschiedenen Personen herangezogen werden, nicht um quantitative kontinuierliche, sondern um qualitative diskrete Variablen, können die Latent-Trait-Modelle nicht angewendet werden. Vielmehr ist die Latent-Class-Analyse (LCA) das geeignete Testmodell. Die LCA ist ein probabilistisches Testmodell für kategoriale latente Variablen. Sie basiert auf der Annahme, dass Personen mit einer gewissen Wahrscheinlichkeit einer von mehreren Klassen (oder Typen) angehören, wobei die Klassenzugehörigkeit aus den Antworten auf die Items eines Tests (Lazarsfeld und Henry 1968) erschlossen wird. Da die Klassenzugehörigkeit nicht direkt beobachtbar ist, spricht man von latenten Klassen. Ziel dieses Kapitels ist es, in die Grundgedanken der LCA einzuführen und anhand von Beispielen zu verdeutlichen, wann und wie die LCA als Testmodell anwendbar ist.

22.1 Einleitung und Überblick

22.1.1 Quantitative vs. qualitative Personenvariablen

Personen unterscheiden sich hinsichtlich einer Vielzahl von Eigenschaften, z. B. ihres Geschlechts, ihrer Körpergröße oder ihres Temperaments. Solche Personenvariablen sind entweder quantitativer oder qualitativer Natur. Die Körpergröße ist beispielsweise eine quantitative Personenvariable: Je größer eine Person ist, desto höher ist ihr „Wert“ auf dem jeweiligen Messinstrument (z. B. einem Zentimetermaßband). Die Körpergröße ist darüber hinaus eine stetige Variable: Zwischen zwei Werten können theoretisch unendlich viele mögliche Werte liegen. Das biologische Geschlecht hingegen ist eine qualitative Personenvariable: Der Unterschied zwischen „männlich“ und „weiblich“ sowie möglichen weiteren biologischen Geschlechterkategorien beschreibt also eine qualitative (und keine quantitative) Andersartigkeit zwischen diesen Kategorien. Qualitative Variablen sind immer nominalskaliert und diskret, d. h., die Ausprägungen sind weder gereiht noch gibt es graduelle Unterschiede zwischen Werten.

Bei den meisten psychologischen Variablen ist es eindeutig, ob sie quantitativer oder qualitativer Natur sind. Intelligenz wird in der Regel als quantitative Variable aufgefasst, die Entscheidung für eine bestimmte politische Partei am Wahltag ist hingegen eine qualitative Variable. Bei anderen Variablen ist die Zuordnung weniger eindeutig. Ob Extraversion bzw. Introversion zwei Ausprägungen einer quantitativen oder doch eher einer qualitativen Personenvariable sind, hängt von der zugrunde liegenden Theorie ab: Während in den Temperamentstheorien von Eysenck (1990) oder Gray (1972) Extraversion und Introversion als zwei Pole eines eindimensionalen quantitativen Kontinuums gelten, fasste C. G. Jung (1921) beide Konstrukte als „Typen“ auf: Ihm zufolge ist eine Person entweder extravertiert oder introvertiert, graduelle Abstufungen in Bezug auf die Merkmalsausprägung sind in seiner Theorie nicht vorgesehen. Interessanterweise sind Typologien in der wissenschaftlichen Psychologie etwas aus der Mode gekommen. Während beispielsweise die Persönlichkeitsforschung in ihren Anfängen von typologischen Modellen dominiert wurde (wie bei der Temperamentstypologie von Hippokrates oder der Körperbautypologie von Ernst Kretschmer), geht man in den meisten modernen Theorien davon aus, dass sich die Persönlichkeit eines Menschen durch ihre Verortung auf mehreren dimensionalen Variablen (beispielsweise den sog. „Big Five“ des Fünf-Faktoren-Modells oder den sechs Dimensionen des HEXACO-Modells) beschreiben lässt.

Für den Fall, dass die latente (d. h. nicht direkt beobachtbare) Personenvariable quantitativer Natur ist (z. B. bei Betrachtung einer Big-Five- oder HEXACO-Dimension) eignen sich die klassischen und die probabilistischen Testmodelle, die in Teil II dieses Lehrbuches beschrieben wurden, beispielsweise Modelle der Item-

**Persönlichkeitstypen oder
Persönlichkeitsdimensionen?**

Latente Klassen

Response-Theorie (IRT, ► Kap. 16), ► Kap. 18 wie das Rasch-Modell oder die Birnbaum-Modelle. Handelt es sich bei der latenten Personenvariablen jedoch um eine qualitative Variable, können diese Testmodelle nicht angewendet werden. An dieser Stelle kommt die LCA ins Spiel. Die LCA ist ein probabilistisches Testmodell, da sie auf der Annahme basiert, dass Personen – gegeben ihre Antworten in einem Test mit mehreren Items – mit einer gewissen Wahrscheinlichkeit einer von mehreren Klassen (oder Typen) angehören (Lazarsfeld und Henry 1968). Da die Klassenzugehörigkeit nicht direkt beobachtbar ist, spricht man von *latenten Klassen*. Kurz gesagt handelt es sich bei Latent-Class-Modellen um Testmodelle für kategoriale latente Variablen. Ziel dieses Kapitels ist es, in die Grundgedanken der LCA einzuführen und anhand von Beispielen zu verdeutlichen, wann und wie die LCA als Testmodell anwendbar ist.

Veranschaulicht werden soll die LCA zunächst anhand einer einfachen Typologie aus der Forschung zu Geschlechterrollen. In einigen Theorien (beispielsweise von Bem 1977; Spence et al. 1975; für einen Überblick s. Athenstaedt und Alfermann 2011) wird davon ausgegangen, dass Maskulinität und Feminität zwei quantitative Dimensionen sind, hinsichtlich derer sich Personen voneinander unterscheiden, dass diese beiden Dimensionen jedoch miteinander unkorreliert sind. Dies impliziert, dass die Kombination von Maskulinität und Feminität in vier möglichen Typen mündet: Als „feminin“ werden Personen bezeichnet, die hohe Werte in Bezug auf die Dimension Feminität und niedrige Werte in Bezug auf die Dimension Maskulinität haben. „Maskuline“ Personen weisen ein umgekehrtes Muster auf. Als „androgyn“ werden Personen bezeichnet, die hohe Werte beider Dimensionen haben; als „undifferenziert“ werden schließlich all jene bezeichnet, die in beiden Dimensionen niedrige Werte aufweisen.

Um eine Typisierung der Testpersonen in eine der vier Klassen vorzunehmen, haben Spence et al. (1975) sowie Bem (1977) vorgeschlagen, die Maskulinitäts- und die Feminitätsskala auf der Basis eines Median-Splits künstlich zu dichotomisieren. Dieses Vorgehen ist jedoch problematisch, u. a. deshalb, weil der Median je nach Stichprobe variiert und daher das gleiche Antwortmuster je nach Lage des Medians zu unterschiedlichen Typenzuordnungen führen kann (für weitere Kritikpunkte am Median-Split s. MacCallum et al. 2002). Als geeigneteres Testmodell erweist sich in diesem Fall die LCA, die im Folgenden erläutert werden soll. Dabei wird von dem einfachen Fall ausgegangen, dass es sich bei den Items zur Erfassung von Maskulinität und Feminität jeweils um Items mit einem dichotomen Antwortformat handelt, also um Aussagen/Statements, die von den Testpersonen entweder bejaht oder verneint werden können (in ► Abschn. 22.5.1 wird kurz auf die Erweiterung der LCA im Falle von polytomous Items, d. h. Items mit mehr als zwei Antwortkategorien, eingegangen).

22.1.2 Die LCA im Überblick

Antwortmuster (Response Pattern)

Gegeben sei das Antwortverhalten y_{vi} einer Menge von Personen ($v \in \{1, \dots, n\}$) bei einer Menge von Testitems ($i \in \{1, \dots, m\}$). Auf jedem Item gibt es prinzipiell k Antwortkategorien, aber für den einfachen Fall eines dichotomischen Antwortformats gibt es nur $k = 2$ Antwortkategorien („nein“ $\Rightarrow y_{vi} = 0$; „ja“ $\Rightarrow y_{vi} = 1$). Die Antworten einer Person v auf allen m Items werden auch als Antwortmuster oder *Response Pattern* a_v bezeichnet. Das ► Beispiel 22.1 zeigt „ideale“ Antwortmuster.

Beispiel 22.1: Ideale Antwortmuster

Ein Test bestehe aus $m = 6$ Items. Drei von ihnen (1, 2, 3) messen typische Feminität (z. B. „Ich bin feinfühlig“, „Ich bin emotional“, „Ich bin verständnis-

voll“), die anderen drei (4, 5, 6) typische Maskulinität (z. B. „Ich bin selbstbewusst“, „Ich gebe nie leicht auf“, „Ich bin unabhängig“). Alle Items haben ein dichotomes Antwortformat: $y_{vi} \in \{0, 1\}$. Die folgende Tabelle zeigt vier hypothetische Antwortmuster a_v (mit $v \in \{1, \dots, 4\}$), die so gewählt wurden, dass sie den vier von Spence et al. (1975) postulierten Typen genau entsprechen:

Item	Femininitätsitems			Maskulinitätsitems		
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Idealmuster a_1 : „femininer Typ“	1	1	1	0	0	0
Idealmuster a_2 : „maskuliner Typ“	0	0	0	1	1	1
Idealmuster a_3 : „androgyner Typ“	1	1	1	1	1	1
Idealmuster a_4 : „undifferenzierter Typ“	0	0	0	0	0	0

Eine perfekte Anpassung der Daten an das Vier-Typen-Modell von Bem (1977) bzw. Spence et al. (1975) wäre gegeben, wenn es in einer Stichprobe mit n Personen lediglich die vier in ► Beispiel 22.1 abgebildeten Antwortmuster gäbe. In der Realität wird man eine solche Eindeutigkeit aber nicht finden; es wird immer Antwortmuster geben, die vom Idealmuster mehr oder weniger abweichen können.

Die Anzahl der *maximal möglichen Antwortmuster* N_a^{\max} ist dabei begrenzt: Sie berechnet sich bei einem dichotomen Antwortformat als $N_a^{\max} = 2^m$. Im Falle von $m = 6$ dichotomen Items gäbe es $N_a^{\max} = 2^6 = 64$ mögliche Antwortmuster. Man sieht, dass die Zahl der möglichen Antwortmuster in Abhängigkeit von der Anzahl der Items rasch in die Höhe schnellt: Schon bei $m = 10$ dichotomen Items gäbe es $2^{10} = 1024$ mögliche Antwortmuster.

Die Anzahl der *empirisch beobachteten Antwortmuster* N_a ist meist kleiner als N_a^{\max} , da nicht jedes mögliche Antwortmuster in der Stichprobe auch tatsächlich vorkommt. Und das ist prinzipiell günstig, da zu viele Antwortmuster in der Stichprobe dazu beitragen würden, dass die Klassenlösung weniger gut interpretierbar wäre. Zudem ist N_a durch die Stichprobengröße begrenzt: Schließlich kann es nur so viele unterschiedliche Antwortmuster wie Personen in der Stichprobe geben: $N_a \leq n$. Wenn N_a zu klein ist, treten allerdings sowohl bei der Parameterschätzung als auch bei der Diagnose der Modellgüte Probleme auf (mehr dazu ► Abschn. 22.3).

Zentral für die LCA ist die theoretische Annahme, dass jede Person (d. h. jedes Element der Population) einer (und *nur* einer) von mehreren latenten Klassen (g) angehört. Unbekannt hierbei ist,

1. mit wie vielen latenten Klassen G ($g \in \{1, \dots, G\}$) die Stichprobendaten angemessen erklärt werden können,
2. wie viele Personen einer jeweiligen Klasse g angehören; der entsprechende Modellparameter, die „relative Klassengröße“, wird mit π_g bezeichnet, und
3. welche Person v welcher Klasse g angehört.

Bei empirischen Anwendungen der LCA ist die Zugehörigkeit einer Person v zu einer latenten Klasse g nicht deterministisch (d. h. eindeutig, mit absoluter Sicherheit zugehörig oder nicht zugehörig), sondern vielmehr probabilistisch. Das bedeutet: Eine Person gehört mit einer mehr oder weniger großen Wahrscheinlichkeit einer bestimmten latenten Klasse an. Für jede Person v – genauer gesagt für das Antwortmuster a_v einer Person – kann eine bedingte Klassenzuordnungswahrscheinlichkeit

Anzahl maximal möglicher Antwortmuster N_a^{\max}

Anzahl empirisch beobachteter Antwortmuster N_a

Theoretische Klassenzuordnung ist deterministisch

Empirische Klassenzuordnung ist probabilistisch

$P(g | a_v)$ berechnet werden, d. h. die Wahrscheinlichkeit, mit der eine Person mit dem Antwortmuster a_v zur Klasse g gehört. Auch die relativen Klassengrößen π_g können innerhalb eines Modells geschätzt werden.

Die einzige Größe, die *nicht* modellimmanent geschätzt werden kann, ist G , die Anzahl der latenten Klassen in der Stichprobe. Diese Größe sollte, falls möglich, theoriegeleitet festgelegt werden. So wird beispielsweise im Modell von Spence et al. (1975) a priori von einer Vier-Klassen-Struktur ausgegangen. Anschließend kann anhand von Modellgüteindizes (► Abschn. 22.4) ermittelt werden, ob die a priori vorgegebene Klassenanzahl zu den Daten passt. In diesem Zusammenhang können mehrere unterschiedliche Modelllösungen hinsichtlich ihrer Passung auf die Daten miteinander verglichen werden oder auch mit konkurrierenden Modellen (z. B. mit einer Drei-Klassen-Lösung, wie etwa in der Untersuchung von Strauß et al. 1996). Ein solches Vorgehen ähnelt der Suche nach der „geeigneten“ Anzahl von Faktoren bei der exploratorischen Faktorenanalyse (EFA, ► Kap. 23).

A-priori-Festlegung der Anzahl latenter Klassen G

Bestimmung der bedingten Klassenzuordnungswahrscheinlichkeit

22.2 Herleitung der Modellgleichung

Im Folgenden wird die allgemeine Modellgleichung für eine LCA mit dichotomen Items hergeleitet. Ziel ist es, für jedes beobachtete Antwortmuster a_v die Wahrscheinlichkeit zu bestimmen, mit der die entsprechende Person einer latenten Klasse angehört. Gesucht ist also die bedingte Wahrscheinlichkeit einer Klasse g bei gegebenem Antwortmuster, d. h. die bedingte Klassenzuordnungswahrscheinlichkeit $P(g | a_v)$. Um diese Wahrscheinlichkeit bestimmen zu können, sind folgende Informationen erforderlich:

- *Bedingte Antwortmusterwahrscheinlichkeit* $P(a_v | g)$, d. h. die Wahrscheinlichkeit eines Antwortmusters a_v unter der Bedingung, dass die Person v zur Klasse g gehört
- *Unbedingte Antwortmusterwahrscheinlichkeit* $P(a_v)$, d. h. die einfache Wahrscheinlichkeit, mit der dieses Antwortmuster überhaupt vorkommt
- *Relative Klassengrößen* π_g , d. h. die unbedingte Wahrscheinlichkeit, mit der eine Person der Klasse g angehört

22.2.1 Bedingte Wahrscheinlichkeiten für einzelne Testitems

Um die Wahrscheinlichkeiten für die Antwortmuster bestimmen zu können, werden zunächst die unbedingten und nachfolgend die bedingten Bejahungs- und Verneinungswahrscheinlichkeiten für die einzelnen Items betrachtet.

Die (unbedingte) Wahrscheinlichkeit, mit der eine Person v ein dichotomes Item i bejaht ($y_{vi} = 1$), wird als $P(y_{vi} = 1) = P_{vi}$ bezeichnet. Die (unbedingte) Wahrscheinlichkeit, mit der die Person v ein dichotomes Item i verneint ($y_{vi} = 0$), ist dementsprechend die Gegenwahrscheinlichkeit: $P(y_{vi} = 0) = 1 - P_{vi}$.

Es gilt die Annahme, dass die *Bejahungswahrscheinlichkeit* P_{vi} je nach Zugehörigkeit einer Person zu einer der latenten Klassen g variiert. In Anlehnung an ► Beispiel 22.1 sollte das Item 1 (z. B. „Ich bin feinfühlig“) eher von Personen des „femininen“ oder „androgynen“ Typs, aber eher nicht von Personen des „maskulinen“ (oder „undifferenzierten“) Typs bejaht werden. Bei Item 4 (z. B. „Ich bin selbstbewusst“) sollte die Bejahungswahrscheinlichkeit hingegen unter „maskulinen“ (und „androgynen“) Typen größer sein als unter „femininen“ (und „undifferenzierten“). Gesucht wird also die bedingte Wahrscheinlichkeit, mit der eine Person v das Item i bejaht ($y_{vi} = 1$) unter der Bedingung, dass sie der Klasse g angehört. Diese bedingte Bejahungswahrscheinlichkeit wird als $P(y_{vi} = 1 | g)$ bezeichnet. Die Gegenwahrscheinlichkeit, also die bedingte *Verneinungswahrscheinlichkeit*, entspricht $P(y_{vi} = 0 | g) = 1 - P(y_{vi} = 1 | g)$.

Bedingte Bejahungs- und Verneinungswahrscheinlichkeit

An dieser Stelle kommt eine erste einfache, aber wichtige Modellannahme der LCA ins Spiel.

Annahme 1: Konstante Bejahungswahrscheinlichkeiten innerhalb einer Klasse

Angenommen wird, dass die Bejahungswahrscheinlichkeit bzw. die Verneinungswahrscheinlichkeit eines Items für alle Personen innerhalb einer latenten Klasse gleich ist.

Als Folge von Annahme 1 kann man den Index v weglassen und die bedingte Bejahungs- und Verneinungswahrscheinlichkeit wie folgt verkürzen: $P(y_{vi} = 1 | g) = P_{ig}$ und $P(y_{vi} = 0 | g) = 1 - P_{ig}$. Beide Antwortwahrscheinlichkeiten lassen sich in einer einzigen Gleichung ausdrücken, wenn man die jeweiligen Itemantworten y_{vi} als Exponent schreibt:

$$P(y_{vi} | g) = P_{ig}^{y_{vi}} \cdot (1 - P_{ig})^{1-y_{vi}} \quad (22.1)$$

Dies ist nichts anderes als eine Reformulierung, denn für den Fall einer „Ja“-Antwort ($y_{vi} = 1$) verkürzt sich die Gleichung auf der rechten Seite zu P_{ig} ; im Falle einer „Nein“-Antwort ($y_{vi} = 0$) verkürzt sich die rechte Seite zu $1 - P_{ig}$.

22.2.2 Bedingte Wahrscheinlichkeiten für Antwortmuster

In Gl. (22.1) wurde die bedingte Antwortwahrscheinlichkeit für ein einzelnes Item formal bestimmt. Nun wird die bedingte Wahrscheinlichkeit für ein ganzes Antwortmuster a_v , formal: $P(a_v | g)$, bestimmt. Damit ist die Wahrscheinlichkeit für ein Antwortmuster a_v (d. h. für die Antworten einer Person v auf allen m Items) unter der Bedingung, dass die Person v der Klasse g angehört, gemeint.

An dieser Stelle wird die Annahme der *lokalen stochastischen Unabhängigkeit* der Items untereinander bedeutsam. Diese Annahme ist zentral für probabilistische Testmodelle; sie wurde bereits im Rahmen der IRT vorgestellt (► Kap. 16). Vereinfacht besagt die Annahme lokaler stochastischer Unabhängigkeit, dass die Wahrscheinlichkeit $P(y_{v1} = 1, y_{v2} = 1 | g)$, mit der eine Person v aus der Klasse g zwei Items (1, 2) eines Tests bejaht, dem Produkt der beiden bedingten Bejahungswahrscheinlichkeiten für diese Items entspricht:

$$P(y_{v1} = 1, y_{v2} = 1 | g) = P(y_{v1} = 1 | g) \cdot P(y_{v2} = 1 | g) \quad (22.2)$$

Lokale stochastische Unabhängigkeit

Das bedeutet: Die Wahrscheinlichkeit, beide Items zu bejahen, hängt – außer von der Bejahungswahrscheinlichkeit der Items selbst – nur von der Klassenzugehörigkeit g ab und nicht etwa von der Reihenfolge, in der die Items beantwortet wurden, oder aber davon, ob ein vorhergehendes Item bereits beantwortet wurde oder nicht. Wendet man diese Logik auf ein Antwortmuster a_v an, das aus den Antworten $y_{vi} = 1$ bzw. $y_{vi} = 0$ auf den m Items besteht, so ergibt sich die bedingte Wahrscheinlichkeit für das gesamte Antwortmuster, indem man die bedingten Antwortwahrscheinlichkeiten über alle Items hinweg aufmultipliziert (dies wird ausgedrückt durch den Produktoperator \prod).

$$P(a_v | g) = \prod_{i=1}^m P(y_{vi} | g) \quad (22.3)$$

Annahme 2: Lokale stochastische Unabhängigkeit innerhalb einer Klasse

Die Wahrscheinlichkeit, zwei oder mehrere unabhängige Testitems gemeinsam zu bejahen, hängt nur davon ab, welcher Klasse g eine Person angehört.

Allgemeine Modellgleichung der LCA bei dichotomen Items

Setzt man nun die rechte Seite der Gl. (22.1) in Gl. (22.3) ein, so erhält man die *allgemeine Modellgleichung der LCA* für dichotome Items:

$$\begin{aligned} P(a_v|g) &= \prod_{i=1}^m P(y_{vi}|g) \\ &= \prod_{i=1}^m (P_{ig}^{y_{vi}} \cdot (1 - P_{ig})^{1-y_{vi}}) \end{aligned} \quad (22.4)$$

Die Modellgleichung besagt, dass sich die Wahrscheinlichkeit eines Antwortmusters a_v unter der Bedingung, dass die Person v der Klasse g angehört, aus dem Produkt der bedingten Bejahungs- bzw. Verneinungswahrscheinlichkeiten über alle m Items ergibt.

22.2.3 Unbedingte Wahrscheinlichkeiten für Antwortmuster**Relative Klassengröße**

Zusätzlich müssen die *relativen Klassengrößen* π_g eingeführt werden, um die unbedingten Wahrscheinlichkeiten für die Antwortmuster a_v bestimmen zu können. Hier muss die Annahme getroffen werden, dass die latenten Klassen exhaustiv und disjunkt (zu den Begriffen s. ▶ Kap. 5) sind:

Annahme 3: Exhaustive und disjunkte Klassen

- Alle Personen können einer Klasse zugeordnet werden (d. h. die Klassen sind exhaustiv).
- Eine Person kann nur einer Klasse angehören, nicht mehreren (d. h. die Klassen sind disjunkt).

Unter dieser Annahme müssen sich die *relativen Klassengrößen* π_g zu 1 aufaddieren. Das bedeutet gleichzeitig, dass sich die bedingten Wahrscheinlichkeiten für ein Antwortmuster a_v (Gl. 22.4) über alle Klassen g hinweg nach Gewichtung mit den Klassengrößen π_g zu einer unbedingten Antwortmusterwahrscheinlichkeit $P(a_v)$ aufaddieren:

$$P(a_v) = \sum_{g=1}^G \left[\pi_g \prod_{i=1}^m (P_{ig}^{y_{vi}} \cdot (1 - P_{ig})^{1-y_{vi}}) \right] \quad (22.5)$$

22.2.4 Bedingte Klassenzuordnungswahrscheinlichkeiten**Bayes-Theorem**

In ▶ Abschn. 22.1.2 wurde bereits darauf hingewiesen, dass es möglich ist, für jedes beliebige Antwortmuster a_v anzugeben, wie groß die Wahrscheinlichkeit ist, mit der sich die entsprechende Person v in der Klasse g befindet. Diese bedingte Klassenzuordnungswahrscheinlichkeit $P(g | a_v)$ lässt sich mithilfe des *Bayes-Theorems* aus der relativen Klassengröße π_g sowie aus der bedingten und der

22.2 · Herleitung der Modellgleichung

unbedingten Wahrscheinlichkeit für das Antwortmuster a_v bestimmen:

$$P(g|a_v) = \frac{\pi_g \cdot P(a_v|g)}{P(a_v)} \quad (22.6)$$

Alle drei Größen auf der rechten Seite der Gleichung sind bereits eingeführt: Die bedingte Wahrscheinlichkeit $P(a_v | g)$ für ein Antwortmuster a_v unter der Bedingung g (Gl. 22.4), die unbedingte Wahrscheinlichkeit $P(a_v)$ für ein Antwortmuster a_v (Gl. 22.5) sowie die relativen Klassengrößen π_g in ► Abschn. 22.2.3.

In ► Beispiel 22.2 wird die Berechnung der bedingten Klassenzuordnungswahrscheinlichkeiten gemäß Gl. (22.6) veranschaulicht.

Beispiel 22.2: Ist Fritz eher dem „femininen“ oder dem „maskulinen“ Typ zuzuordnen?

Unter der Annahme eines Vier-Klassen-Modells ($G = 4$) mit $\pi_1 = 43\%$; $\pi_2 = 28\%$; $\pi_3 = 17\%$ und $\pi_4 = 12\%$ wird für einen Test mit $m = 6$ Items davon ausgegangen, dass folgende bedingte (fiktive) Bejahungswahrscheinlichkeiten P_{ig} gemäß Gl. (22.1) vorliegen:

(P_{ig})	Item 1 (P_{1g})	Item 2 (P_{2g})	Item 3 (P_{3g})	Item 4 (P_{4g})	Item 5 (P_{5g})	Item 6 (P_{6g})
Klasse 1 ($\pi_1 = 43\%$)	0.83	0.77	0.90	0.56	0.24	0.43
Klasse 2 ($\pi_2 = 28\%$)	0.33	0.28	0.45	0.75	0.81	0.69
Klasse 3 ($\pi_3 = 17\%$)	0.90	0.86	0.59	0.77	0.56	0.40
Klasse 4 ($\pi_4 = 12\%$)	0.32	0.22	0.09	0.19	0.31	0.29

Eine bestimmte Person v in der Stichprobe („Fritz“) habe folgendes Antwortmuster produziert: $a_{\text{Fritz}} = \langle 1, 0, 1, 1, 0, 0 \rangle$. Welcher latenter Klasse gehört Fritz unter den angenommenen Bedingungen am ehesten an? Um diese Frage beantworten zu können, muss man die durch sein Antwortmuster a_{Fritz} bedingten Klassenzuordnungswahrscheinlichkeiten $P(g | a_{\text{Fritz}})$ berechnen. Aus Gl. (22.6) wird deutlich, dass hierzu die unbedingten und bedingten Antwortmusterwahrscheinlichkeiten sowie die relativen Klassengrößen (π_g) bekannt sein müssen. Die bedingten Antwortmusterwahrscheinlichkeiten ergeben sich laut Gl. (22.4) aus dem Produkt der Bejahungs- bzw. Verneinungswahrscheinlichkeiten für jedes einzelne Item. Die in der Tabelle angegebenen klassenspezifischen Bejahungswahrscheinlichkeiten (P_{ig}) sind jedoch nur auf die tatsächlich von Fritz bejahten Items (1, 3 und 4) anzuwenden. Für die von Fritz verneinten Items (2, 5 und 6) müssen die klassenspezifischen Verneinungswahrscheinlichkeiten ($1 - P_{ig}$) verwendet werden. Es ergeben sich die folgenden klassenspezifischen (bedingten) Antwortmusterwahrscheinlichkeiten:

$$\begin{aligned} P(a_{\text{Fritz}}|g=1) &= 0.83 \cdot (1 - 0.77) \cdot 0.90 \cdot 0.56 \cdot (1 - 0.24) \cdot (1 - 0.43) \approx 0.042 \\ P(a_{\text{Fritz}}|g=2) &= 0.33 \cdot (1 - 0.28) \cdot 0.45 \cdot 0.75 \cdot (1 - 0.81) \cdot (1 - 0.69) \approx 0.005 \\ P(a_{\text{Fritz}}|g=3) &= 0.90 \cdot (1 - 0.86) \cdot 0.59 \cdot 0.77 \cdot (1 - 0.56) \cdot (1 - 0.40) \approx 0.015 \\ P(a_{\text{Fritz}}|g=4) &= 0.32 \cdot (1 - 0.22) \cdot 0.09 \cdot 0.19 \cdot (1 - 0.31) \cdot (1 - 0.29) \approx 0.002 \end{aligned}$$

Die unbedingte Wahrscheinlichkeit für das Antwortmuster von Fritz kann man nun leicht ermitteln, indem man – Gl. (22.5) folgend – die bedingten Antwortmusterwahrscheinlichkeiten mit den relativen Klassengrößen π_1 bis π_4 (Tabellenspalte 1)

gewichtet und dann über alle vier Klassen hinweg aufaddiert:

$$\begin{aligned} P(a_{\text{Fritz}}) &= \sum_{g=1}^G \pi_g \cdot P(a_{\text{Fritz}}|g) \\ &= 0.43 \cdot 0.042 + 0.28 \cdot 0.005 + 0.17 \cdot 0.015 + 0.12 \cdot 0.002 = 0.022 \end{aligned}$$

Nun können gemäß Gl. (22.6) die bedingten Klassenzuordnungswahrscheinlichkeiten berechnet werden:

$$\begin{aligned} P(g = 1|a_{\text{Fritz}}) &= \frac{0.43 \cdot 0.042}{0.022} = 0.812 \\ P(g = 2|a_{\text{Fritz}}) &= \frac{0.28 \cdot 0.005}{0.022} = 0.060 \\ P(g = 3|a_{\text{Fritz}}) &= \frac{0.17 \cdot 0.015}{0.022} = 0.116 \\ P(g = 4|a_{\text{Fritz}}) &= \frac{0.12 \cdot 0.002}{0.022} = 0.011 \end{aligned}$$

Die Wahrscheinlichkeit, mit der Fritz (gegeben sein Antwortmuster a_v) der ersten Klasse angehört, ist deutlich größer als die Wahrscheinlichkeit, einer der drei übrigen Klassen anzugehören. Unter der Annahme, dass die Klassen tatsächlich in Anlehnung an die von Spence et al. (1975) vorgeschlagene Typologie interpretiert werden dürfen, könnte man also behaupten, Fritz sei eher dem „femininen“ Typ zuzuordnen.

22.3 Parameterschätzung und Überprüfung der Modellgüte

22.3.1 Likelihood-Funktion

Unbekannte Modellparameter

Das Ziel bei empirischen Anwendungen der LCA besteht nun darin, die unbekannten Modellparameter möglichst genau und zuverlässig aus den Daten zu schätzen. Unbekannt sind die

- relativen *Klassengrößen* π_g . Hiervon sind lediglich $G - 1$ Parameter zu schätzen, da sich die Parameter über die Klassen hinweg zu 1 addieren und der g -te Parameter dadurch festliegt;
- *klassenspezifischen Antwortwahrscheinlichkeiten* für jedes Item P_{ig} gemäß Gl. (22.1). Die Anzahl dieser Wahrscheinlichkeiten ergibt sich aus der Anzahl Items (m) mal der Anzahl der Klassen (G).

Die Anzahl der unbekannten Modellparameter wird mit t bezeichnet und hängt von der Anzahl der Klassen und der Anzahl der Items ab. Wenn man also, bezogen auf das ► Beispiel 22.1 in ► Abschn. 22.1.2, weiterhin von $G = 4$ latenten Klassen sowie $m = 6$ dichotomen Items ausgeht, so müssen 3 Klassengrößenparameter und $6 \times 4 = 24$ klassenspezifische Antwortwahrscheinlichkeiten, also insgesamt $t = 27$ Parameter, geschätzt werden.

Anzahl der unbekannten Modellparameter

Die Anzahl der unbekannten Modellparameter t einer LCA berechnet sich als $t = G \cdot (m + 1) - 1$.

Die unbekannten Modellparameter werden iterativ geschätzt. Das bedeutet, dass die Parameter schrittweise angepasst werden mit dem Ziel, ein bestimmtes Optimierungskriterium zu erfüllen. Im ersten Schritt werden für die Modellparameter sog. „Startwerte“ eingesetzt, wobei die Anpassung an das Optimierungskriterium noch relativ schlecht ist. In weiteren Schritten (Iterationen) werden die Parameter dann so lange adjustiert, bis die Anpassung an das Optimierungskriterium nicht mehr bedeutsam verbessert werden kann.

Als Optimierungskriterium bei der LCA dient die *Likelihood* L . Sie ist definiert als das Produkt der unbedingten Antwortmusterwahrscheinlichkeiten $P(a_v)$ über alle beobachteten (Anzahl N_a) Antwortmuster in der Stichprobe hinweg:

$$L = \prod_{v=1}^{N_a} P(a_v) \quad (22.7)$$

Die unbekannten Modellparameter π_g und P_{ig} sollen nun so geschätzt werden, dass die Likelihood unter den gegebenen Voraussetzungen den größtmöglichen (= maximalen) Wert annimmt. Daher wird dieses Verfahren *Maximum-Likelihood-Verfahren* (ML-Verfahren) genannt. Hierbei werden die Parameterschätzungen so lange adjustiert, bis sich die Likelihood nicht mehr weiter erhöht bzw. bis die Adjustierung nur noch zu unbedeutenden Veränderungen in den Parameterschätzungen führt. Dieser Zustand wird *Konvergenz* genannt; die resultierenden Parameterschätzungen sind insoweit „optimal“, als bei den gegebenen Daten keine bessere Lösung erzielt werden kann.

Die Logik hinter dem Gedanken, dass eine Maximierung der Likelihood zum optimalen Modell führt, lässt sich gut an der Frage beschreiben, was eigentlich der Unterschied zwischen einem „guten“ und einem „schlechten“ Modell ist. Bei einem guten Modell ist die Wahrscheinlichkeit sehr hoch, dass sich mit den geschätzten Modellparametern die Verteilung der Antwortmuster in der Stichprobe vollständig rekonstruieren lässt. Anders gesagt: Die Wahrscheinlichkeit, durch Einsetzen der geschätzten Modellparameter in die Modellgleichung Gl. (22.4) bzw. Gl. (22.5) genau jene Daten (Antwortmuster) zu erhalten, die man auch tatsächlich beobachtet hat, ist sehr hoch. Bei einem schlechten Modell hingegen ist die Wahrscheinlichkeit, die empirischen Daten mit den geschätzten Modellparametern rekonstruieren zu können, gering. Das bedeutet: Je höher die Likelihood ist, desto zutreffender sind die geschätzten Modellparameter. Je unzutreffender hingegen die geschätzten Modellparameter sind, desto geringer sind die Antwortmusterwahrscheinlichkeiten und desto kleiner ist die Likelihood.

Obwohl das Prinzip des ML-Verfahrens zunächst sehr einleuchtend klingt, stellen sich einige mathematische Probleme. Was ist z. B., wenn mehrere Kombinationen von Modellparametern zu einer identischen (maximalen) Likelihood führen („multiple Maxima“)? Auf das Problem genauer einzugehen, würde an dieser Stelle zu weit führen (hierzu sei auf Titterington et al. 1985, verwiesen). Wichtig ist in diesem Zusammenhang lediglich der Umstand, dass die Problematik multipler Maxima umso geringer wird, je größer die Stichprobe ist und je mehr beobachtete Antwortmuster vorliegen.

Ein weiteres Problem besteht darin, dass die absolute Höhe der Likelihood nicht nur von der tatsächlichen Passung eines Modells, sondern auch von Aspekten abhängt, die mit der Modellgüte zunächst gar nichts zu tun haben, nämlich der Stichprobengröße bzw. der Anzahl und Verteilung der vorkommenden Antwortmuster sowie der Anzahl der Items. Aus diesem Grund wird zur Prüfung der Modellgüte statt der absoluten eine *standardisierte Likelihood* verwendet. Hierzu vergleicht man die in dem Modell geschätzten unbedingten Antwortmusterwahrscheinlichkeiten $P(a_v)$ mit den empirisch beobachteten relativen Häufigkeiten der Antwortmuster $f(a_v)$. Das Produkt der gemäß Gl. (22.7) *geschätzten* Antwortmusterwahrscheinlichkeiten über alle in der Stichprobe vorkommenden Antwortmuster hinweg wird mit L_1 bezeichnet. Das Produkt der *beobachteten* relativen Antwort-

Iteratives Schätzverfahren

Likelihood

ML-Verfahren

Problem: Multiple Maxima

Standardisierte Likelihood

musterhäufigkeiten über alle in der Stichprobe vorkommenden Antwortmuster hinweg bezeichnet man mit L_0 :

$$L_0 = \prod_{v=1}^{N_a} f(a_v) \quad (22.8)$$

Ein Vergleich zwischen L_1 und L_0 ist also nichts anderes als eine Quantifizierung des Ausmaßes, in dem die Vorhersagen des Modells von den empirisch vorgefundenen Gegebenheiten in der Stichprobe abweichen: je größer diese Abweichung, desto schlechter der „Modellfit“; je kleiner diese Abweichung, desto besser passt das Modell zu den Daten (s. auch ► Kap. 16).

22.3.2 Likelihood-Ratio-Test

Eine Möglichkeit, den Unterschied zwischen L_1 und L_0 zu quantifizieren, besteht in der Quotientenbildung beider Größen. Dieser Quotient wird als *Likelihood Ratio (LR)* bezeichnet:

$$LR = \frac{L_1}{L_0} \quad (22.9)$$

Je näher LR am Wert 1 liegt, desto besser passt das Modell zu den Daten; je weiter LR unter 1 liegt, desto schlechter passt das Modell. Dieser Quotient ist unabhängig von der Größe der Stichprobe, der Anzahl latenter Klassen und der Itemanzahl.

Unter der Voraussetzung, dass die Stichprobe groß genug ist, kann der LR zur inferenzstatistischen Absicherung in eine Prüfgröße L^2 umgerechnet werden, die approximativ einer χ^2 -Verteilung folgt (vgl. Bollen 1989):

$$\begin{aligned} L^2 &= -2 \cdot \log \left(\frac{L_1}{L_0} \right) \\ &= 2 \cdot (\log(L_0) - \log(L_1)) \\ &\sim \chi^2 \end{aligned} \quad (22.10)$$

Die Nullhypothese des Likelihood-Ratio-Tests impliziert, dass das Modell perfekt zu den Daten passt. Die Alternativhypothese impliziert, dass das Modell nicht perfekt zu den Daten passt.

Die *Freiheitsgrade (df)* der χ^2 -Statistik ergeben sich aus der Differenz zwischen der Anzahl gegebener Informationen (s) und der Anzahl zu schätzender Modellparameter (t):

$$df = s - t \quad (22.11)$$

Die Anzahl gegebener Informationen (s) entspricht der Anzahl der möglichen Antwortmuster (► Abschn. 22.1.2) minus eins: $s = N_a^{\max} - 1$, im Falle dichotomer Items also $s = 2^m - 1$. Die Anzahl zu schätzender Modellparameter entspricht $t = G \cdot (m + 1) - 1$ (► Abschn. 22.3.1). Also berechnen sich die Freiheitsgrade der χ^2 -Statistik für dichotome Items wie folgt:

$$df = (2^m - 1) - (G \cdot (m + 1) - 1) = 2^m - G \cdot (m + 1) \quad (22.12)$$

Liegt der Wert der Prüfgröße L^2 im Ablehnungsbereich unter der χ^2 -Verteilung (wobei der Ablehnungsbereich durch das Signifikanzniveau definiert wird, das meist auf 5 % festgesetzt wird), heißt das: Die Nullhypothese muss verworfen werden, das Modell passt nicht zu den Daten. Entscheidend für die Robustheit

des Likelihood-Ratio-Tests ist die Stichprobengröße n , denn nur mit ausreichend großer Stichprobe ist L^2 tatsächlich approximativ χ^2 -verteilt. Als Faustregel wird vorgeschlagen (z. B. Formann 1984), dass die Stichprobe mindestens so viele Personen umfassen sollte, wie es mögliche Antwortmuster gibt, also $n \geq N_a^{\max}$ (besser noch: $n \geq 5 \cdot N_a^{\max}$). Die erforderliche Stichprobengröße wird in Abhängigkeit von der Anzahl der Items (und der Anzahl der Antwortkategorien) daher sehr schnell sehr groß.

Erforderliche Stichprobengröße

22.3.3 Klassischer χ^2 -Test

Eine zweite Möglichkeit, die Passung eines spezifischen LCA-Modells, den *Modellfit*, zu überprüfen, besteht darin, die aus den geschätzten Modellparametern rekonstruierte Häufigkeit eines Antwortmusters a_v direkt – ohne den Umweg über die Likelihood – mit der empirisch beobachteten Häufigkeit dieses Antwortmusters zu vergleichen. Die aus den geschätzten Modellparametern rekonstruierte absolute Häufigkeit eines Antwortmusters a_v ergibt sich aus dem Produkt der unbedingten Wahrscheinlichkeit für dieses Antwortmuster $P(a_v)$ mit dem Stichprobenumfang n . Die empirische relative Häufigkeit eines Antwortmusters $f(a_v)$ ist in den Daten gegeben. Konkret bildet man die quadrierte Differenz zwischen empirischer und geschätzter relativer Antwortmusterhäufigkeit und teilt diese durch die empirische relative Antwortmusterhäufigkeit. Dies macht man für alle beobachteten Antwortmuster ($v = 1, \dots, N_a$) und bildet dann die Summe. Die resultierende Prüfgröße ist annähernd χ^2 -verteilt:

$$\chi^2 = \sum_{v=1}^{N_a} \frac{(f(a_v) - n \cdot P(a_v))^2}{f(a_v)} \quad (22.13)$$

Die Freiheitsgrade dieser χ^2 -Statistik berechnen sich auch hier nach Gl. (22.12). In der Regel führen der Likelihood-Ratio-Test aus Gl. (22.10) und der „klassische“ χ^2 -Test aus Gl. (22.13) zu annähernd gleichen Ergebnissen (Rost 2004).

Modellfit

Ein großer Vorteil beider Tests ist, dass die Diskrepanz eines Modells inferenzstatistisch abgesichert werden kann; die starke Abhängigkeit von der Stichprobengröße stellt hingegen einen Nachteil dar: Ist die Stichprobe „zu klein“, ist der Test nicht exakt, da die χ^2 -Verteilung nicht hinreichend gut approximiert wird (Read und Cressie 1988); in diesem Fall sollte besser ein Bootstrap-Verfahren (► Abschn. 22.3.4) zur Anwendung kommen. Ist die Stichprobe hingegen „zu groß“, so werden auch irrelevante Unterschiede zwischen L_1 und L_0 bzw. zwischen $f(a_v)$ und $P(a_v)$ signifikant; in diesem Fall sollte besser auf die Informationskriterien (► Abschn. 22.3.5) zurückgegriffen werden.

Vor- und Nachteile des χ^2 -Tests

22.3.4 Bootstrap-Verfahren

Für den Fall, dass die Stichprobengröße zu klein ist, um eine hinreichende χ^2 -Approximation zu erreichen, besteht die Möglichkeit, die Prüfverteilung durch simulierte Daten selbst zu erzeugen und mit den 5 %- bzw. 1 %-Quantilen der jeweils resultierenden Verteilung den Ablehnungsbereich der Nullhypothese zu definieren (s. Langeheine et al. 1996; van Kollenburg et al. 2015; von Davier 1997).

Erzeugung einer Prüfverteilung mit simulierten Daten

Ein solches Verfahren wird Bootstrap-Verfahren genannt (engl. „bootstrap“ = Stiefelschlaufen¹). Durch dieses ist es möglich, die Vorteile einer inferenzstatistischen Absicherung der Modelldiskrepanz selbst dann zu nutzen, wenn die Voraussetzungen für einen üblichen χ^2 -Test nicht erfüllt sind.

Im konkreten Fall kann eine solche simulierte Prüfverteilung erzeugt werden, indem man – unter der Annahme der Gültigkeit eines Modells – eine große Zahl künstlicher Datensätze generieren lässt (Resimulation). Für die jeweils zu berechnende Prüfgröße (z. B. L^2 oder χ^2) erhält man eine Wahrscheinlichkeitsverteilung, deren 5%- bzw. 1%-Quantile man zur Bestimmung des Ablehnungsbereichs unter der Nullhypothese verwenden kann. Liegt die Prüfgröße im Ablehnungsbereich (bei einem vorher definierten Signifikanzniveau), so ist das Modell zu verwerfen.

22.3.5 Informationskriterien

Eine Alternative zu den in den vorangegangenen Abschnitten behandelten Verfahren stellt die Inspektion sog. „Informationskriterien“ dar. Auch sie basieren auf der Likelihood L eines Modells (► Abschn. 22.3.1). Der konzeptuelle Vorteil der Informationskriterien liegt jedoch darin, dass die Anzahl der Modellparameter berücksichtigt wird, um Modelle mit zu vielen (unnötigen) Parametern zu „bestrafen“. Diese Logik wird nachvollziehbar, wenn man sich vergegenwärtigt, dass ein Modell mit vielen latenten Klassen trivialerweise besser zu den Daten passt als ein sparsames Modell mit nur wenigen latenten Klassen. Ähnliches gilt auch für andere statistische Verfahren: Ein Regressionsmodell mit vielen Prädiktoren erklärt in der Regel mehr Varianz als ein Modell mit wenigen Prädiktoren; eine Faktorenanalyse, in der viele Faktoren extrahiert wurden, erklärt die Varianz aller Items in der Regel besser als ein Faktormodell mit wenigen Faktoren usw. Akzeptiert man das – wissenschaftstheoretisch begründete – Argument, dass sparsame Modelle belohnt werden sollten („Parsimonitätsprinzip“, vgl. ► Kap. 24), so liegt es nahe, die Likelihood eines (zu) komplexen Modells mithilfe eines Bestrafungsfaktors abzuwerten.

Parsimonitätsprinzip

Informationskriterien AIC, BIC, aBIC und CAIC

Im Allgemeinen unterscheidet man vier Informationskriterien, die sich jedoch alle stark ähneln: das *Akaike Information Criterion* (AIC), das *Bayesian Information Criterion* (BIC), das *adjustierte BIC* (aBIC) und das – etwas seltener verwendete – *Consistent AIC* (CAIC). In die Indizes gehen also die (logarithmierte) Likelihood des Modells (L), die Anzahl der Modellparameter (t) und – bei BIC, aBIC und CAIC – zusätzlich die Stichprobengröße (n) ein. Für alle gilt: Je niedriger der Wert ist, desto besser passt das Modell zu den Daten (s. auch Schermelleh-Engel et al. 2003).

$$\begin{aligned} \text{AIC} &= -2 \cdot \ln(L) + 2 \cdot t \\ \text{BIC} &= -2 \cdot \ln(L) + t \cdot \ln(n) \\ \text{aBIC} &= -2 \cdot \ln(L) + t \cdot \ln\left(\frac{n+2}{24}\right) \\ \text{CAIC} &= -2 \cdot \ln(L) + t \cdot (\ln(n) + 1) \end{aligned} \quad (22.14)$$

Auf der Basis der Informationskriterien würde man ein Modell, das mehr Parameter beinhaltet, aber die gleiche Likelihood wie ein „sparsamer“ Modell besitzt, eher verwerfen. Die Stichprobengröße macht diese „Bestrafung“ noch einmal härter: Je größer die Stichprobe ist, desto stärker schlägt die Bestrafung zu Buche.

¹ Die Anpassung eines Modells anhand von Daten zu prüfen, die unter der Annahme der Gültigkeit des Modells simuliert wurden, hat Ähnlichkeit mit der Münchhausen'schen Fähigkeit, sich am eigenen Zopf – oder eben an den eigenen Stiefelschlaufen – aus dem Sumpf zu ziehen.

Simulationsstudien haben gezeigt, dass das BIC und das aBIC unter allen Informationskriterien am ehesten geeignet sind, ein Modell mit der passenden Klassenzahl zu identifizieren (Hagenaars und McCutcheon 2002; Yang 2006).

22.3.6 Genauigkeit der Klassenzuordnung

Eine weitere – eher deskriptive – Möglichkeit, den Modellfit, d. h. die Modellanpassung eines LCA-Modells, zu überprüfen, besteht in der Analyse der *Treffsicherheit* („hitrate“), d. h. der Anzahl korrekt zugeordneter Fälle. Die Treffsicherheit T kann dabei über die Höhe der bedingten Klassenzuordnungswahrscheinlichkeiten geschätzt werden. Eine Person v mit dem Antwortmuster a_v wird der Klasse mit der höchsten bedingten Klassenzuordnungswahrscheinlichkeit zugewiesen (im ► Beispiel 22.2 wurde die Person „Fritz“ der Klasse 1 zugewiesen, ► Abschn. 22.2.4). Je höher nun diese maximalen bedingten Klassenzuordnungswahrscheinlichkeiten sind, desto treffsicherer dürfte die Klassenzuordnung insgesamt sein. Daher ist die Treffsicherheit T definiert als die durchschnittliche Höhe der höchsten bedingten Klassenzuordnungswahrscheinlichkeit $P^{\max}(g | a_v)$ über alle in der Stichprobe vorkommenden Antwortmuster (N_a) hinweg:

$$T = \frac{\sum_{v=1}^{N_a} P^{\max}(g | a_v)}{N} \quad (22.15)$$

Umgekehrt lässt sich die Wahrscheinlichkeit einer falschen Klassenzuordnung (E) wie folgt berechnen (vgl. Lazarsfeld und Henry 1968):

$$E = 1 - \sum_{v=1}^{N_a} P(a_v) \cdot P^{\max}(g | a_v) \quad (22.16)$$

Die Treffsicherheit (T) und die Wahrscheinlichkeit einer falschen Klassenzuordnung (E) ähneln in ihrer Bedeutung den Begriffen „Reliabilität“ und „Messfehler“ aus der Klassischen Testtheorie (KTT, ► Kap. 13 und 14). Auch bei der LCA ist es so, dass die Treffsicherheit ansteigt, je mehr Items der Test umfasst – vorausgesetzt, dass alle Items das gleiche kategoriale Merkmal messen.

Treffsicherheit T („hitrate“)

Wahrscheinlichkeit einer falschen Klassenzuordnung

22.3.7 Eliminierung nicht trennscharfer Items

Da die Anzahl möglicher Antwortmuster mit wachsender Itemanzahl (m) exponentiell ansteigt, ist es sinnvoll, nicht trennscharfe Items von vornherein zu vermeiden bzw. im Nachhinein aus dem Itempool zu entfernen.

In der deskriptiven Itemanalyse ist die geringe Trennschärfe eines Items ein mögliches Argument dafür, das Item zu entfernen (vgl. ► Kap. 7): Je geringer ein Item mit der Summe der übrigen Testitems korreliert, desto schlechter repräsentiert es den Gesamttest, weshalb auf das Item verzichtet werden kann. Ein ähnliches Vorgehen kann man auch auf die LCA anwenden. Ein trennscharfes Item zeichnet sich dadurch aus, dass sich die bedingte Bejahungswahrscheinlichkeit zwischen den unterschiedlichen latenten Klassen stark unterscheidet. Ein nicht trennscharfes Item hätte hingegen annähernd gleiche bedingte Bejahungswahrscheinlichkeiten in allen latenten Klassen und würde nur wenig oder nichts zur Treffsicherheit beitragen, mit der von dem Antwortmuster einer Person auf ihre Klassenzugehörigkeit geschlossen werden kann.

Trennschärfe eines Items

Diskriminationsindex

Rost (2004) schlägt daher einen *Diskriminationsindex* (D_i) vor, der angibt, wie groß die Unterschiedlichkeit (Varianz) der erwarteten Itemantworten zwischen den verschiedenen latenten Klassen ist, wobei diese „Zwischenklassenvarianz“ an der Unterschiedlichkeit (Varianz) der Itemantworten innerhalb einer Klasse, summiert über die Klassen hinweg, relativiert wird. Man kann zeigen, dass durch die Eliminierung von Items mit geringem Diskriminationsindex die Treffsicherheit der Klassenzuordnung vergrößert werden kann. Auf der anderen Seite führt die Eliminierung von Items unter Umständen dazu, dass die Treffsicherheit (analog quasi die Reliabilität) sinkt. Das Dilemma ist also das gleiche wie bei der Itemselektion im Rahmen der KTT.

22.4 Exploratorische und konfirmatorische Anwendungen der LCA

Exploratorisch vs. konfirmatorisch

Ohne irgendwelche Restriktionen bezüglich der Struktur der Antwortwahrscheinlichkeiten ist die LCA ein struktursuchendes bzw. -entdeckendes Verfahren und insoweit exploratorisch (► Abschn. 22.4.1). Lediglich die Anzahl der Klassen muss von vornherein spezifiziert werden, alle anderen Größen ergeben sich im Zuge der Parameterschätzung aus den Daten. A priori formulierte hypothetische Annahmen darüber, wie sich die Klassen möglicherweise in Bezug auf das Antwortverhalten unterscheiden, sind nicht direkt testbar. Werden bei der Schätzung der bedingten Antwortmusterwahrscheinlichkeiten hingegen bestimmte Randbedingungen (Restriktionen) eingeführt, z. B. theoretisch begründete Erwartungen bezüglich der Klassenlösung bzw. der Modellparameter, so kann auch inferenzstatistisch getestet werden, ob die Randbedingungen zutreffen oder nicht. Solche Anwendungen der LCA sind konfirmatorisch (► Abschn. 22.4.2).

22.4.1 Exploratorische Anwendungen der LCA: Finden des besten Modells

Indirekter Vergleich von Modellen mit unterschiedlicher Klassenanzahl

Bevor die Frage beantwortet werden kann, wie sich die Testpersonen der latenten Klassen hinsichtlich ihres Antwortverhaltens unterscheiden, muss zunächst geklärt werden, wie viele Klassen sinnvollerweise angenommen werden sollen. Diese Frage kann über einen „indirekten“ Vergleich verschiedener Modelle mit unterschiedlicher Klassenanzahl beantwortet werden. Die Werkzeuge, die für einen deskriptiven Modellvergleich zur Verfügung stehen, wurden bereits in ► Abschn. 22.3 vorgestellt (z. B. Informationskriterien, ► Abschn. 22.3.5). Man kann sie auch für einen indirekten Modellvergleich nutzen. Hierbei geht man so vor, dass man die entsprechenden Modellfit-Indizes (also L^2 , χ^2 , Informationskriterien etc.) für jedes Modell bestimmt und sich dann für das Modell mit dem besten Fit (also beispielsweise dem niedrigsten BIC-Wert) entscheidet. Im Falle von L^2 und χ^2 können – ausreichend große Stichproben vorausgesetzt – auch die p -Werte interpretiert werden: Modelle, bei denen die Nullhypothese nicht verworfen werden musste (üblicherweise $p \geq .05$), sind gegenüber Modellen, bei denen sie verworfen wurde (üblicherweise $p < .05$), zu bevorzugen. Simulationsstudien haben gezeigt, dass ein indirekter Modellvergleich anhand des BIC bzw. des aBIC am ehesten geeignet sind, das am besten passende Modell zu identifizieren (s. auch Hagenaars und McCutcheon 2002).

Ein solcher Vergleich von Modellen mit unterschiedlicher Klassenanzahl wird in ► Beispiel 22.3 beschrieben.

Beispiel 22.3: Wie viele „Geschlechtsrollentypen“ lassen sich empirisch unterscheiden:

Die fiktive (!) Forscherin Gisela Kreuzwald habe mit $n = 2350$ Studierenden eine Befragung zum Thema „Geschlechtsrollen“ anhand von $m = 6$ Testitems durchgeführt. Es sollte die Frage geklärt werden, mit wie vielen latenten Klassen die Unterschiede im Antwortverhalten der Testpersonen am zutreffendsten erklärt werden können.

Hierzu wurden vier LCA-Modelle berechnet: Ein Ein-Klassen-Modell, ein Zwei-Klassen-Modell, ein Drei-Klassen-Modell und ein Vier-Klassen-Modell. In folgender Tabelle seien für jedes Modell die Anzahl der zu schätzenden Modellparameter (t), die logarithmierte Likelihood ($\ln(L)$), der L^2 -Wert, der χ^2 -Wert inklusive df , das Ergebnis eines Bootstrap-Tests mit jeweils 100 resimulierten Datensätzen sowie die Informationskriterien AIC, BIC, aBIC und CAIC angegeben:

Modell	Ein-Klassen-Modell	Zwei-Klassen-Modell	Drei-Klassen-Modell	VierKlassen-Modell
t	6	13	20	27
$\ln(L)$	−4499.97	−4148.68	−4080.85	−4077.23
<i>Inferenzstatistische Absicherung der Modellgüte</i>				
L^2	888.33 $p < .001$	185.76 $p < .001$	50.10 $p = .21$	42.85 $p = .20$
χ^2	4770.41 $p < .001$	6142.50 $p < .001$	68.33 $p < .01$	57.96 $p = .01$
df	57	50	43	36
Bootstrap-Test	$p < .001$	$p < .001$	$p = .03$	$p = .10$
<i>Informationskriterien</i>				
AIC	9011.94	8323.37	8201.71	8208.46
BIC	9046.51	8398.28	8316.95	8364.04
aBIC	9027.45	8356.96	8253.40	8278.25
CAIC	9052.51	8411.28	8336.95	8391.04

Anhand der inferenzstatistischen Kriterien sollten das Ein- und das Zwei-Klassen-Modell in jedem Fall verworfen werden, da die jeweiligen Parameter hoch signifikant von null abweichen. Das Drei- und das Vier-Klassen-Modell passen wesentlich besser zu den Daten. Anhand des Bootstrap-Tests müsste man das Drei-Klassen-Modell ebenfalls verwerfen. Die Informationskriterien sind allerdings für das Drei-Klassen-Modell am günstigsten. Gisela Kreuzwald gewichtet das Modellsparsumskriterium höher als das inferenzstatistische Ergebnis und entscheidet sich daher für das Drei-Klassen-Modell.

Das Beispiel zeigt, wie man aufgrund eines indirekten Modellvergleichs die Anzahl der latenten Klassen bestimmen kann. Aber selbst wenn ein Modell hervorragend zu den Daten passt, weiß man noch nicht, was die Klassenzuordnung inhaltlich bedeutet bzw. wie die Unterschiede zwischen den Klassen psychologisch zu interpretieren sind. Man wäre geneigt zu sagen, dass es sich bei einem Test zur Messung von Geschlechtsrollen im Falle einer Drei-Klassen-Lösung um „drei Geschlechtsrollentypen“ handelt. Aber: Möglicherweise sind es völlig andere Merk-

Interpretation der Klassenunterschiede

male, hinsichtlich derer sich die drei Klassen voneinander unterscheiden. So wäre es möglich, dass Klasse 1 aus Personen besteht, die in allen Fragebogen eher „nein“ ankreuzen, während Personen in Klasse 2 aufgrund einer dispositionellen Akquieszenzneigung (vgl. „Antworttendenzen“, s. ▶ Kap. 4, ▶ Abschn. 4.3.3) viele Items eher mit „ja“ ankreuzen. Personen in Klasse 3 könnten sich schließlich durch ein stereotyped Antwortmuster auszeichnen (z. B. $a_v = \langle 0, 1, 0, 1, 0, 1 \rangle$). Dieses Extrembeispiel macht deutlich, dass – ähnlich wie bei der EFA (vgl. ▶ Kap. 23) – die Anzahl der latenten Klassen noch nichts darüber aussagt, ob sich die Klassen in Bezug auf die inhaltlich vermuteten bzw. theoretisch relevanten Merkmale unterscheiden. Im ungünstigsten Fall handelt es sich um Merkmale, die mit der eigentlich interessierenden latenten Personenvariablen (hier: Geschlechtsrollen) überhaupt nichts zu tun haben. Genau wie bei der EFA ist man im Falle der exploratorischen LCA gezwungen, Unterschiede zwischen den latenten Klassen durch Inspektion der bedingten Antwortwahrscheinlichkeiten per Augenschein zu beurteilen und anschließend an externen Kriterien zu validieren (▶ Beispiel 22.4).

Beispiel 22.4: Wie sind Klassenunterschiede zu interpretieren?

In unserem Beispieldatensatz wurden für die Drei-Klassen-Lösung folgende klasse-spezifische Bejahungswahrscheinlichkeiten P_{ig} für Item 1 bis 6 ermittelt bzw. berechnet:

(P_{ig})	Femininitätsitems			Maskulinitätsitems		
	Item 1 (P_{1g})	Item 2 (P_{2g})	Item 3 (P_{3g})	Item 4 (P_{4g})	Item 5 (P_{5g})	Item 6 (P_{6g})
Klasse 1 ($\pi_1 = 73\%$)	.11	.04	.08	.02	.01	.02
Klasse 2 ($\pi_2 = 25\%$)	.85	.44	.65	.02	.02	.00
Klasse 3 ($\pi_3 = 2\%$)	.14	.08	.13	.49	.56	.66

Klasse 1 umfasst mit 73 % die große Mehrheit der Stichprobe. Personen dieser Klasse zeichnen sich durch niedrige Bejahungswahrscheinlichkeiten bei allen sechs Items aus. Personen der Klasse 2 neigen dazu, die Items 1, 2 und 3 eher zu bejahen; die Items 4, 5 und 6 lehnen sie mit großer Wahrscheinlichkeit ab. Klasse 3 besteht aus Personen, die den Items 1, 2 und 3 im Vergleich zu den Items 4, 5 und 6 weniger zustimmen. In Anlehnung an die Typologie von Spence et al. (1975) könnte man Klasse 1 als Personen des „undifferenzierten“, Klasse 2 als Personen des „femininen“ und Klasse 3 als Personen des „maskulinen“ Typs interpretieren. Ob dem so ist, kann streng genommen nur über eine konvergente und diskriminante Validierung mit externen Kriterien überprüft werden (vgl. ▶ Kap. 21).

22.4.2 Konfirmatorische Anwendungen der LCA: Testen von Modellrestriktionen

Restriktionsformen

In konfirmatorischen Anwendungen der LCA werden begründete Annahmen über die Struktur der Unterschiede zwischen den latenten Klassen eingeführt. Diese Strukturannahmen münden in Bedingungen oder Einschränkungen im Wertebereich bei der Schätzung der Modellparameter. Allgemein gebräuchliche Formen der Parameterrestriktion bestehen im

22.4 · Exploratorische und konfirmatorische Anwendungen der LCA

- Fixieren von Parametern auf einen bestimmten Wert,
- Gleichsetzen zweier (oder mehrerer) Parameter und
- Einführen von Ordnungsrestriktionen.

Beim *Fixieren von Parametern* lassen sich sowohl die Klassengrößen als auch die bedingten Antwortwahrscheinlichkeiten auf bestimmte Werte festsetzen. Bezogen auf das ► Beispiel 22.2 könnte man die bedingte Bejahungswahrscheinlichkeit der Items 1, 2 und 3 in Klasse 1 und die der Items 4, 5 und 6 in Klasse 2 jeweils auf .90 festsetzen. Alle anderen Parameter würden dann frei geschätzt.

Beim *Gleichsetzen von Parametern* werden keine konkreten Werte für die Parameter vorgegeben, sondern es wird lediglich verfügt, dass bestimmte Parameter gleich sein müssen. So wäre es möglich, in der LCA zwei gleich große Klassen zu erzwingen ($\pi_1 = \pi_2 = .5$). Eine solche Restriktion könnte man als das qualitative Pendant eines Median-Splits (bei einer kontinuierlichen, unidimensionalen Personenvariablen, ► Abschn. 22.1.1) bezeichnen. Auch bedingte Antwortwahrscheinlichkeiten können innerhalb und zwischen Klassen gleichgesetzt werden (► Beispiel 22.5).

Beispiel 22.5: Gleiche Antwortwahrscheinlichkeiten in den Klassen

Bei einer Zwei-Klassen-Lösung wird folgende Restriktion eingeführt: Zum einen müssen die drei Femininitätsitems in Klasse 1 identische Bejahungswahrscheinlichkeiten haben, zum anderen müssen auch die drei Maskulinitätsitems in Klasse 2 identische Bejahungswahrscheinlichkeiten haben (diese Parameter sind in der nachfolgenden Tabelle jeweils mit * gekennzeichnet). Formal lassen sich die beiden Gleichheitsrestriktionen wie folgt ausdrücken: $\{P_{11} = P_{21} = P_{31}\}$ und $\{P_{42} = P_{52} = P_{62}\}$. Alle anderen Parameter werden frei geschätzt. Eine solche Restriktion führt zu folgender Parameterschätzung:

(P_{ig})	Femininitätsitems			Maskulinitätsitems		
	Item 1 (P_{1g})	Item 2 (P_{2g})	Item 3 (P_{3g})	Item 4 (P_{4g})	Item 5 (P_{5g})	Item 6 (P_{6g})
Klasse 1 ($\pi_1 = 68\%$)	.05*	.05*	.05*	.03	.03	.04
Klasse 2 ($\pi_2 = 32\%$)	.79	.35	.56	.01*	.01*	.01*

Angesichts der Parameter, die in ► Beispiel 22.4 abgebildet waren, verwundert das Ergebnis nicht: Für Klasse 1 werden die Bejahungswahrscheinlichkeiten aller Items sehr niedrig geschätzt. Klasse 2 hat lediglich in Bezug auf die Items 1, 2 und 3 hohe Bejahungswahrscheinlichkeiten.

Bei *Ordnungsrestriktionen* besteht die Möglichkeit, eine bestimmte Ordnungsrelation der Parameter zu erzwingen. Beispielsweise könnte man verfügen, dass die Antwortwahrscheinlichkeiten in Klasse 1 auf jedem Item höher liegen als die Antwortwahrscheinlichkeiten in Klasse 2. In solchen Fällen würde man zwei geordnete Klassen erzwingen. Eine sinnvolle Ordnungsrestriktion könnte – angewandt auf das Geschlechtsrollenbeispiel – etwa darin bestehen, dass die Bejahungswahrscheinlichkeiten für die Femininitätsitems (1, 2 und 3) in Klasse 1 höher sind als in Klasse 2, während für die Maskulinitätsitems (4, 5 und 6) das umgekehrte Muster erwartet wird.

Fixierungsrestriktionen

Gleichheitsrestriktionen

Ordnungsrestriktionen

22.4.3 Modellvergleichstests

Nested Models

Restringierte Modelle

Likelihood-Quotienten-Test

Nullhypothese; Beide Modelle passen gleich gut

Neben der Möglichkeit, verschiedene Klassenlösungen anhand unterschiedlicher Maße deskriptiv miteinander zu vergleichen (► Abschn. 22.3), kann man zwei Modelle auch direkt gegeneinander testen, sofern die beiden Modelle durch Modellrestriktionen ineinander überführt werden können. Man spricht hier auch von „verschachtelten Modellen“ („nested models“). In diesem Fall kann eine restriktivere Modellvariante gegen eine weniger restringierte oder eine unrestringierte Modellvariante getestet werden.

Beispielsweise könnte man ein restringiertes Modell mit einer Fixierungs-, einer Gleichheits- oder einer Ordnungsrestriktion gegen ein unrestringiertes bzw. weniger restriktives Modell testen. Hierzu wird die Likelihood des restringierten Modells (L_1) durch die Likelihood des unrestringierten bzw. des weniger restriktiven Modells (L_2) geteilt. Man erhält also – ähnlich wie bereits in ► Abschn. 22.3.2 beschrieben – einen Likelihood-Quotienten (LR):

$$LR = \frac{L_1}{L_2} \quad (22.17)$$

Restriktionen verschlechtern im Allgemeinen die Modellanpassung, denn sie bilden theoretische Vorstellungen ab, die mit den Daten nur selten vollständig kompatibel sind. Insofern ist die Likelihood eines restringierten Modells L_1 in den allermeisten Fällen kleiner als die Likelihood eines weniger restriktiven Modells L_2 . Würde die Restriktion exakt den empirischen Gegebenheiten entsprechen, so wären L_1 und L_2 identisch und der Likelihood-Quotient wäre $LR = 1$. Je mehr der Likelihood-Quotient (LR) unterhalb von 1 liegt, desto schlechter ist die Anpassung des restringierten Modells. Bereits bekannt ist (Gl. 22.9), dass der Likelihood-Quotient in einen Wert L^2 umgerechnet werden kann, der im Falle großer Stichproben approximativ χ^2 -verteilt ist. Dieser χ^2 -Test stellt die Basis für den inferenzstatistischen Modellvergleich dar: Die Nullhypothese des Tests lautet, dass beide Modelle gleich gut auf die Daten passen. Muss diese Nullhypothese (gegeben eine vorher festgelegte Irrtumswahrscheinlichkeit) abgelehnt werden, ist das restringierte Modell bedeutsam schlechter als das unrestringierte Modell. In diesem Fall sollte die Restriktion verworfen werden (► Beispiel 22.6).

Beispiel 22.6: Direkte Testung eines Modells mit Gleichheitsrestriktionen

Das in ► Beispiel 22.5 (► Abschn. 22.4.2) dargestellte restringierte Zwei-Klassen-Modell hat eine Log-Likelihood von $\ln(L_1) = -4160.21$. Das unrestringierte Zwei-Klassen-Modell hat eine Log-Likelihood von $\ln(L_2) = -4148.68$ (► Beispiel 22.3). Der L^2 -Wert kann nun – in Anlehnung an Gl. (22.9) – wie folgt berechnet werden:

$$\begin{aligned} L^2 &= 2 \cdot (\ln(L_2) - \ln(L_1)) \\ &= 2 \cdot (-4148.68 - (-4160.21)) = 23.06 \end{aligned}$$

Die Anzahl der Freiheitsgrade der χ^2 -Statistik ergibt sich aus der Differenz der Anzahl der zu schätzenden Modellparameter: Im unrestringierten Zwei-Klassen-Modell werden eine Klassengröße und 12 bedingte Antwortwahrscheinlichkeiten geschätzt ($t_1 = 13$). Im restringierten Modell hingegen werden aufgrund der Gleichheitsrestriktion eine Klassengröße, aber nur 8 Antwortwahrscheinlichkeiten geschätzt ($t_2 = 9$), denn mit der Schätzung von P_{11} liegen auch P_{21} und P_{31} fest (man hat also zwei Freiheitsgrade eingespart), und mit der Schätzung von P_{42} liegen automatisch auch P_{52} und P_{62} fest (man hat also zwei weitere Freiheitsgrade eingespart).

Die χ^2 -Statistik des Likelihood-Quotienten-Tests hat demnach $df = t_1 - t_2 = 4$ Freiheitsgrade. Ein Wert von $\chi^2_{df=4} = 23.06$ (oder jeder größere) hat der Nullhypothese (d. h. keine Abweichung zwischen L_1 und L_2) zufolge eine Wahrscheinlichkeit von $p = .0001$. Die Nullhypothese muss abgelehnt werden; die eingeführte Gleichheitsrestriktion verschlechtert die Modellanpassung bedeutsam und sollte demnach verworfen werden.

Abschließend sei darauf hingewiesen, dass ein direkter Vergleich zweier Modelle, die sich in der Anzahl der latenten Klassen unterscheiden, anhand des Likelihood-Quotienten-Tests nicht möglich ist. Zwar ist beispielsweise ein Zwei-Klassen-Modell restriktiver als ein Drei-Klassen-Modell, da alle Parameter der dritten Klasse (also π_3 und $P(y_{vi} | g_3)$) gleich null gesetzt werden (insofern handelt es sich bei zwei Modellen, die sich nur in der Anzahl der latenten Klassen unterscheiden, formal ebenfalls um verschachtelte Modelle), allerdings ist der Likelihood-Quotient dieser beiden Modelle nicht χ^2 -verteilt (McLachlan und Peel 2000). Ein entsprechender Modellvergleich muss daher entweder indirekt vorgenommen werden (► Abschn. 22.4.1) oder anhand einer mittels Bootstrap gewonnenen Prüfverteilung (McLachlan und Peel 2000; s. auch Nylund et al. 2007).

Direkter Vergleich von Modellen mit unterschiedlicher Klassenanzahl

22.5 Erweiterte Anwendungen der LCA

Alle LCA-Modelle, die in den vorangegangenen Abschnitten besprochen wurden, gingen von der Annahme aus, dass die latente Personenvariable qualitativ ist und dass alle Items ein dichotomes Antwortformat haben. Im folgenden Abschnitt werden zwei erweiterte Anwendungen der LCA behandelt. In ► Abschn. 22.5.1 wird gezeigt, wie sich die LCA auf Items mit mehreren nominalen (d. h. polytomen) Antwortkategorien generalisieren lässt. In ► Abschn. 22.5.2 wird gezeigt, wie die LCA mit dem Rasch-Modell kombiniert werden kann. Die resultierenden latenten Mischverteilungsmodelle gehen von der Annahme aus, dass es in der Stichprobe eine endliche Anzahl latenter Klassen gibt, in denen jeweils ein spezifisches Rasch-Modell gilt.

22.5.1 LCA für polytome Antwortformate

Im Falle dichotomer Items gibt es nur zwei Antwortkategorien ($y_{vi} \in \{0, 1\}$). Im Falle polytomer Items gibt es mehr als zwei Antwortkategorien ($y_{vi} \in \{0, \dots, k, \dots, K-1\}$), man kann also nicht mehr von Bejahungs- oder Verneinungswahrscheinlichkeiten sprechen. Vielmehr hat jede der K Antwortkategorien eine eigene Kategorienwahrscheinlichkeit. Die Wahrscheinlichkeit, mit der eine Person v bei Item i die Antwortkategorie k wählt, wird als $P(y_{vi} = k)$ bezeichnet. Die Wahrscheinlichkeit, mit der eine Person v bei Item i die Antwortkategorie k ankreuzt unter der Bedingung, dass die entsprechende Person einer latenten Klasse g angehört (bedingte Kategorienwahrscheinlichkeit), wird als $P(y_{vi} = k | g)$ bezeichnet. Es kann also für jede Antwortkategorie k eine bedingte Kategorienwahrscheinlichkeit berechnet werden. Über alle Kategorien hinweg addieren sich diese bedingten Kategorienwahrscheinlichkeiten zu 1 auf:

$$\sum_{k=0}^{K-1} P(y_{vi} = k | g) = 1 \quad (22.18)$$

Bedingte Kategorienwahrscheinlichkeit

Für einen Test, der aus $m = 4$ Items mit jeweils den drei Antwortkategorien „nein“ ($y_{vi} = 0$), „ja“ ($y_{vi} = 1$) und „vielleicht“ ($y_{vi} = 2$) besteht, wären theoretisch $K^m = 3^4 = 81$ Antwortmuster möglich. Die einzelnen Antwortmuster werden – genau wie bei der dichotomen LCA – mit a_v bezeichnet (► Beispiel 22.7).

Modellannahmen für die polytome LCA

Bedingte Antwortmusterwahrscheinlichkeit

Die zentralen Annahmen, die für die dichotome LCA bereits in ► Abschn. 22.2.1 bis 22.2.3 vorgestellt wurden, gelten analog auch für die polytome LCA:

- Zunächst wird angenommen, dass die bedingten Kategorienwahrscheinlichkeiten für alle Personen innerhalb einer latenten Klasse gleich sind (► Abschn. 22.2.1). Daher kann man für die bedingten Kategorienwahrscheinlichkeiten den Index v weglassen und diese kürzer schreiben als $P(y_{vi} = k | g) = P_{ikg}$.
- Ferner wird angenommen, dass für alle Items innerhalb einer latenten Klasse die Annahme lokaler stochastischer Unabhängigkeit (► Abschn. 22.2.2) erfüllt ist. Daher kann die bedingte Antwortmusterwahrscheinlichkeit in Anlehnung an Gl. (22.3) wie folgt ausgedrückt werden:

$$P(a_v | g) = \prod_{i=1}^m P_{ikg} \quad (22.19)$$

Unbedingte Antwortmusterwahrscheinlichkeit

- Schließlich wird angenommen, dass die latenten Klassen disjunkt und exhaustiv sind (► Abschn. 22.2.3). Damit kann die unbedingte Antwortmusterwahrscheinlichkeit in Anlehnung an Gl. (22.5) wie folgt ausgedrückt werden:

$$P(a_v) = \sum_{g=1}^G \pi_g \prod_{i=1}^m P_{ikg} \quad (22.20)$$

Beispiel 22.7: Berechnung von bedingten Klassenwahrscheinlichkeiten

Für ein Modell mit $G = 2$ Klassen bei einem Test mit $m = 4$ Items mit jeweils $K = 3$ Antwortkategorien seien folgende bedingte Kategorienwahrscheinlichkeiten ermittelt worden:

(P_{ikg})	Item 1 (P_{1kg})	Item 2 (P_{2kg})	Item 3 (P_{3kg})	Item 4 (P_{4kg})
Klasse 1 $(\pi_1 = 31\%)$	$y_{v1} = 0: P_{101} = .65$ $y_{v1} = 1: P_{111} = .22$ $y_{v1} = 2: P_{121} = .13$	$y_{v2} = 0: P_{201} = .83$ $y_{v2} = 1: P_{211} = .07$ $y_{v2} = 2: P_{221} = .10$	$y_{v3} = 0: P_{301} = .21$ $y_{v3} = 1: P_{311} = .66$ $y_{v3} = 2: P_{321} = .13$	$y_{v4} = 0: P_{401} = .02$ $y_{v4} = 1: P_{411} = .12$ $y_{v4} = 2: P_{421} = .86$
Klasse 2 $(\pi_2 = 69\%)$	$y_{v1} = 0: P_{102} = .32$ $y_{v1} = 1: P_{112} = .20$ $y_{v1} = 2: P_{122} = .48$	$y_{v2} = 0: P_{202} = .09$ $y_{v2} = 1: P_{212} = .11$ $y_{v2} = 2: P_{222} = .80$	$y_{v3} = 0: P_{302} = .26$ $y_{v3} = 1: P_{312} = .14$ $y_{v3} = 2: P_{322} = .60$	$y_{v4} = 0: P_{402} = .40$ $y_{v4} = 1: P_{412} = .11$ $y_{v4} = 2: P_{422} = .49$

Eine der Testpersonen in der Stichprobe, Gerlinde, weist das Antwortmuster $a_{\text{Gerlinde}} = \langle 1, 0, 0, 2 \rangle$ auf. Welcher latenten Klasse gehört Gerlinde am ehesten an? Diese Frage wird anhand der bedingten Klassenzuordnungswahrscheinlichkeiten $P(g = 1 | a_{\text{Gerlinde}})$ und $P(g = 2 | a_{\text{Gerlinde}})$ beantwortet. Gerlinde wird jener Klasse zugeordnet, für die ihre bedingte Zuordnungswahrscheinlichkeit maximal ist. Um diese berechnen zu können, benötigt man noch die unbedingte Antwortmusterwahrscheinlichkeit für das Muster $a_{\text{Gerlinde}} = \langle 1, 0, 0, 2 \rangle$. Diese berechnet sich nach Gl. (22.20), wobei die bedingten Kategorienwahrscheinlichkeiten bereits in der oben

stehenden Tabelle abgetragen sind:

$$\begin{aligned}
 P(a_{\text{Gerlinde}}) &= \sum_{g=1}^G \pi_g \cdot P(a_{\text{Gerlinde}}|g) \\
 &= \sum_{g=1}^2 \pi_g \cdot P_{11g} \cdot P_{20g} \cdot P_{30g} \cdot P_{42g} \\
 &= 0.31 \cdot (0.22 \cdot 0.83 \cdot 0.21 \cdot 0.86) + 0.69 \cdot (0.20 \cdot 0.09 \cdot 0.26 \cdot 0.49) \\
 &= 0.0102 + 0.0016 = 0.012
 \end{aligned}$$

Nun können in Anlehnung an Gl. (22.6) (Bayes-Theorem) die bedingten Klassenzuordnungswahrscheinlichkeiten berechnet werden:

$$\begin{aligned}
 P(g = 1|a_{\text{Gerlinde}}) &= \frac{0.31 \cdot 0.03}{0.012} = 0.87 \\
 P(g = 2|a_{\text{Gerlinde}}) &= \frac{0.69 \cdot 0.002}{0.012} = 0.13
 \end{aligned}$$

Die Wahrscheinlichkeit, dass Gerlinde der ersten Klasse angehört, beträgt 87%; die Wahrscheinlichkeit, dass sie der zweiten Klasse angehört, beträgt hingegen nur 13 %. Es liegt also nahe, Gerlinde der ersten Klasse zuzuordnen.

22.5.2 Mischverteilungs-Rasch-Modelle

Bei der LCA ist die latente Personenvariable qualitativ, d. h. kategorial: Gesucht wird die Wahrscheinlichkeit, mit der eine Person v mit dem Antwortmuster a_v einer von G latenten Klassen angehört. Das Rasch-Modell, das von Kelava und Moosbrugger in ► Kap. 16 besprochen wurde, geht hingegen von einer quantitativen kontinuierlichen latenten Personenvariablen aus. Beide Ansätze lassen sich auch kombinieren, wie im Folgenden gezeigt wird.

Das Rasch-Modell macht eine sehr restriktive Annahme, nämlich dass die Schwierigkeit eines Items für alle Personen in der Stichprobe identisch sein muss. Dies ist eine Implikation des Modells, die ihrerseits mit anderen Implikationen des Rasch-Modells, z. B. mit der, dass die Summe der bejahten Items – insofern Modellkonformität nachgewiesen wurde – eine ausreichende (suffiziente) Statistik für die Schätzung des latenten Personenparameters darstellt, untrennbar verbunden ist.

Man kann sich jedoch leicht Fälle vorstellen, bei denen die Annahme identischer Itemschwierigkeiten verletzt ist. Ein Beispiel: Petra und Michael sind in exakt gleicher Weise extravertiert, d. h., sie haben die gleiche Ausprägung in der latenten Personenvariablen Extraversion. Beide sollen zu dem Extraversionsitem „Ich fühle mich in Gesellschaft anderer Leute wohl“ durch Bejahren oder Verneinen Stellung nehmen. Petra denkt hier eher an Situationen mit guten Freunden oder Familienmitgliedern, in deren Gegenwart sie sich wohl fühlt, und bejaht das Item dementsprechend. Michael hingegen denkt eher an Situationen mit einer Menge fremder Leute, also einem Stehempfang des Präsidenten seiner Universität o. Ä., und verneint das Item. Die unterschiedlichen Bejahungswahrscheinlichkeiten haben in diesem Beispiel also nichts mit einem „wahren“ Unterschied in Bezug auf Extraversion zu tun. Vielmehr bedeutet eine Bejahung für Michael etwas ganz anderes als für Petra. Anders gesagt: Das Item ist für Michael schwieriger als für Petra.

Kombination von Modellen mit qualitativen und quantitativen Variablen

Das Rasch-Modell ist sehr restriktiv

Response Sets

Unterschiede im Antwortverhalten zwischen Personen können viele Gründe haben: Einige Personen neigen habituell dazu, Items zu bejahen (Akquieszenzneigung). Andere Personen neigen – im Falle eines ordinalen Antwortformats – dazu, die Extreme der Antwortskala zu vermeiden (Tendenz zur Mitte). Wieder andere Personen haben sich über das erfragte Verhalten noch nie Gedanken gemacht und neigen vorsichtigerweise dazu, das Item zu verneinen (niedrige Konzeptklarheit) usw. Solche Unterschiede bei der Nutzung der Antwortkategorien werden als Antworttendenzen (*Response Sets*) bezeichnet und werden von Moosbrugger und Brandt in ► Kap. 4 besprochen.

Man könnte nun annehmen, dass es in der Stichprobe eine endliche Anzahl (latenter) Klassen gibt, die sich hinsichtlich solcher Antworttendenzen unterscheiden. Innerhalb einer Klasse haben die Personen dabei die gleiche Antworttendenz. Anders gesagt: Innerhalb jeder Klasse wird versucht, jeweils ein eigenes Rasch-Modell anzupassen. Die bisherige nicht klassenspezifische Form des Rasch-Modells lautete (vgl. ► Kap. 16):

$$P(y_{vi} = 1) = \frac{e^{(\eta_v - \beta_i)}}{1 + e^{(\eta_v - \beta_i)}} \quad (22.21)$$

Diese allgemeine Formulierung wird nun um einen weiteren Parameter, die Klassenzugehörigkeit g , erweitert. Die bedingte (klassenspezifische) Wahrscheinlichkeit, mit der eine Person v ein Item i bejaht, wenn sie der Klasse g angehört, lässt sich also wie folgt ausdrücken:

$$P(y_{vi} = 1|g) = \frac{e^{(\eta_{vg} - \beta_{ig})}}{1 + e^{(\eta_{vg} - \beta_{ig})}} \quad (22.22)$$

Allgemeine Modellgleichung des Mischverteilungs-Rasch-Modells

Unter der Annahme, dass die latenten Klassen disjunkt und exhaustiv sind, sich die relativen Klassengrößen also zu 1 aufaddieren, lässt sich für das sog. „Mischverteilungs-Rasch-Modell“ („mixed Rasch model“) für dichotome Antwortformate die folgende allgemeine Modellgleichung formulieren:

$$P(y_{vi} = 1) = \sum_{g=1}^G \left[\pi_g \frac{e^{(\eta_{vg} - \beta_{ig})}}{1 + e^{(\eta_{vg} - \beta_{ig})}} \right] \quad (22.23)$$

Gibt es nur eine latente Klasse in der Stichprobe, entfallen alle Klassenindizes g und das Modell entspricht dem einfachen Rasch-Modell. Sind umgekehrt die Personenparameter innerhalb jeder Klasse für alle Personen gleich, so entspricht das Modell einer einfachen LCA. Insofern stellt das Mischverteilungs-Rasch-Modell ein gemeinsames Obermodell von LCA und dichotomem Rasch-Modell dar (Rost 2004).

Die Kombination von IRT-Modellen mit LCA-Ansätzen ist nicht auf dichotome Itemformate beschränkt: Auch für nominale oder ordinale Formate lassen sich entsprechende Mischverteilungs-Rasch-Modelle konstruieren. Ein besonderer Vorteil der LCA ist hier, dass nicht a priori bekannt sein muss, was die latenten Klassen hinsichtlich ihres Antwortformats genau unterscheidet. Vielmehr lassen sich Mischverteilungs-Rasch-Modelle explorativ dazu nutzen, qualitative Unterschiede zwischen Personengruppen hinsichtlich der Itembeantwortung zu untersuchen (► Beispiele 22.8).

Beispiele 22.8: Beispiele aus der Persönlichkeits- und klinischen Diagnostik**Beispiel A: Persönlichkeitsdiagnostik**

Rost (1997) berichtet beispielsweise von einer Analyse der Skala „Gewissenhaftigkeit“ im NEO-Fünf-Faktoren-Inventar (NEO-FFI; Costa und McCrae 1992), in der zwei Klassen von Personen identifiziert werden konnten ($\pi_1 = 65.2\%$; $\pi_2 = 34.8\%$). Personen der ersten (größeren) Klasse nutzten die Antwortskala eher im Einklang mit der zu messenden latenten Variablen. Personen der zweiten Klasse wiesen sich hingegen dadurch aus, dass sie eher zu extremen Urteilen in Bezug auf alle Items neigten und die zweite Antwortkategorie (bezeichnet mit „stimme nicht zu“) gar nicht erst verwendeten. Die Befunde deuten darauf hin, dass Personen der zweiten Klasse das Antwortformat der Items anders verwendeten als jene der ersten Klasse. Für sie das gleiche Antwortverhalten (Testmodell) anzunehmen wie für die erste Klasse, wäre diagnostisch problematisch.

Beispiel B: Klinische Diagnostik

Auch für die klinische Diagnostik können solche qualitativen Unterschiede bedeutsam sein: Beispielsweise neigen Personen in klinischen Stichproben dazu, sich entweder als besser darzustellen, indem sie ihre Leiden und Probleme herunterspielen („faking good“) oder als wesentlich schlimmer darzustellen, als sie eigentlich sind („faking bad“). So fanden Gollwitzer et al. (2005) für die dispositionelle Neigung, Ärger in sich hineinzufressen (anstatt ihn entweder kontrolliert oder unkontrolliert auszuleben), gemessen durch die Skala „Anger-In“ des State-Trait-Ärgerausdrucks-Inventars (STAXI; Schwenkmezger et al. 1992), bei weiblichen, stationär behandelten Patientinnen eine Drei-Klassen-Struktur: Frauen der ersten Klasse ($\pi_1 = 48\%$) nutzten die gesamte Antwortskala und wiesen keine Auffälligkeiten in ihrem Antwortverhalten auf. Frauen der zweiten Klasse ($\pi_2 = 27\%$) zeichneten sich durch ein Antwortmuster aus, das eher auf eine Neigung zur sozialen Erwünschtheit schließen ließ. Frauen der dritten Klasse schließlich ($\pi_3 = 25\%$) neigten eher zu einer Extremisierung ihrer Anger-In-Tendenz. Dieser Befund hat für die klinische Einzelfalldiagnostik eine wichtige Implikation: Würde man aus jeder der drei Klassen jeweils eine Patientin mit identischen Rohwerten in der Skala „Anger-In“ des STAXI miteinander vergleichen, so wären ihre „wahren“ latenten Merkmalsausprägungen dennoch unterschiedlich, denn die „wahre“ Anger-In-Neigung wäre bei einer Patientin aus der zweiten Klasse höher als bei einer Patientin aus der dritten Klasse. Durch die Schätzung klassenspezifischer Personenparameter werden solche klassenbedingten Unterschiede in den Rohwerten also quasi korrigiert.

**NEO-Fünf-Faktoren-Inventar
(NEO-FFI)****State-Trait-Ärgerausdrucks-Inventar
(STAXI)**

22.6 Zusammenfassung

Während Latent-Trait-Modelle auf der Annahme beruhen, dass es sich bei dem zu messenden latenten Personenmerkmal um eine quantitative Variable handelt, sind Latent-Class-Modelle immer dann geeignet, wenn das latente Personenmerkmal qualitativer Natur ist (beispielsweise die Zugehörigkeit zu einem bestimmten Persönlichkeitstyp). Mit einer LCA kann die Wahrscheinlichkeit ermittelt werden, mit der eine Person v , die auf m Items ein Antwortmuster a_v produziert, einer bestimmten latenten Klasse g angehört. Die Anzahl der latenten Klassen in der Population (G) ist unbekannt und muss theoriegeleitet vorgegeben oder empirisch über einen Vergleich mehrerer Modelle mit unterschiedlicher Anzahl Klassen ermittelt wer-

den. Alle anderen Modellparameter können modellimmanent, d. h. empirisch aus den Daten geschätzt werden.

Im Sinne der Modellannahmen wird Folgendes vorausgesetzt:

1. Die Antwortwahrscheinlichkeiten auf den m Items müssen für alle Personen innerhalb einer latenten Klasse identisch sein;
2. innerhalb einer latenten Klasse muss die Annahme der lokalen stochastischen Unabhängigkeit erfüllt sein und
3. die latenten Klassen müssen disjunkt und exhaustiv sein.

Die Anwendung einer LCA ist insbesondere dann sinnvoll, wenn

1. das Ziel der Analyse die Klassifikation von Personen ist,
2. es nicht möglich oder sinnvoll ist, über Items hinweg Summenwerte zu bilden, sondern lediglich die Antwortmuster (Profile) ausgewertet werden, oder
3. das Ziel der Analyse darin besteht, ein bestimmtes typologisches Modell zu testen (z. B. ob die Unterschiede im Antwortverhalten der Testpersonen auf die theoretisch vermuteten Persönlichkeitstypen zurückführbar sind).

Die Güte eines LCA-Modells kann mithilfe eines Likelihood-Ratio-Tests, eines „klassischen“ χ^2 -Tests, eines Bootstrap-Verfahrens oder anhand von Informationskriterien beurteilt werden.

Einem LCA-Modell können bestimmte Restriktionen (z. B. Fixierungs-, Gleichheits- oder Ordnungsrestriktionen) auferlegt werden; in diesem Fall wird aus der „exploratorischen“ eine „konfirmatorische“ Analyse. Im Fall von „nested models“ kann ein restringiertes Modell mithilfe eines Likelihood-Quotienten-Tests direkt gegen ein unrestringiertes Modell getestet werden.

22.7 EDV-Hinweise

Die Auswertung eines LCA-Modells kann entweder mit speziellen Softwareprogrammen (z. B. WINMIRA; von Davier 2001; oder Latent GOLD; Vermunt und Magidson 2015) oder mit entsprechenden Modulen in Statistikpaketen (z. B. Mplus, Muthén und Muthén 2017) erfolgen. Auch in der kostenlosen Softwareumgebung R gibt es mittlerweile eine Reihe guter und relativ leicht bedienbarer Pakete, mit denen LCA vorgenommen werden können (beispielsweise das Paket „LCA 1.1“; Waller 2004; oder das Paket „poLCA“; Linzer und Lewis 2011). Eine Übersicht über aktuelle LCA-Software findet sich auf folgender Webseite: ► <http://www.john-uebersax.com/stat/soft.htm>.

22.8 Kontrollfragen

?

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was versteht man bei der LCA unter der
 - a. „relativen Klassengröße“ π_g ?
 - b. „bedingten Klassenzuordnungswahrscheinlichkeit“ $P(g | a_v)$?
 - c. „unbedingten Antwortmusterwahrscheinlichkeit“ $P(a_v)$?
 - d. „bedingten Antwortmusterwahrscheinlichkeit“ $P(a_v | g)$?
2. Ein eingesetzter Test enthalte acht dichotome Items.
 - a. Wie viele mögliche Antwortmuster N_a^{\max} gibt es hier?
 - b. Wie viele Freiheitsgrade hätte die χ^2 -Statistik im Falle eines Modells mit vier latenten Klassen?

- c. Sagen wir, der χ^2 -Wert dieses Modells beträgt 287.6: Welche Schlussfolgerungen ziehen Sie auf der Basis des „klassischen“ χ^2 -Tests in Bezug auf die Gültigkeit dieses Modells (auf einem Signifikanzniveau von $\alpha = 5\%$)?
- 3. Konstruieren und erläutern Sie einen Fall, in dem die Annahme der lokalen stochastischen Unabhängigkeit innerhalb der Klassen verletzt wäre.
- 4. Erläutern Sie, wie (und wieso) sich bei den Informationskriterien die Komplexität eines LCA-Modells niederschlägt? Wie nimmt man auf der Basis von Informationskriterien einen Vergleich zwischen verschiedenen LCA-Modellen vor?
- 5. Was versteht man in der LCA unter einer Fixierungs-, einer Gleichheits- und einer Ordnungsrestriktion? Geben Sie jeweils ein Beispiel.

Literatur

- Athenstaedt, U. & Alfermann, D. (2011). *Geschlechterrollen und ihre Folgen: Eine sozialpsychologische Betrachtung*. Stuttgart: Kohlhammer.
- Bem, S. L. (1977). On the utility of alternative procedures for assessing psychological androgyny. *Journal of Clinical and Consulting Psychology*, 45, 196–205.
- Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Costa, P. T. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Eysenck, H. J. (1990). Biological dimensions of personality. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (S. 244–276). New York, NY: Guilford.
- Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in Theorie und Anwendung*. Weinheim: Beltz.
- Gollwitzer, M., Eid, M. & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, 17, 56–69.
- Gray, J. A. (1972). The psychophysiological basis of introversion-extraversion: A modification of Eysenck's theory. In V. D. Nebylitsyn & J. A. Gray (Eds.), *Biological bases of individual behavior* (pp. 182–205). New York: Academic.
- Hagenaars, J. & McCutcheon, A. (Eds.). (2002). *Applied latent class analysis models*. New York: Cambridge University Press.
- Jung, C. G. (1921). *Psychologische Typen*. Zürich: Rascher.
- Langeheine, R., Pannekoek, J. & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24, 492–516.
- Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Linzer, D. A. & Lewis, J. (2011). poLCA: an R Package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42, 1–29.
- MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- McLachlan, G. & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Muthén, B. O. & Muthén, L. K. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Nylund, K. L., Asparouhov, T. & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569.
- Read, T. & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer.
- Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York, NY: Springer.
- Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2. Aufl.). Bern: Huber.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research-Online*, 8, 23–74.
- Schwenkmezger, P., Hodapp, V. & Spielberger, C. D. (1992). *State-Trait Anger Expression Inventory (STAXI)*. Bern: Huber.
- Spence, J. T., Helmreich, R. & Stapp, J. (1975). Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 32, 29–39.
- Strauß, B., Köller, O. & Möller, J. (1996). Geschlechtsrollentypologien – Eine empirische Prüfung des additiven und des balancierten Modells. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 2, 67–83.

- Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.
- van Kollenburg, G. H., Mulder, J. & Vermunt, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology*, 11, 65–79.
- Vermunt, J. K. & Magidson, J. (2015). *Upgrade Manual for Latent GOLD 5.1*. Belmont, MA: Statistical Innovations Inc.
- von Davier, M. (2001). *WINMIRA 2001* (Software). St. Paul, MN: Assessment Systems Corp. Retrieved from <http://208.76.80.46/~svfkluu/wmira/index.html> [29.12.2019]
- von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online*, 2, 29–48.
- Waller, N. G. (2004). LCA 1.1: An R package for exploratory Latent Class Analysis. *Applied Psychological Measurement*, 28, 141–142.
- Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50, 1090–1104.



Exploratorische Faktorenanalyse (EFA)

Holger Brandt

Inhaltsverzeichnis

- 23.1 Einleitung – 577**
 - 23.1.1 Zielsetzung der EFA – 577
 - 23.1.2 Ablaufschritte der EFA – 578
- 23.2 Faktormodell (Fundamentaltheorem) – 578**
 - 23.2.1 Modellformulierung – 578
 - 23.2.2 Modellannahmen – 580
 - 23.2.3 Varianzerlegung – 581
 - 23.2.4 Interpretation der Modellkomponenten – 581
 - 23.2.4.1 Eigenwert – 582
 - 23.2.4.2 Kommunalität – 583
 - 23.2.4.3 Spezifität – 583
 - 23.2.5 Empirisches Beispiel – 583
- 23.3 Methoden der Faktorenantraktion – 585**
 - 23.3.1 Exkurs: Hauptkomponentenanalyse
(Principal Component Analysis, PCA) – 586
 - 23.3.2 Hauptachsenanalyse (Principal Factor Analysis, PFA) – 588
 - 23.3.3 Maximum-Likelihood-Faktorenanalyse (ML-EFA) – 589
- 23.4 Abbruchkriterien der Faktorenantraktion – 590**
 - 23.4.1 Kaiser-Guttman-Kriterium – 590
 - 23.4.2 Elbow-Kriterium (Scree-Test) – 591
 - 23.4.3 Parallelanalyse – 592
 - 23.4.4 Modelldifferenztest (ML-EFA) – 594
- 23.5 Faktorenrotation – 595**
 - 23.5.1 Faktorenindeterminiertheit – 595
 - 23.5.2 Einfachstruktur – 596
 - 23.5.3 Orthogonale Rotation – 597
 - 23.5.3.1 Quartimax-Rotation ($\kappa = 0$) – 598
 - 23.5.3.2 Varimax-Rotation ($\kappa = 1/p$) – 598

23.5.4	Oblique Rotation – 598
23.5.4.1	Promax- und Harris-Kaiser-Rotation (indirekte oblique Rotation) – 599
23.5.4.2	Oblimin-Rotation (direkte oblique Rotation) – 599
23.5.5	Target-Rotation – 600
23.5.6	Geomin-Rotation – 601
23.5.7	Welche Methode sollte verwendet werden? – 601
23.6	Modellevaluation und Itemauswahl – 604
23.6.1	Modellevaluation – 604
23.6.1.1	Residualmatrix – 604
23.6.1.2	Modellfit – 605
23.6.2	Faktoreninterpretation – 605
23.6.3	Faktorwerte (Faktorscores) – 606
23.6.4	Itemauswahl – 606
23.6.5	Korrelations- oder Kovarianzmatrix? – 607
23.7	Neue Verfahren – 608
23.7.1	Verwendung exploratorischer Faktormodelle in Strukturgleichungsmodellen (ESEM) – 608
23.7.2	Alternativen für dichotome und ordinale Daten – 609
23.7.2.1	IRT-Modelle – 609
23.7.2.2	Robuste Schätzverfahren – 609
23.7.2.3	Bayes'sche Schätzverfahren – 610
23.8	Abschließende Bemerkungen – 610
23.9	Zusammenfassung – 611
23.10	EDV-Hinweise – 611
23.11	Kontrollfragen – 611
	Literatur – 612

i Dieses Kapitel liefert einen Überblick über die exploratorische Faktorenanalyse (EFA) mit dem Schwerpunkt auf ihren Einsatz in der Testkonstruktion. Die EFA kann in der Testkonstruktion z. B. der Beurteilung der Dimensionalität der in einem Test enthaltenen Items dienen oder sie kann die Dimensionalität mehrerer Subtests eines Tests zueinander in Beziehung setzen. Die EFA kann verwendet werden, um die Frage zu beantworten, ob die einzelnen Items, die zu einer Facette eines Tests gehören, auch dasselbe messen, ob also eine Zusammenfassung der einzelnen Itemwerte zu einem Testwert gerechtfertigt ist. In der EFA unterscheidet man Extraktionsverfahren und -kriterien sowie Rotationsverfahren. Extraktionsverfahren und -kriterien werden verwendet, um zu entscheiden, wie viele Dimensionen (Faktoren) notwendig sind, um die in den Items enthaltenen multivariaten Informationen ökonomisch zu repräsentieren. Die verschiedenen Rotationsverfahren erlauben außerdem eine bessere Zuordnung der Items zu den Faktoren. Das Kapitel beginnt mit der Einführung der Grundlagen der EFA zur Bestimmung der Anzahl der relevanten Faktoren, gefolgt vom Konzept der Rotation. Anschließend werden Aspekte der Modellüberprüfung und -interpretation vorgestellt. Zuletzt wird auf aktuelle Entwicklungen der EFA eingegangen.

23.1 Einleitung

Von Moosbrugger und Brandt (► Kap. 4 und 5) wurde ausführlich dargelegt, dass es notwendig ist, mehr als nur ein Item zu generieren, um psychologische Konstrukte zu operationalisieren. In diesem Kapitel wird diese Anforderung nun formalisiert, d. h., es wird eine statistische Methode beschrieben, die den Zusammenhang mehrerer Items mit einem latenten Konstrukt – dem Faktor – mathematisch formuliert. Dieser Zusammenhang wird in der exploratorischen Faktorenanalyse (EFA) *Messmodell* genannt. Die Faktoren selbst sind hypothetische Konstrukte, die in der Testkonstruktion mit inhaltlichen Merkmalen der Items in Verbindung gesetzt werden. Das Ziel der Faktorenanalyse ist es letztlich, anhand eines *Faktorladungsmusters* ein Erklärungsmodell für die multivariaten Informationen der Items zu liefern, das eine ökonomische und einfache Interpretation erlaubt. Die EFA ist ein struktursuchendes Verfahren, das sich – im Unterschied zur konfirmatorischen Faktorenanalyse (CFA) (► Kap. 24) – insbesondere zur Hypothesengenerierung eignet, nicht jedoch zur Hypothesenprüfung.

EFA liefert Erklärungsmodell für die in den Items enthaltenen Informationen

23.1.1 Zielsetzung der EFA

Mit der EFA können u. a. folgende Fragen untersucht werden:

1. Wie viele Dimensionen (Faktoren) sind für eine ökonomische Darstellung der Items notwendig, ohne die in den Items erhaltenen Informationen zu sehr zu reduzieren?
2. Messen alle Items eines Tests oder Subtests (einer Facette) unidimensional, d. h., messen die Items, die zu einem Test oder Subtest zusammengefasst werden sollen, deutlich eine Dimension und keine andere?
3. Weist ein hoher (niedriger) Itemwert jedes Items auch auf einen hohen (niedrigen) Test-/Faktorwert hin?
4. Wie können die erhaltenen Faktoren inhaltlich interpretiert werden?
5. Welche Items genügen insbesondere den Anforderungen an einen „guten“ Test?

Fragen, die mit der EFA beantwortet werden können

Etwas mathematischer ausgedrückt besteht das grundsätzliche psychometrische Ziel der EFA darin, ein Erklärungsmodell für die Korrelationen der p Itemvariablen $y_1, y_2, \dots, y_i, \dots, y_p$ durch k latente Faktoren (Faktorwertvariablen) $\eta_1, \eta_2, \dots, \eta_j, \dots, \eta_k$ zu finden. Es wird angenommen, dass unbeobachtete, *latente* Faktoren die wahre Dimensionen des Tests darstellen. Die Items sind die Beobachtungen dieser Faktoren. Die EFA versucht, die Dimensionen so zu bestimmen, dass die Abhängigkeiten zwischen den Items durch die Faktorenanalyse möglichst gut erklärt werden.

Exploratorisches Vorgehen zur Bestimmung der Faktoren

Sparsamkeit und Einfachstruktur als Ziel der EFA

tente Konstrukte (z. B. Intelligenz) die Beantwortung der Items (z. B. Items aus einem IQ-Test) beeinflussen, weshalb ähnliche Items von derselben Person auch ähnlich beantwortet (oder gelöst) werden. Das traditionelle Vorgehen bei einer EFA ist insoweit naiv, als keine Vorannahmen über die Faktorenanzahl oder der Zugehörigkeit von Items zu bestimmten Faktoren getroffen werden. Vielmehr wird diesen Fragen erst im Laufe der EFA nachgegangen und die entsprechenden Entscheidungen werden anhand statistischer Kriterien getroffen.

Das Ziel einer EFA sollte stets sein, eine möglichst geringe Anzahl von Faktoren zu extrahieren (d. h. $k < p$), wobei die Items mit einem einzigen oder zumindest nur einigen wenigen Faktoren möglichst eindeutig zusammenhängen sollen, während zu den verbleibenden Faktoren kein oder nur ein minimaler Zusammenhang bestehen soll („*Einfachstruktur*“). Dies hat zwei Vorteile: Zum einen erlaubt eine geringe Anzahl von Faktoren eine sparsame Erklärung der erhobenen Testdaten; zum anderen erlaubt eine Einfachstruktur eine eindeutige Interpretation der Faktoren (d. h. der inhaltlichen Bedeutung eines hypothetischen Konstrukt). Wie diese Ziele erreicht werden können, wird in den nächsten Abschnitten besprochen.

23.1.2 Ablaufschritte der EFA

Vier Ablaufschritte der EFA

Die EFA beinhaltet mehrere Ablaufschritte. In jedem dieser Schritte erlauben zentrale *Kennwerte* die Beurteilung der Faktoren und der Items. Folgende Schritte sollten für eine Testkonstruktion bei der EFA berücksichtigt werden:

1. Aufstellung eines Faktormodells (► Abschn. 23.2).
2. Ermittlung der Anzahl der notwendigen Faktoren (Faktorenantraktion): Hierbei muss eine Entscheidung getroffen werden, welche *Extraktionsmethode* und welche *Kriterien* zur Bestimmung der Faktorenanzahl verwendet werden sollen (► Abschn. 23.3 und 23.4).
3. *Rotation* der Faktoren, um ein möglichst eindeutiges *Ladungsmuster* zu erzeugen, das eine Interpretation der Faktoren erlaubt. Hierbei muss eine Entscheidung darüber getroffen werden, welches Rotationsverfahren verwendet werden soll (► Abschn. 23.5).
4. *Beurteilung* der Modellgüte und des Ladungsmusters hinsichtlich dessen, wie gut Items die Faktoren messen (► Abschn. 23.6).

In den folgenden Abschnitten wird detailliert auf jeden dieser Ablaufschritte eingegangen. Am Ende des Kapitels wird zudem ein Einblick in neuere Entwicklungen im Rahmen der Faktorenanalyse gegeben (► Abschn. 23.7).

23.2 Faktormodell (Fundamentaltheorem)

In diesem Abschnitt soll zuerst auf die Modellformulierung (► Abschn. 23.2.1), die Modellannahmen (► Abschn. 23.2.2), die Varianzzerlegung (► Abschn. 23.2.3) und die Interpretation der Modellkomponenten (► Abschn. 23.2.4) in der EFA eingegangen werden, die die Grundlage für das Verständnis der Verfahren zur Faktorenantraktion (► Abschn. 23.3), der Abbruchkriterien (► Abschn. 23.4) und der Faktorenrotation (► Abschn. 23.5) bilden.

23.2.1 Modellformulierung

Die grundlegende Idee zur Modellformulierung ist vergleichbar mit der in einer Regressionsanalyse, in der die Ausprägungen einer abhängigen Variablen (hier

23.2 · Faktormodell (Fundamentaltheorem)

Itemwerte, d. h. Antworten der Testpersonen auf die Items) in gewichteter Form zurückgeführt werden auf die Ausprägungen in unabhängigen Variablen (hier Faktoren):

$$\begin{aligned} \text{Itemwert} &= \frac{\text{gewichtete Summe der Einflüsse}}{\text{der berücksichtigten Faktoren}} \\ &\quad + \frac{\text{Summe der Einflüsse der}}{\text{nicht berücksichtigten Faktoren}} \end{aligned} \quad (23.1)$$

Zu berücksichtigende Faktoren stellen hierbei die *gemeinsamen Faktoren* („common factors“) dar, d. h. jene Faktoren, die nicht nur ein einzelnes Item, sondern mehrere Items beeinflussen (z. B. IQ, Ängstlichkeit, Perfektionismus). *Nicht zu berücksichtigende Faktoren* sind alle weiteren Einflüsse, die itemspezifisch sind (z. B. Verständnis des spezifischen Iteminhalts, biografische oder situationsspezifische Komponenten der Beantwortung). Diese werden zu einem sog. „Residuum“ (d. h. dem unerklärten Teil im Modell) zusammengefasst. Einer der Hauptunterschiede zu einer Regressionsanalyse besteht allerdings darin, dass die gemeinsamen Faktoren *latente Variablen* (das Wort „latent“ bedeutet „verborgen“) mit zunächst unbekannten Ausprägungen (Faktorwerten, Faktorscores) darstellen. Erst wenn mehrere Items simultan analysiert werden, können Rückschlüsse auf die Einflussgewichte und auf die Zusammenhänge der Items mit den Faktoren gezogen werden.

Angenommen, es wurden Antworten auf p Items y_1, y_2, \dots, y_p erhoben und jedes Item wurde standardisiert¹. Dann ergibt sich als faktorenanalytische Modellvorstellung für jede Person $v = 1, \dots, N$ mit $k \leq p$ extrahierten (gemeinsamen) Faktoren $\eta_1, \eta_2, \dots, \eta_k$ (vgl. z. B. Mulaik 2010):

$$\begin{aligned} y_{1v} &= \lambda_{11}\eta_{1v} + \lambda_{12}\eta_{2v} + \dots + \lambda_{1j}\eta_{jv} + \dots + \lambda_{1k}\eta_{kv} + \varepsilon_{1v} & (23.2) \\ y_{2v} &= \lambda_{21}\eta_{1v} + \lambda_{22}\eta_{2v} + \dots + \lambda_{2j}\eta_{jv} + \dots + \lambda_{2k}\eta_{kv} + \varepsilon_{2v} \\ &\vdots \\ y_{iv} &= \lambda_{i1}\eta_{1v} + \lambda_{i2}\eta_{2v} + \dots + \lambda_{ij}\eta_{jv} + \dots + \lambda_{ik}\eta_{kv} + \varepsilon_{iv} \\ &\vdots \\ y_{pv} &= \lambda_{p1}\eta_{1v} + \lambda_{p2}\eta_{2v} + \dots + \lambda_{pj}\eta_{jv} + \dots + \lambda_{pk}\eta_{kv} + \varepsilon_{pv} \end{aligned}$$

Hierbei stellt λ_{ij} die *Faktorladung* des i -ten Items auf dem j -ten Faktor dar. Sie ist ein Maß für die Größe des Zusammenhangs zwischen dem jeweiligen Item und dem jeweiligen Faktor. Eine Faktorladung liegt in einem Wertebereich zwischen -1 und $+1$ und kann als Korrelationskoeffizient zwischen Item und Faktor interpretiert werden – solange die Faktoren unkorreliert sind (für korrelierte Faktoren ► Abschn. 23.5.4). Die Fehlerterme $\varepsilon_{1v}, \dots, \varepsilon_{pv}$ stellen für die jeweilige Testperson v den Anteil der Itemwerte dar, der *nicht* durch die gemeinsamen Faktoren erklärt werden kann. Die simultane modelltheoretische Dekomposition aller Items bedeutet, dass die beobachteten individuellen Itemwerte („Antwortmuster“ der jeweiligen Testperson) durch eine endliche Anzahl von gemeinsamen Faktoren und einen verbliebenen, unerklärten Anteil beschrieben werden können. Diese Dekomposition wird als *Fundamentaltheorem* der Faktorenanalyse bezeichnet.

Für die Darstellung der EFA ist es sinnvoll, die Matrixschreibweise zu verwenden (zur Matrixschreibweise und -algebra s. z. B. Moosbrugger 2011). Hierbei werden die Itemantworten, Faktorwerte und Fehlerterme jeder v -ten Person in die entsprechenden Vektoren $\mathbf{y}_v, \boldsymbol{\eta}_v, \boldsymbol{\varepsilon}_v$ zusammengefasst sowie die Faktorladungen in

Gemeinsame Faktoren vs. nicht zu berücksichtigende Faktoren

Intuitive Modellvorstellung

Fundamentaltheorem: Erklärung der beobachteten Itemwerte durch gemeinsame Faktoren

¹ Für eine didaktisch klarere Darstellung der EFA werden vorerst ausschließlich standardisierte Variablen und Korrelationsmatrizen verwendet. Auf die Analyse von unstandardisierten Variablen und Kovarianzmatrizen wird in ► Abschn. 23.6.5 eingegangen.

eine sog. „Faktorladungsmatrix“ λ vom Format ($p \times k$) überführt:

$$\underbrace{\begin{pmatrix} y_{1v} \\ y_{2v} \\ \vdots \\ y_{iv} \\ \vdots \\ y_{pv} \end{pmatrix}}_{\mathbf{y}_v} = \underbrace{\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1j} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2j} & \cdots & \lambda_{2k} \\ \vdots & \vdots & & \vdots & & \vdots \\ \lambda_{i1} & \lambda_{i2} & \cdots & \lambda_{ij} & \cdots & \lambda_{ik} \\ \vdots & \vdots & & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pj} & \cdots & \lambda_{pk} \end{pmatrix}}_{\boldsymbol{\lambda}} \cdot \underbrace{\begin{pmatrix} \eta_{1v} \\ \eta_{2v} \\ \vdots \\ \eta_{iv} \\ \vdots \\ \eta_{pv} \end{pmatrix}}_{\boldsymbol{\eta}_v} + \underbrace{\begin{pmatrix} \varepsilon_{1v} \\ \varepsilon_{2v} \\ \vdots \\ \varepsilon_{iv} \\ \vdots \\ \varepsilon_{pv} \end{pmatrix}}_{\boldsymbol{\varepsilon}_v} \quad (23.3)$$

Faktormodell in Matrixschreibweise

In Matrixschreibweise lautet die Modellgleichung für eine Testperson v somit:

$$\mathbf{y}_v = \boldsymbol{\lambda} \cdot \boldsymbol{\eta}_v + \boldsymbol{\varepsilon}_v \quad (23.4)$$

Die Interpretation der Ergebnisse einer EFA konzentriert sich häufig auf die Faktorladungsmatrix λ . In ihr sind die einzelnen Faktorladungen enthalten, wobei jede Zeile einem Item und jede Spalte einem Faktor zugeordnet ist. Sämtliche Faktorladungen des ersten Items auf allen k Faktoren sind somit in der ersten Zeile aufgeführt ($\lambda_{11}, \dots, \lambda_{1k}$). Entsprechend sind sämtliche Faktorladungen aller p Items auf dem ersten Faktor in der ersten Spalte ($\lambda_{11}, \dots, \lambda_{p1}$) aufgeführt.

23.2.2 Modellannahmen

Annahmen des Faktormodells

Die Modellannahmen für das Standardmodell der EFA beinhalten Annahmen über die latenten Variablen, d. h. über die Faktoren und die Fehlerterme. Konkret wird für die Faktorwerte $\boldsymbol{\eta}_v$ in den k Faktoren für alle Personen angenommen, dass ihre Erwartungswerte null sind, ihre Varianzen eins und dass die Faktoren unkorreliert sind:

$$E[\boldsymbol{\eta}_v] = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{und} \quad Cov(\boldsymbol{\eta}_v) = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}, \quad (23.5)$$

wobei $E[\boldsymbol{\eta}_v]$ die Erwartungswerte und $Cov(\boldsymbol{\eta}_v)$ die Kovarianzmatrix der Faktoren mit den Varianzen in der Hauptdiagonalen bezeichnen. (Bei standardisierten Variablen besteht zwischen der Kovarianzmatrix und der Korrelationsmatrix kein Unterschied.) Für die Fehlerterme $\boldsymbol{\varepsilon}_v$ in den p Items wird ebenfalls für alle Personen angenommen, dass sie einen Erwartungswert von null aufweisen, dass sie unkorreliert sind und dass sie eine für das jeweilige i -te Item spezifische Varianz von $Var(\varepsilon_{iv}) = \psi_{ii}^{\varepsilon}$ haben:

$$E[\boldsymbol{\varepsilon}_v] = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{und} \quad Cov(\boldsymbol{\varepsilon}_v) = \boldsymbol{\Psi}^{\varepsilon} = \begin{bmatrix} \psi_{11}^{\varepsilon} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_{pp}^{\varepsilon} \end{bmatrix}. \quad (23.6)$$

Zudem wird angenommen, dass alle Faktoren und Residuen wechselseitig unkorreliert sind. Dies bedeutet, dass die nicht erklärten Anteile der Itemantworten (also itemspezifische Aspekte) nicht mit den erklärten Anteilen (also jenen Aspekten, die die Beantwortung aller Items systematisch beeinflussen) zusammenhängen. Diese Annahme ist analog zu den Annahmen der Klassischen Testtheorie (KTT, ► Kap. 13).

Das so aufgestellte Faktormodell impliziert, dass alle Korrelationen zwischen den Itemvariablen durch die gemeinsamen Faktoren erklärt werden können. Für Personen, die denselben Faktorwert aufweisen, sind die Itemvariablen unkorreliert. Man spricht dann auch von bedingter Unkorreliertheit bei gegebenem Faktorwert (vgl. hierzu die lokale stochastische Unabhängigkeit in der IRT, ▶ Kap. 16).

23.2.3 Varianzerlegung

Modelltheoretisch (oder auch modellimpliziert, d. h. basierend auf den Modellgleichungen Gl. 23.2) lässt sich die Varianz einer Itemvariablen y_i in einen erklärten Anteil und einen unerklärten Anteil zerlegen:

$$\text{Var}(y_i) = \underbrace{\lambda_{i1}^2 + \lambda_{i2}^2 + \dots + \lambda_{ik}^2}_{\text{erklärte Varianz}} + \underbrace{\psi_{ii}^e}_{\text{nicht erklärte Varianz}}. \quad (23.7)$$

Da es sich bei den Faktorladungen λ_{ij} um die Korrelationen zwischen den Itemvariablen und den Faktoren handelt, geben die quadrierten Ladungen λ_{ij}^2 den jeweiligen erklärten Varianzanteil an (analog zu den Determinationskoeffizienten der Regressionsanalyse, s. z. B. Moosbrugger 2011 und ▶ Exkurs 23.1). Die Itemvarianz ist also die Summe der quadrierten Faktorladungen über die Faktoren hinweg und der unerklärten Varianz (ψ_{ii}^e). Für ein gutes Erklärungsmodell sollten die Faktorladungen groß sein, damit der Anteil der Varianz, der durch die Faktoren erklärt wird, im Vergleich zu dem unerklärten Anteil (ψ_{ii}^e) möglichst groß wird.

Weiterhin ergeben sich die *modellimplizierten Korrelationen* zwischen jeweils zwei Itemvariablen y_i und $y_{i'}$ als:

$$\text{Cor}(y_i, y_{i'}) = \lambda_{i1}\lambda_{i'1} + \lambda_{i2}\lambda_{i'2} + \dots + \lambda_{ik}\lambda_{i'k} \quad (23.8)$$

Sämtliche Korrelationen zwischen den Items werden somit dadurch erklärt, dass sie auf denselben gemeinsamen Faktoren ($1, \dots, k$) laden. Eine hohe Korrelation wäre zu erwarten, wenn die Items auf einem (oder mehreren) Faktor(en) hoch laden (und somit das Produkt der Faktorladungen groß ist); eine geringe Korrelation wäre zu erwarten, wenn keines oder nur eines der Items auf dem gemeinsamen Faktor hoch lädt.

Die Zusammenfassung sämtlicher Korrelationen (und Varianzen) in einer Matrix erlaubt nun die Beschreibung sämtlicher multivariater Informationen. Hierbei liefern die empirischen (beobachteten) Korrelationen eine sog. „empirische Korrelationsmatrix“ S und die auf den Modellgleichungen (Gl. 23.2 bzw. Gl. 23.3) basierenden Korrelationen eine sog. „modelltheoretische“ oder „modellimplizierte Korrelationsmatrix“ $\Sigma(\theta)$. Sie hängt ausschließlich von den zu schätzenden Parametern in den Matrizen λ und ψ ab, die im sog. „Parametervektor“ θ zusammengefasst werden. Der ▶ Exkurs 23.1 liefert den theoretischen Hintergrund für die Varianzzerlegung, während ▶ Beispiel 23.1 ein einfaches Beispiel zur Bestimmung der Varianzkomponenten liefert.

Die modellimplizierten Varianzen und Korrelationen bilden die Grundlage für die Faktorextraktion in der EFA, wie in ▶ Abschn. 23.3 gezeigt wird. Ebenso spielen sie eine wichtige Rolle bei der Modellbeurteilung (▶ Abschn. 23.6.1).

**Faktoren erklären
Korrelationsmuster der Items**

Modellimplizierte Zerlegung der Itemvarianz in (durch Faktoren) erklärte und unerklärte Anteile

Modellimplizierte Korrelation zweier Items

Empirische und modellimplizierte Korrelationsmatrizen

Modellimplizierte Varianzen erlauben eine Modellbeurteilung

23.2.4 Interpretation der Modellkomponenten

Die zentralen Begriffe bei der Interpretation einer EFA sind der *Eigenwert* eines Faktors sowie die *Kommunalität* und die *Spezifität* eines Items.

Exkurs 23.1

Theoretischer Hintergrund zur modellimplizierten Varianzzerlegung in der EFA

Die modellimplizierte Varianz eines Items y_i lässt sich anhand von Gl. (23.2) folgendermaßen beschreiben:

$$\text{Var}(y_i) = \text{Var}(\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + \dots + \lambda_{ik}\eta_k + \varepsilon_i) \quad (23.9)$$

Da die Fehlerterme und die Faktoren wechselseitig unkorreliert sind ($\text{Cov}((\eta_1, \dots, \eta_k)', \varepsilon_i) = 0$) und auch die Faktoren untereinander unkorreliert sind (Gl. 23.5, ► Abschn. 23.2.2), kann diese Varianz aufgeteilt werden in:

$$\text{Var}(y_i) = \text{Var}(\lambda_{i1}\eta_1) + \text{Var}(\lambda_{i2}\eta_2) + \dots + \text{Var}(\lambda_{ik}\eta_k) + \text{Var}(\varepsilon_i) \quad (23.10)$$

Da $\text{Var}(\lambda_{ij}\eta_j) = \lambda_{ij}^2 \text{Var}(\eta_j)$ gilt, nach Gl. (23.5) $\text{Var}(\eta_j) = 1$ und nach Gl. (23.6) $\text{Var}(\varepsilon_i) = \psi_{ii}^e$, kann $\text{Var}(y_i)$ wie folgt geschrieben werden:

$$\text{Var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots + \lambda_{ik}^2 + \psi_{ii}^e \quad (23.11)$$

Ebenso kann die modellimplizierte Korrelation zwischen zwei Items y_i und $y_{i'}$ folgendermaßen dargestellt werden:

$$\begin{aligned} \text{Cor}(y_i, y_{i'}) &= \text{Cor}(\lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + \dots \\ &\quad + \lambda_{ik}\eta_k + \varepsilon_i, \lambda_{i'1}\eta_1 + \lambda_{i'2}\eta_2 + \dots + \lambda_{i'k}\eta_k + \varepsilon_{i'}) \end{aligned} \quad (23.12)$$

Diese Korrelation beinhaltet hier vier Teile (auf eine technische Darstellung soll hier verzichtet werden): Die Korrelationen

1. der (mit der Faktorladung gewichteten) Faktoren mit sich selbst,
2. zwischen verschiedenen Faktoren (die jeweils 0 sind),
3. zwischen Faktoren und Fehlertermen (ebenfalls 0) und
4. zwischen den Fehlertermen (ebenfalls 0).

Es verbleiben die Korrelationen der Faktoren mit sich selbst, die sich als Summe der Produkte der jeweiligen Faktorladungen auf demselben Faktor ergeben:

$$\begin{aligned} \text{Cor}(y_i, y_{i'}) &= \sum_{j=1}^k \text{Cor}(\lambda_{ij}\eta_j, \lambda_{i'j}\eta_j) = \sum_{j=1}^k \lambda_{ij}\lambda_{i'j} \underbrace{\text{Var}(\eta_j)}_{=1} \\ &= \sum_{j=1}^k \lambda_{ij}\lambda_{i'j} = \lambda_{i1}\lambda_{i'1} + \lambda_{i2}\lambda_{i'2} + \dots + \lambda_{ik}\lambda_{i'k} \end{aligned} \quad (23.13)$$

Somit spiegeln hohe empirische Korrelationen zwischen Items die Idee wider, dass diese Items auf denselben Faktoren hoch laden.

23.2.4.1 Eigenwert

Eigenwerte beschreiben die Dimensionalität einer Korrelationsmatrix

Allgemein sind Eigenwerte mathematische Kenngrößen für symmetrische Matrizen. Im Falle der EFA beziehen sie sich auf die Korrelationsmatrix und liefern Kenngrößen für die Anzahl der zugrunde liegenden Faktoren (Dimensionen). Wie die Eigenwerte gewonnen werden, wird im nachfolgenden ► Abschn. 23.3 zur Faktorextraktion erklärt.

23.2 · Faktormodell (Fundamentaltheorem)

Die für die Testkonstruktion praktisch relevante Interpretation ist folgende: Der Eigenwert Λ_j eines Faktors j kann als der Anteil der Varianz angesehen werden, den der j -te Faktor an allen p Items erklärt. Der Anteil ergibt sich für den j -ten Faktor als *Summe der quadrierten Faktorladungen λ_{ij} über alle p Items hinweg*:

$$\Lambda_j = \sum_{i=1}^p \lambda_{ij}^2 = \lambda_{1j}^2 + \lambda_{2j}^2 + \dots + \lambda_{pj}^2 \quad (23.14)$$

Eine Umrechnung in einen prozentualen Anteil kann durch

$$\Lambda_j \% = \frac{\Lambda_j}{\sum_{i=1}^p \text{Var}(y_i)} \cdot 100 \% \quad (23.15)$$

erfolgen, indem die Größe des Eigenwertes an der Summe der Varianzen aller Itemvariablen relativiert wird. Im hier vorliegenden Fall standardisierter Itemvariablen mit $\text{Var}(y_i) = 1$ ist die Summe der Varianzen $\sum_{i=1}^p \text{Var}(y_i) = 1 + \dots + 1 = p$, und entspricht damit der Anzahl der Items. Somit können Aussagen getroffen werden, wie viel Varianz ein Faktor von der Gesamtvarianz aller Items erklären kann. Der Eigenwert eines Faktors spiegelt somit seine Bedeutsamkeit für die Erklärung des Korrelationsmusters der Items wider. Je größer der Eigenwert, desto besser können die Korrelationen der Items durch den Faktor erklärt werden.

Eigenwert erlaubt Beurteilung der Bedeutung eines Faktors

23.2.4.2 Kommunalität

Die Kommunalität h_i^2 ist der Varianzanteil des i -ten Items, der durch alle k extrahierten Faktoren erklärt werden kann. Sie ergibt sich rechnerisch für jedes Item als die *Summe der quadrierten Faktorladungen über alle k Faktoren hinweg*:

$$h_i^2 = \sum_{j=1}^k \lambda_{ij}^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots + \lambda_{ik}^2 \quad (23.16)$$

Je höher die Kommunalität ist, desto besser kann das Item durch die gemeinsamen Faktoren erklärt werden. Die Kommunalität liegt zwischen null und eins, wobei null bedeutet, dass das Item überhaupt nicht durch die Faktoren erklärt werden kann, und eins, dass es vollständig erklärt werden kann.

Kommunalität beschreibt, wie gut ein Item durch die gemeinsamen Faktoren erklärt werden kann

23.2.4.3 Spezifität

Die Spezifität ψ_{ii}^ε („Uniqueness“)² ist der eigenständige Varianzanteil des i -ten Items, der nicht durch die gemeinsamen Faktoren im Modell erklärt werden kann. Sie ergibt sich rechnerisch als Differenz zwischen der Varianz des Items und seiner Kommunalität:

$$\psi_{ii}^\varepsilon = \text{Var}(y_i) - h_i^2 \quad (23.17)$$

Spezifität ist der Anteil der Itemvarianz, der nicht durch gemeinsame Faktoren erklärt werden kann

Analog zur Kommunalität liegt die Spezifität zwischen null und eins, wobei null bedeutet, dass das Item vollständig durch die gemeinsamen Faktoren erklärt werden kann.

23.2.5 Empirisches Beispiel

Zur Illustration der EFA soll ein empirisches Beispiel dienen. Die Daten entstammen der PISA-Studie von 2012 (Programme for International Student Assessment;

² Wie die Synonymsetzung mit „Uniqueness“ zeigt, ist der hier verwendete Begriff „Spezifität“ weder identisch mit dem in der ROC-Analyse (► Kap. 9) noch mit dem in der LST-Theorie (► Kap. 26).

OECD 2013) und sind öffentlich zugänglich (► <http://www.oecd.org/pisa/data/>). Hier soll eine Substichprobe von 835 deutschen Schülerinnen und Schülern näher untersucht werden. Insgesamt stehen acht Items zur Verfügung, die potentiell die Konstrukte *Interesse an Mathematik* (Items 1 bis 4) sowie *Instrumentelle Motivation* (Items 5 bis 8) operationalisieren sollen. Die Iteminhalte sind in □ Tab. 23.2 zu finden (► Abschn. 23.5.7).

Beispiel 23.1: Ein Zweifaktormodell für drei Items (Minimalbeispiel)

Modellgleichungen

Zur Illustration der Modellgleichungen sollen aus didaktischen Gründen lediglich drei der acht Items herangezogen werden (Items 1, 2 und 7). Die Korrelationsmatrix für die drei Items ist:

$$\mathbf{S} = \begin{pmatrix} 1.00 & & \\ 0.60 & 1.00 & \\ 0.39 & 0.45 & 1.00 \end{pmatrix} \quad (23.18)$$

Obwohl in der EFA grundsätzlich keine anfänglichen Annahmen über die Anzahl der Faktoren getroffen werden, wird hier aus didaktischen Gründen vermutet, dass es zwei Faktoren sind, auf denen die Items laden. Während in einer konfirmatorischen Faktorenanalyse (CFA) ein spezifisches Faktorladungsmuster angenommen wird, bleibt das Ladungsmuster auf den beiden Faktoren vorerst offen:

$$\begin{aligned} y_{1v} &= \lambda_{11}\eta_{1v} + \lambda_{12}\eta_{2v} + \varepsilon_{1v} \\ y_{2v} &= \lambda_{21}\eta_{1v} + \lambda_{22}\eta_{2v} + \varepsilon_{2v} \\ y_{3v} &= \lambda_{31}\eta_{1v} + \lambda_{32}\eta_{2v} + \varepsilon_{3v} \end{aligned} \quad (23.19)$$

Faktorladungsmatrix

Die Anwendung der EFA führt zu folgenden Schätzungen für die Faktorladungsmatrix λ und die unerklärten Varianzen Ψ^e :

$$\lambda = \begin{pmatrix} 0.80 & 0.18 \\ 0.69 & 0.28 \\ 0.27 & 0.96 \end{pmatrix} \quad \text{und} \quad \Psi^e = \begin{pmatrix} 0.33 & & \\ 0.00 & 0.45 & \\ 0.00 & 0.00 & 0.01 \end{pmatrix} \quad (23.20)$$

Alle drei Items laden auf beiden Faktoren: Das Faktorladungsmuster deutet hier darauf hin, dass die ersten beiden Items primär auf dem ersten Faktor laden (weil die Faktorladungen von 0.80 und 0.69 groß sind), während das dritte Item insbesondere auf dem zweiten Faktor (0.96) lädt. Basierend auf den Iteminhalten könnte man somit argumentieren, dass der erste Faktor das Konstrukt *Interesse an Mathematik* widerspiegelt und der zweite Faktor die *Instrumentelle Motivation*.

Varianzzerlegung und Korrelationen

Die Varianzzerlegung für die drei Items in diesem Beispiel ist somit:

$$\begin{aligned} 1 &= \text{Var}(y_1) = \lambda_{11}^2 + \lambda_{12}^2 + \psi_{11}^e = 0.80^2 + 0.18^2 + 0.33 = 1 \\ 1 &= \text{Var}(y_2) = \lambda_{21}^2 + \lambda_{22}^2 + \psi_{22}^e = 0.69^2 + 0.28^2 + 0.45 = 1 \\ 1 &= \text{Var}(y_3) = \lambda_{31}^2 + \lambda_{32}^2 + \psi_{33}^e = 0.27^2 + 0.96^2 + 0.01 = 1 \end{aligned} \quad (23.21)$$

23.3 · Methoden der Faktorextraktion

und die modellimplizierten Korrelationen ergeben sich als:

$$\begin{aligned} 0.60 &= \text{Cor}(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} = 0.55 + 0.05 = 0.60 \\ 0.39 &= \text{Cor}(y_1, y_3) = \lambda_{11}\lambda_{31} + \lambda_{12}\lambda_{32} = 0.22 + 0.17 = 0.39 \\ 0.45 &= \text{Cor}(y_2, y_3) = \lambda_{21}\lambda_{31} + \lambda_{22}\lambda_{32} = 0.18 + 0.27 = 0.45 \end{aligned} \quad (23.22)$$

Die gefundene Lösung für die Faktorladungen und für die unerklärten Varianzen ergeben in diesem Fall modellimplizierte Varianzen und Korrelationen, die exakt die empirischen Werte reproduzieren.

Eigenwerte

Aus den Schätzungen für die Faktorladungsmatrix lassen sich folgende Eigenwerte für die ersten beiden Faktoren berechnen:

$$\begin{aligned} \Lambda_1 &= \sum_{i=1}^3 \lambda_{i1}^2 = 0.80^2 + 0.69^2 + 0.27^2 = 1.19 \\ \Lambda_2 &= \sum_{i=1}^3 \lambda_{i2}^2 = 0.18^2 + 0.28^2 + 0.96^2 = 1.03 \end{aligned} \quad (23.23)$$

Somit erklären die beiden Faktoren $1.19/3 \cdot 100\% = 40\%$ und $1.03/3 \cdot 100\% = 34\%$ der Varianz der Items.

Kommunalitäten

Weiterhin ergeben sich die Kommunalitäten der drei Items als

$$\begin{aligned} h_1^2 &= \sum_{j=1}^2 \lambda_{1j}^2 = \lambda_{11}^2 + \lambda_{12}^2 = 0.80^2 + 0.18^2 = 0.67 \\ h_2^2 &= \sum_{j=1}^2 \lambda_{2j}^2 = \lambda_{21}^2 + \lambda_{22}^2 = 0.69^2 + 0.28^2 = 0.55 \\ h_3^2 &= \sum_{j=1}^2 \lambda_{3j}^2 = \lambda_{31}^2 + \lambda_{32}^2 = 0.27^2 + 0.96^2 = 0.99 \end{aligned} \quad (23.24)$$

Spezifitäten

Die Spezifitäten sind in der Matrix Ψ^e gegeben und sind 0.33, 0.45 sowie 0.01. Das dritte Item kann fast vollständig durch die beiden Faktoren erklärt werden.

23.3 Methoden der Faktorextraktion

Ein wichtiger Schritt in der EFA besteht darin, zu entscheiden, wie viele Faktoren notwendig sind, um das Korrelationsmuster der Items angemessen zu erklären. Fast alle Extraktionsverfahren und Entscheidungskriterien basieren auf den *Eigenwerten der Faktoren*.

Zur Extraktion der Faktoren können in der EFA verschiedene Methoden verwendet werden. Die beiden am häufigsten eingesetzten Extraktionsverfahren sind die *Hauptachsenanalyse* (Principal Factor Analysis, PFA) und die *Maximum-Likelihood-Faktorenanalyse* (ML-EFA). Das Ziel beider Verfahren besteht darin,

Entscheidungskriterien für Faktorextraktion basieren auf Eigenwerten

Extraktionsverfahren minimieren Unterschiede zwischen empirischer und modellimplizierter Korrelationsmatrix

die Unterschiede zwischen empirischer Korrelationsmatrix \mathbf{S} und modellimplizierter Korrelationsmatrix $\Sigma(\theta)$ zu minimieren (► Abschn. 23.2.3).

Im Folgenden sollen kurz die Prinzipien der PFA und der ML-EFA skizziert werden. Insbesondere das Prinzip der Faktorextraktion in der PFA ist dem der Hauptkomponentenanalyse (Principal Component Analysis, PCA) sehr ähnlich (und wird manchmal verwechselt), weshalb zunächst die PCA als Exkurs in ► Abschn. 23.3.1 vorgestellt werden soll.

23.3.1 Exkurs: Hauptkomponentenanalyse (Principal Component Analysis, PCA)

PCA extrahiert keine Faktoren, sondern maximiert Varianz der Hauptkomponenten

Sukzessive Varianzabnahme der Hauptkomponenten

Ziel der PCA ist Identifikation einer ökonomischen Anzahl von Komponenten

Wahl der Hauptkomponenten und ihrer Lage

Einer der Hauptunterschiede zwischen PCA und EFA besteht darin, dass die EFA ein statistisches Verfahren ist, während die PCA ein mathematisches Verfahren ist. Ein statistisches Verfahren basiert stets auf Annahmen, wie sie in der EFA über unbeobachtete Residuen und Faktoren getroffen werden. In der PCA werden weder statistische Annahmen getroffen noch Faktoren extrahiert. Die PCA wird verwendet, um Kompositsscores (sog. „Hauptkomponenten“) zu bilden, die eine spezifische Aggregation der Items erlauben. Die Bedeutung einer Hauptkomponente besteht darin, dass interindividuelle Unterschiede von Personen in den Itemvariablen in ihr optimal abgebildet werden, d.h., dass die Varianz der individuellen Scores maximal ist, die die Personen in Bezug auf die Hauptkomponente aufweisen.

Im Prinzip können für p Itemvariablen stets auch p Hauptkomponenten extrahiert werden. Die erste Hauptkomponente ist diejenige, die am meisten Varianz erklärt, die folgende Hauptkomponente erklärt den verbliebenen Teil der Varianz der Itemvariablen, der nicht von der ersten Hauptkomponente erklärt werden konnte, etc. Die Bedeutung der Hauptkomponenten nimmt sukzessive von der ersten hin zur letzten ab. Weiterhin sind die Hauptkomponenten wechselseitig unkorreliert, wodurch die Informationen der Hauptkomponenten nicht redundant sind und jede Hauptkomponente einen anderen Aspekt der interindividuellen Unterschiede abbildet.

Obwohl zur *vollständigen Erklärung* der interindividuellen Unterschiede der Personen in p Itemvariablen auch p Hauptkomponenten extrahiert werden müssen, besteht das Ziel der PCA in der Wahl von $q < p$ Hauptkomponenten (also weniger Komponenten als Itemvariablen), die zumindest einen Großteil dieser Unterschiede abbilden. Die verbleibenden $p - q$ Komponenten tragen dann nur in einem sehr geringen Ausmaß zur Varianzerklärung bei und können unter Umständen unberücksichtigt bleiben.

Um die Idee der PCA zur Wahl der ersten beiden Hauptkomponenten bildlich zu veranschaulichen, ist in ► Abb. 23.1 die bivariate Streuung der Itemwerte von 18 Testpersonen in $p = 2$ Itemvariablen y_1 und y_2 dargestellt. Die Variabilität in den beiden Itemvariablen wird vollständig durch die zwei extrahierten Hauptkomponenten z_1 und z_2 abgebildet (sog. „Projektion“), was exemplarisch für vier Testpersonen durch gepunktete Linien angezeigt wird. Die Wahl der Lage der beiden Komponenten (also des Winkels der Komponenten bezogen auf das Koordinatensystem der Itemvariablen; ► Abb. 23.1a) erfolgt so, dass die erste Komponente z_1 den Haupteil der Variabilität abbildet und maximale Varianz aufweist, während die zweite Komponente z_2 die verbleibende Variabilität erklärt. Hierbei liegen z_1 und z_2 orthogonal zueinander (► Abb. 23.1b), wodurch beide Komponenten nicht redundante Information abbilden. In diesem Beispiel ist intuitiv zu erkennen, dass ein Großteil der Informationen durch die erste Hauptkomponente z_1 abgebildet werden kann. Personen unterscheiden sich primär durch z_1 und weisen nur geringe Unterschiede in der Hauptkomponente z_2 auf. Für eine sparsame Erklärung könnte somit auf z_2 verzichtet werden.

23.3 · Methoden der Faktorenanalyse

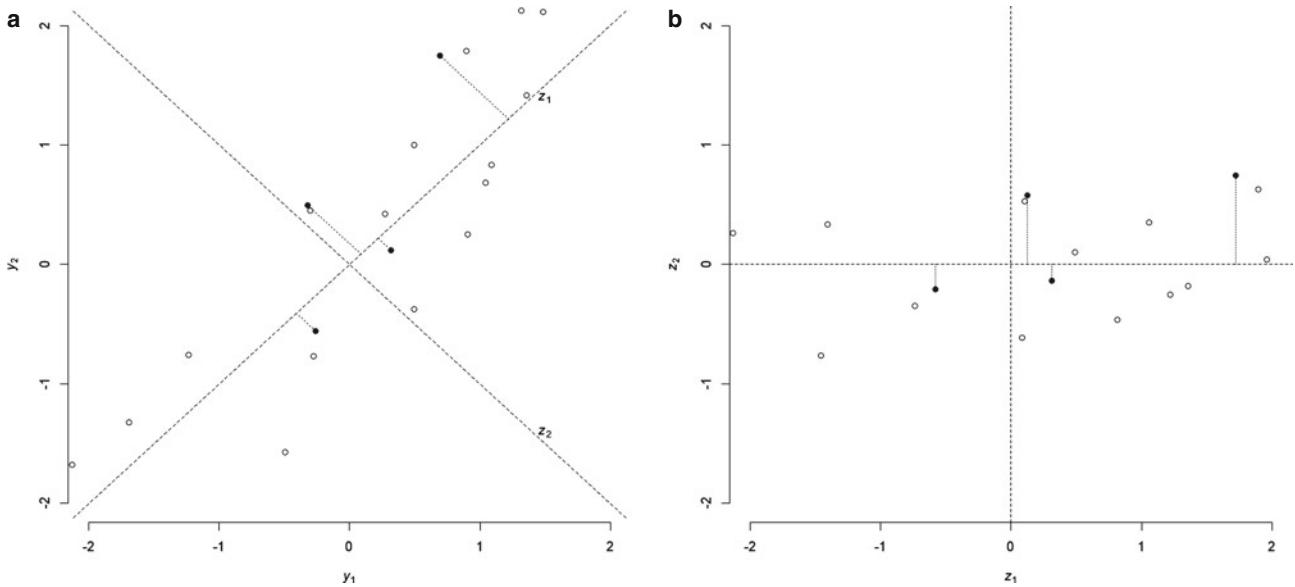


Abb. 23.1 **a** Die bivariate Streuung der Itemvariablen y_1 und y_2 wird durch die Projektion der 18 Datenpunkte auf die Hauptkomponenten z_1 und z_2 erklärt. **b** Die exemplarisch ausgewählten vier Personen unterscheiden sich in dem durch die Komponenten aufgespannten Raum insbesondere in z_1 . Zur Orientierung sind dieselben vier Testpersonen und ihre Projektionen durch *gefüllte Punkte* und *gepunkteten Linien* in der linken und rechten Abbildung hervorgehoben

Die Modellformulierung der PCA in allgemeiner Form für p Variablen y_{1v}, \dots, y_{pv} und q Komponenten z_{1v}, \dots, z_{qv} für $v = 1, \dots, N$ Personen ist durch folgende Linearkombinationen gegeben:

$$\begin{aligned} z_{1v} &= a_{11}y_{1v} + a_{12}y_{2v} + \dots + a_{1p}y_{pv} \\ z_{2v} &= a_{21}y_{1v} + a_{22}y_{2v} + \dots + a_{2p}y_{pv} \\ &\vdots \\ z_{qv} &= a_{q1}y_{1v} + a_{q2}y_{2v} + \dots + a_{qp}y_{pv}, \end{aligned} \quad (23.25)$$

wobei $a_{11}, a_{12}, \dots, a_{pq}$ die Gewichtungskoeffizienten sind, die beschreiben, in welchem Ausmaß eine Itemvariable in eine Komponente eingeht.

Der wesentliche Vorgang in der PCA besteht in der Berechnung der *Eigenwerte* und der *Eigenvektoren* der empirischen Korrelationsmatrix S . Für jede Korrelationsmatrix von p Items können stets p Eigenwerte bestimmt werden. Diese geben an, wie groß die Varianzen der Scores der jeweiligen Hauptkomponenten sind, d. h., sie sind ein Maß für die Bedeutsamkeit der Komponenten zur Abbildung interindividueller Unterschiede zwischen den Itemvariablen.

Für jeden Eigenwert kann ein Eigenvektor bestimmt werden, der Gewichtungskoeffizienten a_{ji} enthält, die die Lage (d. h. den Winkel bezogen auf das Koordinatensystem der Itemvariablen) der jeweiligen Hauptkomponente bestimmen. Die Gewichtungskoeffizienten geben weiterhin an, wie stark die Komponenten mit den einzelnen Itemvariablen zusammenhängen. Die *Bestimmung der Koeffizienten* ist bei der PCA eindeutig, während sie bei der EFA uneindeutig ist, wie in ► Abschn. 23.5 zur Faktorenrotation (und Faktorenindeterminiertheit) besprochen wird. In einer PCA ergibt es keinen Sinn, die Komponenten zu rotieren, da sie dadurch ihre Eigenschaft verlieren würden, maximale Varianz zu erklären.

Modell der PCA

Eigenwerte der Korrelationsmatrix bestimmen die Relevanz der Komponenten

Eigenvektoren beinhalten eindeutige Gewichtungskoeffizienten

Hinweis: Ein häufig genannter Kritikpunkt der PCA besteht darin, dass behauptet wird, dass Itemwerte messfehlerfrei seien und nur aus wahrer Varianz bestünden, wodurch Kommunalitäten von eins (und Spezifitäten von null) für die Items

Messfehler, Kommunalität und Spezifität sind in der PCA nicht definiert

Kritik an der PCA ist eigentlich eine Kritik an ihrer unangemessenen Verwendung

Eigenwertextraktion aus reduzierter Korrelationsmatrix S^*

Kommunalitätenproblem und iterative Schätzung

resultieren würden. Dieser Kritikpunkt ist jedoch unangebracht und auf eine inadäquate Anwendung der PCA zurückzuführen. Ziel der PCA ist nicht die Erklärung von Items und ihrem Korrelationsmuster. Wie in den Modellgleichungen oben dargestellt (Gl. 23.25) werden die Hauptkomponenten als Linearkombination der Itemvariablen formuliert (und nicht andersherum). Sie repräsentieren eine optimale Gewichtung der Items, um interindividuelle Unterschiede abzubilden. Es ist somit ein „mathematisches Verfahren“ (s. Rencher und Christensen 2012). Hingegen ist die EFA (s. u.) ein statistisches Verfahren, bei dem die Itemvariablen durch die Faktoren und einen Messfehler beschrieben werden. Die Begriffe Kommunalität und Spezifität spielen in der PCA keine Rolle und sollten daher nicht für die PCA verwendet werden.

Wird eine PCA durch den Anwender dennoch so verwendet, als wäre es eine EFA, und das Verfahren wird z. B. mit einer Rotation verknüpft, dann ist dies keine PCA mehr, da die Komponenten ihre Bedeutung verlieren. Von dieser Verwendungsart der PCA ist grundsätzlich abzuraten, da aus dieser inadäquaten Anwendung tatsächlich die oben behaupteten Nachteile entstünden.

23.3.2 Hauptachsenanalyse (Principal Factor Analysis, PFA)

Ähnlich wie in der PCA werden in der PFA Eigenwerte und Eigenvektoren einer Matrix berechnet und als Kriterium für die Anzahl der extrahierten Faktoren und der zugehörigen Faktorladungen der Itemvariablen herangezogen. Bei der PFA werden hierbei aber die Spezifitäten der Itemvariablen (Varianzen der Residuen) berücksichtigt, d. h., es wird – in Übereinstimmung mit der Annahme von Messfehlern in der KTT (► Kap. 13) – davon ausgegangen, dass ein Teil der beobachteten Varianz nicht durch die gemeinsamen Faktoren erklärt werden kann (Gl. 23.7). Dies erfolgt in der PFA dadurch, dass die Eigenwerte nicht aus der Korrelationsmatrix S der Items direkt berechnet werden (wie in der PCA, ► Abschn. 23.3.1), sondern aus einer sog. „reduzierten Korrelationsmatrix“ S^* , die eine um die Spezifität korrigierte Form darstellt. Für eine Korrelationsmatrix bedeutet dies, dass die Einsen in der Diagonalen durch Schätzungen für die *Kommunalitäten* der Items ersetzt werden (► Abschn. 23.2.4.2); diese sind entsprechend kleiner als eins und maximal so groß wie die Itemreliabilität.

Da anfänglich weder die Faktorladungen bzw. Kommunalitäten noch die Spezifitäten bekannt sind (sog. „Kommunalitätenproblem“), muss für die PFA ein iteratives Schätzverfahren verwendet werden, das basierend auf Initialschätzungen für die Kommunalitäten eine Berechnung der Eigenwerte von S^* und der Spezifitäten erlaubt (für die Wahl verschiedener Initialschätzungen s. Mulaik 2010). Diese Schätzungen werden dann in einem nächsten Schritt für eine verbesserte Schätzung der Kommunalitäten verwendet, auf die wiederum eine erneute Berechnung der Eigenwerte erfolgt. Dieses iterative Verfahren wiederholt sich so lange, bis sich die Schätzungen der Kommunalitäten von einer Iteration zur nächsten nicht mehr substantiell ändern (sog. „Konvergenzkriterium“).

Die Ergebnisse einer PFA manifestieren sich in den Eigenwerten der Faktoren und in den dazugehörigen Eigenvektoren. Die Eigenwerte liefern eine Information über die Bedeutsamkeit der Faktoren. Die Eigenvektoren beinhalten die Faktorladungen und beschreiben – wie in der PCA – die Bedeutung der einzelnen Itemvariablen für die (zunächst unrotierten) Faktoren (nähtere Ausführungen zur Faktorenrotation folgen in ► Abschn. 23.5).

23.3.3 Maximum-Likelihood-Faktorenanalyse (ML-EFA)

Die ML-EFA ist ein alternatives Schätzverfahren zur PFA. Das ML-Prinzip ist ein sehr allgemeines Schätzprinzip, das in sehr vielen statistischen Verfahren angewendet wird, z. B. in Strukturgleichungsmodellen, Multilevelmodellen oder der logistischen Regression. Ein in der Praxis relevanter Unterschied zwischen PFA und ML-EFA besteht darin, dass die ML-Parameterschätzung auf der Annahme einer *multivariaten Normalverteilung* der Itemvariablen basiert.

In der ML-EFA wird die folgende Diskrepanzfunktion F zwischen der modellimplizierten Korrelationsmatrix $\Sigma(\theta)$ und der empirischen Korrelationsmatrix S minimiert:

$$F(\Sigma(\theta) - S) \rightarrow \min \quad (23.26)$$

Inhaltlich bedeutet dies, dass die Parameter (Faktorladungen und Residualvarianzen) so geschätzt werden, dass sich die resultierende modellimplizierte Korrelationsmatrix (die auf diesen Parametern basiert) und die empirische Korrelationsmatrix möglichst ähnlich sind. Die Schätzung erfolgt anhand einer Log-Likelihood-Funktion LL :

$$LL := F(\Sigma(\theta) - S) = \log |\Sigma(\theta)| - \log |S| + \text{tr}(S\Sigma(\theta)^{-1}) - p, \quad (23.27)$$

wobei $\log |\Sigma(\theta)|$ bzw. $\log |S|$ die natürlichen Logarithmen der Determinanten von $\Sigma(\theta)$ bzw. von S sind, $\text{tr}(S\Sigma(\theta)^{-1})$ die Spur des Matrixprodukts $S\Sigma(\theta)^{-1}$ und p die Anzahl der Items. Die Log-Likelihood-Funktion ist identisch mit jener, die in einer CFA verwendet wird (s. ▶ Kap. 24, ▶ Abschn. 24.5.1). Sie ist die Grundlage für den Modelldifferenztest, wie er in ▶ Abschn. 23.4.4 beschrieben wird.

Ein Vorteil der ML-EFA besteht darin, dass für die Faktorladungen auch *Standardfehler* berechnet werden können, die dann zur Bildung von Konfidenzintervallen und zur Beurteilung der Signifikanz der Faktorladungen herangezogen werden können. Ebenso stehen weitere *Gütekriterien* zur Modellbeurteilung zur Verfügung (▶ Abschn. 23.6.1.2). Es sollte beachtet werden, dass im Falle einer Verletzung der Normalverteilungsannahme (z. B. bei deutlich nicht normalverteilten oder grobstufigen Itemvariablen) insbesondere die Standardfehler und die Gütemaße einer ML-EFA mit Vorsicht interpretiert werden sollten. In solchen Fällen wäre es angemessener, „Robuste“ Schätzverfahren (LRM) einzusetzen (▶ Abschn. 23.7.2.2).

An dieser Stelle soll noch auf einen weiteren praktischen Unterschied zwischen der PFA und der ML-EFA hingewiesen werden. In einer PFA kann man bis zu $p - 1$ Faktoren extrahieren (z. B. Brown 2015), d. h., man kann einen Faktor weniger, als Items vorhanden sind, identifizieren. In der ML-EFA muss die Identifikation eines ML-EFA-Modells anhand der Freiheitsgrade des Modells (df) überprüft werden. Für ein identifiziertes Modell müssen die Freiheitsgrade null oder größer sein. Die Freiheitsgrade berechnen sich als Differenz aus der Anzahl empirischer Informationen (s) und der Anzahl der zu schätzenden Parametern (t):

$$df = s - t \quad (23.28)$$

Die Anzahl der empirischen Informationen entspricht der Anzahl von Varianzen und Kovarianzen im Modell (konkret: $s = p(p + 1)/2$). Die Anzahl der zu schätzenden Parameter ergibt sich in der EFA durch:

$$t = pk + p - k(k - 1)/2, \quad (23.29)$$

wobei p die Anzahl der Items und k die Anzahl der Faktoren sind.³ Für vier Items stehen z. B. $s = 4(4 + 1)/2 = 10$ Informationen zur Verfügung. Ein Zweifaktor-

Annahme: multivariate

Normalverteilung der Itemvariablen

Diskrepanzfunktion wird minimiert

Log-Likelihood-Funktion

**Standardfehler und
Modellgütekriterien**

**Modellidentifikation und
Freiheitsgrade**

³ Für die Berechnung der Freiheitsgrade spielt es hier keine Rolle, ob eine Korrelations- oder Kovarianzmatrix verwendet wird, da sich die Anzahl der Varianzen bei der Differenzbildung wegfürzt.

modell ohne weitere Restriktionen besitzt in diesem Fall $t = 4 \cdot 2 + 4 - 2 \cdot (2-1)/2 = 11$ Parameter und wäre nicht identifiziert, da $df = 10 - 11 = -1 < 0$.

23.4 Abbruchkriterien der Faktorextraktion

Dem Sparsamkeitsprinzip folgend sollte das Ziel in der EFA darin bestehen, ein Modell mit einer möglichst geringen Anzahl von Faktoren zu finden, das aber trotzdem eine gute Modell-Daten-Passung aufweist. Um über die Anzahl der relevanten Faktoren für eine Gruppe von Items zu entscheiden, existieren verschiedene sog. „Abbruchkriterien“, denen (bis auf den Modelldifferenztest) gemeinsam ist, dass sie sich auf die Eigenwerte der (reduzierten) Korrelationsmatrix beziehen. Typischerweise beinhalten die Methoden eine Beurteilung des *Eigenwerteverlaufs*. Eine Entscheidung darüber, ob es sich um relevante Faktoren handelt, erfolgt anhand einer Beurteilung der relativen Bedeutung der Eigenwerte. Faktoren, die keine bedeutsame gemeinsame Varianz erklären können, sollten in der weiteren Analyse nicht mehr berücksichtigt werden.

Die gebräuchlichsten Kriterien sind das *Kaiser-Guttman-Kriterium* (► Abschn. 23.4.1), das *Elbow-Kriterium* (häufig auch als „Scree-Test“ bezeichnet; ► Abschn. 23.4.2) und die *Parallelanalyse* (► Abschn. 23.4.3). Für die ML-EFA steht ein *Modelldifferenztest* zur Verfügung, der zur Beurteilung der „Passung“ zwischen Modell und Daten verwendet werden kann (► Abschn. 23.4.4). Alle diese Kriterien haben ihre Vor- und Nachteile. Üblicherweise sollten mehrere Kriterien verwendet und miteinander verglichen werden, bevor eine finale Entscheidung über die angemessene Anzahl der Faktoren getroffen wird. Es ist vorteilhaft, wenn die verschiedenen Kriterien zu derselben Entscheidung führen. Ist dies nicht der Fall, so sollte anhand der Eigenschaften der Kriterien entschieden werden, welche Faktorenanzahl relevant ist.

23.4.1 Kaiser-Guttman-Kriterium

Gemäß dem Kaiser-Guttman-Kriterium werden diejenigen Faktoren als relevant erachtet, deren Eigenwerte größer sind als die durchschnittliche Varianz $\overline{Var(y)}$ eines Items (Guttman 1954; Kaiser 1960):

$$\Lambda_j > \overline{Var(y)} \quad (23.30)$$

Im Falle einer Korrelationsmatrix ist die durchschnittliche Varianz $\overline{Var(y)} = 1$ und entsprechend werden Faktoren mit einem Eigenwert größer eins als relevant angesehen. Die Logik dieses Kriteriums besteht darin, dass nur solche Faktoren berücksichtigt werden, die mehr Varianz erklären als ein durchschnittliches Item an Varianz aufweist.

Der Vorteil dieses Kriteriums besteht in seiner Einfachheit und Objektivität, da ausschließlich die absolute Größe der Eigenwerte anhand eines klaren Cut-off-Wertes betrachtet werden muss.

Wie Vergleiche mit anderen Kriterien zeigen (z. B. Cattell und Jaspers 1967) besteht der Nachteil dieses Kriteriums darin, dass fast immer zu viele Faktoren als relevant erachtet werden; dieses Problem vergrößert sich bei zunehmender Itemanzahl (s. z. B. Glorfeld 1995; Zwick und Velicer 1986). Die Entscheidungen nach diesem Kriterium sind nur sehr begrenzt belastbar, eine (unkritische) Anwendung ist daher nicht empfehlenswert.

Hinweis: Weiterhin sollte angemerkt werden, dass das Kriterium ursprünglich für die PCA entwickelt wurde, nicht jedoch zur Beurteilung von Eigenwerten einer

Beurteilung der relativen Bedeutung der Eigenwerte

Entscheidung über Dimensionalität basiert auf mehreren Maßen

Eigenwerte größer eins werden als relevant erachtet

Vorteil: Einfachheit

Nachteil: zu viele Faktoren

23.4 · Abbruchkriterien der Faktorextraktion

reduzierten Korrelationsmatrix (Gorsuch 1980; Horn 1969). Bei reduzierten Korrelationsmatrizen sollte deshalb anstelle der Schranke „größer eins“ der Durchschnitt der Eigenwerte aller Faktoren herangezogen werden, da bei diesem Kriterium die Unreliabilität der Items berücksichtigt wird. Dies führt aber nicht zu einer deutlichen Verbesserung der Anwendbarkeit dieses Kriteriums.

Abhilfe: Eigenwerte größer als durchschnittlicher Eigenwert werden als relevant erachtet

23.4.2 Elbow-Kriterium (Scree-Test)

Auch die relative Größe der Eigenwerte zueinander kann als Entscheidungskriterium für die Dimensionalität herangezogen werden: Sind einige Eigenwerte deutlich größer als die übrigen Eigenwerte, so werden diejenigen Faktoren mit relativ großen Eigenwerten als relevant erachtet. Die Idee dieses Kriteriums besteht darin, dass es häufig nur einige wenige Faktoren gibt, die vergleichsweise große Eigenwerte aufweisen und somit von Relevanz sind. Die weiteren Faktoren mit kleineren Eigenwerten werden als „Geröllkomponenten“ („scree“ bedeutet Geröll und weist auf das Geröll unter einer Klippe hin) ohne Relevanz erachtet (Cattell 1966; Cattell und Jaspers 1967).

Die Überprüfung nach dem Elbow-Kriterium erfolgt üblicherweise durch eine grafische Darstellung des Eigenwertverlaufs in einem sog. „Scree-Plot“ (Abb. 23.2). Im Scree-Plot wird (von links nach rechts) die Nummer des Faktors ($1, \dots, k$) gegen den jeweiligen Eigenwert des Faktors abgetragen. Es werden dann diejenigen Faktoren gewählt, die links vom Knick – dem „Elbow“ – des Eigenwerteverlaufs liegen. Im dargestellten Beispiel in Abb. 23.2a werden also zwei Faktoren als relevant erachtet.

Der Vorteil dieses Kriteriums besteht darin, dass es in eindeutigen Situationen zu guten Entscheidungen führt, d. h., wenn ein eindeutiger Knick („Elbow“) zu erkennen ist (Cattell und Vogelmann 1977; Hakstian et al. 1982; Tucker et al. 1969). Dieser Fall tritt häufiger auf, wenn die Dimensionalität klein ist (z. B. ein oder zwei relevante Faktoren).

Ein Nachteil dieses Kriteriums besteht darin, dass in manchen Situationen ein Eigenwerteverlauf entsteht, der keinen eindeutigen Knick aufweist (Abb. 23.2b); in diesen Fällen ist eine Entscheidung nur schwer möglich und ist zumeist subjektiv

Beurteilung der relativen Größe der Eigenwerte

Scree-Plot

Zuverlässig bei eindeutigem Eigenwertverlauf

Bei uneindeutigem Eigenwertverlauf nicht verwendbar

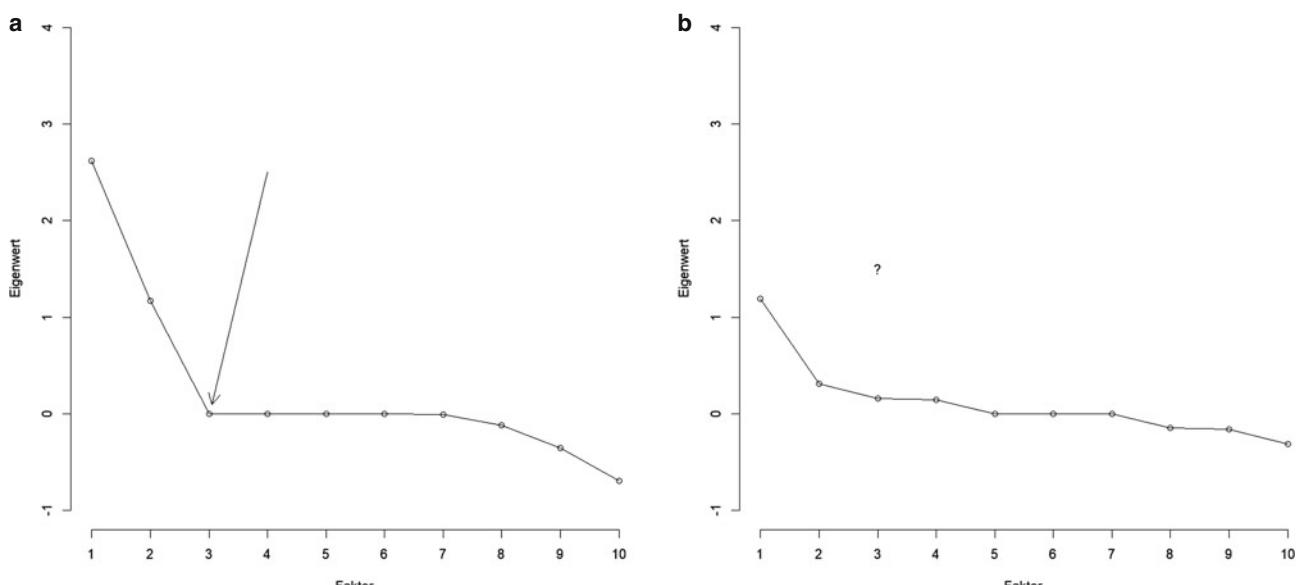


Abb. 23.2 a Eindeutige Entscheidung nach dem Elbow-Kriterium für zwei Faktoren, die links vom Knick liegen (mit Pfeil angedeutet). b Keine eindeutige Entscheidung möglich, da kein deutlicher Knick erkennbar ist

(z. B. Kaiser 1960). Eine solche Situation tritt auf, wenn den Daten keine klare Struktur mit wenigen relevanten Faktoren (dominanten Dimensionen) zugrunde liegt oder wenn sehr viele Facetten simultan untersucht werden (z. B. Items zu acht oder mehr Konstrukten).

Häufig werden fälschlich Eigenwerte der nicht reduzierten Korrelationsmatrix verwendet

Hinweis: Es sollte beachtet werden, dass in den meisten Software-Programmen (wie SPSS) für Scree-Plots die Eigenwerte der Korrelationsmatrix verwendet werden und nicht die der reduzierten Korrelationsmatrix. Hierdurch wird implizit (und wenig plausibel) angenommen, dass die Kommunalität aller Items gleich eins sei, was bedeuten würde, dass alle Items frei von Messfehlern sind und somit sämtliche Varianz durch die Faktoren erklärt werden könnte. Dies führt insbesondere dann zu Unterschieden bei der Beurteilung der Eigenwerte, wenn die tatsächlichen Kommunalitäten klein sind. Wenn korrechterweise eine reduzierte Korrelationsmatrix verwendet wurde (z. B. im R-Paket „psych“; Revelle 2012), können negative Eigenwerte auftreten, da die Matrix nicht positiv definit ist (es handelt sich dann also nicht um einen Berechnungsfehler).

23.4.3 Parallelanalyse

Die Idee der Parallelanalyse (Horn 1965) besteht darin, analog zu den empirischen Werten der Itemvariablen unkorrelierte Zufallsdaten zu erzeugen. Für deren Korrelationsmatrix werden anschließend die Eigenwerte ermittelt und mit den Eigenwerten der empirischen Daten verglichen, um entscheiden zu können, welche Faktoren relevant sind und welche nicht. Als relevant werden jene Faktoren der empirischen Daten erachtet, deren Eigenwerte größer sind als die Eigenwerte der Zufallsdaten.

Das Verfahren besteht aus drei Schritten:

1. *Datengenerierung:* Im Sinne einer Monte-Carlo-Simulation (s. z. B. Kröse et al. 2011) werden mehrere Datensätze mit Zufallsdaten (z. B. 100 Datensätze) anhand eines vorher spezifizierten *Populationsmodells* generiert. Das jeweilige Populationsmodell soll einerseits den empirischen Daten entsprechen, und zwar hinsichtlich der Stichprobengröße, der Anzahl der Items sowie im Falle der Analyse einer Kovarianzmatrix der Varianzen der Itemvariablen. Andererseits soll es sich aber in einem wesentlichen Punkt von den empirischen Daten unterscheiden: Es wird angenommen, dass in der Population alle Variablen unkorreliert sind. Damit stellt das Populationsmodell ein geeignetes Vergleichsmodell im Sinne einer Nullhypothese dar, das einen Eigenwerteverlauf liefert, wie er bei bedeutungslosen Faktoren zu erwarten wäre. Da die Zufallsdaten keine systematische gemeinsame Varianz aufweisen, müssten so viele Faktoren extrahiert werden, wie es Variablen gibt.
2. *Auswertung der Zufallsdaten:* Nach der Generierung der Zufallsdatensätze werden für jeden einzelnen Datensatz die Eigenwerte der reduzierten Korrelationsmatrizen dieser Daten berechnet. Über die Datensätze hinweg erhält man somit eine Verteilung der Eigenwerte (z. B. 100 Eigenwertverläufe). Dadurch können Fehler aufgrund der Stichprobenschwankungen minimiert werden. Aus der Verteilung der Eigenwerte werden die Mittelwerte für jeden der Eigenwerte gebildet. Weiterhin können die zugehörigen bzw. 95 %-igen (Perzentil-)Konfidenzintervalle (Glorfeld 1995) berechnet werden.
3. *Visualisierung der Eigenwertverläufe:* Der gemittelte Verlauf der Eigenwerte der Zufallsdaten (bzw. deren Konfidenzintervalle) kann im Scree-Plot mit dem Eigenwerteverlauf der empirischen Daten verglichen werden (► Abschn. 23.4.2). Es werden diejenigen Faktoren als relevant erachtet, deren empirische Eigenwerte oberhalb des Verlaufs der Eigenwerte der Zufallsdaten liegen (im dargestellten Beispiel in □ Abb. 23.3 also zwei Faktoren). Werden

Generierung künstlicher Zufallsdaten

Berechnung der mittleren Eigenwerte der Zufallsdaten

Vergleich mit Eigenwerten der empirischen Daten

23.4 · Abbruchkriterien der Faktorextraktion

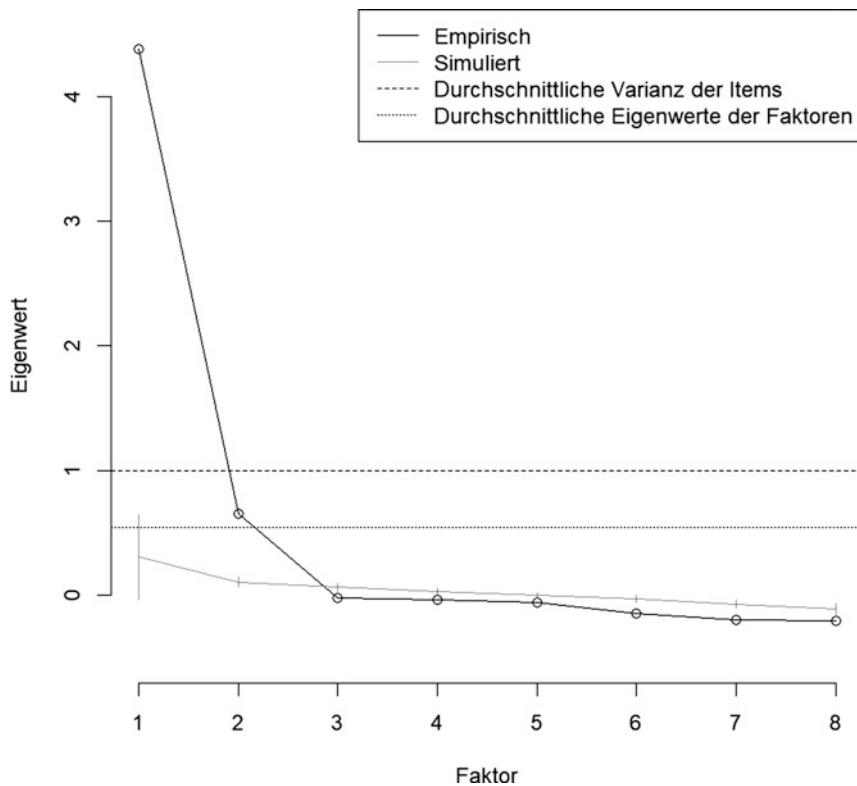


Abb. 23.3 Eigenwerteverlauf für das Beispiel (schwarz) und für die Parallelanalyse (grau; vertikale Balken zeigen die zugehörigen 95 %-Konfidenzintervalle). Das Kaiser-Guttman-Kriterium ist mit einer gestrichelten Linie für die durchschnittliche Varianz der Items (eins) bzw. einer gepunkteten Linie für den durchschnittlichen Eigenwert der Faktoren (0.55) gekennzeichnet

die Konfidenzintervalle verwendet, so werden diejenigen Faktoren gewählt, deren Eigenwerte oberhalb der oberen Intervallgrenzen liegen. Dies führt zu konservativeren Entscheidungen.

Der Vorteil der Parallelanalyse ist darin zu sehen, dass sie zuverlässige Ergebnisse liefert und daher häufig als Methode der Wahl genannt wird (Dinno 2009; Humphreys und Montanelli 1975; Timmermann und Lorenzo-Seva 2011; Zwick und Velicer 1986). In Datensätzen mit unklarem Eigenwerteverlauf (also ohne Knick) sollte die Parallelanalyse auf jeden Fall herangezogen werden.

Der Nachteil der Parallelanalyse besteht darin, dass ihre Anwendung aufwendiger ist als die bisher genannten Kriterien. Dieser Nachteil wird jedoch gerade in letzter Zeit durch direkt verwendbare Software-Skripte (z. B. die R Pakete „paran“ oder „psych“; Dinno 2009; Revelle 2012) abgeschwächt.

Auch bei der Parallelanalyse sollte darauf geachtet werden, dass die Eigenwerte aus einer reduzierten Korrelationsmatrix berechnet wurden.

Die Bestimmung von relevanten Dimensionen wird in ► Beispiel 23.2 exemplarisch aufgezeigt.

Methode der Wahl

Aufwendigere Überprüfung

Beispiel 23.2: Bestimmung der relevanten Dimensionen für das empirische Beispiel

Um zu bestimmen, wie viele Faktoren im empirischen Beispiel zur Erklärung der gemeinsamen Varianz aller acht Items notwendig sind (s. Beschreibung in ► Beispiel 23.1), wurde eine Hauptachsenanalyse (PFA) durchgeführt. Die folgenden Eigenwerte resultierten: 4.38, 0.66, -0.02, -0.03, -0.06, -0.15, -0.20, -0.20.

■ Abb. 23.3 illustriert diese empirischen Eigenwerte sowie die simulierten Eigenwerte der Parallelanalyse. Nach dem Kaiser-Guttman-Kriterium sollten ein bzw. zwei Faktoren extrahiert werden, je nachdem, ob man die durchschnittliche Varianz der Items oder (besser) die durchschnittlichen Eigenwerte verwendet, da diese den Itemmessfehler berücksichtigen. Im Sinne des Elbow-Kriteriums ist ein Knick erkennbar, der eine zweifaktorielle Lösung nahelegt. Die Parallelanalyse weist ebenfalls auf zwei Faktoren hin, da der Verlauf der Parallelanalyse ab dem dritten Faktor oberhalb des empirischen Verlaufs liegt. Zusammengenommen weisen die Kriterien auf zwei Faktoren hin.

23.4.4 Modelldifferenztest (ML-EFA)

In der ML-EFA kann eine Entscheidung über die Anzahl der Faktoren zusätzlich statistisch abgesichert werden, und zwar mit einem *Modelldifferenztest* (oder auch sog. „Likelihood-Quotienten-Test“), in dem jeweils zwei Modelle mit unterschiedlicher Faktorenanzahl verglichen werden. Die Prüfgröße ΔLL errechnet sich aus der Differenz der zwei Log-Likelihood-Werte (► Abschn. 23.3.3, Gl. 23.27) der beiden Modelle als (vgl. Fabrigar und Wegener 2012):

$$\begin{aligned}\Delta LL = & -2LL \text{ (Modell mit weniger Faktoren)} \\ & - (-2LL \text{ (Modell mit mehr Faktoren)})\end{aligned}\quad (23.31)$$

Diese Prüfgröße ist approximativ (d. h. bei großen Stichproben) χ^2 -verteilt mit Δdf Freiheitsgraden:

$$\begin{aligned}\Delta df = & df \text{ (Modell mit weniger Faktoren)} \\ & - df \text{ (Modell mit mehr Faktoren)}\end{aligned}\quad (23.32)$$

Ein signifikanter Test bedeutet, dass die zusätzlichen Faktoren eine signifikant bessere Anpassung der modelltheoretischen Matrix $\Sigma(\theta)$ an die empirische Matrix (S) erlauben. Ein nicht signifikanter Test bedeutet entsprechend, dass das sparsamere Modell mit weniger Faktoren ausreicht, um das Korrelationsmuster der Items zu erklären. ► Beispiel 23.3 illustriert den Modelldifferenztest für das empirische Beispiel.

Es bleibt zu beachten, dass insbesondere dieser inferenzstatistische Test sensibel auf die Verletzung der Normalverteilungsannahme reagiert (z. B. auch bei grobstufig skalierten Items) und in diesen Fällen zu einer falschen Entscheidung hinsichtlich der Dimensionalität führen kann (► Abschn. 23.7.2.2).

Beispiel 23.3: Der Modelldifferenztest für das empirische Beispiel

Für das empirische Beispiel wurde nun ebenfalls eine ML-EFA durchgeführt. ■ Tab. 23.1 zeigt die Ergebnisse für die Modelle mit 1, 2 und 3 Faktoren. Der Modelldifferenztest zwischen je zwei Modellen zeigt einen signifikanten Unterschied zwischen den Modellen mit 1 vs. 2 Faktoren, nicht jedoch zwischen den Modellen mit 2 vs. 3 Faktoren. Dies deutet darauf hin, dass das Dreifaktormodell die Daten nicht signifikant besser abbildet als das Zweifaktormodell; ein dritter Faktor bringt also keine weitere Verbesserung der Modell-Daten-Passung.

23.5 · Faktorenrotation

■ **Tabelle 23.1** Ergebnisse der ML-EFA und des Modelldifferenztests

Faktoren	df	χ^2	Δdf	$\Delta\chi^2$	p
1	20	794.65			
2	13	28.53	7	766.13	<0.01
3	7	16.99	6	11.54	0.07

df = Freiheitsgrade

23.5 Faktorenrotation

23.5.1 Faktorenindeterminiertheit

Eine grundsätzliche Eigenschaft der Faktorenanalyse (unabhängig vom Extraktionsverfahren) besteht darin, dass die Faktorladungsmatrix nicht eindeutig bestimmt ist, sobald mehr als ein Faktor extrahiert wird. Konkret bedeutet dies, dass es eine unendlich große Anzahl äquivalenter Lösungen gibt (bei konstanter gehaliner Zahl von Faktoren), die die Daten gleich gut widerspiegeln. Diese als *Faktorenindeterminiertheit* oder Faktorenunbestimmtheit bezeichnete Eigenschaft wird in ► Exkurs 23.2 näher erklärt.

Faktorenunbestimmtheit: Unendlich viele äquivalente Faktorenlösungen

Exkurs 23.2

Faktorenindeterminiertheit

Formal produziert jede (beliebig gewählte) orthogonale Matrix⁴ \mathbf{T} eine neue Faktorenlösung der Form $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}\mathbf{T}$. Diese ist im Sinne der Modellanpassung äquivalent zur ursprünglichen Faktorenlösung, denn die modelltheoretische Korrelationsmatrix ist in beiden Fällen identisch:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) &= \boldsymbol{\lambda}^*\boldsymbol{\lambda}^{*'} + \boldsymbol{\Psi}^* = \boldsymbol{\lambda}\mathbf{T}(\boldsymbol{\lambda}\mathbf{T}') + \boldsymbol{\Psi}^* \\ &= \boldsymbol{\lambda}\mathbf{T}\mathbf{T}'\boldsymbol{\lambda}' + \boldsymbol{\Psi}^* = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}^* = \boldsymbol{\Sigma}(\boldsymbol{\theta})\end{aligned}\quad (23.33)$$

da für orthogonale Matrizen stets gilt, dass $\mathbf{T}\mathbf{T}' = \mathbf{I}$ ist (eine Einheitsmatrix; vgl. Rencher und Christensen 2012).

Analog ergibt sich für jede der Lösungen derselbe Log-Likelihood-Wert in einer ML-EFA, d. h., aus den äquivalenten Lösungen kann nicht ein Modell (ohne weitere Annahmen) gewählt werden, das eine bessere Modell-Daten-Passung aufweist.

Inhaltlich kann die Multiplikation einer Faktorladungsmatrix $\boldsymbol{\lambda}$ mit einer orthogonalen Matrix \mathbf{T} als eine *rechtwinklige Rotation* der Faktorenlösung verstanden werden. Da ein Faktor seine inhaltliche Bedeutung ausschließlich aus dem Faktorladungsmuster bezieht (dadurch, welche Items auf ihm laden und welche nicht), ist auch die Bedeutung der Faktoren an die jeweilige Faktorenlösung gebunden. Für jede neue Faktorenlösung $\boldsymbol{\lambda}^*$ existiert entsprechend auch ein transformierter Faktoraum $\boldsymbol{\eta}^* = \mathbf{T}'\boldsymbol{\eta}$ mit entsprechend anders zu interpretierenden Faktoren.

4 Eine orthogonale Matrix beinhaltet ausschließlich zueinander orthogonale Vektoren, z. B. (0,1) und (1,0).

Weiteres Kriterium notwendig, um sinnvolle Lösungen zu generieren

Welche Konsequenz sollte man aus der Faktorendeterminiertheit ziehen? Es müssen weitere externe Kriterien formuliert werden, damit eine sinnvolle Lösung aus der unendlichen Lösungsmenge gewählt werden kann. Dieses Kriterium ist die *Einfachstruktur*, die je nach Rotationsverfahren zwar verschieden operationalisiert wird, die aber immer eine sinnvolle Lösung produziert, die im Sinne des gewählten Kriteriums auch eine optimale Lösung darstellt⁵.

23.5.2 Einfachstruktur

Erreichung der Einfachstruktur durch große Primärladungen und geringe Sekundärladungen

Allgemein gesprochen bedeutet Einfachstruktur, dass ein Item möglichst nur eine einzige *Primärladung* auf dem für das Item relevanten Faktor aufweist, aber nur wenige oder keine *Sekundärladungen* auf anderen Faktoren (für die ursprüngliche, spezifischere Definition s. Thurstone 1947). Die Einfachstruktur soll durch die Rotation der Faktoren und der Faktorladungsmatrix erreicht werden. Eine erfolgreiche Rotation erlaubt eine möglichst eindeutige Zuordnung von Items zu Faktoren, wodurch die inhaltliche Interpretation der Faktoren erleichtert wird.

Die Operationalisierung der Einfachstruktur kann anhand einer sog. „Komplexitätsfunktion“ $f(\lambda)$ erfolgen. Die Komplexitätsfunktion berechnet für eine gegebene Faktorenlösung ein Maß basierend auf Paaren von Faktorladungen. Crawford und Ferguson (1970) beschrieben verschiedene Rotationsmethoden basierend auf folgendem Komplexitätsmaß:

$$f(\lambda) = (1 - \kappa) \underbrace{\sum_{i=1}^p \sum_{j=1}^k \sum_{j' \neq j}^k \lambda_{ij}^2 \lambda_{ij'}^2}_{\text{Zeilenkomplexität (Items)}} + \kappa \underbrace{\sum_{j=1}^k \sum_{i=1}^p \sum_{i' \neq i}^p \lambda_{ij}^2 \lambda_{i'j}^2}_{\text{Spaltenkomplexität (Faktor)}} \quad (23.34)$$

Einfachstruktur entsteht durch Minimierung der Komplexitätsfunktion

Das Komplexitätsmaß ist eine mit $(1 - \kappa)$ bzw. κ gewichtete Summe aus Zeilen- und Spaltenkomplexität, wobei κ zwischen null und eins liegt. Das Ziel ist es, die Komplexitätsfunktion zu minimieren, wodurch eine Faktorladungsstruktur entsteht, die dem (mit κ) gewählten Einfachstrukturmuster entspricht.

Anhand der oben gegebenen Gleichung kann für jede Faktorladungsmatrix eine Komplexität berechnet werden. Hierfür werden zunächst für jedes Item i die Faktorladungen quadriert ($\lambda_{i1}^2, \lambda_{i2}^2, \dots, \lambda_{ip}^2$), anschließend werden sämtliche Kombinationen aus jeweils zwei Faktorladungen multipliziert (z. B. $\lambda_{i1}^2 \lambda_{i2}^2, \lambda_{i1}^2 \lambda_{i3}^2, \dots, \lambda_{i1}^2 \lambda_{ip}^2$). Alle diese Produkte werden sodann zunächst für das Item i aufsummiert; anschließend erfolgt die gleiche Prozedur für alle verbliebenen Items. Die Summe über alle Items hinweg ergibt dann die *Zeilenkomplexität* (eine konkrete Berechnung im Minimalbeispiel erfolgt in ► Beispiel 23.4). Für die *Spaltenkomplexität* erfolgt eine analoge Berechnung für die Faktorladung jedes Faktors j . In dem Fall, in dem ein Item nur auf einem einzigen Faktor lädt und alle Ladungen auf anderen Faktoren null sind, ist die (Zeilen-)Komplexität dieses Items null, da alle Produkte von Faktorladungen null ergeben. Wenn ein Item eine Doppelladung aufweist, wird die Komplexität des Items größer als null, weil in diesem Fall mindestens ein Produkt der quadrierten Faktorladungen nicht null ist. Dasselbe gilt für die Komplexität der Faktoren: Sind die Faktorladungen aller Items außer einem einzigen Item auf einem Faktor gleich null, so ist die Komplexität dieses Faktors null. Die Zielfigur besteht darin, durch die Rotation eine Minimierung der Spalten-/Faktorenkomplexität und/oder der Zeilen-/Itemkomplexität zu erreichen,

Komplexitätsfunktion als Maß für Einfachstruktur

⁵ In der Literatur wird mit Faktorendeterminiertheit noch ein weiteres Problem diskutiert: Selbst in einer Situation mit nur einem einzelnen Faktor sind die Faktorwerte (und der Faktor selbst) nicht eindeutig bestimmt (Details s. Maräu 1996; Mulaik 2010; Steiger 1979). Das Problem vergrößert sich in Situationen, in denen Items hohe Spezifitäten aufweisen.

23.5 · Faktorenrotation

wobei mit dem Gewicht κ die relative Bedeutung der beiden Komplexitätsarten variiert wird.

Beispiel 23.4: Illustration der Komplexität am Minimalbeispiel

Die folgende Faktorladungsmatrix stammt aus ► Beispiel 23.1.

$$\lambda = \begin{pmatrix} 0.80 & 0.18 \\ 0.69 & 0.28 \\ 0.27 & 0.96 \end{pmatrix} \quad (23.35)$$

Die Itemkomplexität (Zeilenkomplexität) ist gegeben durch:

$$\sum_{i=1}^3 \sum_{j=1}^2 \sum_{j' \neq j}^2 \lambda_{ij}^2 \lambda_{ij'}^2 = \underbrace{2 \cdot 0.80^2 \cdot 0.18^2}_{\text{Item 1}} + \underbrace{2 \cdot 0.69^2 \cdot 0.28^2}_{\text{Item 2}} + \underbrace{2 \cdot 0.27^2 \cdot 0.96^2}_{\text{Item 3}} = 0.25 \quad (23.36)$$

Die Faktoreinkomplexität (Spaltenkomplexität) ist gegeben durch:

$$\begin{aligned} \sum_{j=1}^2 \sum_{i=1}^3 \sum_{i' \neq i}^3 \lambda_{ij}^2 \lambda_{i'j}^2 &= \underbrace{2 \cdot (0.80^2 \cdot 0.69^2 + 0.80^2 \cdot 0.27^2 + 0.69^2 \cdot 0.27^2)}_{\text{Faktor 1}} \\ &\quad + \underbrace{2 \cdot (0.18^2 \cdot 0.28^2 + 0.18^2 \cdot 0.96^2 + 0.28^2 \cdot 0.96^2)}_{\text{Faktor 2}} \\ &= 0.98 \end{aligned} \quad (23.37)$$

Wie man aus der Berechnung der Zeilen- und Spaltenkomplexität erkennen kann, weisen Item 1 und 2 geringe Komplexitäten auf (im Vergleich zu Item 3). Beide Faktoren weisen ebenfalls eine Spaltenkomplexität verschieden von null auf, da Item 3 auf beiden Faktoren lädt.

Die Gesamtkomplexität ergibt sich dann als Summe der mit $(1 - \kappa)$ gewichteten Zeilenkomplexität und der mit κ gewichteten Spaltenkomplexität. Das Gewicht κ hängt von dem gewählten Rotationskriterium ab.

Die im Folgenden vorgestellten Methoden der Rotation unterscheiden sich darin, wie sie die Einfachstruktur definieren. Sie lassen sich grundsätzlich in orthogonale und oblique Verfahren unterteilen. *Orthogonale* Rotationsverfahren behalten die Unkorreliertheit der Faktoren bei, wohingegen *oblique* Verfahren eine Korrelation zwischen den rotierten Faktoren erlauben. Je nach Wahl des Gewichts κ erfolgt eine verschiedene Gewichtung der Zeilen- und Spaltenkomplexität, die – insbesondere für orthogonale Rotationen – eine Strukturierung der Verfahren erlaubt.

Strukturierung der Rotationsverfahren

23.5.3 Orthogonale Rotation

Bei einer orthogonalen Rotation werden Faktoren so rotiert, dass die Unkorreliertheit der Faktoren bestehen bleibt. Dies hat den Vorteil, dass Faktoren unabhängig voneinander interpretiert werden können. Die bekanntesten orthogonalen Rotationsverfahren sind das Quartimax- und das Varimax-Verfahren, die der sog. „Orthomax-Familie“ angehören (weitere Familienmitglieder wie das Equamax- und Parsimax-Verfahren sind z. B. bei Browne 2001 zu finden).

Bei orthogonaler Rotation bleibt die Unkorreliertheit der Faktoren erhalten

23.5.3.1 Quartimax-Rotation ($\kappa = 0$)

Im Sinne der oben besprochenen Komplexitätsfunktion erfolgt bei der Quartimax-Rotation keine Minimierung der Faktorenkomplexität ($\kappa = 0$); vielmehr wird ausschließlich die Zeilenkomplexität der Items minimiert ($(1 - \kappa) = 1$).

Die Quartimax-Rotation ist das zuerst entwickelte Rotationsverfahren (Carroll 1953; Ferguson 1954; Neuhaus und Wrigley 1954; Saunders 1953). Der Name „Quartimax“ leitet sich davon ab, dass die Summe der Faktorladungen in der viersten Potenz maximiert werden soll.

Unter Benutzung der Komplexitätsfunktion wird die Einfachstruktur in diesem Verfahren dadurch erzielt, dass für alle Items die Summe der paarweisen Produkte der Faktorladungen minimiert wird. Dies ist dann der Fall, wenn die Rotation dazu führt, dass ein Item nur auf einem einzigen Faktor hoch lädt und auf allen anderen Faktoren niedrig (z. B. für drei Faktoren 0.1, 0.1. und 0.8). In dieser Situation sind die beiden Faktorladungen entweder nahe null (z. B. $0.1 \cdot 0.1 = 0.01$), oder eine Faktorladung ist hoch und die andere ist niedrig (z. B. $0.8 \cdot 0.1 = 0.08$). Sind hingegen mehrere Faktorladungen groß (z. B. 0.5, 0.5, und 0.3), so ist die Summe der Produkte größer und nicht optimal im Sinne des Kriteriums. Das Kriterium minimiert ausschließlich die Itemkomplexität.

Der Hauptnachteil dieses Verfahrens besteht darin, dass häufig ein einzelner Globalfaktor resultiert, auf dem alle Items laden, da das Kriterium für jedes Item (jede Zeile in der Faktorladungsmatrix) separat minimiert wird. Auf den verbleibenden, in der Extraktion eigentlich als relevant erachteten Faktoren laden die Items nach der Rotation nur noch minimal.

23.5.3.2 Varimax-Rotation ($\kappa = 1/p$)

Im Sinne der Komplexitätsfunktion von oben wird bei der Varimax-Rotation eine Gewichtung von $\kappa = 1/p$ (mit p = Anzahl der Items) vorgenommen, d. h., es wird sowohl die Faktoren- als auch die Itemkomplexität berücksichtigt. Bei zunehmender Itemanzahl (d. h. bei größerem p) nimmt die Bedeutung der Itemkomplexität zu (da die Faktorenkomplexität mit $1/p$ gewichtet wird).

Die Idee der Einfachstruktur in der Varimax-Rotation ist eine Weiterentwicklung der Quartimax-Rotation (Kaiser 1958). Sie besteht darin, dass die Varianz der quadrierten Faktorladungen insgesamt maximiert wird. Diese Varianzmaximierung führt im Idealfall dazu, dass Items entweder sehr hoch oder sehr niedrig auf einem Faktor laden. Die hohen Ladungen können dann als *Primärladung* interpretiert werden, die verbliebenen kleineren Ladungen als *Sekundärladungen*. Eine Rotation, die gleichermaßen hohe Faktorladungen eines Items auf allen Faktoren produziert, würde hingegen nicht der Idee der Einfachstruktur folgen. Die Varimax-Rotation kann als die Standardmethode für orthogonale Rotationen angesehen werden.

23.5.4 Oblique Rotation

Oblique Rotation führt zu korrelierten Faktoren

Vorteil: Entspricht eher der Vorstellung psychologisch relevanter Konstrukte

Nachteil: Interpretation erschwert

Bei der obliquen Rotation wird die Orthogonalität der Faktoren zugunsten einer besseren Interpretierbarkeit aufgegeben. Als Konsequenz können die rotierten Faktoren miteinander korrelieren. Dies hat Vor- und Nachteile.

Die Vorteile der obliquen Rotation bestehen darin, dass die (korrelierten) Faktoren häufig eher dem Konzept von psychologischen Konstrukten entsprechen (die auch häufig als interkorreliert angenommen werden, z. B. verschiedene Facetten der Intelligenz, die auf einem Globalfaktor laden, s. z. B. Browne 2001; Fabrigar et al. 1999). Weiterhin ist es einfacher, eine Einfachstruktur zu erzielen, wenn die zugrunde liegenden Konstrukte korreliert sind.

Ein Nachteil besteht in der erschwerten Interpretation der Faktorladungen, weil die Ladungen nicht mehr wie Korrelationen zwischen Faktoren und Items inter-

23.5 · Faktorenrotation

pretiert werden können, wie das im Fall der orthogonalen Rotation möglich ist. Im Fall der obliquen Lösung werden die Faktorladungen wie standardisierte Regressionskoeffizienten in einer multiplen Regression interpretiert. Sie sind somit proportional zu *Semipartialkorrelationen* (zum Begriff s. z. B. Bortz und Schuster 2010), bei denen die Korrelationen zwischen den Faktoren Berücksichtigung finden, indem deren Einfluss herauspartialisiert wird.

Die oblique Rotation beabsichtigt, eine klarere Einfachstruktur hinsichtlich der Semipartialkorrelationen zu erzielen; diese sind in der sog. „Mustermatrix“ („pattern matrix“) enthalten und sollten zur Interpretation der Ergebnisse herangezogen werden. Die Korrelationen zwischen Faktoren und Items können hingegen in der sog. „Strukturmatrix“ („structure matrix“) gefunden werden (Mulaik 2010). Eine Interpretation der Elemente der Strukturmatrix ergibt inhaltlich allerdings nur wenig Sinn, weil auch gute oblique Lösungen mit nicht eindeutigen Strukturmatrizen einhergehen können (z. B. wenn Faktoren hoch miteinander korrelieren).

In einer Situation, in der die zugrunde liegenden Faktoren tatsächlich unkorreliert sind, ist das Ergebnis einer orthogonalen und einer obliquen Rotation sehr ähnlich (Harman 1976). Deshalb kann eine oblique Rotation, die zu unkorrelierten Faktoren führt, als Argument herangezogen werden, anschließend eine orthogonale Rotation durchzuführen, die sparsamer ist (weil sie keine Korrelationen zwischen den Faktoren beinhaltet). Für die Interpretation obliquer Lösungen sollten zudem die Korrelationen zwischen den Faktoren stets berücksichtigt werden.

Die obliquen Rotationen lassen sich in indirekte und in direkte oblique Ansätze unterteilen. Die wichtigsten indirekten Rotationsverfahren sind die Promax- und die Harris-Kaiser-Rotation, das wichtigste direkte Verfahren ist die Oblimin-Rotation.

23.5.4.1 Promax- und Harris-Kaiser-Rotation (indirekte oblique Rotation)

Die *Promax*- (Hendrickson und White 1964) und die *Harris-Kaiser-Rotation* (Harris und Kaiser 1964), die auch als orthoblique Rotation bezeichnet wird, gehören zu den wichtigsten Verfahren der indirekten obliquen Rotation. Bei diesen Verfahren wird mit einer orthogonalen Varimax-Rotation begonnen und die Faktorladungen werden anschließend anhand einer Funktion so transformiert, dass eine oblique Lösung entsteht. Für die Promax-Rotation erfolgt dies anhand einer Power-Transformation (also λ^k , wobei $k \geq 2$); Für die orthoblique Rotation muss ein sog. „Harris-Kaiser-Power-Faktor“ (HKP-Faktor) gewählt werden, der die mögliche Höhe der Faktorenkorrelation kontrolliert (z. B. eins für orthogonale Lösungen oder Werte zwischen null und eins für oblique Lösungen). Die Wahl der Faktoren ist vom Anwender zu bestimmen. Zumeist wird $k = 4$ für die Promax- und 0.5 für die orthoblique Rotation verwendet (Fabrigar und Wegener 2012).

Ein Nachteil dieser Verfahren besteht darin, dass nicht alle gewählten Größen für k bzw. die H KP-Faktoren zu guten Lösungen führen. Auch kann es passieren, dass Faktoren kollabieren. Dies ist dann der Fall, wenn sie perfekt miteinander korrelieren und somit zu einem Faktor zusammenfallen.

23.5.4.2 Oblimin-Rotation (direkte oblique Rotation)

Die relevanteste direkte oblique Rotation stellt die Familie der *Oblimin-Rotationen* dar. Diese Verfahren streben die unmittelbare oblique Rotation an, ohne dass in einem ersten Schritt eine orthogonale Lösung produziert wird. Ein Vorteil der direkten Rotation besteht darin, dass Faktoren nicht mehr kollabieren können (wie in den indirekten Rotationen).

Das optimierte Rotationskriterium stellt eine mit dem Parameter „Delta“ gewichtete Summe aus einem obliquen Kriterium und dem Varimax-Verfahren dar. Jedes Mitglied der Familie der Oblimin-Rotationen ist dadurch gekennzeichnet,

Interpretation der Mustermatrix

Unkorrelierte Faktoren in einer obliquen Rotation erlauben die Verwendung von orthogonalen Verfahren

Indirekte Verfahren transformieren orthogonale Rotationen

Faktenkollaps

Direkte Verfahren liefern unmittelbar oblique Lösungen

Delta bestimmt den Grad der Obliqueheit

Standardeinstellung: Direct Quartimin und Biquartimin

Spezifikation einer Target-Matrix

Minimierung von Komplexitätsfunktion

Target-Rotation besonders empfehlenswert bei komplexen Ladungsmustern

dass es einen bestimmten Delta-Faktor aufweist, der die Obliqueheit (Korreliertheit) der Faktoren bestimmt. Nicht alle Delta führen zu guten Ergebnissen (Fabrigar und Wegener 2012), was ein Problem darstellt, da für die Wahl des Delta-Parameters nur unzureichende Richtlinien existieren. Somit können mit der Oblimin-Rotation sowohl gute als auch schlechte Lösungen generiert werden, deren Qualität schwer beurteilt werden kann, weil die jeweilige Rotationslösung stark von subjektiven Entscheidungen abhängt.

Als Standardeinstellung für dieses Verfahren empfohlen wird häufig die Direct-Quartimin- mit einem Delta von 0 (Fabrigar und Wegener 2012) oder die Biquartimin-Variante mit einem Delta von 0.5 (Carroll 1957). Das Delta von 0 in der Direct-Quartimin-Rotation führt beispielsweise zu einer Gleichgewichtung von korrelierten und unkorrelierten Faktoren. Prinzipiell stellen die empfohlenen Werte für Delta Ad-hoc-Empfehlungen dar, die nicht garantieren können, die beste Einfachstruktur für den jeweiligen Datensatz zu generieren (Fabrigar und Wegener 2012; Mulaik 2010).

23.5.5 Target-Rotation

Die Target-Rotation (s. Browne 2001; ursprüngliche Idee entwickelt von Tucker 1944) kann als ein Hybrid aus EFA und CFA (► Kap. 24) angesehen werden. Bei ihr wird eine bestimmte Zielvorstellung oder Referenzstruktur („Target“) in Form einer Faktorladungsmatrix („Target-Matrix“) formuliert. Ähnlich wie in der CFA können einige der Faktorladungen auf null gesetzt werden (Sekundärladungen), während andere Faktorladungen als unbekannt angegeben werden. Anders als in der CFA erfolgt dann eine Rotation, und zwar so, dass die resultierende Ladungsmatrix möglichst der Zielvorstellung entspricht, ohne dass jedoch eine echte Restriktion der Sekundärladungen auf null stattfindet. Als Konsequenz können diese Sekundärladungen kleine Werte annehmen (s. Browne 2001).

Die folgende Target-Matrix \mathbf{B} stellt eine solche Referenzstruktur dar, bei der das Faktorladungsmuster mit Nullladungen zum Teil spezifiziert wurde; das Fragezeichen (?) zeigt unspezifizierte Faktorladungen an:

$$\mathbf{B} = \begin{pmatrix} ? & 0 & 0 \\ ? & 0 & ? \\ 0 & ? & ? \\ 0 & 0 & ? \end{pmatrix} \quad (23.38)$$

Basierend auf dieser Struktur wird bei der Rotation folgende Komplexitätsfunktion minimiert:

$$f(\boldsymbol{\lambda}) = \sum_{j=1}^k \sum_{i \in I_j} (\lambda_{ij} - b_{ij})^2, \quad (23.39)$$

wobei λ_{ij} die Faktorladungen darstellen und b_{ij} die Elemente aus der Referenzstruktur \mathbf{B} von oben. Die Notation „ $i \in I_j$ “ zeigt an, dass nur diejenigen Elemente aus \mathbf{B} zur Minimierung verwendet werden, die mit einem „?“ für den entsprechenden Faktor j spezifiziert wurden.

Die Target-Rotation kann sowohl für eine orthogonale als auch für eine oblique Rotation spezifiziert werden (Browne 1972a, 1972b). Das Verfahren eignet sich insbesondere in Situationen mit komplexen Ladungsmustern (mehrere Sekundärladungen; s. Moore et al. 2015).

23.5.6 Geomin-Rotation

Im Unterschied zu den bisher beschriebenen Rotationsverfahren wird in der Geomin-Rotation (Yates 1987) eine andere Komplexitätsfunktion als z.B. in Gl. (23.34) verwendet und stattdessen das geometrische Mittel der Produkte der quadrierten Faktorladungen pro Item aufsummiert:

$$f(\lambda) = \sum_{i=1}^p \left[\prod_{j=1}^k (\lambda_{ij}^2 + \epsilon) \right]^{1/k}, \quad (23.40)$$

wobei mit einer kleinen positiven Zahl ϵ (z.B. 0.01) die Modellidentifikation sichergestellt wird (technische Details s. z.B. in Browne 2001; Yates 1987). Der Vorteil dieser Rotation besteht darin, dass sie die erste Regel der Einfachstruktur von Thurstone sicherstellt (jedes Item hat mindestens eine Nullladung; vgl. Browne 2001). Das Verfahren wurde für oblique Rotationen entwickelt, kann aber auch für orthogonale Rotationen verwendet werden. Es liefert insbesondere dann gute Lösungen, wenn die Items tatsächlich einer Einfachstruktur folgen und sich in Cluster aufteilen lassen (s. z.B. Sass und Schmitt 2010).

Es sollte beachtet werden, dass jede Wahl von ϵ ein anderes Rotationskriterium produziert. Hierbei gibt es keine richtige oder falsche Lösung. Asparouhov und Muthén (2009) schlagen vor, mehrere ϵ (0.01, 0.001) zu verwenden, um die Sensibilität bzw. Stabilität der Faktorenlösung bei verschiedenen ϵ zu überprüfen.

Einsatz bei Items, die einer Einfachstruktur folgen

Überprüfung der Sensitivität gegenüber (arbiträrer) Wahl von ϵ

23.5.7 Welche Methode sollte verwendet werden?

Zusammengefasst existiert eine ganze Reihe von Rotationsmethoden.

Die Entscheidung für eine Methode sollte anhand folgender Gesichtspunkte erfolgen (vgl. auch ▶ Beispiel 23.5):

1. Sollen korrelierte Faktoren zugelassen sein?
Wenn ja, sollten oblique Verfahren gewählt werden; wenn nein, sollten orthogonale Verfahren gewählt werden (Varimax ist hier Standard).
2. Wird angenommen, dass (zumindest theoretisch) eine echte Einfachstruktur existiert, d.h. weder Methodenfaktoren (dazu ▶ Kap. 25) noch andere substantielle Doppelladungen, die auch inhaltlich Sinn machen, existieren?
Wenn ja, ist das Geomin-Verfahren empfehlenswert.
3. Wird aus inhaltlichen Gründen (d.h. aufgrund konkreter Vorannahmen) nicht nach einer Einfachstruktur gesucht, weil inhaltlich begründete Sekundärladungen Sinn ergeben? Konkrete Vorannahmen liegen z.B. vor, wenn Multitrait-Multimethod-Designs (MTMM-Designs), Bifaktormodelle oder Methodenfaktoren, bei denen substantielle Sekundärladungen immanent sind, untersucht werden (s. z.B. Sass und Schmitt 2010).
Wenn ja, ist die Target-Rotation empfehlenswert.

Beispiel 23.5: Verschiedene Rotationslösungen für das empirische Beispiel

Nach der Entscheidung, zwei Faktoren für diesen Datensatz zu extrahieren (▶ Beispiel 23.2), sollen nun verschiedene Rotationen angewandt werden.

Tab. 23.2 zeigt die unrotierte Faktorladungsmatrix sowie Faktorladungsmatrizen einer orthogonalen Rotation (Varimax), einer obliquen Rotation (Direct Quartimin) und einer Target-Rotation. Für die (orthogonale) Target-Rotation wurde eine Target-Matrix verwendet, die ein Bifaktormodell operationalisieren soll, in dem alle Faktorladungen auf dem ersten Faktor frei geschätzt („?“) und nur für die ersten vier Items auf dem zweiten Faktor eine Faktorladung von null angenommen wurde.

Tabelle 23.2 Faktorladungsmuster (und Itemkommunalitäten) für eine unrotierte sowie drei verschiedene rotierte Lösungen

	Iteminhalt	unrotiert		Varimax		Direct Quartimin		Target		Kommunalität
		F1	F2	F1*	F2*	F1*	F2*	F1*	F2*	
1	Ich habe Spaß daran, über Mathematik zu lesen.	0.65	-0.23	0.64	0.26	0.65	0.06	0.69	0.02	0.48
2	Ich freue mich auf den Matheunterricht.	0.81	-0.37	0.85	0.26	0.90	-0.02	0.89	-0.06	0.79
3	Ich mache Mathematik, weil ich daran Freude habe.	0.83	-0.40	0.89	0.25	0.95	-0.04	0.92	-0.08	0.85
4	Ich bin daran interessiert, Dinge in Mathematik zu lernen.	0.81	-0.17	0.72	0.41	0.67	0.22	0.82	0.13	0.69
5	Es lohnt sich für mich, mich in Mathematik anzustrennen, weil es mir bei meiner späteren Arbeit helfen wird.	0.76	0.23	0.42	0.67	0.19	0.66	0.63	0.48	0.63
6	Es lohnt sich für mich, Mathematik zu lernen, weil es meine Berufschancen verbessern wird.	0.71	0.43	0.24	0.79	-0.08	0.88	0.51	0.66	0.69
7	Mathematik ist ein wichtiges Fach für mich weil ich es für mein späteres Studium benötige.	0.66	0.22	0.35	0.60	0.14	0.60	0.54	0.44	0.48
8	Ich lerne viele Dinge in Mathematik, die mir später helfen werden, einen Job zu finden.	0.75	0.39	0.30	0.79	0.00	0.84	0.56	0.63	0.71
Faktorkorrelation		$r = 0$		$r = 0$		$r = 0.64$		$r = 0$		

Anmerkung: Faktorladungen > 0.3 sind fett gedruckt. Übersetzungen der Items aus dem Englischen durch den Autor. Für die oblique Rotation sind die Elemente der Mustermatrix abgebildet. F1 und F2 repräsentieren die unrotierten Faktoren 1 und 2, F1* und F2* jeweiligen rotierten Faktoren.

Faktorenanalyse für orthogonale und oblique Rotationen

Abb. 23.4 illustriert die Ergebnisse der verschiedenen zweifaktoriellen Lösungen. In den Teileabbildungen sind die unrotierten Faktoren jeweils mit gestrichelten Linien dargestellt, die jeweiligen rotierten Faktoren mit durchgezogenen Linien. Die jeweiligen Faktorladungen eines Items auf den zwei Faktoren zeigen an, wie stark das Item mit dem jeweiligen Faktor zusammenhängt. Für die unrotierte Lösung (s. Tab. 23.2) ist eine Interpretation der Faktoren nur schwer möglich. Alle acht Items laden hoch auf Faktor 1 und entweder positiv (Items 5 bis 8) oder negativ (Items 1 bis 4) auf Faktor 2. Die Varimax-Rotation erlaubt eine bessere Interpretation: Die Items 1 bis 4 laden hoch auf Faktor 1 und die Items 5 bis 8 laden erwartungsgemäß auf Faktor 2. Es sollte jedoch beachtet werden, dass vier Items substantielle Doppeladlungen aufweisen (Items 4, 5, 7 und 8), was daran zu erkennen ist, dass die Items nicht nahe an den Faktoren liegen, sondern „im Raum“ zwischen den Faktoren. Die oblique Direct Quartimin-Rotation erzeugt eine eindeutige Einfachstruktur, die Items laden eindeutig auf jeweils nur einem der beiden Faktoren und weisen Sekundärladungen nahe null auf (und liegen sehr nahe an den Faktoren). Der Nachteil der obliquen Rotation ist eine Korrelation zwischen den beiden Faktoren von 0.64. Diese kann als recht hoch eingeordnet werden und be-

23.5 · Faktorenrotation

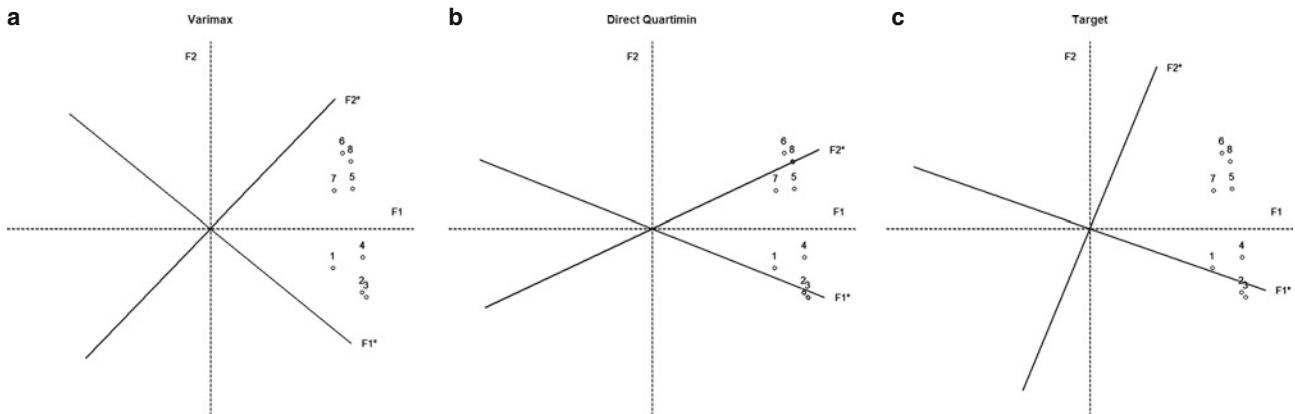


Abb. 23.4 Item- oder Ladungsplots der Items 1 bis 8 für die drei unterschiedlich rotierten Faktorenlösungen (a–c). Die Ladungen sind für die a Varimax-, b Direct Quartimin- und c Target-Rotation abgebildet. Die gestrichelten Linien zeigen die unrotierten Ausgangslösungen an (Faktoren F1, F2); die durchgezogenen Linien die jeweiligen rotierten Faktoren F1*, F2*. Näheres siehe Text

deutet, dass die Faktoren nicht mehr inhaltlich getrennt voneinander interpretiert werden können. Die beiden Faktoren können durchaus als „Interesse an Mathematik“ (Item 1 bis 4) sowie „Instrumentelle Motivation“ (Item 5 bis 8) interpretiert werden. Schülerinnen und Schüler, die Interesse an Mathematik haben, weisen jedoch auch eine hohe instrumentelle Motivation für Mathematik auf.

Für die Target-Rotation wurde eine Target-Matrix (► Abschn. 23.5.5) formuliert, bei der auf dem Faktor 1 sämtliche Items laden. Der Faktor 2 ist ein davon unabhängiger orthogonaler Faktor, auf dem nur die letzten vier Items laden, die ersten vier hingegen nicht. Die Rotation mit dieser Target-Matrix als Zielfigur resultiert in einer Lösung, bei der der Faktor 1 als ein allgemeiner Interessenfaktor für Mathematik angesehen werden kann. Der Faktor 2 kann interpretiert werden als ein Konstrukt, das die sprachliche Besonderheit der Items abbildet; alle vier Items verwenden eine Begründung, die auf eine Bedeutsamkeit in der Zukunft anspielt. Schülerinnen und Schüler können dieser Bedeutsamkeit unabhängig von ihren Mathematikinteressen zustimmen oder sie ablehnen.

Die Kommunalitäten aller Items sind hoch und liegen über 0.48, d. h., es werden mindestens 48 % der Itemvarianzen durch die beiden Faktoren gemeinsam erklärt (unabhängig von der Rotation).⁶ Item 1 hat wohl eine geringere Kommunalität als die Items 2 bis 4, da es sprachlich von „lesen“ und nicht „lernen“ oder „selbst machen“ spricht. Item 7 unterscheidet sich von den Items 5 bis 8 insbesondere dadurch, dass es nicht von Berufen, sondern vom Studium handelt, weshalb Item 5, 6 und 8 höher miteinander korrelieren, was hier höhere Faktorladungen auf dem Faktor 2 produziert.

Welche der Rotationen sollte nun gewählt werden? Jede der Rotationen liefert eine Einfachstruktur im Sinne ihrer Konzeption. Sie unterscheiden sich nicht in ihrer Anpassungsgüte (weil sich z. B. die Kommunalitäten nicht ändern). Eine Entscheidung muss aus inhaltlichen Überlegungen getroffen werden.

Weitere Rotationen wie eine Geomin-Rotation sowie eine oblique Target-Rotation mit einem Target, dem zufolge die ersten bzw. die letzten vier Items auf separaten Faktoren laden, führen zu nahezu identischen Lösungen wie die oblique Rotation (diese zusätzlichen Berechnungen sind in einem eigenen Skript unter ► <http://www.lehrbuch-psychologie.springer.com>, Projekt Testtheorie und Fragebogenkonstruktion, zu finden).

Target-Rotation liefert hier eine andere Interpretation der Faktoren

23.6 Modellevaluation und Itemauswahl

23.6.1 Modellevaluation

Die Dimensionalitätsanalyse eines Datensatzes anhand einer EFA sollte nicht mit der Interpretation der Eigenwerte, der Komunalitäten und der Faktorladungsmatrix enden. Vielmehr stellt erst eine Überprüfung der Modell-Daten-Passung den letzten Schritt in der EFA dar. Traditionellerweise fungiert die Inspektion der Residualmatrix hier als zentrales Kriterium. In modernen Implementierungen können zudem Modellfitmaße herangezogen werden, die für die CFA (► Kap. 24) entwickelt wurden.

23.6.1.1 Residualmatrix

Die Residualmatrix \mathbf{D} beinhaltet die Differenzen zwischen der empirischen (\mathbf{S}) und der modellimplizierten Korrelationsmatrix $\Sigma(\boldsymbol{\theta})$:

$$\mathbf{D} = \mathbf{S} - \Sigma(\boldsymbol{\theta}) \quad (23.41)$$

Vergleich von empirischen und modelltheoretischen Korrelationen

Positive-, negative- und Null-Differenzen

Zwei Arten von standardisierten Residualmatrizen

Interpretation wird mit zunehmender Itemzahl komplizierter

Die empirische Korrelationsmatrix \mathbf{S} enthält (wie in ► Abschn. 23.2.3 beschrieben) die in den Daten beobachteten Korrelationen (z. B. die empirische Korrelation zwischen jeweils zwei Itemvariablen); die modellimplizierte Korrelationsmatrix $\Sigma(\boldsymbol{\theta})$ hingegen enthält die modelltheoretischen Schätzungen der Korrelationen, die sich aus den Faktorladungen und den Fehlervarianzen ergeben.

Positive Differenzen in der Residualmatrix bedeuten, dass die empirischen Korrelationen durch das Modell unterschätzt wurden, während negative Einträge eine Überschätzung anzeigen. Einträge nahe null deuten auf eine gute Modellanpassung hin. Bei einer EFA treten positive Werte üblicherweise dann auf, wenn zu wenige Faktoren extrahiert wurden, weil dann die modellimplizierten Korrelationen niedriger sind als die empirischen. Eine Lösung mit mehr Faktoren kann in diesem Fall zu einer besseren Modellanpassung führen.

Im Allgemeinen sollten standardisierte Residualmatrizen interpretiert werden. Eine Standardisierung kann hierbei (je nach Software-Implementierung) auf zwei verschiedene Arten erfolgen: Sie kann sich darauf beziehen, dass empirische und modelltheoretische Korrelationsmatrizen verglichen wurden; in diesem Fall deuten (per Daumenregel) Elemente größer als 0.1 (oder kleiner als -0.1) auf eine Abweichung zwischen Modell und Daten hin (z. B. Raykov und Marcoulides 2010). Eine Standardisierung kann aber auch erst nach Berechnung der Differenzen erfolgen (wie in Mplus; Muthén und Muthén 2017), wobei die Elemente dann standard-normalverteilt (z-verteilt) sind; Elemente größer als 1.96 (bzw. kleiner als -1.96) weisen auf substantielle Unterschiede zwischen empirischen und modelltheoretischen Korrelationen hin.

Die Interpretation der Residualmatrix wird komplexer, wenn die Anzahl p der Items zunimmt. Die Matrix beinhaltet $p(p - 1)/2$ Elemente, d. h., bei 10 Items sind es schon 45 Elemente. In diesen Fällen sind weitere Maße zur Einschätzung der Modellpassung hilfreich. Im ► Beispiel 23.6 erfolgt eine Interpretation der Residualmatrix für das empirische Beispiel.

6 Die Komunalitätsberechnung ändert sich leicht für die oblique Rotation, da die Korrelation der Faktoren berücksichtigt werden muss. Beispielsweise ergibt sich die Komunalität für das erste Item bei einer Faktorenkorrelation von 0.64 als $0.65^2 + 0.06^2 + 2 \cdot 0.65 \cdot 0.06 \cdot 0.64 = 0.48$.

Beispiel 23.6: Residualmatrix für das empirische Beispiel

Da die Anpassung mit dem Zweifaktormodell eine sehr gute Lösung liefert, soll hier stattdessen die Residualmatrix aus der Einfaktorenlösung dargestellt werden, da sie sehr schön illustriert, wie man Fehlspezifikationen erkennen kann. Die Matrix \mathbf{D} ist gegeben durch:

$$\mathbf{D} = \begin{pmatrix} 0.00 & & & & & & & \\ 0.09 & 0.00 & & & & & & \\ 0.09 & \mathbf{0.20} & 0.00 & & & & & \\ 0.06 & 0.06 & 0.08 & 0.00 & & & & \\ -0.03 & -0.06 & -0.09 & -0.05 & 0.00 & & & \\ -0.09 & \mathbf{-0.13} & \mathbf{-0.13} & -0.06 & \mathbf{0.13} & 0.00 & & \\ -0.04 & -0.07 & -0.07 & -0.05 & 0.03 & \mathbf{0.11} & 0.00 & \\ -0.10 & \mathbf{-0.11} & \mathbf{-0.11} & -0.04 & \mathbf{0.10} & \mathbf{0.19} & \mathbf{0.11} & 0.00 \end{pmatrix} \quad (23.42)$$

Es ergeben sich insbesondere zwei Gruppen von Korrelations-Differenzen. Erstens werden die Korrelationen zwischen den Items 5 bis 8 fast alle unterschätzt (positive Werte > 0.1 in \mathbf{D}). Dies deutet darauf hin, dass ein weiterer Faktor notwendig ist, um die Korrelationen dieser Items untereinander zu erklären. Weiterhin werden die Korrelationen der Items 1 bis 4 mit den Items 5 bis 8 überschätzt (negative Werte). Dies deutet darauf hin, dass sich bei einer Einfaktoriellösung zu große Korrelationen zwischen den beiden Itemgruppen ergeben würden.

23.6.1.2 Modellfit

In aktuellen Implementierungen der (ML-)EFA, beispielsweise in „R“ (R Core Team 2016) oder „Mplus“ (Muthén und Muthén 2017), können weitere Modellfitmaße berechnet werden, die eine Beurteilung der Modellpassung erlauben (Fabrigar et al. 1999; Lorenzo-Seva et al. 2011). Die wichtigsten absoluten Maße sind hierbei der RMSEA (Root Mean Squared Error of Approximation; Browne und Cudeck 1992; Steiger und Lind 1980) und der SRMR (Standardized Root Mean Residual). Wichtige sog. „relative Gütemaße“ sind der CFI (Comparative Fit Index) oder auch TLI (Tucker Lewis Index). Details zu den Maßen sind in der Darstellung der CFA (► Kap. 24) zu finden.

Bei der Verwendung der Modellfitmaße sollte beachtetet werden, dass sie sensitiv gegenüber Nichtnormalität sind, was z. B. bei der Analyse von (grobstufigen) Items zu Problemen führen kann, da diese häufig nicht normalverteilt sind (vgl. ► Abschn. 23.3.3).

RMSEA und SRMR sind wichtige Gütemaße

Sensitivität gegenüber Nichtnormalität

23.6.2 Faktoreninterpretation

Die inhaltliche Deutung der Faktoren stellt in der EFA immer eine Post-hoc-Interpretation dar und sollte entsprechend vorsichtig erfolgen. Die Zuschreibung eines Begriffs (z. B. Intelligenz) zu einem Faktor erfolgt anhand des Faktorladungsmusters der Items. Hierbei muss sowohl die Richtung der Faktorladungen (positiv/negativ) als auch die absolute Größe berücksichtigt werden. Items mit niedrigen Ladungen tragen nicht zur inhaltlichen Diskriminierung im Sinne des Faktors bei. Es ist es wichtig, dass der Iteminhalt repräsentativ für die Namensgebung ist und keine überzogenen Generalisierungen vorgenommen werden (beispielsweise sollte ein Faktor mit Items zur ausschließlichen Feststellung der Additionsfähigkeit nicht „Matheleistung“ genannt werden).

Inhaltliche Faktoreninterpretation basiert auf Faktorladungsmuster

Korrelationsmuster der Faktoren berücksichtigen

Wie in ► Abschn. 23.5 zur Faktorenrotation beschrieben, kann sich die inhaltliche Bedeutung der Faktoren ändern, je nachdem welche Rotationsform gewählt wurde. Bei einer obliquen Rotation sollte ebenfalls das Interkorrelationsmuster der Faktoren berücksichtigt werden, d. h., Faktoren, die hoch (bzw. niedrig) miteinander korrelieren, sollten auch entsprechend verwandte (bzw. unabhängige) theoretische Konstrukte repräsentieren. Je höher Faktoren korrelieren, desto schwerer fällt die Argumentation, dass es inhaltlich verschiedene Konstrukte sind.

Invertierung der Faktoren

Letztlich sollte beachtet werden, dass die Richtung der Faktorenlösung arbiträr ist und eine Umkehrung der Vorzeichen (d. h. $-\lambda = \lambda$) jederzeit erfolgen kann. Dies heißt z. B., dass ein Faktor „Furchtlosigkeit“ nach Vorzeichenumkehr ebenso einen Faktor „Ängstlichkeit“ repräsentieren könnte.

23.6.3 Faktorwerte (Faktorscores)

Faktorenindeterminiertheit führt zu Unbestimmtheit der Faktorwerte

Nachdem eine EFA durchgeführt wurde, können sog. „Faktorwerte“ bzw. „Faktorscores“ („factor scores“) geschätzt werden, die angeben, welche Werte (Ausprägungen) die Testpersonen in jedem extrahierten (und rotierten) Faktor η_j aufweisen. Ein Problem stellt hierbei die Faktorenindeterminiertheit dar, die oben diskutiert wurde (► Abschn. 23.5.1). Sie besagt, dass Faktoren uneindeutig sind – je nach Extraktions- und Rotationsmethode (und Anzahl der Faktoren) können dabei andere Faktoren entstehen und somit andere Faktorwerte für die Personen (z. B. Mulaik 2010).

Verschiedene Methoden der Faktorwertbestimmung

Neben diesem Problem muss beachtet werden, dass es verschiedene Methoden der Faktorwertschätzung gibt. Die beiden bekanntesten Methoden sind die *Regressions-Faktorwerte* (Thomson 1934; Thurstone 1935) und die *Bartlett-Faktorwerte* (Bartlett 1937). Die Regressions-Faktorwerte können sehr verschieden von den Bartlett-Faktorwerten sein, und zwar umso mehr, je geringer die Kommunalitäten der einzelnen Items sind (Acito und Anderson 1986). Weiterhin unterscheiden sich beide Arten von Faktorwerten in weiteren Punkten, z. B. ob die Faktorwerte aus orthogonalen Rotationen auch tatsächlich unkorreliert sind. Werden Faktorwerte in weiteren Analysen (z. B. in einer Regressionsanalyse) verwendet, so müssen bestimmte Korrekturmaßnahmen verwendet werden (s. z. B. Croon 2002; Skrondal und Laake 2001). Diese Korrekturen berücksichtigen den Schätzfehler der Faktorwerte und korrigieren die Analysen entsprechend.

Summenscores vs. Faktorwerte

Die Bildung von Test-Summenwerten (*Summenscores*, ► Kap. 8), die in der Praxis häufig verwendet werden, nimmt implizit eine Gleichgewichtung aller Items an. Sind die Faktorladungen für die Items auf einem Faktor tatsächlich gleich groß, so werden die Faktorwerte und die Summenscores sehr ähnlich sein. Ist dies nicht der Fall, dann können die Summenscores zu Verzerrungen führen, d. h., Personen erhalten Werte, die nicht ihren tatsächlichen Ausprägungen im jeweiligen Konstrukt entsprechen. Als Konsequenz könnten Personen, die eigentlich verschiedene Ausprägungen im interessierenden Merkmal aufweisen, ähnliche Summenscores erhalten oder umgekehrt. Wenn die Summenwerte im Anschluss z. B. für eine Diagnostik oder die Berechnung von Korrelationen (im Rahmen einer Konstruktvalidierung) verwendet werden, kann dies entsprechend zu Problemen führen.

23.6.4 Itemauswahl

Itemauswahl in der Testkonstruktion

In diesem Abschnitt soll auf den Aspekt der Itemauswahl eingegangen werden, der spezifisch für die Testkonstruktion ist. Die EFA kann dazu verwendet werden, Items zu identifizieren, die besonders gut geeignet dazu sind, einen bestimmten Faktor zu messen. Vor der Zusammenfassung solcher Items zu einer Skala sollten *mindestens* die folgenden drei Aspekte berücksichtigt werden:

Erstens kann eine Auswahl dadurch erfolgen, dass Items mit besonders hohen Faktorladungen auf den intendierten Faktoren ausgewählt werden. Ein typischer Cut-off-Wert für die Beurteilung eines Items als „gut“ liegt nach Brown (2015) bei Faktorladungen größer als 0.3 oder 0.4 (was einer durch den Faktor erklärten Varianz von 9 bzw. 16 % entspricht). Eine negative Faktorladung bedeutet, dass hohe Ausprägungen im Item mit einer niedrigen Ausprägung im Faktor einhergehen; dies kann natürlich auftreten, wenn (dem Konstrukt entsprechend) negativ formulierte Items vor der Analyse nicht invertiert wurden. Ist dies nicht der Fall, so muss davon ausgegangen werden, dass das Item nur schwer im Sinne des Konstruktts interpretiert werden kann.

Bei der Itemauswahl sollte beachtet werden, dass die Breite der inhaltlichen Validität (► Kap. 2) nicht eingeschränkt wird. Häufig kann dies der Fall sein, wenn Items mit besonders hohen Faktorladungen gewählt werden, während Items mit mittleren bis niedrigen Faktorladungen ausgeschlossen werden. Der Grund besteht darin, dass die Items mit besonders hohen Faktorladungen auch sehr hoch miteinander korrelieren, weil sie einen sehr ähnlichen Iteminhalt aufweisen. Die Auswahl allein aufgrund hoher Faktorladungen kann somit zu einer inhaltlichen Verengung des Konstruktts führen.

Zweitens sollte die Auswahl berücksichtigen, dass die verwendeten Items möglichst wenige/keine Sekundärladungen (Doppelladungen) aufweisen. Sekundärladungen bedeuten stets, dass die Items auf mehr als einem Faktor laden. Dadurch wird die Interpretation der Zugehörigkeit des Items zu einem Faktor erschwert. Gleichzeitig ist aber auch die Interpretation von Faktoren selbst erschwert, da die Bedeutung nun nicht mehr auf einer Einfachstruktur basiert. Dies ist im Falle einer obliquen Rotation nochmals komplizierter, da bei der Interpretation der Doppelladungen auch das Korrelationsmuster der Faktoren berücksichtigt werden muss. Konfirmatorische Modelle erlauben eine konkrete Untersuchung und auch Interpretation von Doppelladungen, z. B. im Rahmen von Bifaktormodellen (► Kap. 24), die explizit weitere Faktoren modellieren (wie Methodenfaktoren, s. Reise et al. 2010).

Drittens sollte man darauf achten, die Anzahl der Items nicht zu stark zu reduzieren. Dies erfolgt zum einen aus Gründen der inhaltlichen Validität, zum anderen aber auch aus Gründen der Modellidentifikation und einer zuverlässigen Schätzung (► Abschn. 23.3). MacCallum et al. (1999) schlagen eine Mindestanzahl von drei bis fünf Items pro Faktor/Konstrukt vor.

Auswahl von Items mit hohen Faktorladungen

Auf inhaltliche Validität achten

Doppelladungen erschweren Interpretation

Mindestens drei bis fünf Items pro Konstrukt

23.6.5 Korrelations- oder Kovarianzmatrix?

In allen bisherigen Darstellungen wurde stets davon ausgegangen, dass standardisierte Variablen (Items) verwendet werden bzw. dass eine Korrelationsmatrix analysiert wird. Sämtliche vorgestellten Methoden können jedoch auch für die Analyse von unstandardisierten Itemvariablen verwendet werden, deren Zusammenhänge nicht in Korrelations-, sondern in Kovarianzmatrizen gemessen wurden. Es ist dann jedoch zu beachten, dass die meisten berechneten Koeffizienten (Faktorladungen, Kommunalitäten, Eigenwerte) nicht mehr automatisch standardisiert sind (weshalb beispielsweise für das Kaiser-Guttman-Kriterium selbst in einer PCA die durchschnittliche Varianz der Items berechnet werden sollte). Statistik-Programme liefern üblicherweise standardisierte Koeffizienten zusätzlich zu den unstandardisierten.

Vorteile der Verwendung einer Kovarianzmatrix können dann erwartet werden, wenn unterschiedliche Varianzen der Variablen tatsächlich inhaltliche Unterschiede abbilden, z. B. wenn alle Items zwar mit demselben Itemformat (z. B. sechsstufige Antwortskala) gemessen wurden, aber einige Items große, andere hingegen kleine Itemvarianzen (s. dazu die Ausführungen von Kelava und Moosbrugger in

Bei Verwendung von Kovarianzmatrizen Interpretation beachten

Kovarianzmatrizen sind sinnvoll, wenn Varianzunterschiede der Items auch inhaltlich sinnvolle Unterschiede abbilden

Kovarianzmatrizen sind nicht sinnvoll, wenn Items auf verschiedenen Skalen gemessen wurden

► Kap. 7, ► Abschn. 7.4) aufweisen. Sollen diese in den Faktoren abgebildet werden, so sollte eine Kovarianzmatrix analysiert werden. In Situationen, in denen die Varianzen aller Items sehr ähnlich sind, ist kaum zu erwarten, dass Unterschiede zwischen einer EFA mit Korrelations- und einer mit Kovarianzmatrix auftreten.

Nachteile bei der Verwendung einer Kovarianzmatrix entstehen dann, wenn die Varianzunterschiede artifizieller Natur sind, z. B. wenn verschiedene Metriken (z. B. Schuhgröße und IQ) oder verschiedene Itemformate (z. B. ein Gemisch aus sechs- und zehnstufigen Antwortskalen) verwendet wurden. In diesem Fall ist von einer Kovarianzmatrix abzuraten, da Items mit großen Varianzen häufig die Bildung artifizieller Faktoren begünstigen (durch eine Überschätzung der Bedeutsamkeit von Items mit großen Varianzen).

23.7 Neue Verfahren

Im letzten Abschnitt soll noch auf neuere Verfahrensweisen verwiesen werden. Diese beinhalten die exploratorischen Strukturgleichungsmodelle, Alternativen für dichotome bzw. ordinale Daten sowie Bayes'sche Implementierungen.

23.7.1 Verwendung exploratorischer Faktormodelle in Strukturgleichungsmodellen (ESEM)

Exploratorische Strukturgleichungsmodelle (Exploratory Structural Equation Modeling, ESEM; Asparouhov und Muthén 2009) kombinieren die Vorteile der EFA mit denen der Strukturgleichungsmodelle. EFA-Modelle sind dadurch beschränkt, dass komplexere Modellspezifikationen – wie Residualkorrelationen oder spezifischere Strukturen zwischen den Faktoren z. B. in der Form, dass Korrelationen nicht zwischen allen Faktoren zugelassen werden – nicht untersucht werden können. Die Kombination von exploratorischen (anstelle von konfirmatorischen) Faktormodellen in Strukturgleichungsmodellen erlaubt es mit den ESEM, solche Erweiterungen zu untersuchen. Weitere Einsatzmöglichkeiten ergeben sich bei der Analyse von längsschnittlichen Datensätzen („Bleibt die Faktorenstruktur über die Zeit gleich?“) oder bei Mehrgruppenanalysen („Lässt sich dieselbe Faktorenstruktur in verschiedenen Gruppen oder Stichproben wiederfinden?“).

Aus der Sichtweise der Strukturgleichungsmodelle beinhalten konfirmatorische Messmodelle, wie sie in ► Kap. 24 vorgestellt werden, manchmal zu strikte Annahmen über die Faktorenstruktur (Asparouhov und Muthén 2009), die ebenso zu Artefakten bei der Interpretation führen können (wie z. B. Sekundärladungen, die fälschlich auf null fixiert wurden; Bollen et al. 2007).

Jedoch sollte beachtet werden, dass die Schwächen der EFA – insbesondere die Indeterminiertheit der Faktorenlösung und die damit einhergehende sich ändernde Interpretation der Faktoren – auch bei der Analyse der Strukturgleichungsmodelle bestehen bleibt, wenn ESEM verwendet werden. Während ein konfirmatorisches Messmodell zu Faktoren führt, die eindeutig interpretierbar sind, kann dies bei ESEM erschwert sein, z. B. wenn eine Doppelladung in einer Substichprobe auftritt, aber nicht in der anderen: Dann müssen kritische Fragen über die Vergleichbarkeit der Schätzungen beantwortet werden, insbesondere hinsichtlich der veränderten Bedeutung der Faktoren oder ob die gefundenen Zusammenhänge überhaupt inhaltlich identisch zu interpretieren sind.

ESEM erlauben komplexere Modellierung

Konfirmatorische Modelle sind manchmal zu strikt

Faktorendeterminiertheit bleibt bei ESEM-Strukturen bestehen

23.7.2 Alternativen für dichotome und ordinale Daten

Der Einsatz der oben beschriebenen Verfahren (PFA und ML-EFA, ▶ Abschn. 23.3.2 und 23.3.3) setzt kontinuierliche beobachtete Itemvariablen voraus. Dies ist jedoch häufig nicht gegeben, z. B. wenn das Antwortformat aus einer dreistufigen Skala besteht. In einem solchen Fall kann im Prinzip eine der folgenden drei Möglichkeiten gewählt werden:

- der Einsatz von Verfahren aus dem Bereich der Item-Response-Theorie (IRT)
- der Einsatz von sog. „robusten Schätzverfahren“ im Kontext der (ML-)EFA
- der Einsatz von Bayes'schen Verfahren

Drei Alternativen bei nicht kontinuierlich gestuften Items

23.7.2.1 IRT-Modelle

Auf den Einsatz von IRT-Modellen wurde von Kelava und Moosbrugger in ▶ Kap. 16 sowie von Kelava, Noventa und Robitzsch in ▶ Kap. 18 ausführlich eingegangen, weshalb hier auf eine Beschreibung verzichtet wird.

23.7.2.2 Robuste Schätzverfahren

Die am häufigsten eingesetzten robusten Verfahren, die als Alternative z. B. zur ML-EFA verwendet werden können, sind die Robuste ML-Schätzung (z. B. MLR, Robust Maximum Likelihood, in *Mplus*) und die robuste (ungewichtete/gewichtete) Kleinstquadratschätzung (z. B. ULSMV, Mean- and Variance-adjusted Unweighted Least Squares, oder WLSMV, Mean-Variance Adjusted Weighted Least Squares, in *Mplus*). Obwohl die meisten dieser Verfahren ursprünglich für konfirmatorische Faktormodelle entwickelt wurden, liegen sie heute ebenso für exploratorische Modelle vor.

Robuste Verfahren umfassen verschiedene Techniken

Robuste ML-Schätzer (MLR) ignorieren die Ordinalität der Items für die Schätzung der Itemparameter. Jedoch werden Korrekturen für die Standardfehlerschätzung und die Teststatistiken (Modelldifferenztest oder auch RMSEA) angewendet, die häufig zu guten Schätzergebnissen führen (Schmitt 2011).

Robuste ML-Schätzer

Für den Modelldifferenztest ist es notwendig, eine Korrektur durchzuführen (Satorra und Bentler 2010). Diese Korrektur basiert auf einer sog. „Scaling Correction“ c , die für jedes der Vergleichsmodelle berechnet werden kann (durch die Nichtnormalität der Daten entsteht eine nicht zentrale χ^2 -Verteilung, d. h., Testwerte können nicht mehr mit der üblichen χ^2 -Verteilung getestet werden. Die Korrektur verwendet Informationen aus der Rohdatenverteilung, um die nicht zentrierte χ^2 -Verteilung wieder zu zentrieren; Details s. Satorra und Bentler 2001). Sei das Modell mit weniger Faktoren das Modell 1 und das mit mehr Faktoren das Modell 2, dann ist der Korrekturfaktor für den Modelldifferenztest gegeben als:

Korrigierter Modelldifferenztest

$$\Delta c = (df_1 c_1 - df_2 c_2) / (df_1 - df_2) \quad (23.43)$$

Die korrigierte Teststatistik ergibt sich als:

$$\Delta LL^* = (-2LL_1^* - (-2LL_2^*)) / \Delta c, \quad (23.44)$$

wobei LL^* die unkorrigierten Log-Likelihood-Werte der Modelle 1 und 2 sind. Die Beurteilung der Teststatistik erfolgt analog zu der Beschreibung in ▶ Abschn. 23.4.4.

Kleinstquadratschätzer

Die Kleinstquadratschätzer verwenden polychorische Korrelationsmatrizen (s. dazu z. B. Bortz und Schuster 2010), die adäquat für ordinale oder kategoriale Daten sind (s. z. B. Flora und Curran 2004; Holgado-Tello et al. 2010). Sie modellieren die Ordinalität der Items direkt. Robuste Kleinstquadratschätzer verwenden zusätzlich korrigierte Standardfehler und Teststatistiken. Ein Nachteil der Verfahren ist, dass der Modelldifferenztest nicht mehr wie in ▶ Abschn. 23.4.4 beschrieben durchführbar ist. Zum derzeitigen Stand (02/2020) verfügt ausschließlich

die kommerzielle Software *Mplus* (Muthén und Muthén 2017) über eine adäquate Teststatistik, die mit dem Befehl „DIFFTEST“ angefordert werden kann. Verglichen mit ML-Schätzern sind Kleinstquadratschätzer weniger sparsam, da sie für jede Antwortkategorie (minus eins) eines Items einen eigenen Parameter schätzen, während in einer ML- oder MLR-EFA nur eine einzelne Faktorladung pro Item benötigt wird.

MLR häufig bevorzugt

Im Allgemeinen kann empfohlen werden, dass bei einem Antwortformat mit fünf oder mehr Stufen MLR verwendet werden sollte (z. B. Rhemtulla et al. 2012). Für weniger als fünf Stufen ist WLSMV zu empfehlen. In Situationen, in denen das Hauptaugenmerk auf dem Korrelationsmuster der Faktoren liegen soll, wird häufig ebenfalls MLR empfohlen, selbst wenn weniger als fünf Stufen im Antwortformat verwendet wurden.

23.7.2.3 Bayes'sche Schätzverfahren

Bayes-Schätzer erlauben die flexible Implementierung von EFA-Modellen sowohl für kontinuierliche wie auch für kategoriale oder ordinale Daten. Bayes-Schätzer haben eine Reihe von Vorteilen, z. B. liefern sie unter vielen Bedingungen auch gute Lösungen für kleine Stichproben oder können in Situationen, in denen Daten nicht normalverteilt sind (was zumindest ein Vorteil gegenüber einer ML-Schätzung ist), verwendet werden. Einer der wesentlichen Unterschiede zwischen Bayes'schen und sog. „frequentistischen Verfahren“ (darunter fallen alle bisher in diesem Kapitel besprochenen Verfahren) besteht darin, dass *A-priori-Verteilungen* spezifiziert werden müssen, die eine Annahme über die Verteilung der Parameter beinhalten (z. B. Mittelwerte und Varianzen; Details zu Bayes'schen Verfahren sind z. B. bei Gelman et al. 2013, zu finden). In vielen Situationen können diese Spezifikationen durch den Einbezug von Wissen aus vorausgegangenen empirischen Untersuchungen oder Simulationsstudien erfolgen (vgl. dazu ▶ Kap. 19).

23.8 Abschließende Bemerkungen

EFA ist struktursuchendes, hypothesesgenerierendes Verfahren

Die EFA ist ein struktursuchendes Verfahren, das sich – im Unterschied zur CFA (▶ Kap. 24) – nicht zur Hypothesenprüfung, wohl aber zur Hypothesengenerierung eignet.

Allerdings ist es wichtig zu verstehen, dass eine Entscheidung über Extraktions- oder Rotationsmethoden und -kriterien auf substantiellen und theoretischen Überlegungen beruhen muss. Verschiedene Methoden führen zu einer unterschiedlichen Interpretation der Faktoren. Dies kann auch zu verschiedenen Schlüssen hinsichtlich einer Itemselektion oder der Beurteilung der konvergenten und der diskriminanten Validität führen. Da jedwede Rotation zu demselben Modellfit und zu derselben erklärten Varianz führt (z. B. Mulaik 2005), muss eine Entscheidung über das Rotationskriterium im jeweiligen Kontext der Fragestellung erfolgen.

Der Einsatz der EFA wird von manchen Forschern als sehr kritisch erachtet. Dies liegt zum einen an der Faktorendeterminiertheit (▶ Exkurs 23.2; z. B. Steiger 1979, 1994). Zum anderen liegt es an den subjektiven Eingriffsmöglichkeiten bei der Datenanalyse an fast allen Entscheidungspunkten der EFA (Extraktionsmethode, Rotationsmethoden, Entscheidungskriterien über die Anzahl der Faktoren etc.), womit Faktorenlösungen im engeren Sinne nicht mehr objektiv (sondern eher beliebig) sind. Ein Hauptproblem besteht darin, dass für viele dieser Entscheidungen keine bindenden oder überprüfbarer Richtlinien existieren, die eine objektive Beurteilung der Qualität einer durchgeföhrten EFA erlauben (d. h., die Falsifizierbarkeit von EFA-Modellen ist nur beschränkt möglich). Von einigen Autoren wird die EFA daher als wissenschaftlich nicht sinnvoll einsetzbares Verfahren angesehen und von ihrer Verwendung grundsätzlich abgeraten (z. B. Rencher und Christensen 2012).

23.9 · Zusammenfassung

Wenn man die EFA in einen Kontext mit der CFA setzt (► Kap. 24), so kann man jedoch auch die Ansicht vertreten, dass die CFA in manchen Fällen zu restriktiv ist (da sie hypothesenprüfende Nullladungen von einigen Items annimmt), was zu fehlerhaften Schätzungen führen kann (Asparouhov und Muthén 2009). Weiterhin erfolgt die Anwendung einer CFA häufig in einer explorativen Weise, z. B. wenn Modifikationsindizes schrittweise zur Modellmodifikation (z. B. Freisetzung von Sekundärladungen) verwendet werden. In manchen dieser Situationen liefert die EFA plausiblere Lösungen, da alle potentiellen Doppelladungen simultan und nicht in separaten Modellen geschätzt werden (Asparouhov und Muthén 2009; Browne 2001; Gorsuch 1983; MacCallum et al. 1992).

Letztendlich hängt die Beurteilung, was eine Einfachstruktur in einem konkreten Zusammenhang bedeutet, vom Untersuchungsgegenstand und der Fragestellung ab. Von einer unkritischen Verwendung von Standardeinstellungen in Softwareprogrammen (z. B. häufig das Kaiser-Guttman-Kriterium als Abbruchkriterium und Varimax als Rotationsmethode) ist daher dringend abzuraten.

Andere Forscher unterstreichen ihre Bedeutung und praktische Anwendbarkeit

Kritische Beurteilung der Ergebnisse notwendig

23.9 Zusammenfassung

Die EFA ist ein struktursuchendes Verfahren, das sich – im Unterschied zur CFA (► Kap. 24) – nicht zur Hypothesenprüfung, wohl aber zur Hypothesengenerierung eignet. In diesem Kapitel wurde auf die wichtigsten Aspekte bei der Durchführung einer EFA eingegangen. Es wurde mit der allgemeinen Modellvorstellung in der Faktorenanalyse begonnen (Fundamentaltheorem) und die darauf basierende Varianzzerlegung in durch gemeinsame Faktoren erklärte und unerklärte Teile dargestellt. Anschließend wurden die zentralen Begriffe in der EFA eingeführt, d. h. der Eigenwert eines Faktors sowie die Kommunalität und Spezifität eines Items. Die wichtigsten Extraktionsmethoden (PFA und ML-EFA) sowie Rotationskriterien (orthogonal vs. oblique) wurden diskutiert, bevor auf weitere Aspekte wie die Beurteilung der Modellgüte, alternative Schätzverfahren und die Berechnung von Faktorwerten eingegangen wurde.

23.10 EDV-Hinweise

Anhand des Skripts sowie der Daten unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion) können sämtliche hier vorgestellten Ergebnisse zum empirischen Beispiel mit der Statistik-Software R nachvollzogen werden. Weitere Details sind als Anmerkungen im Skript selbst zu finden.

23.11 Kontrollfragen

- ?
- Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
 1. Was sind die zentralen Kennwerte einer exploratorischen Faktorenanalyse (EFA)?
 2. Worin bestehen die Hauptunterschiede zwischen einer Hauptkomponentenanalyse (PCA) und einer Hauptachsenanalyse (PFA)?
 3. Welche Vor- und Nachteile hat eine mit Maximum Likelihood geschätzte exploratorische Faktorenanalyse (ML-EFA) verglichen mit einer Hauptachsenanalyse (PFA)?

4. Was sind die wichtigsten Abbruchkriterien in einer exploratorischen Faktorenanalyse (EFA)?
5. Was ist die Faktorenindeterminiertheit? Welches Kriterium wird verwendet, um eine Lösung auszuwählen?
6. Welche Rotationsverfahren gibt es und worin unterscheiden sie sich?
7. Benennen Sie die wichtigsten Beurteilungsmaße der Güte einer exploratorischen Faktorenanalyse (EFA).

Literatur

- Acito, F. & Anderson, R. D. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*, 23, 111–118.
- Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397–438.
- Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology. General Section*, 28, 97–104.
- Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M. & Chen, F. (2007). Latent variable models under misspecification: two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research*, 36, 48–86.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Heidelberg: Springer.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2. Aufl.). New York, NY: Guilford Press.
- Browne, M. W. (1972a). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, 25, 115–120.
- Browne, M. W. (1972b). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, 25, 207–212.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150.
- Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21, 230–258.
- Carroll, J. B. (1957). Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science*, 126, 1114–1115.
- Carroll, J. B. (1953). An analytic solution for approximating simple structure in factor analysis. *Psychometrika*, 18, 23–28.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.
- Cattell, R. B. & Jaspers, J. (1967). A general plamode (No. 30-10-5-2) for the factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, 67–63.
- Cattell R. B. & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, 12, 289–325.
- Crawford, C. B. & Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35, 321–332.
- Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–223). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, 44, 362–388.
- Fabrigar, L. R. & Wegener, D. T. (2012). *Exploratory factor analysis*. New York NY: Oxford University Press.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of factor analysis in psychological research. *Psychological Methods*, 4, 281–290.
- Ferguson, G. A. (1954). The concept of parsimony in factor analysis. *Psychometrika*, 19, 347–362.
- Flora, D. B. & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall.
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377–393.
- Gorsuch, R. L. (1980). Factor score reliabilities and domain validities. *Educational and Psychological Measurement*, 40, 895–897.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Guttman, L. (1954). "Best possible" systematic estimates of communality. *Psychometrika*, 21, 273–285.

- Hakstian, A. R., Rogers, W. T. & Cattell, R. B. (1982). The behavior of number of factor rules with simulated data. *Multivariate Behavioral Research*, 17, 193–219.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.
- Harris, C. W. & Kaiser, H. F. (1964). Oblique factor analytic transformations by orthogonal transformations. *Psychometrika*, 29, 347–362.
- Hendrickson, A. E. & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17, 65–70.
- Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I. & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity*, 44, 153–166.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.
- Horn, J. L. (1969). On the internal consistency reliability of factors. *Multivariate Behavioral Research*, 4, 115–125.
- Humphreys, L. G. & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193–206.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.
- Kröse, D. P., Taimre, T. & Botev, Z. I. (2011). *Handbook of Monte Carlo methods*. Hoboken, NJ: Wiley.
- Lorenzo-Seva, U., Timmermann, M. E. & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, 46, 340–364.
- MacCallum, R. C., Roznowski, M. & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.
- MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.
- Maraun, M. D. (1996). Metaphor taken as math: indeterminacy in the factor analysis model. *Multivariate Behavioral Research*, 31, 517–538.
- Moore, T. M., Reise, S. P., Depaoli, S. & Haviland, M. G. (2015) Iteration of partially specified target matrices: Applications in exploratory and Bayesian confirmatory factor analysis, *Multivariate Behavioral Research*, 50, 149–161.
- Moosbrugger, H. (2011). *Lineare Modelle: Regressions- und Varianzanalysen* (4. Aufl.). Bern: Huber.
- Mulaik, S. A. (2005). Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 173–206). Mahwah, NJ: Erlbaum.
- Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.) Los Angeles, CA: Muthén & Muthén.
- Neuhaus, J. O. & Wrigley, C. (1954). The quartimax method: an analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology*, 7, 187–191.
- Organisation for Economic Co-operation and Development (OECD). (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264190511-en>
- R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/> [29.12.2019]
- Raykov, T. & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- Reise, S. P., Moore, T. M. & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544–559.
- Rencher, A. C. & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken, NJ: Wiley.
- Revelle, W. (2012). *Psych: Procedures for personality and psychological research*. R package version 1.0-88. Retrieved from <https://CRAN.R-project.org/package=psych> [29.12.2019]
- Rhemtulla, M., Brosseau-Liard, P. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373.
- Sass, D. A. & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, 45, 73–103.
- Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507–514.
- Satorra, A. & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75, 243–248.
- Saunders, D. R. (1953). *An analytic method for rotation to orthogonal simple structure* (Research Bulletin RB-53-10). Princeton, NJ: Educational Testing Service.

- Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, 29, 304–321.
- Skrondal, A. & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66, 563–575.
- Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika*, 44, 157–167.
- Steiger, J. H. (1994). SEPATH – A Statistica for Windows structural equations modeling program. In F. Faulbaum (Ed.), *Softstat '93: Advances in statistical software 4*. Stuttgart: Gustav Fischer.
- Steiger, J. H. & Lind, J. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.
- Timmermann, M. E. & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16, 209–220.
- Thomson, G. H. (1934). The meaning of 'i' in the estimate of 'g'. *British Journal of Psychology. General Section*, 25, 92–99.
- Thurstone, L. L. (1935). *The vectors of mind* (pp. 226–231). Chicago, IL: University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.
- Tucker, L. R. (1944). A semi-analytical method of factorial rotation to simple structure. *Psychometrika*, 9, 43–68.
- Tucker, L. R., Koopman, R. F. & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 34, 421–459.
- Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany, NY: State University of New York Press.
- Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–442.



Konfirmatorische Faktorenanalyse (CFA)

Jana C. Gäde, Karin Schermelleh-Engel und Holger Brandt

Inhaltsverzeichnis

- 24.1 Grundlagen – 617**
- 24.2 Spezifikation eines Messmodells – 619**
 - 24.2.1 Hypothesenbildung – 619
 - 24.2.2 Bezüge zur Klassischen Testtheorie (KTT) – 620
 - 24.2.2.1 Die gemeinsame latente Variable η_j – 621
 - 24.2.2.2 Die additive Konstante α_i (Interzept) – 622
 - 24.2.2.3 Die multiplikative Konstante λ_{ij} (Faktorladung) – 623
 - 24.2.3 Elemente der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$ – 624
 - 24.2.4 Elemente des modellimplizierten Mittelwertevektors $\hat{\mu}$ – 625
 - 24.2.5 Skalierung der Faktoren – 625
 - 24.2.6 Modellidentifikation – 626
 - 24.2.7 Empirisches Beispiel – 628
- 24.3 Eindimensionale Modelle: Stufen der Messäquivalenz – 629**
 - 24.3.1 τ -Kongenerität – 629
 - 24.3.2 Essentielle τ -Äquivalenz – 632
 - 24.3.3 Essentielle τ -Parallelität – 632
 - 24.3.4 Messäquivalenz und Reliabilität – 633
 - 24.3.5 Fehlerkovarianzen und ihre Beurteilung – 634
- 24.4 Mehrdimensionale Modelle – 634**
 - 24.4.1 Modell mit korrelierten Faktoren – 635
 - 24.4.2 Faktormodell höherer Ordnung – 636
 - 24.4.3 Bifaktormodell – 639
 - 24.4.4 Elemente der modellimplizierten Kovarianzmatrix in mehrdimensionalen Modellen – 642
- 24.5 Parameterschätzung – 643**
 - 24.5.1 Kontinuierliche, normalverteilte Indikatorvariablen – 644
 - 24.5.2 Kontinuierliche, nicht normalverteilte Indikatorvariablen – 645
 - 24.5.3 Kategoriale Indikatorvariablen – 647

24.6	Modellevaluation – 648
24.6.1	Beurteilung der Güte des Gesamtmodells – 648
24.6.2	Beurteilung der Modellparameter – 650
24.7	Modifikation der Modellstruktur – 651
24.8	Modellvergleiche – 652
24.8.1	Geschachtelte Modelle – 652
24.8.2	Nicht geschachtelte Modelle – 653
24.9	Messinvarianztestung – 653
24.10	Zusammenfassung – 656
24.11	EDV-Hinweise – 656
24.12	Kontrollfragen – 656
	Literatur – 657

24.1 · Grundlagen

i Die konfirmatorische Faktorenanalyse (CFA, confirmatory factor analysis) unterscheidet sich von der exploratorischen Faktorenanalyse (EFA) dadurch, dass mit ihr die Passung theoretisch begründeter Modelle zu empirischen Daten überprüft werden kann, während mit der EFA nach unbekannten zugrunde liegenden Strukturen gesucht wird. Die theoretischen Annahmen eines Modells wie die Anzahl der Faktoren und die Zuordnung der Testitems zu den Faktoren werden explizit als Hypothesen aufgestellt und getestet. Deskriktiv- und inferenzstatistische Gütekriterien bieten eine Entscheidungsgrundlage, ob ein Modell angenommen oder verworfen werden sollte. Neben der Güte des Gesamtmodells können auch die einzelnen Zusammenhänge zwischen Items und Faktoren auf Signifikanz geprüft werden. Aufgrund ihrer Eigenschaften stellt die CFA ein wichtiges Instrument zur Überprüfung der Dimensionalität und damit der faktoriellen Validität eines Tests dar. Die CFA ist anwendbar für Fragestellungen, in der ein- oder mehrdimensionale latente Strukturen vermutet werden. Im Rahmen der Testkonstruktion liefert die CFA modellbasierte Kennwerte, die mit den deskriptiven Kennwerten Trennschärfe und Itemschwierigkeit der Itemanalyse vergleichbar sind, und bietet die Möglichkeit, die Messäquivalenz von Items zu testen, die als Voraussetzung für Reliabilitätsschätzungen geprüft werden sollte. Darüber hinaus lassen sich konkurrierende Faktormodelle sowie die Messinvarianz eines Tests über Gruppen bzw. Messzeitpunkte hinweg überprüfen. Die CFA ist außerdem Grundlage für komplexere Analysemodelle.

24.1 Grundlagen

Die CFA stellt neben der EFA (► Kap. 23) ein weiteres faktorenanalytisches Verfahren dar, anhand dessen psychometrische Modelle überprüft werden können.

Die CFA zählt zu den Verfahren der Strukturgleichungsmodelle (*Structural Equation Modeling*, SEM; vgl. z. B. Werner et al. 2016), in denen Zusammenhänge zwischen beobachtbaren Variablen (Indikatorvariablen, z. B. Antworten auf Items einer Perfektionismuskala) und latenten Variablen (nicht direkt beobachtbaren Merkmalen, z. B. das Persönlichkeitsmerkmal Perfektionismus) als testbare Annahmen (Hypothesen) formuliert und überprüft werden. Die Hypothesen werden in Messmodellen und Strukturmodellen dargestellt: Im Messmodell betreffen sie im Wesentlichen die Anzahl der latenten Variablen (Faktoren) und die Zuordnung der Indikatorvariablen zu den Faktoren; im Strukturmodell die Beziehungen zwischen den Faktoren.

■■ Einsatzgebiete der CFA

Im Gegensatz zur EFA werden in der CFA Hypothesen explizit spezifiziert und getestet. Ein Vorteil der CFA gegenüber der EFA besteht in der Möglichkeit zur inferenzstatistischen Überprüfung der Modellgüte, um zu entscheiden, ob das Modell beibehalten oder verworfen werden soll. Hier wird ein wesentlicher Unterschied zwischen EFA und CFA deutlich: Die EFA wird als struktursuchendes Verfahren insbesondere zur Hypothesengenerierung eingesetzt, die CFA dagegen als strukturprüfendes Verfahren zur Hypothesenprüfung. Allerdings können unerwartete Ergebnisse einer CFA wiederum zur Generierung neuer Hypothesen führen, so dass auch die CFA als exploratives Instrument genutzt werden kann.

Die CFA bietet verschiedene Möglichkeiten zur psychometrischen Evaluation eines Tests, z. B. die Untersuchung folgender Fragen:

- Sind die Items einer Skala eindimensional?
- Lässt sich die theoretisch begründete faktorielle Struktur eines Tests empirisch nachweisen?
- Sind die einzelnen Items und der gesamte Test reliabel?
- Liegt konvergente und diskriminante Validität vor?

Messmodell und Strukturmodell

Inferenzstatistischer Modelltest

Typische Ausgangsfragestellungen einer CFA

**Kovarianzmatrix als
Datengrundlage bei kontinuierlichen
Indikatorvariablen**

Die nachfolgenden Ausführungen beziehen sich im Wesentlichen auf die Anwendung der CFA bei kontinuierlichen Indikatorvariablen (d. h. mindestens Intervallskalenniveau), können aber konzeptuell auf ordinale Indikatorvariablen übertragen werden, indem zur Erklärung des Zusammenhangs zwischen einer kontinuierlichen latenten Variablen und den kategorialen Indikatorvariablen ein sog. „Schwellenwertmodell“ zugrunde gelegt wird (vgl. z. B. Reinecke 2014, S. 198). Bei kontinuierlichen Indikatorvariablen werden deren Varianzen und Kovarianzen sowie bei bestimmten Fragestellungen auch deren Mittelwerte als wesentliche empirische Informationen genutzt.

■ ■ **Globale Nullhypothese**

Globale Nullhypothese

Ob das Gesamtmodell die empirischen Daten gut beschreiben kann, wird in der CFA über die Prüfung der globalen Nullhypothese beurteilt. Hierbei wird geprüft, ob die Kovarianzmatrix Σ der Indikatorvariablen in der Population übereinstimmt mit der modellimplizierten Kovarianzmatrix $\Sigma(\theta)$, d. h. jener Kovarianzmatrix, die sich aus dem aufgestellten Modell ergibt und ausschließlich von den Modellparametern abhängt (► Abschn. 24.5). Eignet sich das Gesamtmodell zur Beschreibung der empirischen Daten, werden die Modellannahmen als bestätigt („konfirmiert“) angesehen und die globale Hypothese der Passung des Gesamtmodells kann beibehalten werden. Neben der Möglichkeit, die Kovarianzstruktur der empirischen Daten zu prüfen, kann zusätzlich die Hypothese formuliert werden, dass der Mittelwertevektor der Indikatorvariablen in der Population mit dem modellimplizierten Mittelwertevektor übereinstimmt.

Im Folgenden werden die Grundlagen der CFA in einzelnen Ablaufschritten erläutert und für eindimensionale Modelle (► Abschn. 24.3) und ausgewählte mehrdimensionale Modelle (► Abschn. 24.4) exemplarisch dargestellt:

1. *Spezifikation eines Messmodells* (► Abschn. 24.2):
 - Die Hypothesen eines Messmodells können
 - a. in einem linearen Gleichungssystem (*Modellgleichungen*) ausgedrückt und
 - b. in einem *Pfaddiagramm* grafisch veranschaulicht werden (► Abschn. 24.2.1).
 - Die Modellparameter können zur *Klassischen Testtheorie* (KTT, s. ► Kap. 13) in direkten Bezug gestellt werden (► Abschn. 24.2.2).
 - Das Gleichungssystem der CFA muss zur *Modellidentifikation* (► Abschn. 24.2.6) lösbar sein; die *Skalierung der Faktoren* muss dazu festgelegt werden (► Abschn. 24.2.5).
2. Die *Messäquivalenz* mehrerer Messungen (Indikatorvariablen), die für die Reliabilitätsschätzung von Bedeutung ist, lässt sich mittels CFA überprüfen (eindimensional ► Abschn. 24.3, mehrdimensional ► Abschn. 24.4).
3. *Parameterschätzung* (► Abschn. 24.5): Zur Schätzung der Modellparameter stehen je nach Verteilung und Skalenniveau der Indikatorvariablen verschiedene Schätzmethoden zur Verfügung.
4. *Modellevaluation* (► Abschn. 24.6): Die globale Nullhypothese der Passung des Gesamtmodells wird anhand eines Modelltests und verschiedener deskriptiver Gütekriterien geprüft. Bei gutem Modellfit können auch die einzelnen Modellparameter auf Signifikanz getestet und inhaltlich interpretiert werden.
5. *Modifikation der Modellstruktur* (► Abschn. 24.7): Die Modellstruktur kann bei Bedarf (z. B. bei mangelnder Modellgüte) geändert werden. In diesem Fall verliert die CFA allerdings ihren konfirmatorischen Charakter.
6. *Modellvergleiche* (► Abschn. 24.8): Modellvergleiche ermöglichen eine Entscheidung zwischen konkurrierenden Modellen.
7. *Messinvarianz* (► Abschn. 24.9): Die Messinvarianz einer Skala über verschiedene Gruppen von Testpersonen oder über mehrere Messzeitpunkte hinweg sollte nachgewiesen sein.

24.2 Spezifikation eines Messmodells

In Messmodellen wird festgelegt, wie die beobachtbaren Indikatorvariablen mit den nicht direkt beobachtbaren Faktoren (latenten Variablen) regressionsanalytisch verknüpft sind. Die Indikatorvariablen stellen hier abhängige Variablen dar, die durch jeweils einen Faktor als unabhängige Variable erklärt werden. Den Messmodellen liegt die allgemeine Modellvorstellung zugrunde, dass die latente Merkmalsausprägung einer Person ihre beobachteten Indikatorwerte verursacht. Nehmen wir beispielsweise an, das latente Merkmal *Leistungsbezogene Zweifel* wird über mehrere Items eines Fragebogens gemessen, so ist zu erwarten, dass eine hohe Ausprägung des Merkmals Leistungsbezogene Zweifel auch zu hohen Werten der Indikatorvariablen (d. h. den Variablen der Itemantworten, Itemvariablen) führen sollte. Wenn die Items dasselbe Konstrukt messen, resultieren Kovarianzen zwischen den Indikatorvariablen. Ist der Faktor die einzige Ursache für die Kovarianzen zwischen den Indikatorvariablen, so werden diese null, wenn der Faktor als erklärende Variable ins Modell aufgenommen wird. Werden die Kovarianzen null, so gibt es keine weiteren Faktoren, die die Zusammenhänge zwischen den Indikatorvariablen erklären können. Die grundlegende Annahme, dass die Items ein gemeinsames Konstrukt messen, kommt in der Globalhypothese der CFA zum Ausdruck. Zusätzlich lassen sich die Zusammenhänge zwischen einem Faktor und jeder einzelnen beobachteten Indikatorvariablen als Einzelhypothesen des Modells verstehen.

Faktoren erklären empirische Kovarianzen der Indikatorvariablen

24.2.1 Hypothesenbildung

Im Rahmen der Hypothesenbildung wird die Anzahl der Faktoren festgelegt, die Beziehungen zwischen den Faktoren (d. h., welche Faktoren miteinander korrelieren sollen und welche nicht) und außerdem, welche Indikatorvariablen auf welchen Faktoren laden sollen. In den Hypothesen werden lineare Zusammenhänge zwischen den Indikatorvariablen und den Faktoren postuliert.

■■ Messmodellgleichung

Die Hypothesen lassen sich mathematisch in Form von Messmodellgleichungen ausdrücken. Die Faktoren und Modellparameter werden nachfolgend mit griechischen Buchstaben bezeichnet. Entsprechend Gl. (24.1) setzt sich jede Indikatorvariable y_i ($i = 1, \dots, p$) zusammen aus dem mit der Faktorladung λ_{ij} gewichteten Einfluss des Faktors η_j ($j = 1, \dots, q$), aus einer itemspezifischen Konstanten α_i (Interzept) und einem Fehleranteil ε_i . Die Messmodellgleichung für y_i lautet somit:

$$y_i = \alpha_i + \lambda_{ij} \cdot \eta_j + \varepsilon_i \quad (24.1)$$

Messmodellgleichung

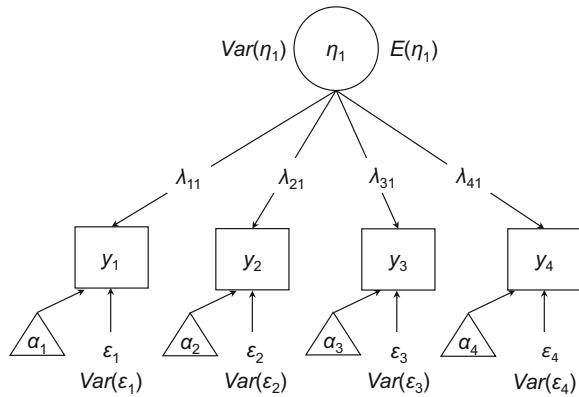
■■ Pfaddiagramm

Die Hypothesen lassen sich grafisch in einem Pfaddiagramm veranschaulichen. Faktoren werden durch Kreise und Indikatorvariablen durch Rechtecke symbolisiert; gerade Pfeile stehen für gerichtete Beziehungen (z. B. der Effekt des Faktors auf eine Indikatorvariable). Abb. 24.1 zeigt ein eindimensionales Modell mit dem Faktor η_1 , der durch vier Items y_1 bis y_4 gemessen wird. Die vier Faktorladungen λ_{11} bis λ_{41} geben jeweils die Stärke des Zusammenhangs zwischen den Indikatorvariablen und dem Faktor an. Für jede Indikatorvariable wird das jeweilige Interzept α_1 bis α_4 (Dreiecke) und zusätzlich der Einfluss des jeweiligen Messfehlers ε_1 bis ε_4 berücksichtigt.

Pfaddiagramm

Unbekannte Parameter in diesem Modell sind die Faktorladungen λ_{11} bis λ_{41} , die Interzepte α_1 bis α_4 , die Varianz des Faktors $Var(\eta_1)$, der Erwartungswert des Faktors $E(\eta_1)$ sowie die Varianzen der Messfehlervariablen $Var(\varepsilon_1)$ bis $Var(\varepsilon_4)$.

Unbekannte Modellparameter



■ Abb. 24.1 Eindimensionales Messmodell für vier Indikatorvariablen y_1 bis y_4

Empirisch beobachtbar sind die individuellen Ausprägungen der Indikatorvariablen y_1 bis y_4 bzw. deren Varianzen, Kovarianzen und Mittelwerte.

Symbole und Variablenbezeichnungen in der CFA

y_i	Indikatorvariable; manifeste oder beobachtbare Variable; Messwertvariable; Itemvariable; Variable mit den Itemantworten
η_j	Faktor; latente oder nicht beobachtbare Variable; Konstrukt; Merkmal
α_i	Interzept; Leichtigkeitsparameter
λ_{ij}	Faktorladung; Diskriminationsparameter; Effekt des Faktors auf die Indikatorvariable
ε_i	Fehlervariable; Messfehleranteil
$Var(.)$	Varianz einer Variablen
$E(.)$	Erwartungswert einer Variablen

24.2.2 Bezüge zur Klassischen Testtheorie (KTT)

Die psychometrischen Modelle der KTT (► Kap. 13) sind Spezialfälle eines konfirmatorischen Faktormodells. Da die KTT-Modelle wesentlich für die Testkonstruktion sind, sollen die Grundlagen der KTT hier kurz wiederholt werden.

Die Grundannahme der KTT ist, dass sich jeder beobachtete Messwert y_{vi} einer Person v auf der Itemvariablen des Items i zusammensetzt aus einem wahren Wert (*True-Score*) τ_{vi} und einem Messfehler ε_{vi} :

$$y_{vi} = \tau_{vi} + \varepsilon_{vi} \quad (24.2)$$

Grundgleichung der KTT

Definition des wahren Wertes als personenbedingtem Erwartungswert

Der wahre Wert einer Person ist dabei unbekannt. Er ist definiert als der personenbedingte Erwartungswert der Variablen y_{vi} (Eid und Schmidt 2014; Guttman 1945; Lord und Novick 1968; Zimmerman 1975). Würde der Messwert y_{vi} einer Person v theoretisch unendlich oft mit demselben Messinstrument (Item) i erfasst, ergäbe sich eine intraindividuelle Verteilung der Messwerte y_{vi} , deren Erwartungswert dem wahren Wert τ_{vi} dieser Person entsprechen würde. Hierbei stellt V die Personenvariable dar, die den Wert einer Person v annimmt:

$$\tau_{vi} := E(y_{vi} | V = v) \quad (24.3)$$

Eigenschaften der True-Score- und Messfehlervariablen

Aus dieser Definition und der Grundgleichung der Messfehlertheorie (Gl. 24.2) folgen für die Messfehler- und True-Score-Variablen einige wichtige Eigenschaf-

24.2 · Spezifikation eines Messmodells

ten (vgl. Eid et al. 2017b; Eid und Schmidt 2014; Lord und Novick 1968; Steyer und Eid 2001; Zimmerman 1975), die bei Moosbrugger, Gädé, Schermelleh-Engel und Rauch in ► Kap. 13 nachzulesen sind.

Ein einzelner Messwert y_{vi} kann als Schätzwert für den wahren Wert τ_{vi} dienen, jedoch kann bei nur einer einzigen Messung keine Angabe über die Präzision der Messung gemacht werden. Erst bei Mehrfachmessungen kann für jede Messwertvariable der True-Score- und der Messfehleranteil bestimmt werden. Werden latente Merkmale anhand mehrerer Items gemessen, entspricht dies einer Mehrfachmessung desselben Merkmals mit unterschiedlichen Messinstrumenten (hier Items). Folglich können die True-Score- und Messfehleranteile an der jeweiligen Itemvarianz bestimmt werden, die für die Schätzung der Item- und Testreliabilität der Messungen nötig sind (vgl. ► Kap. 14 und 15). Die messtheoretischen Grundlagen und ihre Bezüge zur CFA sind ausführlich dargestellt bei Eid und Schmidt (2014), Eid et al. (2017b) sowie Steyer und Eid (2001).

Mehrfachmessungen erlauben Aussagen über True-Score- und Messfehleranteile

24.2.2.1 Die gemeinsame latente Variable η_j

Die Varianzen der Itemvariablen lassen sich jeweils als Summe aus True-Score- und Fehlervarianz ausdrücken. Wird neben der Unkorreliertheit der Messfehler- und True-Score-Variablen zusätzlich angenommen, dass die Fehlervariablen zweier Itemvariablen y_i und $y_{i'}$ unkorreliert sind, d. h. $Cov(\varepsilon_i, \varepsilon_{i'}) = 0$, so ist die Kovarianz der Itemvariablen allein auf die Kovarianz ihrer True-Score-Variablen zurückzuführen:

$$Cov(y_i, y_{i'}) = Cov(\tau_i + \varepsilon_i, \tau_{i'} + \varepsilon_{i'}) = Cov(\tau_i, \tau_{i'}) \quad (24.4)$$

Geht man von einem eindimensionalen Modell aus, bei dem die True-Score-Variablen nur eine einzige gemeinsame latente Variable η_j messen, korrelieren die True-Score-Variablen mit einer Korrelation von $r = 1$ (vgl. Eid und Schmidt 2014, S. 326).

Korrelation der wahren Werte

Unter der Annahme der Eindimensionalität lässt sich somit eine gemeinsame latente Variable η_j definieren, die die Ausprägungen der True-Score-Variable τ_i und damit auch die Ausprägungen der Variable y_i der beobachteten Werte bestimmt (vgl. □ Abb. 24.2). Die True-Score-Variable τ_i , die sich hinsichtlich ihres Interzepts α_i und ihrer Faktorladung λ_{ij} von den anderen True-Score-Variablen unterscheiden kann, lässt sich als lineare Funktion der latenten Variablen η_j ausdrücken:

$$\tau_i = \alpha_i + \lambda_{ij} \cdot \eta_j \quad (24.5)$$

Definition einer gemeinsamen latenten Variablen

Neben den messspezifischen Konstanten α_i und λ_{ij} werden die wahren Werte allein von der Ausprägung der latenten Variablen η_j bestimmt.

Jeder Messwert y_{vi} setzt sich aus einem wahren Wert und einem Fehlerwert zusammen (Gl. 24.2). Dieser Zusammenhang lässt sich von der Personenebene auf die Ebene der Itemvariablen übertragen, sodass für jede Itemvariable y_i gilt, dass sie sich additiv zusammensetzt aus der True-Score-Variablen τ_i und der Fehlervariablen ε_i :

$$y_i = \tau_i + \varepsilon_i \quad (24.6)$$

Durch Einsetzen von Gl. (24.5) in Gl. (24.6) wird deutlich, dass die beobachteten Messwerte jeweils von der True-Score-Variablen τ_i und damit von der gemeinsamen latenten Variablen η_j abhängen:

$$y_i = \tau_i + \varepsilon_i = \alpha_i + \lambda_{ij} \cdot \eta_j + \varepsilon_i \quad (24.7)$$

Gl. (24.7) entspricht der Messmodellgleichung der CFA (Gl. 24.1). In der CFA werden durch Einführung der gemeinsamen latenten Variablen η_j die Messfehler-

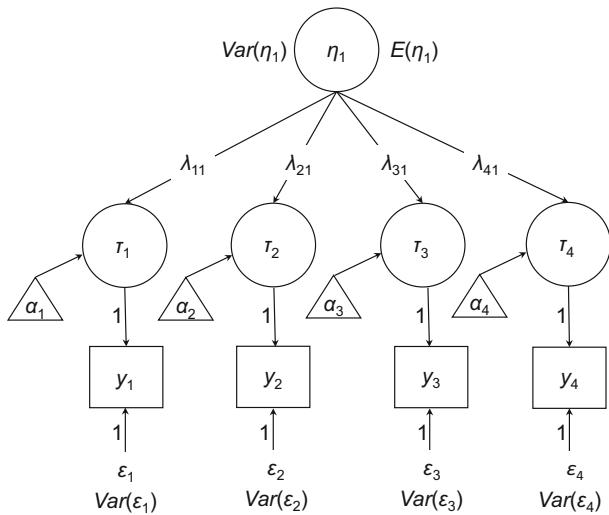


Abb. 24.2 Schematische Darstellung des Zusammenhangs zwischen einer latenten Variable η_1 , vier True-Score-Variablen τ_1 bis τ_4 , vier Indikatorvariablen y_1 bis y_4 und vier Messfehlervariablen ϵ_1 bis ϵ_4 . Die Indikatorvariablen setzen sich aus den mit dem Wert 1 gewichteten Einflüssen der True-Score-Variablen und der Fehlervariablen zusammen (vgl. Gl. 24.7)

und True-Score-Anteile mehrerer Messungen modelliert. Die latente Variable stellt einerseits das zu messende Merkmal als theoretisches Konstrukt dar und dient andererseits als Referenzgröße, um für die beobachteten Messwerte die True-Score- und Messfehleranteile zu bestimmen.

Unkorreliertheit der Fehlervariablen bei Eindimensionalität

Stellt die latente Variable die alleinige Ursache für die Ausprägungen der beobachteten Messwerte dar, geht damit die Unkorreliertheit der Fehlervariablen einher. Dies bedeutet, dass nach Auspartialisierung des Einflusses der latenten Variablen die Partialkorrelation zwischen zwei Indikatorvariablen y_i und $y_{i'}$ null wird (vgl. dazu die lokale stochastische Unabhängigkeit im Rasch-Modell, s. ► Kap. 16, ► Abschn. 16.3.7).

Abb. 24.2 verdeutlicht, dass die latente Variable η_1 die Ausprägungen der True-Score-Variablen τ_1 bis τ_4 bestimmt, die sich in den Parametern α_i und λ_{ij} voneinander unterscheiden können. Die Werte der Indikatorvariablen y_1 bis y_4 setzen sich zusammen aus den Werten der True-Score-Variablen τ_1 bis τ_4 und den Werten der Messfehler ϵ_1 bis ϵ_4 . Da sich die Indikatorvariablen direkt als Funktion der latenten Variablen darstellen lassen (vgl. Gl. 24.7), wird deutlich, dass sich das Modell mit True-Score-Variablen (Abb. 24.2) zu dem Modell ohne True-Score-Variablen (Abb. 24.1) vereinfachen lässt.

24.2.2.2 Die additive Konstante α_i (Interzept)

Unterscheiden sich Messungen mit verschiedenen Messinstrumenten um einen additiven, konstanten Wert α_i , lässt sich dies auf eine unterschiedliche Metrik und/oder Kalibrierung der Messinstrumente zurückführen. Bei psychologischen Tests wäre dies auch dann zu beobachten, wenn sich Testitems hinsichtlich ihrer Schwierigkeit (bzw. Leichtigkeit) unterscheiden.

Beispielsweise könnte eine Person v bei gleichbleibender Ausprägung des zu messenden latenten Merkmals η_j bei Messung mit Item 1 den wahren Wert $\tau_{v1} = E(y_{v1} | V = v)$, und bei Messung mit Item 2 einen anderen, um den Betrag α verschobenen, wahren Wert $\tau_{v2} = E(y_{v2} | V = v)$ haben (Gl. 24.33). Dieser Unterschied wäre dann auf die unterschiedliche Schwierigkeit der Items und nicht auf Unterschiede im latenten Merkmal zurückzuführen.

Leichtigkeitsparameter

Der Parameter α_i kann auch als Leichtigkeitsparameter bezeichnet werden (Eid und Schmidt 2014; Eid et al. 2017b). Eine Person gibt beispielsweise auf Item 1

24.2 · Spezifikation eines Messmodells

zur Messung eines latenten Merkmals η_j auf einer fünfstufigen Ratingskala den Wert 5 an (stark ausgeprägt) und auf Item 2 zur Messung desselben Merkmals nur den Wert 3 (mittelmäßig ausgeprägt). Da Item 2 ein extremeres Beispiel des latenten Merkmals abfragt als Item 1, ist es „schwieriger“, auf Item 2 eine hohe Ausprägung zu erreichen. Item 1 ist somit das leichtere Item mit einem höheren „Sockelbetrag“ α_1 : Auch bei einer Person mit einer geringeren Ausprägung des latenten Merkmals wird mit Item 1 ein höherer Wert gemessen als mit Item 2.

Die Itemschwierigkeit (bzw. die Itemleichtigkeit) ist als Kennwert aus der deskriptiven Itemanalyse bekannt (► Kap. 7); im Rahmen der CFA kann der Itemleichtigkeitsparameter als Interzept modellbasiert geschätzt werden (vgl. hierzu auch die Itemschwierigkeitsparameter in der Item-Response-Theorie [IRT], ► Kap. 16).

24.2.2.3 Die multiplikative Konstante λ_{ij} (Faktorladung)

Der multiplikative Faktor λ_{ij} , die Faktorladung, wird auch als Diskriminationsparameter bezeichnet (vgl. Eid und Schmidt 2014). Er gibt an, wie eine Änderung der latenten Variablen η_j übersetzt wird in eine Änderung der True-Score-Variablen τ_i . Damit gibt er ebenfalls an, welche Änderung der Itemvariablen y_i mit einer Änderung der latenten Variablen η_j einhergeht, und stellt eine spezifische Eigenart jeder Messung dar. Die numerische Höhe der Faktorladung hängt zusätzlich von der Skalierung der Messinstrumente (Items) ab.

Der Faktorladungsparameter stellt ein Äquivalent zum Trennschärfeleffizienten aus der Itemanalyse dar (► Kap. 7) und ist dem Itemdiskriminationsparameter des 2PL-Modells der IRT (zweiparametrisches logistisches Modell, ► Abschn. 16.4) vergleichbar. Allerdings beruhen die Trennschärfeleffizienten auf den empirischen Korrelationen zwischen den messfehlerbehafteten Itemvariablen und den ebenfalls messfehlerbehafteten Testwerten (part-whole-korrigiert, d. h. ohne das entsprechende Item). Die Faktorladungen hingegen sind Schätzungen der Korrelationen zwischen den Itemvariablen und der messfehlerfreien latenten Variablen auf Basis eines klar definierten Messmodells, z. B. des Modells τ -kongenerischer Variablen (► Abschn. 24.3.1), dessen Passung auf die Daten anhand der Modellgüte getestet werden kann. Eine hohe positive Faktorladung bedeutet, dass ein hoher Anteil der Messwertvarianz durch den Faktor erklärt wird.

Diskriminationsparameter

Faktorladung als modellbasierte Itemtrennschärfe

Standardisierung von Faktorladungen

So wie in einer einfachen Regression ein standardisierter Regressionskoeffizient der Korrelation zwischen Prädiktor- und Kriteriumsvariable entspricht, so entspricht eine standardisierte Faktorladung jeweils der Korrelation zwischen dem Faktor und einer Indikatorvariablen. Zur Standardisierung wird die unstandardisierte Faktorladung $\lambda_{ij}(\text{unstand.})$ an der Standardabweichung des Faktors σ_{η_j} und der Standardabweichung der Indikatorvariablen σ_{y_i} relativiert:

$$\lambda_{ij}(\text{stand.}) = \frac{\lambda_{ij}(\text{unstand.}) \cdot \sigma_{\eta_j}}{\sigma_{y_i}} \quad (24.8)$$

Die quadrierte standardisierte Faktorladung entspricht dem Anteil der durch den Faktor erklärten Varianz an der Gesamtvarianz der Indikatorvariablen und gibt die Reliabilität (Messgenauigkeit) der Itemvariablen y_i an.

$$\begin{aligned} \lambda_{ij}^2(\text{stand.}) &= \left(\frac{\lambda_{ij}(\text{unstand.}) \cdot \sigma_{\eta_j}}{\sigma_{y_i}} \right)^2 = \frac{\lambda_{ij}^2(\text{unstand.}) \cdot \sigma_{\eta_j}^2}{\sigma_{y_i}^2} = \frac{\overbrace{\lambda_{ij}^2(\text{unstand.}) \cdot \text{Var}(\eta_j)}^{\text{erklärte Varianz}}}{\underbrace{\text{Var}(y_i)}_{\text{Gesamtvarianz}}} \\ &= \text{Rel}(y_i) \end{aligned} \quad (24.9)$$

Itemreliabilität

24.2.3 Elemente der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$

Varianzzerlegung

Die wesentliche Grundlage zur Bestimmung der Modellparameter einer CFA sind die Varianzen und Kovarianzen der Indikatorvariablen. Die Varianz einer Indikatorvariablen lässt sich additiv zerlegen in die Varianz der True-Score-Variablen und die Varianz der Messfehlervariablen (s. Eigenschaften der True-Score- und Messfehlervariablen, ▶ Kap. 13). Die Varianz der Indikatorvariablen y_i lässt sich unter Verwendung von Gl. (24.7) in die folgenden Modellparameter zerlegen und entsprechend vereinfachen, da die Varianz einer Konstanten (hier α_i) null ist:

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\tau_i + \varepsilon_i) = \text{Var}(\alpha_i + \lambda_{ij} \cdot \eta_j) + \text{Var}(\varepsilon_i) \\ &= \lambda_{ij}^2 \cdot \text{Var}(\eta_j) + \text{Var}(\varepsilon_i) \end{aligned} \quad (24.10)$$

Die mit λ_{ij}^2 gewichtete Varianz des Faktors entspricht der Varianz der True-Score-Variablen.

Empirische Varianzen und Kovarianzen als Funktionen der Modellparameter

Auch die Kovarianz zwischen zwei Indikatorvariablen y_i und $y_{i'}$ lässt sich durch Einsetzen ihrer Messmodellgleichungen in Modellparametern ausdrücken (Gl. 24.11). Da die Kovarianz einer Variablen mit einer Konstanten gleich null ist (d. h., alle Kovarianzen mit α_i und $\alpha_{i'}$) und die Fehlervariablen ε_i und $\varepsilon_{i'}$ untereinander sowie mit dem Faktor η_j unkorreliert sind, lässt sich Gl. (24.11) entsprechend vereinfachen:

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \text{Cov}(\tau_i + \varepsilon_i, \tau_{i'} + \varepsilon_{i'}) \\ &= \text{Cov}(\alpha_i + \lambda_{ij} \cdot \eta_j + \varepsilon_i, \alpha_{i'} + \lambda_{i'j} \cdot \eta_j + \varepsilon_{i'}) \\ &= \lambda_{ij} \cdot \lambda_{i'j} \cdot \text{Var}(\eta_j) \end{aligned} \quad (24.11)$$

Die Varianzen und Kovarianzen der Indikatorvariablen stellen die Elemente der empirischen Kovarianzmatrix \mathbf{S} dar; sie dienen als Schätzer für die Populationskovarianzmatrix Σ . Aus Gln. (24.10) und (24.11) geht hervor, dass sich die Varianzen und Kovarianzen der Indikatorvariablen (zusammengefasst in der Matrix \mathbf{S}) als Funktionen der Modellparameter darstellen lassen. Die derart durch die Modellparameter reproduzierten Varianzen und Kovarianzen stellen die Elemente der vom Modell implizierten Kovarianzmatrix $\hat{\Sigma}(\hat{\theta})$ dar, die zur Schätzung der modellimplizierten Populationskovarianzmatrix $\Sigma(\theta)$ dient. Sie hängt ausschließlich von den Modellparametern ab, die im Vektor θ zusammengefasst sind. Die modellimplizierte Kovarianzmatrix $\hat{\Sigma}(\hat{\theta})$ wird hier vereinfacht als $\hat{\Sigma}$ bezeichnet.

Für das Modell mit vier Indikatorvariablen (vgl. □ Abb. 24.1) sind die Elemente der Kovarianzmatrizen \mathbf{S} und $\hat{\Sigma}$ in □ Tab. 24.1 aufgeführt.

■ Tabelle 24.1 Elemente der empirischen Kovarianzmatrix \mathbf{S} als Schätzer von Σ und der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$ als Schätzer von $\Sigma(\theta)$

Empirische Kovarianzmatrix \mathbf{S}				Modellimplizierte Kovarianzmatrix $\hat{\Sigma}$			
$\text{Var}(y_1)$				$\lambda_{11}^2 \text{Var}(\eta_1) + \text{Var}(\varepsilon_1)$			
$\text{Cov}(y_2, y_1)$	$\text{Var}(y_2)$			$\lambda_{21} \lambda_{11} \text{Var}(\eta_1)$	$\lambda_{21}^2 \text{Var}(\eta_1) + \text{Var}(\varepsilon_2)$		
$\text{Cov}(y_3, y_1)$	$\text{Cov}(y_3, y_2)$	$\text{Var}(y_3)$		$\lambda_{31} \lambda_{11} \text{Var}(\eta_1)$	$\lambda_{31} \lambda_{21} \text{Var}(\eta_1)$	$\lambda_{31}^2 \text{Var}(\eta_1) + \text{Var}(\varepsilon_3)$	
$\text{Cov}(y_4, y_1)$	$\text{Cov}(y_4, y_2)$	$\text{Cov}(y_4, y_3)$	$\text{Var}(y_4)$	$\lambda_{41} \lambda_{11} \text{Var}(\eta_1)$	$\lambda_{41} \lambda_{21} \text{Var}(\eta_1)$	$\lambda_{41} \lambda_{31} \text{Var}(\eta_1)$	$\lambda_{41}^2 \text{Var}(\eta_1) + \text{Var}(\varepsilon_4)$

λ = Faktorladung, η = Faktor, ε = Fehlervariable, Var = Varianz, Cov = Kovarianz

24.2.4 Elemente des modellimplizierten Mittelwertevektors $\hat{\mu}$

Auch die Erwartungswerte $E(y_i)$ der Indikatorvariablen lassen sich in Abhängigkeit der Modellparameter ausdrücken:

$$E(y_i) = E(\tau_i) + E(\varepsilon_i) = E(\alpha_i + \lambda_{ij} \cdot \eta_j) + 0 = \alpha_i + \lambda_{ij} \cdot E(\eta_j) \quad (24.12)$$

Der Erwartungswert $E(y_i)$ einer Indikatorvariablen setzt sich additiv zusammen aus dem Interzept α_i und dem mit der Faktorladung λ_{ij} gewichteten Erwartungswert des Faktors η_j . Der Erwartungswert des Messfehlers ε_i ist null und entfällt somit (vgl. Eigenschaften der True-Score- und Messfehlervariablen, ▶ Kap. 13). Wenn der Erwartungswert des Faktors η_j , $E(\eta_j)$, zur Skalierung des Faktors auf den Wert null gesetzt wird (▶ Abschn. 24.2.5), vereinfacht sich Gl. (24.12) zu $E(y_i) = \alpha_i$. In diesem Fall entsprechen die Interzepte α_i den Erwartungswerten der jeweiligen Indikatorvariablen. Da der wahre Wert definiert ist als personenbedingter Erwartungswert (vgl. Gl. 24.3), gilt weiter, dass α_i in dem Fall τ_i entspricht ($E(y_i) = \tau_i = \alpha_i$).

Erwartungswerte der Indikatorvariablen als Funktionen der Modellparameter

- ❶ Der Erwartungswert entspricht dem Wert, den eine Zufallsvariable bei theoretisch unbegrenzt wiederholter Stichprobeneziehung im Mittel annehmen würde. Nach dem zentralen Grenzwerttheorem konvergieren Stichprobenmittelwerte bei wachsender Stichprobengröße gegen diesen Erwartungswert. Stichprobenmittelwerte werden daher als Schätzer für den Erwartungswert einer Variablen verwendet. Die Varianz einer Zufallsvariablen wird durch die Stichprobenvarianz geschätzt.

Für das Modell mit vier Indikatorvariablen (vgl. □ Abb. 24.1) werden die Populations-Mittelwertevektoren μ und $\mu(\theta)$ über den empirischen Mittelwertevektor \bar{y} und den modellimplizierten Mittelwertevektor $\mu(\hat{\theta})$ geschätzt, wobei $\mu(\hat{\theta})$ hier vereinfacht als $\hat{\mu}$ bezeichnet wird (□ Tab. 24.2).

24.2.5 Skalierung der Faktoren

Faktoren sind als latente Variablen hinsichtlich ihrer Metrik nicht definiert, sodass eine Skalierung vorgenommen werden muss. Die gängigsten Skalierungsarten zur Festlegung der Varianz und des Erwartungswertes latenter Variablen werden nachfolgend dargestellt, daneben sind weitere Skalierungsarten möglich (vgl. Brown 2015).

Skalierung zur Festlegung der Maßeinheiten der Faktoren

■■ Festlegung der Faktorvarianz

Zur Festlegung der Faktorvarianz wird meist eine der folgenden Skalierungsmethoden verwendet. Die Faktorvarianz kann auf den Wert eins fixiert werden, wodurch

□ Tabelle 24.2 Elemente des empirischen und des modellimplizierten Mittelwertevektors

Empirischer Mittelwertevektor \bar{y}	Modellimplizierter Mittelwertevektor $\hat{\mu}$
\bar{y}_1	$\alpha_1 + \lambda_{11} \cdot E(\eta_1)$
\bar{y}_2	$\alpha_2 + \lambda_{21} \cdot E(\eta_1)$
\bar{y}_3	$\alpha_3 + \lambda_{31} \cdot E(\eta_1)$
\bar{y}_4	$\alpha_4 + \lambda_{41} \cdot E(\eta_1)$

λ = Faktorladung, η = Faktor, α = Interzept, E = Erwartungswert

die latente Variable standardisiert wird. Alternativ kann die Faktorvarianz festgelegt werden, indem pro Faktor je eine Faktorladung auf den Wert eins fixiert wird. Damit dient jeweils eine Indikatorvariable als „Skalierer“ (Referenzitem) und bestimmt die Varianz des Faktors, die mit der erklärten Varianz der Skaliervariablen gleichgesetzt wird. Es wird empfohlen, möglichst das reliabelste Item als Referenzitem zu verwenden. Die Wahl der Skalierungsmethode hängt von der Fragestellung ab. Ist die latente Varianz ein bedeutsamer Parameter, weil z. B. zwei Gruppen miteinander verglichen werden sollen, die sich in den latenten Varianzen unterscheiden können, so würde man eine Faktorladung auf eins fixieren. Wird hingegen die Dimensionalität eines Tests überprüft, wird häufig die Varianz des Faktors auf eins fixiert.

■ ■ Festlegung der Erwartungswerte der Faktoren

Analog zur Varianz ist auch der Erwartungswert eines Faktors nicht eindeutig bestimmt und muss durch eine Normierung festgelegt werden. Eine Möglichkeit besteht darin, den Erwartungswert des Faktors auf den Wert null zu fixieren. In dem Fall sind die Interzepte α_i eindeutig interpretierbar, sie entsprechen dann den Erwartungswerten der jeweiligen Indikatorvariablen, die durch die empirischen Mittelwerte der Messwerte geschätzt werden (vgl. Gl. 24.12).

Alternativ kann der Erwartungswert des Faktors auch dadurch festgelegt werden, dass pro Faktor ein Interzept auf null fixiert wird. Damit dient das Interzept dieser Indikatorvariablen als Referenzwert. Wird dasselbe Item als Skalierer für die Varianz und den Erwartungswert eines Faktors verwendet, dann (und nur dann) entspricht der Erwartungswert des Faktors dem Mittelwert des Referenzitems (vgl. Gl. 24.12, wenn $\alpha_i = 0$ und $\lambda_{ij} = 1$). Auch hier hängt die Wahl der Skalierungsmethode von der Fragestellung ab. In der Regel wird der Erwartungswert des Faktors auf null fixiert und nur dann frei geschätzt, wenn z. B. geprüft werden soll, ob sich die Erwartungswerte in verschiedenen Stichproben unterscheiden.

24.2.6 Modellidentifikation

Lösbarkeit der Messmodellgleichungen

Ein Messmodell gilt als „identifiziert“, wenn das lineare Gleichungssystem des Messmodells gelöst werden kann. Damit das Gleichungssystem lösbar ist, muss zunächst die Skalierung der Faktoren festgelegt werden (► Abschn. 24.2.5). Außerdem muss die Anzahl empirisch verfügbarer Informationen (s) die Anzahl der zu schätzenden Modellparameter (t) übersteigen, da ansonsten keine Freiheitsgrade zur Überprüfung des Modells („Modelltest“) vorliegen.

Anzahl der Freiheitsgrade (df)

Die Differenz zwischen der Anzahl s der empirisch verfügbaren Informationen und der Anzahl zu schätzender Modellparameter t ergibt die Anzahl der Freiheitsgrade (degrees of freedom, df) des Modells: $df = s - t$.

Prinzipiell sind drei Ergebnisse dieser Differenzbildung möglich (vgl. Bollen 1989; Brown 2015):

- $s < t$: Das Modell ist nicht identifiziert, das Gleichungssystem ist nicht lösbar und die Parameter nicht schätzbar: $df < 0$.
- $s = t$: Das Modell ist genau identifiziert, das Gleichungssystem ist eindeutig lösbar, die Parameter reproduzieren die empirischen Zusammenhänge perfekt, ein Modelltest ist nicht möglich: $df = 0$.
- $s > t$: Das Modell ist überidentifiziert, das Gleichungssystem ist lösbar, alle Parameter sind schätzbar, ein Modelltest kann durchgeführt werden: $df > 0$.

■■ Anzahl der zu schätzenden Modellparameter

Zur Bestimmung der Anzahl zu schätzender Parameter t eines Modells können die frei zu schätzenden Parameter im Pfadmodell gezählt werden. Hierzu gehören sämtliche Faktorladungen, Fehlervarianzen, Faktorvarianzen und -kovarianzen sowie ggf. die Interzepte und Erwartungswerte der Faktoren.

Benötigte Parameter für die Analyse der Kovarianzstruktur: Um die Kovarianzstruktur der Indikatoren anhand der modellimplizierten Kovarianzmatrix zu reproduzieren, müssen die Parameter λ_{ij} , $Var(\eta_i)$ und $Var(\varepsilon_i)$ auf Basis der empirischen Varianzen und Kovarianzen der Indikatorvariablen, d. h. der empirischen Kovarianzmatrix S , geschätzt werden.

■ Anzahl der verfügbaren empirischen Informationen für die Analyse der Kovarianzstruktur

Um die Modellparameter schätzen zu können, muss die Anzahl der empirischen Varianzen und Kovarianzen groß genug sein. Die Anzahl s der empirisch verfügbaren Varianzen und Kovarianzen lässt sich ermitteln als

$$s = \frac{p(p+1)}{2}, \quad (24.13)$$

wobei p die Anzahl der Indikatoren ist und s die Anzahl der nicht redundanten Elemente der Kovarianzmatrix.

Für das Beispiel eines Messmodells mit vier Indikatoren würden als Informationen in der empirischen Kovarianzmatrix zehn Elemente zur Verfügung stehen: $s = 4(4 + 1)/2 = 10$ (vier Varianzen und sechs Kovarianzen, □ Tab. 24.1).

Benötigte Parameter für die zusätzliche Analyse der Mittelwertestruktur: Ist neben der Kovarianzstruktur auch die Mittelwertestruktur von Interesse, müssen die Parameter α_i und $E(\eta_j)$ auf Basis der empirischen Mittelwerte der Indikatorvariablen geschätzt werden.

■ Anzahl der verfügbaren empirischen Informationen für die zusätzliche Analyse der Mittelwertestruktur

Als empirisch verfügbare Informationen dienen in diesem Fall die p Mittelwerte der Indikatorvariablen (Elemente des empirischen Mittelwertevektors). Insgesamt liegen somit neben den $p(p+1)/2$ Varianzen und Kovarianzen zusätzlich die p Mittelwerte als empirisch verfügbare Informationen s vor:

$$s = p + \frac{p(p+1)}{2} = \frac{2p}{2} + \frac{p(p+1)}{2} = \frac{p(p+3)}{2} \quad (24.14)$$

Für das Beispiel eines Messmodells mit vier Indikatoren würden als empirische Informationen vier Varianzen, sechs Kovarianzen sowie vier Mittelwerte zur Verfügung stehen: $s = 4(4 + 3)/2 = 14$.

Erhöhung der Freiheitsgrade durch Modellrestriktionen: Wird ein Parameter auf einen bestimmten Wert fixiert (z. B. eine Faktorladung für die Skalierung auf den Wert eins), so zählt er nicht mehr zu den frei zu schätzenden Parametern. Die Anzahl der unbekannten Parameter reduziert sich und die Anzahl der Freiheitsgrade steigt entsprechend an. Werden außerdem Modellparameter gleichgesetzt (z. B. zwei Faktorladungen: $\lambda_{11} = \lambda_{12}$), so wird nur ein Wert für beide Parameter geschätzt und die Anzahl der zu schätzenden Parameter im Modell reduziert sich entsprechend durch derartige Modellrestriktionen. Die Fixierung einzelner Parameter sowie die Einführung von Modellrestriktionen ergeben sich einerseits aus der Notwendigkeit der Skalierung und andererseits aus theoretischen Überlegungen (Beispiele für Modelle mit Modellrestriktionen finden sich in ▶ Abschn. 24.3 und 24.9).

Modellrestriktionen erhöhen die Anzahl der Freiheitsgrade

Anmerkung zur Analyse der Mittelwertestruktur: Die Anzahl der zu schätzenden Parameter der Mittelwertestruktur (Interzepte, Erwartungswerte der Faktoren) reduziert sich um den Parameter, der für die Skalierung fixiert wird. Somit entspricht die Anzahl der zu schätzenden Parameter genau der Anzahl empirisch verfügbarer Informationen des Mittelwertevektors. Die Mittelwertestruktur ist daher genau identifiziert, sodass sich weder die Anzahl der Freiheitsgrade noch die Passung des Gesamtmodells (Modellfit) bei Berücksichtigung der Mittelwertestruktur ändert im Vergleich zur Analyse desselben Modells ohne Mittelwertestruktur. Erst wenn explizite Annahmen hinsichtlich der Mittelwertestruktur geprüft werden sollen und dazu entsprechende Modellrestriktionen eingeführt werden, wird die Mittelwertestruktur für die Modellschätzung relevant. Dies ist z. B. bei Invarianztestung über Gruppen oder Messzeitpunkte hinweg der Fall (► Abschn. 24.9).

Für die Erklärung der Zusammenhangsstruktur sind die Interzepte irrelevant, da Verschiebungen um eine additive Konstante keinen Einfluss auf die Kovarianzstruktur haben. Die Mittelwertestruktur wird daher nur für bestimmte Fragestellungen explizit analysiert und ansonsten nicht weiter berücksichtigt. Auch die Bestimmung der Freiheitsgrade bezieht sich daher meist nur auf die Kovarianzstruktur.

Analyse der Mittelwertestruktur nur bei bestimmten Fragestellungen relevant

Daumenregel

Wenn ein Modell aus mehreren Faktoren besteht (► Abschn. 24.4), kann jedes einzelne Messmodell als ein Teilmodell angesehen werden. Hier muss zunächst geprüft werden, ob das Gesamtmodell mit allen Faktoren identifiziert ist. Zusätzlich sollte auch die Identifikation der Teilmodelle geprüft werden. Als Daumenregel gilt, dass mindestens drei Indikatoren pro Faktor zur Identifikation benötigt werden.

24.2.7 Empirisches Beispiel

Um die eindimensionalen und mehrdimensionalen Modelle in den nachfolgenden ► Abschn. 24.3 und 24.4 anhand von empirischen Daten veranschaulichen zu können, wird ein Datensatz von $N = 250$ Personen aus einer Untersuchung von Amend (2015) herangezogen. Die Daten beziehen sich auf die Mehrdimensionale Perfektionismus-Skala (MPS-F) von Frost et al. (1990) in der deutschen Fassung von Stöber (1995), die hinsichtlich ihrer faktoriellen Struktur überprüft werden soll.

Von den sechs Subskalen der MPS-F werden nachfolgend jene drei verwendet, die zentrale Aspekte des Perfektionismus darstellen (vgl. u. a. Altstötter-Gleich und Bergemann 2006). Die Items wurden auf einer fünfstufigen Ratingskala beantwortet ($1 = \text{„trifft überhaupt nicht zu“}$ bis $5 = \text{„trifft ganz genau zu“}$). Aus didaktischen Gründen wird jeweils nur eine Auswahl an Items pro Subskala verwendet.

Für vier Items der Subskala *Concern over Mistakes* (CM; Fehlersensibilität), drei Items der Subskala *Doubts about Actions* (DA; Leistungsbezogene Zweifel) und drei Items der Subskala *Personal Standards* (PS; Hohe Standards) werden in den nachfolgenden Abschnitten verschiedene Modelle überprüft:

- Eindimensionales Modell der Subskala CM (► Abschn. 24.3.1, ► Beispiel 24.1)
- Mehrdimensionales Modell mit drei korrelierten Faktoren PS, DA, CM (► Abschn. 24.4.1, ► Beispiel 24.2)
- Mehrdimensionales Modell mit einem Faktor höherer Ordnung (► Abschn. 24.4.2, ► Beispiel 24.3)
- Mehrdimensionales Bifaktormodell mit einem Generalfaktor (η_{Gen}) und drei subskalenspezifischen Residualfaktoren ξ_{PS} , ξ_{DA} und ξ_{CM} (► Abschn. 24.4.3, ► Beispiel 24.4)

Alle Analysen wurden mit dem Programm *Mplus*, Version 8 (Muthén und Muthén 2017) durchgeführt. Da die Indikatorvariablen nicht normalverteilt sind, wurde die robuste Maximum-Likelihood-Schätzmethode (MLR) verwendet (► Abschn. 24.5).

24.3 Eindimensionale Modelle: Stufen der Messäquivalenz

Die Überprüfung der Dimensionalität und somit der faktoriellen Validität einer Skala (eines Tests/Fragebogens) stellt einen wesentlichen Bestandteil der Validitätsprüfung eines psychologischen Tests dar. Eindimensionalität lässt sich im Rahmen der CFA formal beschreiben und prüfen und bedeutet, dass eine Skala genau *ein* latentes Merkmal misst bzw. dass den Items einer Skala genau *ein* Faktor zugrunde liegt. Sind die Items eines Fragebogens eindimensional, so lassen sich die systematischen Zusammenhänge (Kovarianzen/Korrelationen) der Indikatorvariablen durch einen einzigen gemeinsamen Faktor erklären (► Abschn. 24.2.2).

Liegt ein Messmodell einer über p Indikatoren gemessenen latenten Variablen η_j vor, so lassen sich verschiedene Stufen der Messäquivalenz der Indikatoren unterscheiden. Die Stufen der Messäquivalenz gehen mit bestimmten Modellannahmen einher. Entsprechende auf der KTT aufbauende Modelle (vgl. Eid und Schmidt 2014; Raykov und Marcoulides 2011) können definiert und mithilfe der CFA getestet werden. Die Stufe der Messäquivalenz hängt davon ab, ob sich die Indikatoren in Bezug auf die Parameter α_i , λ_{ij} oder $Var(\varepsilon_i)$ unterscheiden.

Messäquivalenz spielt z. B. bei der Reliabilitätsschätzung eines Tests oder einer Skala eine wichtige Rolle. Den verschiedenen Reliabilitätsmaßen liegen unterschiedlich strenge Annahmen hinsichtlich der Messäquivalenz zugrunde, die erfüllt sein müssen, um eine zuverlässige Schätzung der Reliabilität zu erhalten (► Abschn. 24.3.4, s. auch ► Kap. 14).

Im Wesentlichen werden drei unterschiedlich strenge Formen der Messäquivalenz unterschieden, die mit der CFA getestet werden können und die für die Reliabilitätsschätzung einer Skala von Bedeutung sind:

- τ -Kongenerität
- Essentielle τ -Äquivalenz
- Essentielle τ -Parallelität

Stufen der Messäquivalenz

! Um zu entscheiden, welches Modell zur Beschreibung der empirischen Daten verwendet werden soll, können die Modelle τ -kongenerischer Variablen, essentiell τ -äquivalenter Variablen und essentiell τ -paralleler Variablen als geschachtelte Modelle mit zunehmend strengeren Restriktionen mittels χ^2 -Differenztest im Modellvergleich gegeneinander getestet werden (► Abschn. 24.8).

24.3.1 τ -Kongenerität

Die Eindimensionalität einer Skala ohne weitere Zusatzannahmen wird durch ein τ -kongenerisches Messmodell (Jöreskog 1971) geprüft. Für zahlreiche als eindimensional konzipierte Tests dürfte dieses Modell das passendste sein, da es die – vergleichsweise – schwächsten Modellannahmen trifft. Das τ -kongenerische Modell lässt zu, dass sich verschiedene Messungen einer latenten Variablen sowohl in Bezug auf die Parameter α_i als auch in Bezug auf λ_{ij} unterscheiden dürfen. Das bedeutet, dass sich der Faktor η_j auf die Indikatorvariablen y_i unterschiedlich stark auswirkt (unterschiedliche Faktorladungen λ_{ij}) und dass die Variablen y_i unterschiedliche Mittelwerte (Interzepte α_i) aufweisen können. Zusätzlich dürfen die Indikatorvariablen unterschiedliche Messfehlervarianzen $Var(\varepsilon_i)$ aufweisen. Da die quadrierte Faktorladung bei standardisierten manifesten und latenten Variablen den Anteil der erklärten (wahren) Varianz an der Gesamtvarianz einer Indikatorva-

Keine Modellrestriktionen
hinsichtlich α_i , λ_{ij} und $Var(\varepsilon_i)$

riablen angibt (vgl. Gl. 24.9), unterscheiden sich im τ -kongenerischen Messmodell aufgrund unterschiedlicher Faktorladungen auch die erklärten Varianzanteile der Indikatorvariablen.

Das in ► Abschn. 24.2.1 behandelte Modell (► Abb. 24.1) stellt ein τ -kongenerisches Modell dar, welches in ► Beispiel 24.1 anhand des Perfektionismus-Datensatzes (► Abschn. 24.2.7) auf Eindimensionalität überprüft werden soll. Die vier Indikatoren y_1 bis y_4 messen ein gemeinsames latentes Merkmal (Faktor η_j). Für jede Indikatorvariable lässt sich eine lineare Gleichung entsprechend Gl. (24.1) aufstellen. Zur Skalierung des Faktors η_j wird dessen Varianz auf eins und dessen Erwartungswert auf null fixiert. Als Datengrundlage liegen $s = 14$ empirische Informationen vor (4 Varianzen, 6 Kovarianzen und 4 Mittelwerte bzw. $s = 4(4 + 3)/2$), bei $t = 12$ zu schätzenden Parametern (4 Faktorladungen, 4 Fehlervarianzen und 4 Interzepte). Das Modell hat somit zwei Freiheitsgrade ($df = 14 - 12 = 2$).

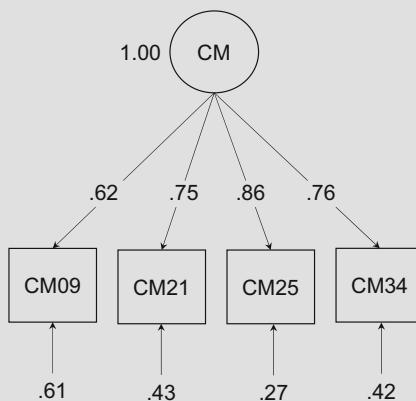
Beispiel 24.1: Eindimensionales Modell

Für die Items der Skala CM der Perfektionismusdaten (► Abschn. 24.2.7) wird exemplarisch geprüft, ob ihnen ein eindimensionales Messmodell ohne weitere Annahmen, d. h. ein τ -kongenerisches Messmodell zugrunde liegt (vgl. ► Abb. 24.1). Die Hypothesen des Modells lassen sich entsprechend Gl. (24.1) in vier Messmodellgleichungen ausdrücken.

Die Modellparameter werden auf Grundlage der empirischen Informationen der vier Indikatoren geschätzt: $s = 4(4 + 3)/2 = 14$, bestehend aus zehn nicht redundanten Elementen in der Kovarianzmatrix und vier Elementen des Mittelwertvektors. Zur Skalierung des Faktors CM wird dessen Varianz auf eins und der Erwartungswert des Faktors auf null fixiert.

Zu schätzen sind somit noch $t = 12$ Parameter (4 Interzepte + 4 Faktorladungen + 4 Fehlervarianzen), das Modell hat somit $df = 14 - 12 = 2$ Freiheitsgrade. Man erhält hier die Interzepte als Schätzwerte der Itemschwierigkeiten, da der Erwartungswert des Faktors auf null fixiert wurde.

Wird das eindimensionale (τ -kongenerische) Messmodell mittels MLR-Schätzung (► Abschn. 24.5) analysiert, resultieren für die vier ausgewählten Items der Skala CM folgende Parameterschätzungen (komplett standardisierte Lösung):



Das Modell zeigt eine sehr gute Passung (► Abschn. 24.6.1). Der χ^2 -Wert ist nicht signifikant ($\chi^2(2) = 1.11, p = .57$) und auch die deskriptiven Gütekriterien weisen auf einen sehr guten Modellfit hin (Root Mean Square Error of Approximation, RMSEA = .00, 90 %-Konfidenzintervall [.00; .11], Comparative Fit Index, CFI = 1.00; Standardized Root Mean Square Residual, SRMR = .01).

Da der Modellfit sehr gut ist und auch die Faktorladungen der vier Indikatoren auf dem gemeinsamen Faktor mit Werten zwischen .62 und .86 zufriedenstellend

hoch sind, ist davon auszugehen, dass den Indikatorvariablen ein gemeinsamer Faktor zugrunde liegt.

Der empirische Mittelwertevektor \bar{y} der vier CM-Items entspricht hier genau dem modellimplizierten Mittelwertevektor $\hat{\mu}$, weil der Erwartungswert der latenten Variable $E(CM)$ auf null fixiert wurde. Die Interzepte (s. Elemente des modellimplizierten Mittelwertevektors $\hat{\mu}$) stellen mit Werten zwischen 2.25 (CM34) und 2.60 (CM09) mittlere Itemschwierigkeiten bzw. Itemleichtigkeiten dar.

Die Analyseprogramme (► Abschn. 24.11) bieten die Möglichkeit, sich die modellimplizierte Kovarianzmatrix $\hat{\Sigma}$ ausgeben zu lassen. Die Werte der Varianzen und Kovarianzen lassen sich entsprechend Gln. (24.10) und (24.11) aus den geschätzten Modellparametern berechnen:

- Empirische Kovarianzmatrix S der vier ausgewählten CM-Items (CM09, CM21, CM25, CM34):

$$S = \begin{pmatrix} 1.48 & & & \\ .64 & 1.23 & & \\ .80 & .85 & 1.44 & \\ .61 & .72 & .87 & 1.22 \end{pmatrix}$$

- Modellimplizierte Kovarianzmatrix $\hat{\Sigma}$:

$$\hat{\Sigma} = \begin{pmatrix} 1.48 & & & \\ .64 & 1.23 & & \\ .78 & .86 & 1.44 & \\ .64 & .71 & .87 & 1.22 \end{pmatrix}$$

- Empirischer Mittelwertevektor \bar{y} und modellimplizierter Mittelwertevektor $\hat{\mu}$:

$$\bar{y} = \begin{pmatrix} 2.60 \\ 2.59 \\ 2.44 \\ 2.25 \end{pmatrix}, \hat{\mu} = \begin{pmatrix} 2.60 \\ 2.59 \\ 2.44 \\ 2.25 \end{pmatrix}$$

Deutlich erkennbar ist, dass die empirische Kovarianzmatrix S durch die modellimplizierte Kovarianzmatrix $\hat{\Sigma}$ mit nur geringen Abweichungen rekonstruiert wird. Das Modell passt somit gut zu den Daten.

Für die modellimplizierte Varianz der Indikatorvariablen y_i folgt aus Gl. (24.10) und mit $Var(\eta_j) = 1$:

$$Var(y_i) = \lambda_{ij}^2 + Var(\varepsilon_i), \quad \text{mit } \lambda_{ij}^2 = Var(\tau_i) \quad (24.15)$$

Für die modellimplizierte Kovarianz zwischen zwei Indikatorvariablen y_i und $y_{i'}$ folgt aus Gl. (24.11) und mit $Var(\eta_j) = 1$:

$$Cov(y_i, y_{i'}) = \lambda_{ij} \cdot \lambda_{i'j} \cdot Var(\eta_j) = \lambda_{ij} \cdot \lambda_{i'j} \quad (24.16)$$

- !** Die τ -Kongenerität der Items ist eine Mindestvoraussetzung für die Eindimensionalität der Items. Wenn das τ -kongenerische Modell aufgrund mangelnden Modelfits (► Abschn. 24.6) verworfen werden muss, kann die Annahme der Eindimensionalität der Skala nicht aufrechterhalten werden.

24.3.2 Essentielle τ -Äquivalenz

Modellrestriktionen hinsichtlich λ_{ij}

Eindimensionalität mit der Zusatzannahme, dass sich Messungen zwar in Bezug auf den additiven Parameter α_i und die Fehlervarianzen, nicht jedoch bezüglich der Faktorladungen λ_{ij} unterscheiden, entspricht dem strengeren Modell essentiell τ -äquivalenter Messungen. Indikatorvariablen dürfen in diesem Modell unterschiedliche Mittelwerte und Fehlervarianzen haben, müssen jedoch gleiche Anteile wahrer Varianz aufweisen.

Für das in Abb. 24.1 dargestellte Modell ergibt sich zur Überprüfung der essentiellen τ -Äquivalenz entsprechend die Zusatzhypothese, dass alle Faktorladungen gleich sind: $\lambda_{11} = \lambda_{21} = \lambda_{31} = \lambda_{41} = \lambda_1$. Zur Skalierung des Faktors η_1 wird dessen Varianz auf eins und dessen Erwartungswert auf null fixiert.

Die Anzahl der zu schätzenden Parameter reduziert sich durch die Gleichheitsrestriktion der Faktorladungen auf $t = 9$ Parameter (4 Interzepte + 1 Faktorladung + 4 Fehlervarianzen). Somit hat das Modell fünf Freiheitsgrade ($df = 14 - 9 = 5$).

Das Modell der essentiellen τ -Äquivalenz lässt sich in vier Modellgleichungen ($i = 1, \dots, 4$) der folgenden Form ausdrücken:

$$y_i = \lambda_j \cdot \eta_j + \varepsilon_i \quad (24.17)$$

Mit $Var(\eta_j) = 1$ und $\lambda_{ij} = \lambda_{i'j} = \lambda_j$ folgt aus Gl. (24.10) für die modellimplizierte Varianz der Indikatorvariablen:

$$Var(y_i) = \lambda_j^2 + Var(\varepsilon_i), \quad \text{mit } \lambda_j^2 = Var(\tau_i) = Var(\tau_j) = Var(\tau) \quad (24.18)$$

Indikatorvariablen mit gleicher wahrer Varianz

Aus Gl. (24.18) geht hervor, dass in diesem Modell die Annahme gleicher wahrer Varianzen der Indikatorvariablen geprüft wird. Aufgrund der Gleichheitsrestriktion der Faktorladungen unterscheiden sich die modellimplizierten Varianzen der Indikatorvariablen nur aufgrund ihres jeweiligen Fehleranteils.

Aus Gl. (24.11) folgt für die modellimplizierte Kovarianz zwischen zwei Indikatorvariablen y_i und $y_{i'}$:

$$Cov(y_i, y_{i'}) = \lambda_j^2 \quad (24.19)$$

Die modellimplizierten Kovarianzen zwischen jeweils zwei Indikatorvariablen y_i und $y_{i'}$ sind aufgrund der Gleichheitsrestriktion für alle Paare von Indikatorvariablen identisch. Die Kovarianzen sind somit identisch mit den modellimplizierten wahren Varianzen.

Führt man in diesem Modell zusätzlich eine Gleichheitsrestriktion bezüglich α_i ein, entspricht dies dem strengeren Modell τ -äquivalenter Messungen. Unterschiede der additiven Konstanten spielen jedoch für viele Fragestellungen keine Rolle, da häufig nur die Kovarianz-/Varianzinformationen von Interesse sind, beispielsweise bei der Voraussetzungsprüfung verschiedener Reliabilitätsmaße (► Abschn. 24.3.4). Daher wird häufig das weniger strenge Modell der essentiellen τ -Äquivalenz geprüft.

24.3.3 Essentielle τ -Parallelität

Modellrestriktionen hinsichtlich λ_{ij} und $Var(\varepsilon_i)$

Eindimensionalität mit den Zusatzannahmen gleicher Faktorladungen *und* gleicher Fehlervarianzen für alle Indikatorvariablen wird durch das noch strengere essentiell τ -parallele Messmodell geprüft. Auch hier dürfen sich die Mittelwerte der Indikatorvariablen (d. h. die Interzepte α_i) unterscheiden.

Für das in Abb. 24.1 dargestellte Modell ergeben sich zur Überprüfung der essentiellen τ -Parallelität entsprechend die Zusatzhypthesen, dass alle Faktorladungen und alle Fehlervarianzen gleich sind: $\lambda_{11} = \lambda_{21} = \lambda_{31} = \lambda_{41} = \lambda_1$ und

$\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2) = \text{Var}(\varepsilon_3) = \text{Var}(\varepsilon_4) = \text{Var}(\varepsilon)$. Zur Skalierung des Faktors η_1 wird dessen Varianz auf eins und dessen Erwartungswert auf null fixiert.

Die Anzahl der zu schätzenden Parameter reduziert sich durch die Gleichheitsrestriktionen der Faktorladungen und Fehlervarianzen auf $t = 6$ Parameter (4 Interzepte + 1 Faktorladung + 1 Fehlervarianz). Somit hat das Modell acht Freiheitsgrade ($df = 14 - 6 = 8$).

Das Modell der essentiellen τ -Parallelität lässt sich in vier Messmodellgleichungen ausdrücken, die identisch zu den Messmodellgleichungen des Modells der essentiellen τ -Äquivalenz sind (vgl. Gl. 24.17). Jedoch unterscheiden sich die modellimplizierten Varianzen zwischen diesen Modellen.

Mit $\text{Var}(\eta_j) = 1$, $\lambda_{ij} = \lambda_{i'j} = \lambda_j$ und $\text{Var}(\varepsilon_i) = \text{Var}(\varepsilon_{i'}) = \text{Var}(\varepsilon)$ folgt aus Gl. (24.10) für die modellimplizierte Varianz der Indikatorvariablen y_i :

$$\text{Var}(y_i) = \lambda_j^2 + \text{Var}(\varepsilon) \quad (24.20)$$

Da die Fehlerkovarianzen gleich null sind und keinen Einfluss auf die modellimplizierten Kovarianzen der Indikatorvariablen haben, folgt aus Gl. (24.11), dass die modellimplizierten Kovarianzen der Indikatorvariablen im Modell der essentiellen τ -Parallelität identisch mit den modellimplizierten Kovarianzen im Modell der essentiellen τ -Äquivalenz sind (vgl. Gl. 24.19).

Aufgrund der Gleichheitsrestriktion der Faktorladungen und der Fehlervarianzen unterscheiden sich in diesem Modell nun weder die modellimplizierten Varianzen noch die Kovarianzen der Indikatorvariablen, was einer sehr strengen Modellannahme entspricht.

Führt man in diesem Modell zusätzlich eine Gleichheitsrestriktion bezüglich α_i ein, entspricht dies dem strengeren Modell τ -paralleler Messungen. Unterschiede der additiven Konstanten spielen jedoch für viele Fragestellungen keine Rolle, da häufig nur die Kovarianz-/Varianzinformationen von Interesse sind, beispielsweise bei der Voraussetzungsprüfung verschiedener Reliabilitätsmaße (► Abschn. 24.3.4). Daher wird häufig das weniger strenge Modell der essentiellen τ -Parallelität geprüft.

Zusätzliche Modellrestriktion hinsichtlich α_i

24.3.4 Messäquivalenz und Reliabilität

Bekannte Reliabilitätsmaße wie Cronbachs Alpha (α) oder McDonalds Omega (ω) setzen jeweils unterschiedliche Stufen der Messäquivalenz voraus. So beruht McDonalds Omega auf dem Modell τ -kongenerischer Messungen, das Unterschiede in den Parametern λ_{ij} und $\text{Var}(\varepsilon_i)$ erlaubt; Interzepte sind für die Reliabilitätsschätzungen dagegen irrelevant. Cronbachs Alpha dagegen basiert auf dem strengeren Modell essentiell τ -äquivalenter Messungen, das von gleichen Faktorladungen ausgeht (vgl. Eid und Schmidt 2014, S. 288; ► Kap. 14). Cronbachs Alpha darf somit nur bei essentieller τ -Äquivalenz verwendet werden, andernfalls würden verzerrte Reliabilitätsschätzungen resultieren (vgl. Cho und Kim 2015; Raykov und Marcoulides 2011). Im Unterschied dazu kann McDonalds Omega sowohl bei τ -Kongenerität als auch bei Gültigkeit des strengeren Modells essentiell τ -äquivalenter Messungen als Reliabilitätsmaß verwendet werden.

Die Dimensionalität und Messäquivalenz der Items einer Skala sollten daher vor der Reliabilitätsschätzung überprüft werden. Das Reliabilitätsmaß sollte mindestens der geforderten Stufe der Messäquivalenz entsprechen, wie von Gäde, Schermelleh-Engel und Werner in ► Kap. 14 sowie von Schermelleh-Engel und Gäde in ► Kap. 15 genauer ausgeführt wird.

24.3.5 Fehlerkovarianzen und ihre Beurteilung

Unkorrelierte Messfehler als Voraussetzung für Eindimensionalität

Konstruktirrelevante Fehlerkovarianzen weisen auf Methodeneffekte hin

Konstruktrelevante Fehlerkovarianzen weisen auf weitere Faktoren hin

Durch die Modellierung eines gemeinsamen Faktors wird dessen Anteil aus den Beziehungen der Indikatorvariablen auspartialisiert. Die Partialkorrelationen der Indikatorvariablen sollten somit null sein. Bleiben jedoch Restkorrelationen bestehen (sichtbar in Fehlerkovarianzen), liegen weitere systematische Zusammenhänge zwischen zwei oder mehreren Indikatorvariablen vor, die nicht durch den gemeinsamen Faktor erklärt werden. Um von Eindimensionalität ausgehen zu können, sollten daher keine substantiellen Fehlerkovarianzen zwischen den Indikatorvariablen vorliegen. Die Unkorreliertheit der Fehlervariablen ist somit eine zentrale Voraussetzung für die Eindimensionalität einer Skala.

Liegen empirische Fehlerkovarianzen vor, sollten diese kritisch geprüft werden. Sofern es sich dabei um konstruktirrelevante Methodeneffekte handelt, können die Fehlerkovarianzen als Methodeneffekte ins Modell aufgenommen werden (Näheres zu Methodeneffekten und der Bestimmung der Reliabilität einer Skala mit Fehlerkovarianzen ► Kap. 15).

Substantielle, konstruktrelevante Fehlerkovarianzen würden hingegen eine Modifikation des Modells erforderlich machen, indem für Indikatorvariablen, die mit anderen Indikatorvariablen substantielle Fehlerkovarianzen aufweisen, zusätzliche Faktoren spezifiziert werden müssten. Hierdurch könnte ein eindimensionales Modell zu einem mehrdimensionalen Modell erweitert werden (► Abschn. 24.4).

! Die CFA verliert durch nachträgliche Modifikationen ihren konfirmatorischen Charakter. Bei mehreren Fehlerkovarianzen und unklarem Muster kann unter Umständen die Durchführung einer EFA (► Kap. 23) hilfreich sein, um festzustellen, wie viele Faktoren zur Beschreibung der beobachteten Zusammenhänge notwendig sind. Derart explorativ gefundene faktorielle Strukturen sollten anschließend an einer weiteren unabhängigen Stichprobe wiederum konfirmatorisch überprüft werden.

24.4 Mehrdimensionale Modelle

Zuordnung der Items zu Dimensionen

Viele psychologische Merkmale sind mehrdimensional definiert, z. B. Perfektionismus oder Intelligenz. Die verschiedenen Dimensionen werden meist über Subskalen eines Tests erfasst und anhand der Items der Subskalen wird jeweils ein Subskalenwert pro Person gebildet. In diesem Fall wird davon ausgegangen, dass die einzelnen Items Indikatoren für jeweils nur eine der Dimensionen sind, sodass die Items in einer CFA genau einem Faktor zugeordnet werden können. In anderen Fällen wird zusätzlich über alle Items hinweg ein Gesamttestwert als Maß für das übergeordnete Konstrukt gebildet. In diesem Fall wird zusätzlich davon ausgegangen, dass die einzelnen Items auch jeweils einen gemeinsamen übergeordneten Faktor messen. Für mehrdimensionale Tests kann die Dimensionalität und faktorielle Validität mithilfe der CFA geprüft werden.

Zur Überprüfung, in welcher Beziehung die Dimensionen bzw. Faktoren zueinander stehen, können theoriegeleitete Hypothesen aufgestellt werden. In den meisten Anwendungsfällen der empirischen Sozialforschung wird davon ausgegangen, dass die Faktoren miteinander korrelieren (► Abschn. 24.4.1). Die Höhe der Faktorkorrelationen hängt von der (erwarteten) diskriminanten und konvergenten Validität der Konstrukte ab. Beispielsweise ist Perfektionismus als mehrdimensionales Konstrukt definiert, dessen Dimensionen (u. a. *Doubts about Actions*, *Concern over Mistakes* und *Personal Standards*, ► Abschn. 24.2.7) miteinander korrelieren. Die Höhe der Korrelationen zwischen den Faktoren lässt sich mittels CFA schätzen und ggf. durch übergeordnete Faktoren erklären (► Abschn. 24.4.2).

Hierarchische Modelle wie das Bifaktormodell (► Abschn. 24.4.3) werden zum einen konfirmatorisch für theoretisch von vornherein als mehrdimensional defi-

nierte Merkmale verwendet, zum anderen können sie aber auch exploratorisch wertvolle Informationen zur Theoriebildung liefern: Erweist sich ein als eindimensional konzipiertes Merkmal als mehrdimensional, so muss häufig nicht das gesamte Konstrukt verworfen werden. Vielmehr können hierarchische Modelle hilfreich sein, die Merkmalsstruktur durch Spezifikation eines Generalfaktors und mehrerer Residualfaktoren zu erklären. Hierdurch lässt sich abschätzen, wie viel Varianz der Indikatorvariablen durch den Generalfaktor und wie viel Varianz durch einen spezifischen Residualfaktor erklärt wird.

Die Klärung der faktoriellen Struktur ist dabei nicht nur theoretisch, sondern auch diagnostisch relevant: Die Bildung von *Gesamttestwerten* (pro Person als Summe über die Messwerte aller Items eines Tests) bzw. *Subskalenwerten* (pro Person als Summe über die Messwerte aller Items einer Subskala) und die anschließende Reliabilitätsschätzung der Test- bzw. Subskalenwerte sollten auf einer validen Zuordnung der Items zu Faktoren basieren (zur Reliabilitätsschätzung für mehrdimensionale Modelle ► Kap. 15). Laden alle Itemvariablen im Wesentlichen auf dem Generalfaktor, so ist die Bildung eines Testwertes über alle Items gerechtfertigt. Werden die Messwerte der Items dagegen im Wesentlichen durch ihren jeweiligen spezifischen Faktor erklärt, so ist die Bildung von Subskalenwerten angebracht, nicht jedoch die Bildung eines Testwertes für den Gesamttest.

Analog zu den eindimensionalen Modellen besteht auch bei den mehrdimensionalen Modellen das Ziel, die Varianzen und Kovarianzen bestmöglich durch die geschätzten Modellparameter zur reproduzieren. Im ► Abschn. 24.4.4 werden die Varianzen und Kovarianzen als Funktionen der geschätzten Modellparameter für die hier vorgestellten mehrdimensionalen Modelle vergleichend dargestellt, um die Unterschiede der Modelle zu verdeutlichen.

Für die Darstellung der mehrdimensionalen Modelle soll nachfolgend nur die Analyse der Kovarianzstruktur betrachtet werden, während die Mittelwertestruktur und damit die Interzepte nicht weiter berücksichtigt werden. Die Analyse der Mittelwertestruktur ist bei Bedarf aber auch für mehrdimensionale Modelle möglich. Die nachfolgenden Gleichungen lassen sich analog zu den eindimensionalen Modellen um die Interzepte erweitern.

- !** Die hier vorgestellten Modelle können zu komplexeren Modellen erweitert werden. So ist zum Nachweis der Konstruktvalidität eine Kombination von mehreren Merkmalen, gemessen mit mehreren Messmethoden, möglich (vgl. ► Kap. 25) und die Stabilität und Veränderlichkeit von Merkmalen kann systematisch untersucht werden (vgl. ► Kap. 26). Durch eine Erweiterung einer CFA auf ein vollständiges Strukturgleichungsmodell (SEM) können weitere Fragen der Validität eines Tests untersucht werden, indem auch gerichtete Zusammenhänge zwischen mehreren Konstrukten auf latenter Ebene analysiert werden (vgl. Werner et al. 2016).

24.4.1 Modell mit korrelierten Faktoren

In ► Beispiel 24.2 wird ein Modell mit korrelierten Faktoren gezeigt. Die Items y_1 bis y_3 messen Faktor η_1 (PS), die Items y_4 bis y_6 messen Faktor η_2 (DA) und die Items y_7 bis y_{10} messen Faktor η_3 (CM). Die Korrelationen zwischen den Faktoren sind als gebogene Doppelpfeile dargestellt. In diesem Modell ist jedes Item genau einem Faktor zugeordnet (*Einfachstruktur*), während die Faktorladungen auf den jeweils anderen Faktoren auf null fixiert sind. Derartige Restriktionen werden in der CFA als Hypothesen explizit getestet. Dies verdeutlicht einen wesentlichen Unterschied zur EFA, bei der für jedes Item Ladungen auf allen Faktoren zugelassen sind.

In Analogie zum eindimensionalen Modell setzt sich jede Indikatorvariable y_i zusammen aus einem systematischen (wahren) Anteil, d. h. dem mit der Faktorladung λ_{ij} gewichteten Einfluss des Faktors η_j , und einem Fehleranteil ε_i . Die Hypo-

Informationsgehalt hierarchischer Modelle

Bildung von Gesamttestwerten und Subskalenwerten in hierarchischen Modellen

Eindeutige Zuordnung der Items zu den Faktoren

Messmodellgleichungen

thesen des Modells lassen sich in zehn Modellgleichungen entsprechend Gl. (24.1) ausdrücken. Die Modellgleichungen für das eindimensionale Modell lassen sich somit auf das Modell mit korrelierten Faktoren übertragen (► Beispiel 24.2).

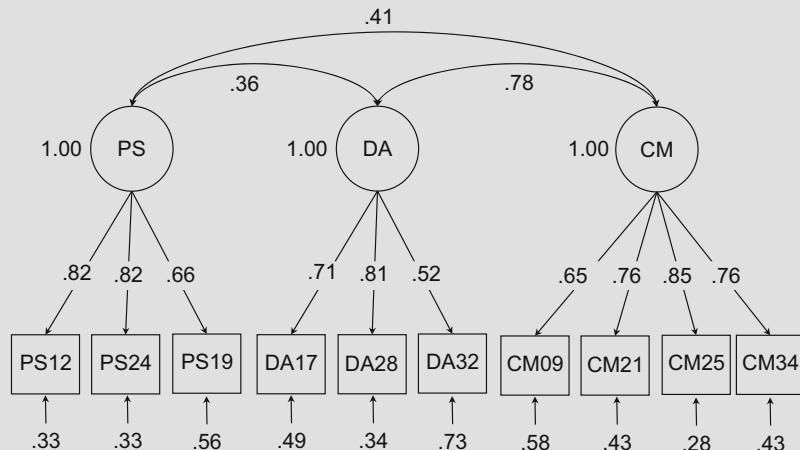
In ► Abschn. 24.4.4 werden die Varianzen und Kovarianzen als Funktionen der geschätzten Modellparameter dargestellt.

Beispiel 24.2: Modell mit korrelierten Faktoren

Für die zehn ausgewählten Items der MPS-F (► Abschn. 24.2.7) wird exemplarisch ein Modell mit drei korrelierten Faktoren, die den Dimensionen der Subskalen PS, DA und CM entsprechen, überprüft.

Die empirische Kovarianzmatrix \mathbf{S} der zehn Items besteht aus 55 Elementen ($s = 10(10 + 1)/2$; vgl. Gl. 24.13). Zur Skalierung der Faktoren η_1 , η_2 und η_3 werden die Varianzen der Faktoren auf eins fixiert. Zur Analyse der Kovarianzstruktur sind in diesem Modell $t = 23$ Parameter zu schätzen (10 Faktorladungen + 10 Fehlervarianzen + 3 Kovarianzen der Faktoren). Das Modell hat damit 32 Freiheitsgrade ($df = 55 - 23$).

Zur Parameterschätzung wird die MLR-Methode (► Abschn. 24.5.3) verwendet, die für kategoriale Variablen mit mindestens fünf Kategorien häufig gut geeignet ist (Rhemtulla et al. 2012). Die Ergebnisse der Parameterschätzung sind im folgenden mehrdimensionalen Modell mit drei korrelierten Faktoren für zehn ausgewählte Items der MPS-F dargestellt (komplett standardisierte Lösung):



Das Modell zeigt eine sehr gute Passung (► Abschn. 24.6). Der robuste χ^2 -Wert ist nicht signifikant ($\chi^2(32) = 39.80, p = .16$). Auch die deskriptiven Gütekriterien weisen auf einen sehr guten Modellfit hin (RMSEA = .03, 90 %-Konfidenzintervall [.00; .06], CFI = .99, SRMR = .04), sodass die Parameter interpretiert werden dürfen. Die standardisierten Faktorladungen sind ausreichend hoch und liegen zwischen .52 (DA32) und .85 (CM25). Die Faktoren korrelieren signifikant miteinander mit Werten zwischen .36 (PS, DA) und .78 (CM, DA).

24.4.2 Faktormodell höherer Ordnung

In einem Modell mit korrelierten Faktoren werden neben ungerichteten Beziehungen zwischen den Faktoren keine weiteren Hypothesen formuliert, um die Zusammenhänge der Faktoren näher zu beschreiben. Mitunter können jedoch theoretische Begründungen dafür vorliegen, dass die korrelativen Zusammenhänge zwischen

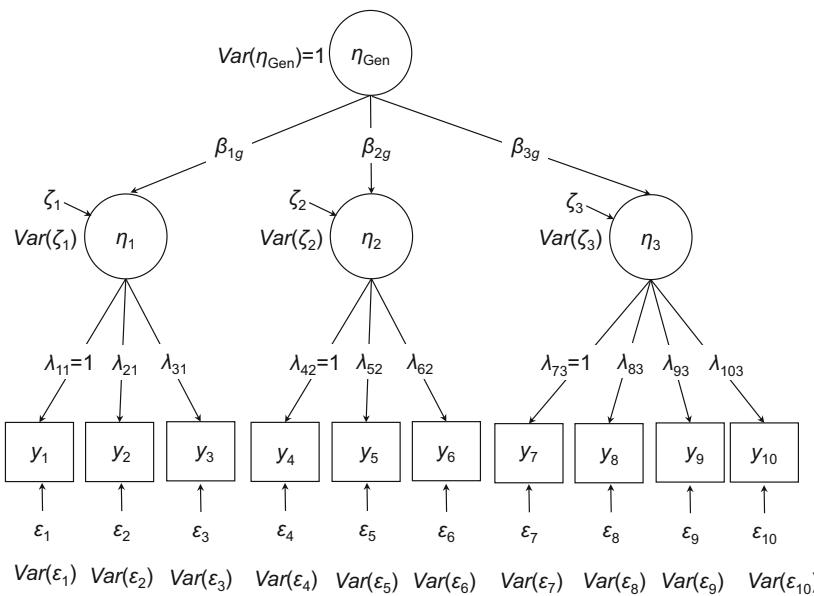


Abb. 24.3 Faktormodell höherer Ordnung mit einem Generalfaktor η_{Gen} und drei Faktoren erster Ordnung η_1 , η_2 , und η_3 (komplett standardisierte Lösung)

Faktoren wiederum durch übergeordnete Merkmale erklärt werden können. So ließe sich die Hypothese aufstellen, dass das breitere, allgemeinere Merkmal *Perfectionismus* die Korrelationen zwischen den enger definierten, spezifischen Merkmalen *Personal Standards* (PS), *Doubts about Actions* (DA) und *Concern over Mistakes* (CM) erklärt. Eine solche hierarchische Modellannahme eines übergeordneten Generalfaktors zeigt Abb. 24.3. In diesem Faktormodell höherer Ordnung sind die Items wieder je einem Faktor erster Ordnung (η_1 , η_2 oder η_3) zugeordnet. Zusätzlich wirkt der Generalfaktor η_{Gen} als übergeordneter Faktor auf die Faktoren erster Ordnung.

Die Strukturkoeffizienten β_{jg} geben den Effekt des Generalfaktors auf die Faktoren erster Ordnung an, wobei der Index g für den Generalfaktor steht, mit dem die Faktoren erster Ordnung η_j in Beziehung stehen. Index j ($j = 1, \dots, q$) steht für die Faktoren erster Ordnung. Die Residuen ζ_j entsprechen dem nicht durch den Generalfaktor erklärten Anteil der Faktoren erster Ordnung. In dem Modell wird angenommen, dass die Residuen ζ_j unkorreliert sind.

In Analogie zum eindimensionalen Modell und zum Modell korrelierter Faktoren setzt sich jede Indikatorvariable y_i zusammen aus einem systematischen (wahren) Anteil, d. h. dem mit der Faktorladung λ_{ij} gewichteten Einfluss des Faktors η_j , und einem Fehleranteil ε_i . Die Hypothesen des Modells lassen sich in zehn Modellgleichungen entsprechend Gl. (24.1) ausdrücken. Die Modellgleichungen für das eindimensionale Modell lassen sich somit ebenfalls auf das Faktormodell höherer Ordnung übertragen.

Die Darstellung der Beziehungen zwischen den Faktoren wird in SEM (Structural Equation Modeling) allgemein als *Strukturmodell* bezeichnet. Um das Strukturmodell der Beziehungen zwischen dem Generalfaktor und den Faktoren erster Ordnung formal zu beschreiben, werden die Hypothesen des Modells durch drei Strukturmodellgleichungen für die Faktoren erster Ordnung ergänzt. Die Faktoren erster Ordnung setzen sich zusammen aus dem mit dem Strukturkoeffizienten β_{jg} gewichteten Einfluss des Generalfaktors η_{Gen} und einem Residuum ζ_j :

$$\eta_j = \beta_{jg} \cdot \eta_{\text{Gen}} + \zeta_j \quad (24.21)$$

Generalfaktor erklärt Kovarianzen der Faktoren erster Ordnung

Messmodellgleichungen

Strukturmodellgleichung

Residualvarianz auf Ebene der Faktoren erster Ordnung

Dabei sollen die Kovarianzen der Faktoren erster Ordnung durch den Generalfaktor komplett erklärt werden, sodass ihre Partialkorrelationen null sind (dies entspricht unkorrelierten Residualvariablen ζ_j). Da die Varianzen der Faktoren erster Ordnung im Allgemeinen nicht vollständig durch den Generalfaktor erklärt werden, wird nun auch auf Ebene der Faktoren unterschieden zwischen einem Varianzanteil, der durch den Generalfaktor erklärt wird, und einem nicht erklären Varianzanteil (subskalenspezifische Varianz, Residualvarianz $Var(\zeta_j)$).

Zur Erklärung der Varianz einer Indikatorvariablen lässt sich in diesem Modell der Anteil, der auf den Generalfaktor (η_{Gen}) zurückgeht, von dem Anteil, der auf den spezifischen Faktor erster Ordnung (ζ_j) zurückgeht, unterscheiden und anhand der Modellparameter ermitteln. Der Generalfaktor wirkt dabei indirekt, also über die Faktoren erster Ordnung vermittelt, auf die Indikatorvariablen. Faktormodelle höherer Ordnung werden daher gelegentlich auch als indirekt hierarchische Modelle bezeichnet.

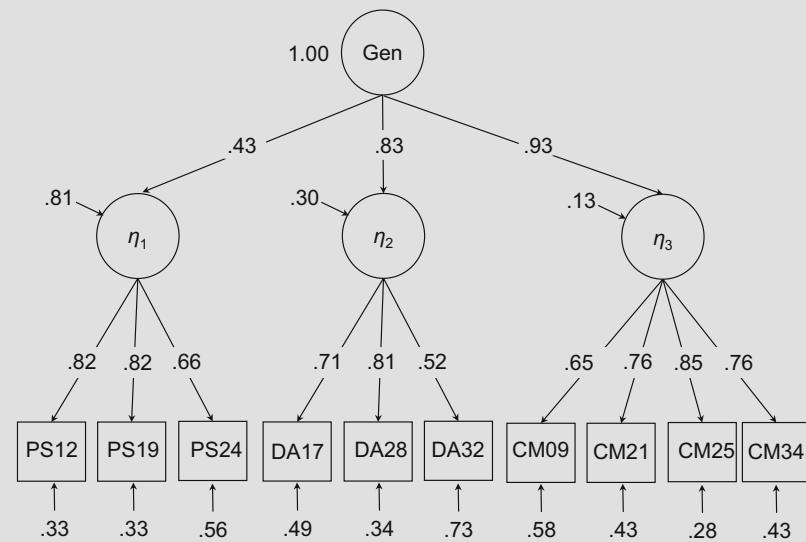
In die Berechnung des indirekten Effekts des Generalfaktors auf eine Indikatorvariable geht der Pfadkoeffizient β_{jg} mit ein. Da dieser Anteil für alle Indikatorvariablen eines Faktors erster Ordnung identisch ist, liegt hier eine Gleichheitsrestriktion (*equality constraint*) vor. Die Stärke des Zusammenhangs zwischen Generalfaktor und Indikatorvariable ergibt sich – für standardisierte Variablen und Faktoren – aus dem jeweiligen Produkt $\beta_{jg} \cdot \lambda_{ij}$ (► Beispiel 24.3).

Beispiel 24.3: Faktormodell höherer Ordnung

Für die zehn ausgewählten Items der MPS-F (► Abschn. 24.2.7) wird exemplarisch ein Faktormodell höherer Ordnung mit einem Faktor höherer Ordnung und drei Faktoren erster Ordnung überprüft.

Die empirische Kovarianzmatrix S der zehn Items besteht aus 55 Elementen ($s = 10 \cdot (10 + 1)/2$; vgl. Gl. 24.13). Zur Skalierung der Faktoren erster Ordnung η_j wird jeweils die Faktorladung eines Indikators auf eins fixiert (► Abb. 24.3). Zur Skalierung des Generalfaktors wird dessen Varianz auf eins fixiert. Zur Analyse der Kovarianzstruktur sind in diesem Modell $t = 23$ Parameter zu schätzen (7 Faktorladungen, 10 Fehlervarianzen, 3 Strukturkoeffizienten, 3 Residualvarianzen). Das Modell hat damit 32 Freiheitsgrade ($df = 55 - 23$).

Die Ergebnisse der MLR-Schätzung (► Abschn. 24.5.3) sind in folgendem Faktormodell höherer Ordnung für zehn ausgewählte Items der MPS-F dargestellt (komplett standardisierte Lösung):



Das Modell zeigt eine sehr gute Passung (► Abschn. 24.6). Der robuste χ^2 -Wert ist nicht signifikant ($\chi^2(32) = 39.80, p = .16$). Auch die deskriptiven Gütekriterien weisen auf einen sehr guten Modellfit hin (RMSEA = .03, 90 %-Konfidenzintervall [.00; .06], CFI = .99, SRMR = .04). Die Faktorladungen sind hoch und liegen standardisiert zwischen .52 (DA32) und .85 (CM25). Der Modellfit des Faktormodells höherer Ordnung ist identisch zum Fit des Modells korrelierter Faktoren. Die beiden Modelle sind äquivalent, da das Strukturmodell mit drei Faktoren erster Ordnung als Indikatoren des Faktors höherer Ordnung genau identifiziert ist.

Die Faktoren CM und DA werden zu einem großen Anteil durch den gemeinsamen Generalfaktor erklärt. So wird die Varianz des Faktors CM zu 87 % durch den Generalfaktor erklärt ($\beta_{CMg}^2 \cdot Var(\eta_{Gen}) = .93^2 \cdot 1 = .87$). Der nicht erklärte Varianzanteil beträgt entsprechend 13 % ($Var(\zeta_{CM}) = .13$). Der Faktor PS weist dagegen einen schwächeren Zusammenhang mit dem Generalfaktor auf ($\beta_{PSg} = .43$). Der Generalfaktor erklärt damit nur 19 % der Varianz von PS, während 81 % der Varianz spezifisch für die Subskala sind ($Var(\zeta_{PS}) = .81$).

In ► Abschn. 24.4.4 werden die Varianzen und Kovarianzen als Funktionen der geschätzten Modellparameter dargestellt.

24.4.3 Bifaktormodell

Ein alternatives hierarchisches Modell ohne die Gleichheitsrestriktionen des Pfadkoeffizienten β_{jg} für die Indikatoren eines Faktors (vgl. ► Abschn. 24.4.2) stellt das Bifaktormodell dar (Eid et al. 2017a; Holzinger und Swineford 1937). Im Bifaktormodell wird ebenfalls ein Generalfaktor modelliert, der jedoch auf derselben Ebene mit den spezifischen Faktoren steht. Da der Generalfaktor direkt auf die Indikatoren wirkt, werden Bifaktormodelle auch als direkt hierarchische Modelle bezeichnet. Im klassischen orthogonalen Bifaktormodell sind die spezifischen Faktoren untereinander sowie mit dem Generalfaktor unkorreliert.

Bezogen auf das empirische Beispiel zum Perfektionismus (► Abschn. 24.2.7) kann geprüft werden, ob alle Items einen Generalfaktor *Perfektionismus* messen und ob die Items, die jeweils einer Subskala zugeordnet sind, zusätzlich noch einen spezifischen Residualfaktor messen, der unabhängig vom Generalfaktor ist. In dem Fall gäbe es so viele spezifische Residualfaktoren, wie es Subskalen gibt.

■ Abb. 24.4 zeigt ein Bifaktormodell mit einem Generalfaktor und drei subskalenspezifischen Faktoren. Alle Items y_1 bis y_{10} messen den Generalfaktor η_{Gen} . Zusätzlich messen die Items y_1 bis y_3 den spezifischen Residualfaktor ζ_1 , die Items y_4 bis y_6 den spezifischen Residualfaktor ζ_2 und die Items y_7 bis y_{10} den spezifischen Residualfaktor ζ_3 . In diesem Modell ist jedes Item dem Generalfaktor und genau einem spezifischen Residualfaktor zugeordnet; die Faktorladungen auf alle anderen spezifischen Faktoren sind auf null fixiert.

Jede Indikatorvariable y_i setzt sich folglich zusammen aus dem mit der Faktorladung λ_{ig} gewichteten Einfluss des Generalfaktors η_{Gen} , dem mit der Faktorladung λ_{ij} gewichteten Einfluss eines spezifischen Faktors ζ_j und einem Fehleranteil ε_i . Die Messmodellgleichung für y_i lautet:

$$y_i = \lambda_{ig} \cdot \eta_{Gen} + \lambda_{ij} \cdot \zeta_j + \varepsilon_i \quad (24.22)$$

Zu beachten ist, dass sich in Bifaktormodellen die inhaltliche Interpretation der spezifischen Faktoren ζ_j , auf denen die Indikatorvariablen einer Subskala laden, im Vergleich zu den Faktoren η_j der bisher behandelten Modelle ändert. Sie stellen

Direkt hierarchisches Modell

Zuordnung der Items zum Generalfaktor und einem Residualfaktor

Messmodellgleichung des Bifaktormodells

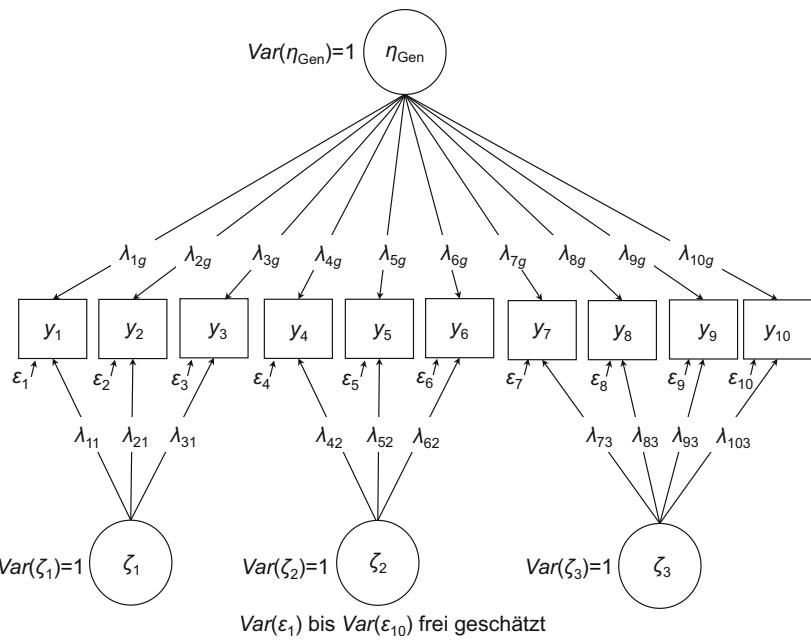


Abb. 24.4 Bifaktormodell mit einem Generalfaktor η_{Gen} und drei spezifischen Residualfaktoren ζ_1 , ζ_2 , und ζ_3

im Bifaktormodell Residualfaktoren dar, aus denen der Anteil des Generalfaktors auspartialisiert wurde. Deshalb werden sie als spezifische Faktoren oder Residualfaktoren bezeichnet und sind in etwa vergleichbar mit den spezifischen Faktoren ζ_j der Modelle höherer Ordnung (► Abschn. 24.4.2).

Da die Indikatorvariablen im Vergleich zu den bisher behandelten Modellen hier jeweils auf zwei Faktoren laden, lassen sich auch zwei Quellen der Varianz unterscheiden: Der Anteil erklärter Varianz, der auf den Generalfaktor zurückgeht, und der Anteil erklärter Varianz, der auf den Residualfaktor zurückgeht. Die erklärten Varianzanteile können anhand der Modellparameter ermittelt werden (► Beispiel 24.4).

Im Vergleich zum Faktormodell höherer Ordnung liegt hier keine Gleichheitsrestriktion (► Abschn. 24.4.2) vor. Da der Generalfaktor auf derselben Ebene modelliert wird wie die Residualfaktoren, wird auch der Zusammenhang zwischen Generalfaktor und jeder Indikatorvariablen als direkter Effekt geschätzt.

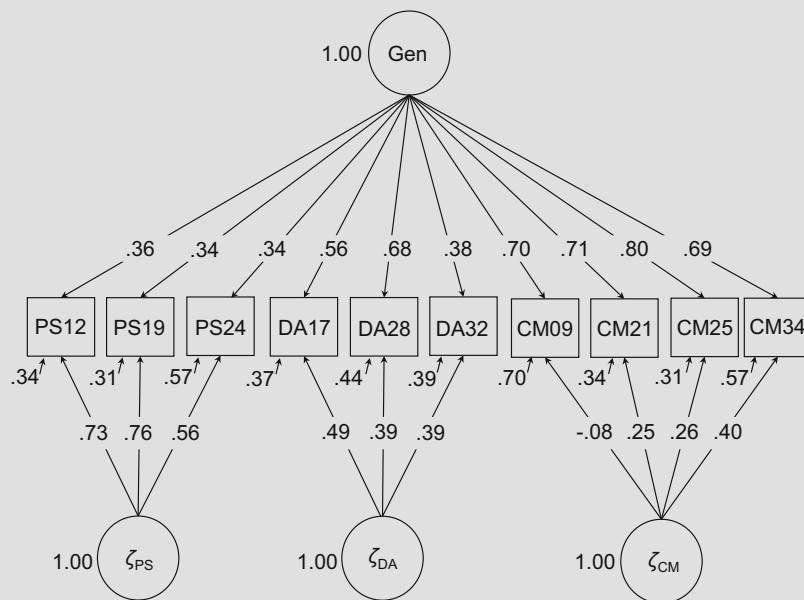
In ► Abschn. 24.4.4 werden die Varianzen und Kovarianzen als Funktionen der geschätzten Modellparameter dargestellt.

Beispiel 24.4: Bifaktormodell

Für die zehn ausgewählten Items der MPS-F (► Abschn. 24.2.7) wird exemplarisch ein Bifaktormodell mit einem Generalfaktor und drei spezifischen Faktoren, die den Dimensionen der Subskalen PS, DA und CM entsprechen, überprüft.

Die empirische Kovarianzmatrix S der zehn Items besteht aus 55 Elementen ($s = 10 \cdot (10 + 1)/2$; vgl. Gl. 24.13). Zur Skalierung der Faktoren η_{Gen} und ζ_j werden die Varianzen der Faktoren jeweils auf eins fixiert. Zur Analyse der Kovarianzstruktur sind in diesem Modell $t = 30$ Parameter zu schätzen (10 Faktorladungen auf dem Generalfaktor + 10 Faktorladungen auf den spezifischen Faktoren + 10 Fehlervarianzen). Das Modell hat damit 25 Freiheitsgrade ($df = 55 - 30$).

Die Ergebnisse der MLR-Schätzung (► Abschn. 24.5.3) sind im folgenden Bifaktormodell für zehn ausgewählte Items der MPS-F dargestellt (komplett standardisierte Lösung):



Das Modell zeigt eine sehr gute Passung (► Abschn. 24.6). Der robuste χ^2 -Wert ist nicht signifikant ($\chi^2(25) = 32.44, p = .15$). Auch die deskriptiven Gütekriterien weisen auf einen sehr guten Modellfit hin ($\text{RMSEA} = .03$, 90 %-Konfidenzintervall [.00; .07], $\text{CFI} = .99$, $\text{SRMR} = .03$).

Die standardisierten Faktorladungen auf dem Generalfaktor liegen für die CM-Items zwischen .69 und .80, für die DA-Items zwischen .38 und .68 und für die PS-Items zwischen .34 und .36.

Die Faktorladungen der PS-Items auf dem spezifischen PS-Faktor sind hoch und liegen zwischen .56 und .76. Diejenigen der DA-Items auf dem DA-Residualfaktor sind dagegen nur moderat, aber statistisch signifikant mit Werten zwischen .39 und .49. Die Faktorladungen der CM-Items auf dem CM-Residualfaktor sind jedoch nicht signifikant. Dies stellt ein in Bifaktormodellen häufig auftretendes Phänomen dar (vgl. Eid et al., 2017a; Geiser et al. 2015). Durch den Generalfaktor können ein oder mehrere spezifische Faktoren überflüssig werden, da ein Großteil der Indikatorvarianz bereits durch den Generalfaktor erklärt wird. Empfohlen wird hier der $(S - 1)$ -Ansatz, bei dem ein spezifischer Residualfaktor S weniger modelliert wird, als Subskalen vorhanden sind. Die zu diesem Faktor gehörigen Indikatorvariablen laden somit nur noch auf dem Generalfaktor (Eid et al., 2017a). Für das Beispiel bedeutet dies, dass die Ladungen der CM-Items auf dem Residualfaktor CM auf null fixiert werden könnten.

Insgesamt zeigen die Ergebnisse, dass die empirischen Zusammenhänge der Indikatorvariablen durch das Bifaktormodell gut reproduziert werden können. Jedoch verdeutlichen die Faktorladungen, dass der Generalfaktor nur einen geringen Anteil der Indikatorvarianz der Skala PS, jedoch einen bedeutenden Teil der Indikatorvarianz der Skala DA und fast den gesamten Anteil der Indikatorvarianz der Skala CM erklärt. Die PS-Items bilden demnach einen relativ eigenständigen Faktor. Daraus sollte hier nicht unkritisch ein Generalfaktor für alle Items angenommen werden. Diese Interpretation wird auch durch die Schätzung der mehrdimensionalen Omega-Koeffizienten gestützt (► Kap. 15).

24.4.4 Elemente der modellimplizierten Kovarianzmatrix in mehrdimensionalen Modellen

Die folgende Gegenüberstellung verdeutlicht, dass sich in Abhängigkeit von der zugrunde liegenden Modellstruktur die modellimplizierten Varianzen und Kovarianzen ändern. Das Modell, dessen modellimplizierte Kovarianzmatrix die empirische Kovarianzmatrix am besten reproduziert, zeigt den besten Modellfit.

■■ Modell korrelierter Faktoren

Die Varianzaufteilung einer Indikatorvariablen in einen Anteil erklärter Varianz und einen Anteil Fehlervarianz erfolgt analog zum eindimensionalen Modell τ -kongenerischer Messungen (vgl. Gl. 24.10). Auch die Kovarianz zweier Indikatorvariablen y_i und $y_{i'}$, die auf demselben Faktor η_j laden, ist identisch wie im eindimensionalen Modell τ -kongenerischer Messungen (vgl. Gl. 24.11). Der relevante Unterschied zeigt sich darin, dass zur Erklärung der Kovarianz zweier Indikatorvariablen y_i und y_k , die auf verschiedenen Faktoren η_j und $\eta_{j'}$ laden, die Kovarianz der beteiligten Faktoren berücksichtigt wird:

$$\text{Cov}(y_i, y_k) = \lambda_{ij} \cdot \lambda_{kj'} \cdot \text{Cov}(\eta_j, \eta_{j'}) \quad (24.23)$$

■■ Modell höherer Ordnung

Die erklärte Varianz einer Indikatorvariablen y_i lässt sich im Modell höherer Ordnung weiter differenzieren in einen durch den Generalfaktor erklärten Varianzanteil und einen durch den Faktor erster Ordnung erklärten Varianzanteil, die sich durch Einsetzen von Gl. (24.21) in Gl. (24.10) anhand der Modellparameter ausdrücken lassen:

$$\begin{aligned} \text{Var}(y_i) &= \lambda_{ij}^2 \cdot \text{Var}(\eta_j) + \text{Var}(\varepsilon_i) = \lambda_{ij}^2 \cdot \text{Var}(\beta_{jg} \cdot \eta_{\text{Gen}} + \xi_j) + \text{Var}(\varepsilon_i) \\ &= \underbrace{\lambda_{ij}^2 \cdot \beta_{jg}^2 \cdot \text{Var}(\eta_{\text{Gen}})}_{\substack{\text{durch den Generalfaktor} \\ \text{erklärte Varianz}}} + \underbrace{\lambda_{ij}^2 \cdot \text{Var}(\xi_j)}_{\substack{\text{durch den Residualfaktor} \\ \text{1. Ordnung erklärte Varianz}}} + \underbrace{\text{Var}(\varepsilon_i)}_{\text{Fehlervarianz}} \end{aligned} \quad (24.24)$$

Auch bei der Kovarianz zwischen zwei Variablen y_i und $y_{i'}$, die auf demselben Faktor η_j laden, wird zwischen den Anteilen differenziert, die jeweils auf den Generalfaktor und den Residualfaktor zurückgehen. So folgt durch Einsetzen von Gl. (24.21) in Gl. (24.11) wobei alle Kovarianzen mit Fehlervariablen null sind:

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \text{Cov}(\lambda_{ij} \cdot \eta_j + \varepsilon_i, \lambda_{i'j} \cdot \eta_{j'} + \varepsilon_{i'}) \\ &= \text{Cov}(\lambda_{ij} \cdot [\beta_{jg} \cdot \eta_{\text{Gen}} + \xi_j] + \varepsilon_i, \lambda_{i'j} \cdot [\beta_{jg} \cdot \eta_{\text{Gen}} + \xi_{j'}] + \varepsilon_{i'}) \\ &= \underbrace{\lambda_{ij} \cdot \lambda_{i'j} \cdot \beta_{jg}^2 \cdot \text{Var}(\eta_{\text{Gen}})}_{\substack{\text{durch den Generalfaktor} \\ \text{erklärte Kovarianz}}} + \underbrace{\lambda_{ij} \cdot \lambda_{i'j} \cdot \text{Var}(\xi_j)}_{\substack{\text{durch den Residualfaktor} \\ \text{1. Ordnung erklärte Kovarianz}}} \end{aligned} \quad (24.25)$$

Die Kovarianz zwischen zwei Variablen y_i und y_k , die auf verschiedenen spezifischen Faktoren η_j und $\eta_{j'}$ laden, wird allein durch den Generalfaktor erklärt, da die Residualfaktoren unabhängig voneinander sind ($\text{Cov}(\xi_j, \xi_{j'} = 0)$). Durch Einsetzen von Gl. (24.21) in Gl. (24.11) folgt:

$$\begin{aligned} \text{Cov}(y_i, y_k) &= \text{Cov}(\lambda_{ij} \cdot \eta_j + \varepsilon_i, \lambda_{kj'} \cdot \eta_{j'} + \varepsilon_k) \\ &= \text{Cov}(\lambda_{ij} \cdot [\beta_{jg} \cdot \eta_{\text{Gen}} + \xi_j] + \varepsilon_i, \lambda_{kj'} \cdot [\beta_{j'g} \cdot \eta_{\text{Gen}} + \xi_{j'}] + \varepsilon_k) \\ &= \lambda_{ij} \cdot \lambda_{kj'} \cdot \beta_{jg} \cdot \beta_{j'g} \cdot \text{Var}(\eta_{\text{Gen}}) + \lambda_{ij} \cdot \lambda_{kj'} \cdot \text{Cov}(\xi_j, \xi_{j'}) \\ &= \underbrace{\lambda_{ij} \cdot \lambda_{kj'} \cdot \beta_{jg} \cdot \beta_{j'g} \cdot \text{Var}(\eta_{\text{Gen}})}_{\text{durch den Generalfaktor erklärte Kovarianz}} \end{aligned} \quad (24.26)$$

■■ Bifaktormodell

Die erklärte Varianz einer Indikatorvariablen y_i lässt sich im Bifaktormodell weiter differenzieren in einen durch den Generalfaktor erklärten Varianzanteil und einen durch den Residualfaktor erklärten Varianzanteil; beide Varianzanteile lassen sich durch Einsetzen von Gl. (24.22) in Gl. (24.10) anhand der Modellparameter ausdrücken:

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\lambda_{ig} \cdot \eta_{\text{Gen}} + \lambda_{ij} \cdot \xi_j + \varepsilon_i) \\ &= \underbrace{\lambda_{ig}^2 \cdot \text{Var}(\eta_{\text{Gen}})}_{\substack{\text{durch den Generalfaktor} \\ \text{erklärte Varianz}}} + \underbrace{\lambda_{ij}^2 \cdot \text{Var}(\xi_j)}_{\substack{\text{durch den Residualfaktor} \\ \text{erklärte Varianz}}} + \underbrace{\text{Var}(\varepsilon_i)}_{\text{Fehlervarianz}} \end{aligned} \quad (24.27)$$

Die Kovarianz zweier Indikatorvariablen y_i und $y_{i'}$, die auf demselben spezifischen Faktor laden, berechnet sich als Summe der Varianzanteile, die jeweils auf den Generalfaktor η_{Gen} und den spezifischen Faktor ξ_j zurückzuführen sind. Durch Einsetzen von Gl. (24.22) in Gl. (24.11) folgt (ohne Darstellung der Zwischenschritte):

$$\begin{aligned} \text{Cov}(y_i, y_{i'}) &= \underbrace{\lambda_{ig} \cdot \lambda_{i'g} \cdot \text{Var}(\eta_{\text{Gen}})}_{\substack{\text{durch den Generalfaktor} \\ \text{erklärte Kovarianz}}} + \underbrace{\lambda_{ij} \cdot \lambda_{i'j} \cdot \text{Var}(\xi_j)}_{\substack{\text{durch den Residualfaktor} \\ \text{erklärte Kovarianz}}} \end{aligned} \quad (24.28)$$

Die Kovarianz zwischen zwei Indikatorvariablen y_i und y_k , die auf verschiedenen spezifischen Faktoren ξ_j und $\xi_{j'}$ laden, die voneinander und vom Generalfaktor unabhängig sind, wird allein durch den Generalfaktor erklärt. Durch Einsetzen von Gl. (24.22) in Gl. (24.11) folgt (ohne Darstellung der Zwischenschritte):

$$\text{Cov}(y_i, y_k) = \underbrace{\lambda_{ig} \cdot \lambda_{kg} \cdot \text{Var}(\eta_{\text{Gen}})}_{\substack{\text{durch den Generalfaktor} \\ \text{erklärte Kovarianz}}} \quad (24.29)$$

24.5 Parameterschätzung

Im Rahmen einer CFA wird die Nullhypothese $H_0: \Sigma = \Sigma(\theta)$ getestet, die von der Gleichheit der Populationskovarianzmatrix Σ der Indikatorvariablen mit der modellimplizierten Kovarianzmatrix $\Sigma(\theta)$ ausgeht (θ bezeichnet den Vektor der Modellparameter). Die Nullhypothese lässt sich auch so formulieren, dass die modellimplizierte Kovarianzmatrix, die sich aus den geschätzten Modellparametern des Parametervektors θ ergibt, mit der Populationskovarianzmatrix Σ der Indikatorvariablen identisch ist (vgl. Browne und Cudeck 1992). Da die Populationswerte nicht bekannt sind, werden sie auf Grundlage der Stichprobendaten geschätzt: Die Kovarianzmatrix Σ wird durch die empirische Kovarianzmatrix S geschätzt, und $\Sigma(\theta)$ durch die modellimplizierte Kovarianzmatrix $\hat{\Sigma}(\hat{\theta})$, die auf dem geschätzten Parametervektor $\hat{\theta}$ beruht und vereinfachend mit $\hat{\Sigma}$ bezeichnet wird.

Die Elemente der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$ ergeben sich aus den geschätzten Modellparametern. Das Ziel der Schätzung besteht darin, die Modellparameter so zu schätzen, dass die Differenzen zwischen den Elementen der empirischen Kovarianzmatrix S und der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$ möglichst gering ausfallen. Dieses Ziel haben alle Schätzverfahren gemeinsam.

Zur Schätzung der optimalen Modellparameter werden die Abweichungen zwischen den Elementen der empirischen Kovarianzmatrix S und der modellimplizierten Kovarianzmatrix $\hat{\Sigma}$ bei der Parameterschätzung anhand einer Diskrepanzfunktion (Fit-Funktion) minimiert.

Nullhypothese der CFA

Diskrepanzfunktion

Der Vergleich zwischen empirischer und modellimplizierter Kovarianzmatrix stellt außerdem die Grundlage zur Beurteilung dar, wie gut das Modell die empirischen Zusammenhänge erklärt (► Abschn. 24.6).

Für die CFA stehen verschiedene Schätzmethoden zur Verfügung, die sich hinsichtlich der zugrunde liegenden Diskrepanzfunktion sowie ihrer Voraussetzungen bezüglich des Skalenniveaus und der Verteilungseigenschaften der Daten unterscheiden. Unter praktischen Gesichtspunkten lassen sich insbesondere drei Gruppen unterscheiden:

1. Verfahren für multivariat normalverteilte, kontinuierliche Indikatorvariablen
2. Verfahren für nicht normalverteilte, kontinuierliche Indikatorvariablen
3. Verfahren für kategoriale Indikatorvariablen

Obwohl im Prinzip eine Kovarianzmatrix (und bei Bedarf ein Mittelwertevektor) die Datengrundlage der CFA bildet, werden bei den meisten modernen Schätzverfahren Rohdaten verwendet. Dies ist beispielsweise notwendig, wenn fehlende Werte berücksichtigt oder robuste Verfahren verwendet werden sollen. In diesen Fällen werden weitere Informationen aus den Rohdaten benötigt, die nicht in den Varianzen und Kovarianzen abgebildet sind (z. B. Verteilungseigenschaften).

Sofern die Mittelwertestruktur für die Fragestellung relevant ist, kann diese in die Fit-Funktion aufgenommen werden. In diesem Fall wird zusätzlich die Differenz zwischen den Elementen des empirischen und des modellimplizierten Mittelwertevektors minimiert. Die Mittelwertestruktur ist beispielsweise von Interesse, wenn Gruppenvergleiche oder Invarianztestungen vorgenommen werden. In vielen CFA-Modellen ist die Mittelwertestruktur jedoch von untergeordnetem Interesse, da deren Struktur saturiert ist. Das bedeutet, dass keine Restriktionen bezüglich der Interzeptparameter formuliert und diese frei geschätzt werden (vgl. Rhemtulla et al. 2012). Dies gilt analog für die Schwellenwerte, die bei kategorialen Daten anstelle der Mittelwerte geschätzt werden.

24.5.1 Kontinuierliche, normalverteilte Indikatorvariablen

ML-Schätzmethode

Die am häufigsten verwendete Schätzmethode ist die Maximum-Likelihood-Methode (ML-Methode), die in großen Stichproben zu ML-Schätzungen mit erwartungstreuen Ergebnissen führt (vgl. Bollen 1989). Voraussetzungen sind ein korrekt spezifiziertes Modell, multivariat normalverteilte Daten und lineare Beziehungen zwischen kontinuierlichen manifesten und latenten Variablen.

Minimierung der Diskrepanzfunktion

Die ML-Schätzung maximiert die Likelihood („Wahrscheinlichkeit“), dass bei Gültigkeit des Modells in der Population die empirische Kovarianzmatrix resultiert. Die Likelihood ist entsprechend maximal, wenn die Differenzen zwischen empirischer und modellimplizierter Kovarianzmatrix möglichst gering sind, die entsprechende Diskrepanzfunktion also minimal wird. Die zu minimierende *Diskrepanzfunktion* lautet (vgl. Bollen 1989; Jöreskog 1971):

$$F_{\text{ML}} = \log |\Sigma(\theta)| - \log |S| + \text{tr}(S\Sigma(\theta)^{-1}) - p \quad (24.30)$$

Hierbei bezeichnet F_{ML} den Funktionswert der ML-Methode, \log den natürlichen Logarithmus, $|\Sigma(\theta)|$ die Determinante der modellimplizierten Kovarianzmatrix, die abhängig ist vom Vektor θ der Modellparameter, $|S|$ die Determinante der empirischen Kovarianzmatrix, $\text{tr}(S\Sigma(\theta)^{-1})$ die Spur des Produkts der Matrix S mit der inversen Matrix $\Sigma(\theta)^{-1}$ und p die Anzahl der Items. Die Funktion wird hinsichtlich θ minimiert. Das bedeutet, dass geschätzte Parameter resultieren (zusammengefasst im Vektor der geschätzten Parameter $\hat{\theta}$), für die die Differenzen zwischen der modellimplizierten Kovarianzmatrix $\Sigma(\hat{\theta})$ und der empirischen Kovarianzmatrix S möglichst gering sind. Das Ergebnis der Schätzung ist die modellimplizierte Ko-

24.5 · Parameterschätzung

varianzmatrix $\Sigma(\hat{\theta})$, für die vereinfachend die Bezeichnung $\hat{\Sigma}$ verwendet werden kann.

Die Log-Likelihood-Funktion erfüllt die Bedingung, dass der Funktionswert nur dann null wird ($F_{ML} = 0$), wenn die empirische und die modellimplizierte Kovarianzmatrix identisch sind ($S = \hat{\Sigma}$), andernfalls ist F_{ML} größer als null. Bei steigenden Differenzen vergrößert sich der Funktionswert. Aus dem Funktionswert werden die meisten Modellgütekriterien abgeleitet (► Abschn. 24.6).

ML-Diskrepanzfunktion

Zum besseren Verständnis der ML-Diskrepanzfunktion (Gl. 24.30) helfen einige Grundlagen der Matrixalgebra (s. z. B. Moosbrugger 2011). Multipliziert man die quadratische Matrix S mit ihrer Inversen S^{-1} , so erhält man die Einheitsmatrix I . In einer Einheitsmatrix sind alle Hauptdiagonalelemente gleich eins und die Elemente unterhalb und oberhalb der Hauptdiagonalen gleich null: $SS^{-1} = I$. Ist die modellimplizierte Kovarianzmatrix identisch mit der empirischen Kovarianzmatrix, d.h. $S = \Sigma(\theta)$, so ergibt das Produkt $S\Sigma(\theta)^{-1}$ die Einheitsmatrix: $S\Sigma(\theta)^{-1} = SS^{-1} = I$.

Die Spur einer Matrix ($tr = \text{trace}$) ist die Summe ihrer Hauptdiagonalelemente. Für den Fall, dass die modellimplizierte und die empirische Kovarianzmatrix identisch sind, entspricht $tr(S\Sigma(\theta)^{-1})$ der Spur der Einheitsmatrix $tr(I)$. Die Anzahl der Elemente auf der Hauptdiagonalen der Einheitsmatrix I entspricht dann der Anzahl der Items p . Somit ist die Spur der Matrix I gleich der Anzahl der Items ($tr(I) = p$). Damit würde sich $tr(S\Sigma(\theta)^{-1}) - p = 0$ ergeben.

Da in diesem Fall außerdem $S = \Sigma(\theta)$ und damit auch $\log |S| = \log |\Sigma(\theta)|$ ist, ergibt die Differenz dieser Werte ebenfalls null. Dies verdeutlicht, dass die ML-Diskrepanzfunktion genau dann null wird ($F_{ML} = 0$), wenn die modellimplizierte und die empirische Kovarianzmatrix identisch sind.

Damit die Funktion geschätzt werden kann, müssen die Determinanten der empirischen Kovarianzmatrix und der modellimplizierten Kovarianzmatrix, $|S|$ und $|\Sigma(\theta)|$, ungleich null sein, denn nur dann existiert die inverse Matrix $\Sigma(\theta)^{-1}$. Aus einer Determinanten von null kann auf lineare Abhängigkeiten der Spaltenvektoren dieser singulären Matrix geschlossen werden. Dies ist in der Praxis z. B. dann der Fall, wenn Variablen sehr hoch korreliert sind (Kollinearität). Deutlich wird auch, dass keine der Determinanten gleich oder kleiner als null sein darf, da der Logarithmus für Werte ≤ 0 nicht definiert ist.

ML-Schätzungen können auf Datensätze mit fehlenden Werten angewendet werden, sofern Rohdaten vorliegen und die Daten zufällig und nicht systematisch fehlen (*missing completely at random*, MCAR, bzw. *missing at random*, MAR; zur Klassifikation fehlender Werte s. z. B. Schafer und Graham 2002). Die am häufigsten angewendete Methode ist die robuste ML-Methode (MLR), bei der alle verfügbaren Daten für die Parameterschätzung verwendet werden (vgl. z. B. Reinecke 2014, S. 241 ff.).

24.5.2 Kontinuierliche, nicht normalverteilte Indikatorvariablen

Sind die empirischen Daten nicht normalverteilt, so werden die Modellparameter mittels ML-Schätzung zwar unverzerrt („unbiased“) geschätzt, jedoch ist der Schätzer ineffizient. Dies bedeutet, dass Standardfehler überschätzt werden und somit eine größere Stichprobe benötigt wird, um Effekte aufzudecken zu können. Diese Ineffizienz wirkt sich auch auf den χ^2 -Wert (► Abschn. 24.6.1) aus, der ebenfalls

Falsche Schlussfolgerungen bei Verletzung der Verteilungsvoraussetzungen

erhöht ist (vgl. Curran et al. 1996; Olsson et al. 2000; Savalei 2014). Die Überschätzung der Standardfehler kann zu falschen Schlussfolgerungen hinsichtlich der Relevanz einzelner Parameter sowie zu fehlerhaften Entscheidungen hinsichtlich der Gültigkeit des Gesamtmodells führen, da zu viele Modelle abgelehnt werden, die bei normalverteilten Variablen akzeptiert werden (vgl. Curran et al. 1996).

Bei Verletzung der Normalverteilungsvoraussetzung können verschiedene Schätzmethoden verwendet werden, die eine Korrektur der Teststatistik und der Standardfehler vornehmen (Satorra und Bentler 1994; Yuan und Bentler 1998, 2000). Die wohl bekannteste Methode ist die robuste ML-Schätzmethode MLR, die z. B. vom Programm *Mplus* (Muthén und Muthén 2017) und dem R-Paket lavaan (Rosseel 2012) zur Verfügung gestellt wird. Bei MLR wird die Teststatistik so korrigiert, dass der Erwartungswert der Teststatistik der Anzahl der Freiheitsgrade entspricht. Die Standardfehler werden mithilfe eines sog. „Sandwich-Schätzers“ (vgl. Savalei 2014) so korrigiert, dass sie den Standardfehlern von Parameterschätzungen entsprechen, die auf normalverteilten Daten beruhen. MLR ist ebenfalls anwendbar auf Datensätze mit fehlenden Werten (vgl. Reinecke 2014).

Die nach Yuan und Bentler (2000) korrigierte χ^2 -Teststatistik und die korrigierten Standardfehler der ML-Schätzung (*Mean and Variance adjusted ML*) sind beispielsweise verfügbar in *Mplus* und im R-Paket lavaan (Rosseel 2012). Die Parameterschätzungen sind identisch mit der ML-Schätzung; die korrigierten χ^2 -Werte und Standardfehlerschätzungen sind jedoch bei Nichtnormalität der Daten robust (vgl. Reinecke 2014, S. 103).

Robuste ML-Schätzmethode (MLR)

WLS-Methode ULS-Methode

Kleinste-Quadrat-Schätzung

Für nicht normalverteilte Daten können prinzipiell auch verteilungsfreie (gewichtete/ungewichtete) Kleinste-Quadrat-Schätzmethoden verwendet werden, z. B. die Methoden *Weighted Least-Squares* (WLS) oder *Unweighted Least-Squares* (ULS).

Die Fit-Funktion der WLS-Methode (Browne 1982, 1984; vgl. auch Reinecke 2014, S. 105) lautet:

$$F_{WLS} = [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})]' \mathbf{W}^{-1} [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})] \quad (24.31)$$

Hierbei enthält der Vektor \mathbf{s} die nicht redundanten Elemente der empirischen Kovarianzmatrix und der Vektor $\boldsymbol{\sigma}(\boldsymbol{\theta})$ die Elemente der modellimplizierten Kovarianzmatrix, die abhängig ist vom Vektor $\boldsymbol{\theta}$ der Modellparameter.

Für verschiedene Kleinste-Quadrat-Schätzmethoden wird die Gewichtungsmatrix \mathbf{W} unterschiedlich gewählt. Im Fall von WLS wird eine vollständige Gewichtungsmatrix \mathbf{W} verwendet, die asymptotische Varianzen und Kovarianzen enthält und somit Momente höherer Ordnung (Schiefe und Kurtosis, ► Kap. 7) berücksichtigt. Sie besteht aus $[p(p+1)/2]^2$ Elementen, das wären z. B. bei fünf Indikatoren $15 \cdot 15 = 225$ Elemente, bei 10 Indikatoren bereits $55 \cdot 55 = 3025$ Elemente (für ausführliche Informationen s. Reinecke 2014, S. 105 ff.). In kleinen Stichproben kommt es deshalb häufig zu Schätzproblemen, da die Gewichtungsmatrix singulär, d. h. nicht invertierbar ist. WLS hat zusätzlich den Nachteil, dass diese Methode weniger effizient ist als die ML-Methode und somit sehr große Stichproben von mehreren 1000 Personen benötigt werden (Boomsma und Hoogland 2001). Unter praktischen Gesichtspunkten ist WLS daher meistens nicht zu empfehlen.

In der ULS-Methode wird keine Gewichtungsmatrix verwendet, wodurch im Vergleich zur WLS-Methode weniger Schätzprobleme auftreten. Sie ist aber ebenfalls weniger effizient als die ML-Methode und hat zudem den Nachteil, nicht skaleninvariant zu sein, wodurch die Analyse einer Kovarianzmatrix zu anderen Ergebnissen führt als die einer Korrelationsmatrix (vgl. Bollen 1989).

24.5.3 Kategoriale Indikatorvariablen

Liegen kategoriale Daten (z. B. Antworten auf Items) vor, können verschiedene Schätzmethoden verwendet werden. Für Ratingskalen von Items werden in der Regel zwischen 2 und 7 Kategorien verwendet, sodass die resultierenden Indikatorvariablen ordinalskaliert sind. Bei kategoriale Variablen sind sowohl die Voraussetzung linearer Beziehungen zwischen kontinuierlichen manifesten und latenten Variablen als auch die Normalverteilungsvoraussetzung der ML-Schätzung verletzt. Werden kategoriale Daten als Basis für eine CFA verwendet, führt die ML-Methode zu verzerrten Parameter- sowie inkorrekten Standardfehlerschätzungen und Modelltests.

Diese Probleme verringern sich jedoch mit steigender Anzahl der verwendeten Kategorien. Werden kategoriale Variablen mit hinreichender Anzahl der Kategorien als grob gestufte kontinuierliche Variablen aufgefasst, so kann unter bestimmten Bedingungen die robuste ML-Methode (MLR) verwendet werden. In einer Simulationsstudie erzielte die MLR-Methode für kategoriale Variablen mit mindestens fünf Kategorien und gleichzeitig nicht zu starker Asymmetrie der Schwellenwerte unverzerrte und effiziente Schätzungen (Rhemtulla et al. 2012).

Werden kategoriale Variablen als ordinalskalierte Variablen behandelt, können verteilungsfreie Methoden zur Schätzung der Modellparameter verwendet werden. Im Wesentlichen lassen sich zwei Ansätze unterscheiden (vgl. Rhemtulla et al. 2012, S. 354). Eine Möglichkeit ist die Annahme einer normalverteilten Variablen y^* , die der kategoriale y zugrunde liegt. In der CFA werden dann lineare Beziehungen zwischen den y^* -Variablen und dem latenten Faktor η geschätzt (vgl. auch Muthén et al. 1997). Schätzmethoden für ein solches Modell beruhen auf einer polychorischen Korrelationsmatrix; sie werden als *Limited Information Methods* bezeichnet.

Geeignete Methoden sind robuste Kleinste-Quadrat-Schätzmethoden für kategoriale Variablen. Die Fit-Funktion von WLS für kategoriale Daten lautet (vgl. Bandalos 2014, S. 103; Moshagen und Musch 2014, S. 60):

$$F_{WLS} = [\mathbf{r} - \boldsymbol{\rho}(\boldsymbol{\theta})]' \mathbf{W}^{-1} [\mathbf{r} - \boldsymbol{\rho}(\boldsymbol{\theta})] \quad (24.32)$$

Hierbei enthält der Vektor \mathbf{r} die nicht redundanten empirischen polychorischen Korrelationen und die geschätzten Schwellenwerte und der Vektor $\boldsymbol{\rho}(\boldsymbol{\theta})$ die modellimplizierten polychorischen Korrelationen und Schwellenwerte, die abhängig sind vom Vektor $\boldsymbol{\theta}$ der Modellparameter.

Als Gewichtungsmatrix \mathbf{W} wird jeweils eine leichter zu invertierende Matrix gewählt, wodurch das Invertieren der asymptotischen Kovarianzmatrix entfällt. Im Falle der *ULS-Methode* wird als Gewichtungsmatrix eine Einheitsmatrix verwendet (Browne 1984, S. 65; Muthén 1993, S. 228). Die Methode *Diagonally Weighted Least Squares* (DWLS) verwendet dagegen die asymptotischen Varianzen (nicht jedoch die Kovarianzen) der asymptotischen Kovarianzmatrix für die Gewichtungsmatrix, wodurch sich die Gewichtungsmatrix auf eine Diagonalmatrix mit den asymptotischen Varianzen in der Hauptdiagonalen reduziert (Muthén et al. 1997).

Da beide Verfahren weniger effizient sind als die WLS-Methode, gibt es Korrekturen der Standardfehler und der Teststatistik, die zu effizienten Schätzern führen. Derartige robuste Schätzverfahren wie z. B. Mean- and Variance-adjusted Unweighted Least Squares (ULSMV) oder Mean- and Variance-adjusted Diagonally Weighted Least Squares (DWLS; in Mplus und lavaan als WLSMV bezeichnet) zeigen auch in kleinen Stichproben günstige Schätzeigenschaften bei Verwendung von kategoriale Variablen mit bis zu sieben Kategorien (vgl. Beauducel und Herzberg 2006; Rhemtulla et al. 2012). ULSMV zeigt insgesamt noch etwas bessere Schätzeigenschaften als WLSMV.

Robuste ML-Schätzmethoden

Limited Information Methods

Robuste Kleinste-Quadrat-Schätzmethoden

Full Information Methods

Die zweite Möglichkeit besteht in der Verwendung von nicht-linearen Linkfunktionen, bei denen die Wahrscheinlichkeit der beobachteten Antworten abhängig von der latenten Variablen η geschätzt wird. Bei diesen Methoden handelt es sich in der Regel um Modelle der IRT (vgl. ► Kap. 18), die die kategorialen Rohdaten verwenden und deswegen zu den *Full Information Methods* gezählt werden.

Ein Vorteil der *Full Information Methods* gegenüber den *Limited Information Methods* sind etwas effizientere Parameterschätzer. Jedoch haben Simulationsstudien gezeigt, dass die *Limited Information Methods* zu etwas präziseren Parameterschätzungen führen (Forero und Maydeu-Olivares 2009). Für die empirische Praxis spielen diese geringen Unterschiede zumeist keine Rolle (Rhemtulla et al. 2012), sodass häufig eine weniger komplizierte „limited information method“ verwendet wird.

! Rhemtulla et al. (2012) empfehlen bei Items mit weniger als fünf Antwortkategorien die robuste ULS-Schätzmethode ULSMV oder die robuste WLS-Schätzmethode WLSMV, die meist vergleichbare Ergebnisse liefern. Ab fünf Kategorien kann die robuste ML-Schätzmethode MLR verwendet werden, wenn die Variablen nicht zu stark von der Normalverteilung abweichen.

24.6 Modellevaluation

Die Nullhypothese, in der die Übereinstimmung der Populationskovarianzmatrix der Indikatorvariablen mit der modellimplizierten Kovarianzmatrix postuliert wird, kann inferenzstatistisch mittels χ^2 -Test überprüft werden. Wird die Nullhypothese nicht verworfen, wird das Modell als zu den empirischen Daten passend eingestuft. Zusätzlich liefern deskriptive Gütekriterien Hinweise auf die Passung des Gesamtmodells (Modellfit). Sprechen die Ergebnisse für einen guten Modellfit, können auch die einzelnen Modellparameter interpretiert und die zugehörigen Einzelhypthesen geprüft werden.

24.6.1 Beurteilung der Güte des Gesamtmodells

χ^2 -Test

χ^2 -Test

Der χ^2 -Test ist der einzige Test zur *inferenzstatistischen Beurteilung* der Modellgüte (vgl. Bollen 1989; Jöreskog 1971; Schermelleh-Engel et al. 2003) und sollte deshalb immer berichtet werden.

Mit dem χ^2 -Test wird die Nullhypothese H_0 getestet, dass in der Population kein Unterschied zwischen der Kovarianzmatrix Σ der Indikatorvariablen und der modellimplizierten Matrix $\Sigma(\theta)$ besteht. Je geringer die Differenz $S - \hat{\Sigma}$ zwischen der empirischen Kovarianzmatrix S und der modellimplizierten Matrix $\hat{\Sigma}$ in der Stichprobe ist, desto besser passt das Modell zu den Daten und desto kleiner ist der χ^2 -Wert. Die H_0 wird beibehalten, wenn der χ^2 -Wert nicht signifikant ist. Ein signifikanter χ^2 -Wert führt dagegen zur Ablehnung der H_0 . Der χ^2 -Wert ergibt sich im Rahmen der Parameterschätzung aus dem Funktionswert der Diskrepanzfunktion (► Abschn. 24.5). Die χ^2 -verteilte Prüfgröße (mit $df = s - t$ Freiheitsgraden) ergibt sich, indem der ML-Funktionswert mit $(N - 1)$ multipliziert wird.

Deskriptive Gütekriterien

Zur *deskriptiven Beurteilung* der Modellgüte sind folgende Gütekriterien gebräuchlich:

- Der *Root Mean Square Error of Approximation (RMSEA)* (Steiger und Lind 1980; s. auch Browne und Cudeck 1993; Steiger 2016) ist ein populationsbasiertes Gütekriterium zur Beurteilung der approximativen Passung von Modell und Daten und basiert auf der nicht zentralen χ^2 -Verteilung. Ein 90 %-Konfidenzintervall gibt Auskunft darüber, wie präzise die Schätzung des RMSEA ist.
- Der *Comparative Fit Index (CFI)* (Bentler 1990) und der *Tucker Lewis Index (TLI)*, auch *Nonnormed Fit Index, NNFI*, genannt (Bentler und Bonett 1980; Tucker und Lewis 1973), beruhen auf einem Vergleich des untersuchten Modells mit dem *Unabhängigkeitsmodell*. Im Unabhängigkeitsmodell wird angenommen, dass alle Indikatorvariablen unkorreliert sind und messfehlerfrei gemessen werden. Es passt entsprechend schlecht zu den Daten, wenn tatsächlich Zusammenhänge zwischen den Variablen bestehen. CFI und TLI beruhen auf einem Vergleich der Teststatistiken des untersuchten Modells mit dem Unabhängigkeitsmodell und sind Maße dafür, wie viel besser das untersuchte Modell die Daten erklärt als das Unabhängigkeitsmodell.
- Der *Standardized Root Mean Square Residual (SRMR)* (Jöreskog und Sörbom 1989; s. auch Bentler 1995) basiert auf den mittleren standardisierten Differenzen zwischen den Elementen der empirischen Kovarianzmatrix S und der modellimplizierten Kovarianzmatrix \hat{S} und ist ein Maß dafür, wie gut die empirischen Kovarianzen durch die Modellparameter reproduziert werden können (vgl. Bentler 2006; Zhang und Savalei 2016).

RMSEA**CFI und NNFI****SRMR****Cut-off-Werte der Gütekriterien**

Simulationsstudien haben zu unterschiedlichen Cut-off-Werten der deskriptiven Gütekriterien zur Beurteilung der Modellgüte geführt. Dennoch bieten folgende „Daumenregeln“ (► Tab. 24.3) zusätzlich zum χ^2 -Test eine Orientierung zur Beurteilung der Modellgüte anhand der deskriptiven Gütekriterien.

Generell gilt: Es gibt kein Gütekriterium, dessen Cut-off-Wert universelle Gültigkeit besitzt, da die Gütekriterien von zahlreichen Einflussfaktoren abhängig sind (► Exkurs 24.1).

Sollen mehrere Modelle hinsichtlich ihrer Güte miteinander verglichen werden, können dazu beispielsweise der χ^2 -Differenztest als inferenzstatistischer Test und das Akaike-Informationskriterium (AIC) bzw. das Bayes'sche Informationskriterium (BIC) als deskriptive Gütemaße genutzt werden (► Abschn. 24.8).

■ Tabelle 24.3 Daumenregeln zur Beurteilung der Modellgüte, differenziert nach homogenen und heterogenen Items (homogene Items messen ähnliche Informationen und sind hoch korreliert, heterogene Items messen weniger einheitliche Informationen und sind geringer korreliert)

Gütekriterium	Guter Modellfit	Akzeptabler Modellfit
χ^2 -Wert	$\chi^2/df \leq 2$	$\chi^2/df \leq 3$
RMSEA	$RMSEA \leq .05$	$RMSEA \leq .08$
CFI	$CFI \geq .97$ (homogene Items) $CFI \geq .95$ (heterogene Items)	$CFI \geq .95$ $CFI \geq .90$
TLI	$TLI \geq .97$ (homogene Items) $TLI \geq .95$ (heterogene Items)	$TLI \geq .95$ $TLI \geq .90$
SRMR	$SRMR \leq .05$	$SRMR \leq .10$

Exkurs 24.1**Gütekriterien unter der Lupe**

Die resultierenden Werte der Modellgütekriterien sind von zahlreichen Einflussfaktoren abhängig. In kleinen Stichproben werden häufig zu viele passende Modelle aufgrund eines hohen χ^2 -Wertes fälschlicherweise abgelehnt (Boomsma und Herzog 2009; Moshagen 2012). Bei zunehmender Stichprobengröße werden zwar die Modellparameter genauer geschätzt, jedoch wird gleichzeitig der χ^2 -Wert aufgrund höherer Power wiederum sensibler für Fehlspezifikationen, sodass Modelle schon bei geringen Abweichungen zwischen empirischer und modellimplizierter Kovarianzmatrix abgelehnt werden. Der SRMR hingegen verringert sich, sodass mehr Modelle trotz starker Fehlspezifikation akzeptiert werden. Im Gegensatz dazu scheint die Stichprobengröße keinen Einfluss auf den CFI und den RMSEA zu haben (vgl. Bentler 1990; Bollen 1990; Heene et al. 2011; Hu und Bentler 1998, 1999).

Eine steigende Anzahl manifester Variablen und damit einhergehend eine zunehmende Modellkomplexität führen dazu, dass Modelle durch überhöhte χ^2 -Werte eher abgelehnt werden. In der Konsequenz werden auch durch deskriptive Maße, die auf dem χ^2 -Wert beruhen, wie RMSEA und CFI in vergleichbaren Fällen zu viele Modelle abgelehnt (vgl. Moshagen 2012).

Ebenso können auch hohe Faktorladungen schon bei leichten Fehlspezifikationen zur Ablehnung von Modellen durch den χ^2 -Wert, RMSEA und SRMR führen. Hingegen führen geringe Faktorladungen dazu, dass fehlspezifizierte Modelle oftmals nicht aufgedeckt werden (vgl. Heene et al. 2011; Savalei 2012). Daher wird zum Teil angeraten, bei niedrigen Faktorladungen für den RMSEA einen geringeren Cut-off-Wert als .05 zu wählen (Chen et al. 2008).

Die verschiedenen Gütekriterien hängen außerdem ab vom Ausmaß der Fehlspezifikation des Modells, der Abweichung der Daten von der Normalverteilung und der Anzahl der Abstufungen der verwendeten Ratingskalen. Die Daumenregeln bieten daher nur eine grobe Orientierung und können im Einzelfall sogar zu falschen Entscheidungen führen. Es wird empfohlen, immer verschiedene Gütekriterien simultan zur Beurteilung der Modellgüte heranzuziehen. Weitere Gütekriterien und detailliertere Informationen finden sich z. B. bei Kline (2016), Reinecke (2014) und Schermelleh-Engel et al. (2003).

24.6.2 Beurteilung der Modellparameter

Wurde zur Skalierung der Faktoren entweder die Faktorladung einer Indikatorvariablen oder die Faktorvarianz auf eins fixiert, basieren alle Parameterschätzungen in der *unstandardisierten Lösung* einer CFA auf der Originalmetrik der Indikatorvariablen und die Faktorladungen können als unstandardisierte Regressionskoeffizienten interpretiert werden.

In der *komplett standardisierten Lösung* sind Faktoren und Indikatorvariablen standardisiert und Faktorladungen können als standardisierte Regressionskoeffizienten – bzw. wie in einer einfachen Regression – als Korrelation zwischen Faktor und Indikatorvariablen interpretiert werden. Die jeweilige quadrierte Faktorladung entspricht dann dem durch den Faktor erklärten Varianzanteil einer Indikatorvariable (vgl. Gl. 24.9).

Jeder Parameter kann einzeln auf Signifikanz geprüft werden, indem der geschätzte Parameterwert durch seinen geschätzten Standardfehler geteilt wird. Für den resultierenden z -Wert gilt ± 1.96 als kritischer Wert (zweiseitige Signifikanztestung, $\alpha = .05$).

Signifikanzprüfung

24.7 · Modifikation der Modellstruktur

Die Beurteilung des Modells anhand der einzelnen Modellparameter kann neben den Modellgütekriterien äußerst relevant sein. So kann ein Modell zwar einen guten Modellfit aufweisen, aber dennoch ein schlechtes Erklärungsmodell sein, wenn z. B. alle Faktorladungen nahe null liegen.

Auch ist eine Faktorvarianz nahe null ein Hinweis dafür, dass ein Faktor irrelevant ist. Dies kann dann auftreten, wenn Indikatorvariablen einer Skala tatsächlich unkorreliert sind. Kritisch können auch hohe Korrelationen zwischen Faktoren (z. B. $> .90$) sein, die auf eine fehlende diskriminante Validität hinweisen. Ein Modell sollte daher auch bei gutem Modellfit zusätzlich immer anhand der Modellparameter hinsichtlich der Plausibilität und des Erklärungswertes beurteilt werden.

Modellparameter zur Beurteilung der Plausibilität eines Modells

24.7 Modifikation der Modellstruktur

Weist die Modellevaluation auf eine Fehlspezifikation hin, so kann durch eine Modifikation der Modellstruktur ggf. eine Verbesserung des Modellfits erzielt werden. Modifikationsindizes geben an, um wie viel sich der χ^2 -Wert ungefähr verringern würde, wenn ein vorher fixierter oder restringierter Parameter freigesetzt wird, während das restliche Modell gleichzeitig unverändert bleibt. Ein Modifikationsindex kann als χ^2 -Wert mit einem Freiheitsgrad interpretiert werden. Werte größer als $\chi^2_{\text{krit}}(1) = 3.84$ ($\alpha = .05$) können somit als Anhaltspunkt für signifikante Modellveränderungen dienen. Aufgrund von hohen Modifikationsindizes können entweder Faktorladungen oder Fehlerkovarianzen hinzugefügt werden.

- !** Nachträgliche Modellanpassungen sollten jedoch nicht unkritisch vorgenommen werden. Die CFA verliert durch Modifikationen ihren konfirmatorischen Charakter und wird zu einem exploratorischen Analyseverfahren. Auch sind zufällige Stichprobeneffekte in der Regel nicht auszuschließen, sodass sich rein „datengetriebene“ Modellmodifikationen unter Umständen in anderen Stichproben nicht replizieren lassen, wenn das Modell zu stark an die Daten angepasst wird.

Modifikationsindex

Fehlerkovarianzen resultieren aus substantiellen korrelativen Zusammenhängen zwischen verschiedenen Indikatorvariablen, die unabhängig vom Einfluss des gemeinsamen Faktors bestehen. Die Ursache für Fehlerkovarianzen können Methodeneffekte sein, die z. B. aus invers formulierten Items resultieren. Unter dem Begriff „Methode“ werden unterschiedliche systematische Varianzquellen subsumiert, die sich über den Faktor hinaus auf die Messungen auswirken können (s. dazu ▶ Kap. 15, 25 und 26).

Methodeneffekte

Auch wenn ein hoher Modifikationsindex die Aufnahme einer Fehlerkovarianz in das Modell nahelegt, so kann dennoch nicht ausgeschlossen werden, dass sich hinter dem vermeintlichen Methodeneffekt ein inhaltlich bedeutsamer Effekt verbirgt, der eher für die Hinzunahme eines weiteren, für das Konstrukt relevanten Faktors sprechen würde. Dies könnte z. B. dann der Fall sein, wenn alle Items das Konstrukt *Depressivität* messen sollen, zwei Items jedoch *Lebenszufriedenheit* erfassen, die nicht zwingend als invers kodierte Depressivitätsitems interpretiert werden können. In solchen Fällen ließe sich durch das Freisetzen einzelner Parameter womöglich ein guter Modellfit erzielen, obwohl tatsächlich die Modellstruktur falsch und beispielsweise statt eines einfaktoriellen Modells ein zweifaktorielles Modell angemessener wäre.

Änderung der Modellstruktur

Bei Verwendung der Modifikationsindizes zur Modellmodifikation ist zu beachten, dass diese immer nur anzeigen können, ob das Freisetzen einzelner Parameter bei Konstanthaltung aller anderen Strukturen zu einer Verbesserung des Modellfits führen würde. Eine grundsätzliche Fehlspezifikation kann damit jedoch nicht aufgedeckt werden. Im Allgemeinen empfiehlt sich die Kreuzvalidierung eines modifizierten Modells an einer weiteren, unabhängigen Stichprobe, um die neuen Modellhypothesen zu bestätigen.

Kreuzvalidierung modifizierter Modelle

24.8 Modellvergleiche

In der Praxis ergibt sich häufig die Situation, dass zwei konkurrierende Modelle miteinander verglichen werden sollen. Dabei werden Modelle, die ineinander geschachtelt sind (häufig als „hierarchisch geschachtelt“ bezeichnet), unterschieden von Modellen, die nicht geschachtelt sind. Modelle sind geschachtelt, wenn sie auf denselben Variablen und derselben Modellstruktur beruhen und ein Modell aus dem anderen durch zusätzlich zu schätzende Parameter hervorgeht. Modelle sind nicht geschachtelt, wenn sie auf einer unterschiedlichen Anzahl von Variablen oder einer unterschiedlichen Modellstruktur beruhen.

24.8.1 Geschachtelte Modelle

Restringiertes vs. unrestringiertes Modell

Bei geschachtelten Modellen („nested models“) wird geprüft, ob sich der Modellfit eines restringierten (restriktiveren) Modells (Modell A) mit weniger frei geschätzten Parametern und daher mehr Freiheitsgraden (df_A) von dem Modellfit eines unrestringierten Modells (Modell B) mit mindestens einem Freiheitsgrad weniger (df_B), in dem also mindestens ein Parameter mehr frei geschätzt wird, signifikant unterscheidet. Beispielsweise könnten im restringierten Modell A zwei Faktorladungen gleichgesetzt oder eine Faktorkovarianz oder eine Fehlerkovarianz auf null fixiert werden, während der jeweilige Parameter im unrestringierten Modell B frei geschätzt werden würde.

Sind die Modelle geschachtelt, kann der χ^2 -Differenztest verwendet werden (vgl. Bollen 1989; Schermelleh-Engel et al. 2003). Der Test prüft die Nullhypothese, dass sich die beiden Modelle in ihren modellimplizierten Kovarianzmatrizen nicht unterscheiden. Die Nullhypothese für den Modellvergleich lautet somit: $H_0: \Sigma_A = \Sigma_B$.

Für den Modellvergleich wird die Differenz zwischen den χ^2 -Werten des restringierten Modells A und des unrestringierten Modells B sowie die Differenz der dazugehörigen Freiheitsgrade gebildet:

$$\Delta\chi^2 = \chi^2_A - \chi^2_B \quad \text{mit} \quad \Delta df = df_A - df_B \quad (24.33)$$

χ^2 -Differenztest

Diese Differenz ist wiederum χ^2 -verteilt und kann auf Signifikanz geprüft werden. Die Differenz der χ^2 -Werte kann entweder mittels einer χ^2 -Tabelle auf Signifikanz geprüft oder vom verwendeten Analyseprogramm berechnet und ausgegeben werden.

Aus Gründen der Parsimonie als allgemeinem Forschungsprinzip (vgl. Reiß und Sarris 2012) spricht ein nicht signifikanter Differenzwert ($p > .01$) für das restringierte, sparsamere Modell mit weniger frei geschätzten Parametern (Modell A), da sich der Modellfit durch die Fixierung eines oder mehrerer Parameter im Vergleich zum unrestringierten Modell (Modell B) nicht signifikant verschlechtert. In diesem Fall sollte Modell A beibehalten werden. Ist die Differenz aber signifikant ($p < .01$), dann passt das restriktivere Modell A schlechter zu den Daten als Modell B und das Modell B sollte beibehalten werden.

Zwingende Voraussetzung für den χ^2 -Differenztest ist, dass die zu vergleichenden Modelle ineinander geschachtelt sind, d. h., dass ein Modell durch Fixieren eines oder mehrerer Parameter aus dem anderen Modell hervorgeht. Zusätzlich sollte darauf geachtet werden, dass zumindest das weniger restriktive Modell einen guten Modellfit aufweist.

Basiert der Modellvergleich auf χ^2 -Statistiken, die mittels robuster Schätzmethoden gewonnen wurden (► Abschn. 24.5), ist die resultierende Differenz nicht mehr χ^2 -verteilt. Erst die Gewichtung der robusten χ^2 -Werte der beiden Modelle mit speziellen Skalierungsfaktoren, die von Statistikprogrammen bereitgestellt

Robuster χ^2 -Differenztest

24.9 · Messinvarianztestung

werden, führt wieder zu einer χ^2 -verteilten Differenz, die auf Signifikanz geprüft werden kann (vgl. ► Kap. 23; Satorra und Bentler 2010).

Dazu wird zunächst anhand der Skalierungsfaktoren (c) und der Freiheitsgrade df beider Modelle ein Korrekturfaktor Δc berechnet:

$$\Delta c = (df_A \cdot c_A - df_B \cdot c_B) / (df_A - df_B) \quad (24.34)$$

Im nächsten Schritt kann die korrigierte χ^2 -Differenz berechnet werden:

$$\Delta\chi^2_{\text{korrigiert}} = (\chi^2_A \cdot c_A - \chi^2_B \cdot c_B) / \Delta c \quad (24.35)$$

In ► Beispiel 24.5 wird der Modellvergleich anhand des empirischen Beispiels verdeutlicht.

Beispiel 24.5: Modellvergleich

Für die vier ausgewählten Items der Skala CM (► Abschn. 24.3, ► Beispiel 24.1) kann geprüft werden, ob das τ -kongenerische Modell besser zu den Daten passt als das restriktivere Modell essentiell τ -äquivalenter Messungen mit gleichgesetzten Faktorladungen. Da der Modellvergleich auf χ^2 -Statistiken beruht, die mittels robuster Schätzmethoden gewonnen wurden, muss der robuste χ^2 -Differenztest durchgeführt werden. Die korrigierte χ^2 -Differenz ist mit $\Delta\chi^2_{\text{korrigiert}} = 13.93$ bei $\Delta df = 3$ signifikant ($p < .01$). Dies bedeutet, dass das τ -kongenerische Modell mit frei geschätzten Faktorladungen eine signifikant bessere Anpassung an die Daten aufweist als das Modell essentiell τ -äquivalenter Messungen mit gleichgesetzten Faktorladungen.

24.8.2 Nicht geschachtelte Modelle

Sollen Modelle miteinander verglichen werden, die *nicht geschachtelt* sind, weil sie z. B. auf einer unterschiedlichen Anzahl von Indikatorvariablen, einer unterschiedlichen Modellstruktur oder auf derselben Anzahl an Freiheitsgraden beruhen, so können zwei deskriptive, unstandardisierte Maße verwendet werden, nämlich das Akaike-Informationskriterium (AIC; Akaike 1987) und das Bayes'sche Informationskriterium (BIC; Schwarz 1978).

Bei diesen Maßen handelt es sich um *Informationskriterien*, die sich aus einer Gewichtung aus Modellfit und Modellsparsamkeit (Anzahl der Freiheitsgrade bzw. Anzahl der zu schätzenden Parameter) berechnen (nähere Informationen zu den Maßen finden sich z. B. bei Mulaik 2009, oder Schermelleh-Engel et al. 2003; s. auch ► Kap. 22). Die beiden Maße unterscheiden sich darin, dass das BIC die Stichprobengröße berücksichtigt und bei komplexeren Modellen mit mehr zu schätzenden Parametern höhere Werte annimmt als das AIC, d. h. die Modellkomplexität stärker „bestraft“ (vgl. Brown 2015, S. 153). Kleinere AIC- und BIC-Werte deuten darauf hin, dass das Verhältnis von Modellfit und Modellsparsamkeit in dem entsprechenden Modell besser ist als im Vergleichsmodell. Von zwei Modellen wird also dasjenige bevorzugt, das den kleineren AIC- bzw. BIC-Wert aufweist.

Deskriptiver Modellvergleich

24.9 Messinvarianztestung

Messinstrumente werden häufig in Stichproben verschiedener Populationen (z. B. Therapie- vs. Kontrollgruppe, Frauen vs. Männer) oder in Längsschnittuntersuchungen zu mehreren Messzeitpunkten eingesetzt. Typische Forschungsfragen beziehen sich dabei z. B. darauf, ob mit dem entwickelten Test Merkmalsunterschiede

Ziel der Messinvarianztestung

zwischen den Gruppen oder die Stabilität/Veränderung des Merkmals über die Zeit festgestellt werden können. Eine zwingende Voraussetzung für derartige Analysen ist die Invarianz der Messung, was bedeutet, dass das Messinstrument (der Test) in allen Gruppen bzw. über die Zeit hinweg dasselbe Merkmal in gleicher Weise (= invariant) misst. Nur bei gegebener Messinvarianz lassen sich sinnvolle Schlussfolgerungen aus den Analysen ziehen. Die Messinvarianztestung hat das Ziel, diese Voraussetzung zu überprüfen. Im Unterschied zu den dimensionalen Fragestellungen, bei denen die Analyse der Kovarianzstruktur in der Regel ausreichend ist, muss bei Gruppenvergleichen von latenten Mittelwerten (z. B.: Unterscheiden sich Männer und Frauen auf latenter Ebene hinsichtlich des Merkmals „Perfektionistische Sorge“?) oder bei Verlaufsfragestellungen (z. B.: Ändert sich die Ausprägung des Merkmals „Perfektionistische Sorge“ auf latenter Ebene über die Zeit?) auch die Mittelwertestruktur beachtet werden. Daher werden für derartige Fragestellungen neben den Faktorladungen und Fehlervarianzen auch die Interzepte und Erwartungswerte der Faktoren in die Invarianzbestimmung einbezogen.

Die Frage, ob ein Messinstrument in verschiedenen Gruppen oder über mehrere Messzeitpunkte hinweg dasselbe Merkmal misst und dabei vergleichbare psychometrische Eigenschaften aufweist, stellt eine Erweiterung der Validitäts- und Reliabilitätsprüfung eines Tests dar und kann mithilfe der CFA getestet werden.

Dabei werden verschiedene, aufeinander aufbauende Stufen der Messinvarianz unterschieden, die für ein Messinstrument über Gruppen bzw. Messzeitpunkte hinweg überprüft werden können. Je nach Invarianzstufe eines Messinstruments/Tests sind unterschiedliche Vergleiche zwischen Gruppen bzw. Messzeitpunkten auf manifesten und latenter Ebene zulässig. Im Folgenden ist zur besseren Lesbarkeit nur von der Gleichheit über Gruppen hinweg die Rede, die Ausführungen gelten analog für die Gleichheit über Messzeitpunkte hinweg (Meredith 1993; vgl. auch Gregorich 2006; Kline 2016).

■ ■ Stufen der Messinvarianz

- *Konfigurale Invarianz* setzt nur dieselbe Anzahl von Faktoren sowie dieselbe Zuordnung der Items zu den Faktoren über verschiedene Gruppen hinweg voraus. Die Gleichheit der Struktur stellt eine Mindestvoraussetzung für nachfolgende Invarianzstufen dar.
- *Schwache (metrische) Invarianz* setzt neben konfiguraler Invarianz voraus, dass die Faktorladungen für jedes Item i auf einem Faktor η_j über die Gruppen A und B hinweg gleich sind: $\lambda_{ij}^A = \lambda_{ij}^B$. Sind die Faktorladungen invariant, haben die Faktoren in beiden Gruppen dieselbe Bedeutung und Gruppenvergleiche bezüglich latenter Varianzen und Kovarianzen sind zulässig. Metrische Invarianz erlaubt dagegen nicht, von Gruppenunterschieden in beobachteten Varianzen und Kovarianzen auf Gruppenunterschiede auf latenter Ebene zu schließen, da auf manifesten Ebene wahre Varianz und Residualvarianz konfundiert sind (Gregorich 2006). Die Interzepte der Items werden nicht berücksichtigt, da sich eine Verschiebung um einen konstanten Term nicht auf die Kovarianzstruktur auswirkt.
- *Starke (skalare) Invarianz* setzt neben metrischer Invarianz voraus, dass die Interzepte für jedes Item i über die Gruppen hinweg gleich sind: $\alpha_i^A = \alpha_i^B$. Sind auch die Interzepte invariant, hängen Gruppenunterschiede der beobachteten Mittelwerte direkt von der Ausprägung des Faktors in den Gruppen bzw. den Erwartungswerten der Faktoren in den Gruppen, $E(\eta_j^A), E(\eta_j^B)$, ab und gehen nicht (zusätzlich) auf andere, vom Faktor unabhängige Unterschiede zwischen den Gruppen zurück. Andernfalls sind tatsächliche Unterschiede in den Faktoren mit davon unabhängigen Einflüssen (z. B. gruppenspezifischen Antworttendenzen) in den Mittelwertunterschieden auf manifesten Ebene konfundiert. Manifeste Mittelwertunterschiede sind ohne skalare Invarianz

24.9 · Messinvarianztestung

nicht als Unterschiede in der Ausprägung des latenten Merkmals interpretierbar.

- **Strikte Invarianz** setzt neben skalarer Invarianz voraus, dass die Fehlervarianzen für jedes Item i über die Gruppen hinweg gleich sind: $\lambda_{ij}^A = \lambda_{ij}^B, \alpha_i^A = \alpha_i^B, \text{Var}(\varepsilon_i)^A = \text{Var}(\varepsilon_i)^B$. Sind auch die Fehlervarianzen invariant, dann sind manifeste Varianzunterschiede zwischen den Gruppen nur auf Varianzunterschiede der Faktoren zurückzuführen. Andernfalls sind tatsächliche Unterschiede in den Faktoren mit davon unabhängigen Einflüssen in den Varianzunterschieden auf manifester Ebene konfundiert.

Diese sukzessiv restriktiveren Annahmen der Messinvarianz können anhand multipler Gruppenvergleiche schrittweise getestet werden (vgl. Kline 2016; Reinecke 2014). Dazu wird zunächst das zu prüfende Faktormodell für alle Gruppen gleichzeitig spezifiziert. Weist das Modell ohne zusätzliche Gleichheitsrestriktionen zwischen den Gruppen einen guten Modellfit auf, ist von konfiguraler Invarianz auszugehen und die weiteren Invarianzstufen können sukzessive geprüft werden.

Die Modellparameter werden entsprechend der zu testenden Invarianzstufe als invariant über die Gruppen definiert und das Modell analysiert. Da Modelle mit verschiedenen Invarianzstufen ineinander geschachtelt sind, kann das Modell mit mehr Gleichheitsrestriktionen mit dem Modell mit weniger Gleichheitsrestriktionen mittels χ^2 -Differenztest verglichen werden (► Abschn. 24.8.1). Passt das restiktivere Modell genauso gut bzw. nicht signifikant schlechter zu den Daten als das weniger restiktive Modell mit den in allen Gruppen frei geschätzten Parametern, so gelten die Annahmen der strenger Invarianzstufe als erfüllt.

Treffen die Restriktionen einer Invarianzstufe auf einzelne Items nicht zu, so kann die partielle Invarianz getestet werden. Dies bedeutet, dass für einzelne Items die Parameterrestriktionen aufgehoben werden (vgl. Byrne et al. 1989; Steenkamp und Baumgartner 1998). Sind nur wenige Items betroffen, können die Gruppen bezüglich der Erwartungswerte und der Varianzen/Kovarianzen der Faktoren näherungsweise miteinander verglichen werden.

Invarianztestungen für kategoriale Daten werden u. a. von Millsap und Yun-Tein (2004) für Gruppenvergleiche und von Liu et al. (2017) für Längsschnittdaten erläutert. Messinvarianz kann nicht nur über Gruppen oder Messzeitpunkte getestet werden, sondern auch über kontinuierliche Kovariaten, z. B. über das Alter. Hierfür kann u. a. die moderierte nichtlineare Faktorenanalyse (MNLFA) verwendet werden (Bauer 2017; Curran et al. 2014).

Gruppenvergleiche auf manifester Ebene (beispielsweise Mittelwertvergleiche der Testwerte in zwei Gruppen mittels t -Test) können nur bei strikter Invarianz als Unterschiede in der latenten Variablen interpretiert werden, da die Teststatistik auf manifester Ebene neben den Mittelwertunterschieden auch die Varianzen (bzw. Standardfehler) in den Gruppen berücksichtigt.

Die Erwartungswerte der Faktoren lassen sich jedoch auch direkt im Rahmen der CFA für verschiedene Gruppen vergleichen. Voraussetzung dafür ist nur die starke, nicht aber die strikte Invarianz. Für den Gruppenvergleich können die Erwartungswerte der Faktoren im ersten Schritt in allen Gruppen, z. B. in Gruppe A und Gruppe B, frei geschätzt und im zweiten Schritt über die Gruppen hinweg gleichgesetzt werden ($E(\eta_j^A) = E(\eta_j^B)$). Die beiden Modelle können mittels χ^2 -Differenztest verglichen werden. Verschlechtert sich der Modellfit durch die zusätzliche Modellrestriktion gleicher latenter Erwartungswerte signifikant, so kann man davon ausgehen, dass sich die Erwartungswerte der Faktoren in der Population unterscheiden.

Schrittweise Prüfung mittels multipler Gruppenvergleiche

Vergleich der Invarianzstufen mittels χ^2 -Differenztests

Partielle Invarianz

Messinvarianztestung für kategoriale Daten und über kontinuierliche Kovariaten möglich

Gruppenvergleich auf latenter Ebene

24.10 Zusammenfassung

Die CFA stellt ein wichtiges statistisches Instrument zur psychometrischen Evaluation eines Testverfahrens dar. Mit ihrer Hilfe lassen sich zentrale Aspekte der Validität und der Reliabilität eines Tests untersuchen. So erlaubt die CFA die Überprüfung der Dimensionalität zum Nachweis der faktoriellen Validität eines Tests. Die Überprüfung der Dimensionalität und Messäquivalenz eines Tests ist dabei relevant für die spätere Testwertbildung und die Reliabilitätsschätzung. Dazu sind die Modellannahmen der τ -Kongeneritität, der essentiellen τ -Äquivalenz und der essentiellen τ -Parallelität von Messungen zu unterscheiden. Die Grundlagen der CFA lassen sich mit Bezug zur KTT darstellen und die Parameter eines CFA-Modells als modellbasierte Itemkennwerte interpretieren.

Die Wahl der Schätzmethoden zur Parameterschätzung hängt insbesondere vom Skalenniveau und der Verteilung der Indikatorvariablen ab. Zur Beurteilung der Güte des Gesamtmodells stehen der χ^2 -Test als inferenzstatistischer Test sowie weitere deskriptive Gütekriterien zur Verfügung.

Neben der Möglichkeit, die vorgestellten Modelle zu komplexeren Modellen zu erweitern, wird die CFA in der empirischen Forschung für weitere Fragestellungen genutzt und kann bei Modifikation der Modellstruktur auch als exploratives Instrument verwendet werden. Der Vergleich konkurrierender Modelle stellt ein hilfreiches Werkzeug bei der Modellauswahl dar. Auch die Frage nach der Messinvarianz eines Testverfahrens über mehrere Gruppen oder Messzeitpunkte hinweg ist für zahlreiche Forschungsfragen zentral und lässt sich mithilfe der CFA empirisch überprüfen.

24.11 EDV-Hinweise

Die Anwendung einer CFA erfordert entsprechende Software wie *Mplus* oder das R-Paket *lavaan*. Eine Dokumentation der vorgestellten Beispiele und deren Umsetzung in *Mplus* und R findet sich unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

24.12 Kontrollfragen

- ?
- Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Beschreiben Sie den Zusammenhang zwischen empirischer und modellimplizierter Kovarianzmatrix im Rahmen der konfirmatorischen Faktorenanalyse (CFA).
 2. Was versteht man unter Messäquivalenz? Welche Formen der Messäquivalenz können unterschieden werden und wozu ist dieses Konzept wichtig?
 3. Nach welchen Aspekten sollte eine Methode zur Schätzung der Modellparameter ausgewählt werden? Welche Methoden kennen Sie und wie lassen sich diese nach praktischen Gesichtspunkten klassifizieren?
 4. Welche Gütekriterien zur Modellevaluation werden unterschieden? Warum sollten diese nicht unkritisch angewendet werden?
 5. Wie lassen sich konkurrierende Modelle vergleichen?
 6. Erklären Sie das Konzept der Messinvarianz und dessen Relevanz im Rahmen der Testtheorie.

Literatur

- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–322.
- Altstötter-Gleich, C. & Bergemann, N. (2006). Testgüte einer deutschsprachigen Version der Mehrdimensionalen Perfektionismus Skala von Frost, Marten, Lahart und Rosenblatt (MPS-F). *Diagnostica*, 52, 105–118.
- Amend, N. (2015). *Who's perfect? Pilotstudie zur Untersuchung potenzieller Korrelate des Merkmals Perfektionismus*. Unveröffentlichte Bachelorarbeit, Institut für Psychologie, Goethe Universität, Frankfurt am Main.
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling*, 21, 102–116.
- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22, 507–526.
- Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood vs. means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, 13, 186–203.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P. M. (1995). *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software.
- Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.
- Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, 107, 256–259.
- Boomsma, A. & Herzog, W. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling*, 16, 1–27.
- Boomsma, A. & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.
- Browne, M. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). New York, NY: Cambridge University Press.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230–258.
- Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J. & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36, 462–494.
- Cho, E. & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18, 207–230.
- Curran, P. J., West, S. G. & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29.
- Curran, P. J., McGinley, J. S., Bauer, D. J., Husong, A. M., Burns, A., Chassin, L., Sher, K. & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, 49, 214–231.
- Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Eid, M., Geiser, C., Koch, T. & Heene, M. (2017a). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22, 541–562.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2017b). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
- Forero, C. G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full-information methods. *Psychological Methods*, 14, 275–299.
- Frost, R. O., Marten, P., Lahart, C. & Rosenblatt, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research*, 14, 449–468.
- Geiser, C., Bishop, J. & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology*, 6, 946. <https://doi.org/10.3389/fpsyg.2015.00946>

- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44, 78–94.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Heene, M., Hilbert, S., Draxler, C., Ziegler, M. & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16, 319–336.
- Holzinger, K. J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41–54.
- Hu, L.-T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.
- Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria vs. new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7 user's reference guide*. Chicago, IL: SPSS Publications.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.
- Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R. & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, 22, 486–506.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Pub. Co.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.
- Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Journal of Multivariate Behavioral Research*, 39, 479–515.
- Moosbrugger, H. (2011). *Lineare Modelle* (4. Aufl.). Bern: Huber.
- Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, 19, 86–98.
- Moshagen, M. & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, 10, 60–70.
- Mulaik, S. A. (2009). *Linear Causal Modeling with Structural Equations*. Boca Raton, FL: Chapman & Hall/CRC.
- Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, B. O., du Toit, S. H. C. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes. Unpublished technical report. Retrieved from <http://www.statmodel.com/wlscv.shtml> [29.12.2019]
- Olsson, U. H., Foss, T., Troye, S. V. & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7, 557–595.
- Raykov, T. & Marcoulides, G. A. (2011). *Psychometric Theory*. New York, NY: Routledge.
- Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften* (2. Aufl.). München: Oldenbourg Wissenschaftsverlag.
- Reiß, S. & Sarris, V. (2012). *Experimentelle Psychologie – Von der Theorie zur Praxis*. München: Pearson.
- Rhemtulla, M., Brosseau-Liard, P. E. & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48, 1–36.
- Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Satorra, A. & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75, 243–248.
- Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, 72, 910–932.
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21, 149–160.
- Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.

Literatur

- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8, 23–74.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Steenkamp, J. B. E. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–90.
- Steiger, J. H. (2016). Notes on the Steiger-Lind (1980) Handout. *Structural Equation Modeling*, 23, 777–781.
- Steiger, J. H. & Lind, J. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA. Retrieved from <https://www.statpower.net/Steiger%20Biblio/Steiger-Lind%201980.pdf> [10.07.2020]
- Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin, Heidelberg: Springer.
- Stöber, J. (1995). *Frost Multidimensional Perfectionism Scale-Deutsch (FMPS-D)*. Unveröffentlichtes Manuskript. Freie Universität Berlin, Institut für Psychologie.
- Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Werner, C. S., Schermelleh-Engel, K., Gerhard, C. & Gäde, J. C. (2016). Strukturgleichungsmodelle. In N. Döring & J. Bortz (Hrsg.), *Forschungsmethoden und Evaluation* (5. Aufl., S. 945–973). Berlin: Springer.
- Yuan, K. H. & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, 51, 289–309.
- Yuan, K. H. & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological Methodology 2000* (pp. 165–200). Washington, DC: ASA.
- Zhang, X. & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation models. *Structural Equation Modeling*, 23, 392–408.
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, 40, 395–412.



Multitrait-Multimethod-Analysen (MTMM-Analysen)

Karin Schermelleh-Engel, Christian Geiser und G. Leonard Burns

Inhaltsverzeichnis

- 25.1 Einleitung – 663**
- 25.2 Konvergente und diskriminante Validität – 663**
- 25.3 Methodeneffekte – 664**
 - 25.3.1 Erfassungsmethoden als Alternativerklärung für Zusammenhänge zwischen Traits – 664
 - 25.3.2 Typische Entstehungsquellen von Methodeneffekten – 665
 - 25.3.3 Untereinander austauschbare vs. strukturell unterschiedliche Methoden – 665
 - 25.3.4 Aufdeckung von Methodeneffekten – 666
- 25.4 Das MTMM-Design – 666**
 - 25.4.1 Aufbau der MTMM-Matrix – 667
 - 25.4.2 Unterscheidung und Interpretation der Blöcke und Koeffizienten – 668
- 25.5 Korrelationsbasierte Analyse der MTMM-Matrix – 669**
 - 25.5.1 Nachweiskriterien für die konvergente und diskriminante Validität – 669
 - 25.5.2 Empirisches Anwendungsbeispiel – 670
 - 25.5.3 Kritik an der korrelationsbasierten MTMM-Analyse – 672
- 25.6 Faktorenanalytische Ansätze:
Klassische CFA-MTMM-Modelle – 672**
 - 25.6.1 Das CTCM-Modell – 673
 - 25.6.2 Empirisches Anwendungsbeispiel – 674
 - 25.6.3 Kritik am CTCM-Modell – 678
- 25.7 Faktorenanalytische Ansätze:
Neuere CFA-MTMM-Modelle – 678**
 - 25.7.1 Das CTC($M-1$)-Modell – 679
 - 25.7.2 Empirisches Anwendungsbeispiel – 680

25.7.3 Kritische Bewertung und Modellerweiterungen – 682
25.7.4 Weitere neue CFA-MTMM-Ansätze – 683

25.8 Zusammenfassung – 683

25.9 EDV-Hinweise – 684

25.10 Kontrollfragen – 684

Literatur – 684

- i** Unter der Bezeichnung „Multitrait-Multimethod-Analyse“ (MTMM-Analyse) wird eine Gruppe von Verfahren zum Nachweis der Konstruktvalidität von Messungen eines Tests oder Fragebogens verstanden. Charakteristischerweise wird zum Nachweis der Konstruktvalidität eine systematische Kombination von mehreren Traits („Multitrait“, MT), gemessen mit mehreren Messmethoden („Multimethod“, MM), vorgenommen; diese Maße werden miteinander korreliert. Konvergente Validität und diskriminante Validität können einerseits anhand eines Vergleichs der Korrelationskoeffizienten und andererseits anhand verschiedener Modelle der konfirmatorischen Faktorenanalyse (CFA) beurteilt werden.

25.1 Einleitung

Die MTMM-Analyse umfasst verschiedene Verfahren zum Nachweis der Konstruktvalidität von Messungen eines Tests oder Fragebogens. Durch die systematische Kombination von mehreren Traits (Merkmale), die mit mehreren Messmethoden erfasst werden, ist es möglich, die konvergente und die diskriminante Validität dieser Messungen zu bestimmen. Für die Analyse von MTMM-Matrizen macht es dabei keinen Unterschied, ob die systematischen Beziehungen zwischen Traits und Methoden als Konstruktvalidität im Sinne einer Eigenschaft eines Tests (vgl. Campbell und Fiske 1959) oder aber als Quelle der Evidenz für die Zulässigkeit der aus den Testwerten gezogenen Schlüsse (vgl. ► Kap. 21; Messick 1995) angesehen wird. Nachfolgend sollen zwei Facetten der Konstruktvalidität, die konvergente und die diskriminante Validität, näher betrachtet werden. Diese Aufteilung der Konstruktvalidität hat eine lange Tradition; sie unterscheidet sich jedoch von den neueren Konzepten der konvergenten und diskriminanten Evidenz insofern nicht, als beide Ansätze darauf beruhen, dass Messungen verschiedener Tests, die dasselbe Konstrukt messen, hoch miteinander korrelieren und Messungen verschiedener Tests, die unterschiedliche Konstrukte messen, gering oder gar nicht miteinander korrelieren (vgl. Eid und Schmidt 2014).

25.2 Konvergente und diskriminante Validität

Zum Nachweis der Konstruktvalidität (vgl. ► Kap. 2 und 21) von Messungen eines neu entwickelten Tests oder Fragebogens wird häufig die MTMM-Analyse verwendet (Campbell und Fiske 1959), mit der es möglich ist, sowohl die konvergente als auch die diskriminante Validität zu bestimmen. Generell wird die Konstruktvalidität eines neu entwickelten Tests oder Fragebogens – neben anderen Methoden, z. B. experimentellen Untersuchungen – durch Korrelationen der Testwertvariablen mit denjenigen von anderen Tests bestimmt, die dasselbe Merkmal erfassen sollen. Ein solches Vorgehen fokussiert die *konvergente Validität*, indem hohe Korrelationen zwischen Testwertvariablen von verschiedenen Tests, die dasselbe Merkmal messen sollen, als Beleg für die Konstruktvalidität der Messungen des neuen Tests interpretiert werden. Wird z. B. ein neuer Fragebogen zur Erfassung des Traits „Extraversion“ konzipiert, so sollten die mit ihm erhobenen Testwerte hoch mit den Testwerten von anderen Verfahren korrelieren, die dasselbe Konstrukt messen.

Erst seit dem wegweisenden Artikel von Campbell und Fiske (1959) wurde auch die *diskriminante Validität* neben der konvergenten Validität als ein wesentlicher Teil der Konstruktvalidität erkannt. Dem Konzept der diskriminanten Validität liegt die Überlegung zugrunde, dass Messungen von *verschiedenen* Merkmalen (Traits) miteinander geringer korrelieren sollten als Messungen *dieselben* Merkmals. Entsprechend sollten die Testwerte eines neu entwickelten Fragebogens zur Erfassung des Traits „Extraversion“ z. B. mit einem Verfahren zur Messung des

Konvergente Validität

Diskriminante Validität

trait-fremden Konstrukts „Neurotizismus“ nicht oder zumindest geringer korrelieren als mit einem anderen Extraversionstest.

Konvergente und diskriminante Validität

- *Konvergente Validität* liegt vor, wenn Messungen eines Konstrukt, die mit verschiedenen Methoden erfasst werden, hoch miteinander korrelieren.
- *Diskriminante Validität* liegt vor, wenn Messungen verschiedener Konstrukte, die mit derselben Methode oder mit unterschiedlichen Methoden erfasst werden, nicht oder nur gering miteinander korrelieren.

25.3 Methodeneffekte

25.3.1 Erfassungsmethoden als Alternativerklärung für Zusammenhänge zwischen Traits

Trait-Methoden-Einheit

Über die konvergente und die diskriminante Validität hinausgehend ist es ein besonderes Verdienst von Campbell und Fiske (1959), erstmals thematisiert zu haben, dass die Methoden, die für die Erfassung eines Traits verwendet werden, für die Bestimmung der Validität ebenfalls eine bedeutsame Rolle spielen können. Campbell und Fiske wiesen nämlich darauf hin, dass Traits nicht unabhängig von der verwendeten Methode erfasst werden können, sondern dass sich jede Messung aus einer systematischen *Trait-Methoden-Einheit* und einem unsystematischen Fehleranteil zusammensetzt. Deshalb sollte nicht nur der gemessene Trait (z. B. Extraversion), sondern darüber hinaus auch die verwendete Erfassungsmethode (z. B. Selbsteinschätzung eines Schülers oder Fremdeinschätzung durch die Mutter oder einen Lehrer) als Bestandteil der Messung berücksichtigt werden.

Die Korrelationen von Messwerten können somit einerseits auf einen gemeinsamen Trait zurückgeführt werden, andererseits aber auch auf systematische Effekte der verwendeten Erfassungsmethode, wodurch die Richtigkeit von Schlussfolgerungen erheblich beeinträchtigt sein kann (vgl. Podsakoff et al. 2003).

Beispiel 25.1: Systematische Verzerrungen durch Methodeneffekte

Werden die Traits Depressivität und Angst von Schülern sowohl von einem Freund/ einer Freundin als auch von einer Lehrerin/einem Lehrer beurteilt, so können die Korrelationen zwischen den Traits eine systematische Verzerrung (*Bias*) aufweisen. Ein solcher Bias könnte sich dahingehend auswirken, dass

- die Korrelationen zwischen den Traits höher ausfallen, wenn die Traits von engen Bezugspersonen (Freundin/Freund) eingeschätzt werden, da diese sehr differenziert urteilen können;
- sie jedoch geringer ausfallen, wenn sie von Lehrkräften (Lehrerin/Lehrer) eingeschätzt werden, die Unterschiede zwischen den Traits hauptsächlich am Verhalten erkennen und deshalb möglicherweise weniger differenziert urteilen können.

Methodeneffekte können somit alternative Erklärungen für beobachtete Zusammenhänge zwischen Konstrukten liefern.

Methodeneffekt

Der Begriff „Methodeneffekt“ ist ein Sammelbegriff für verschiedene systematische Varianzquellen, die sich über den Trait hinausgehend auf die Validität der

25.3 · Methodeneffekte

Messung auswirken können. Früher wurden Methodeneffekte als systematische Messfehler betrachtet, die eliminiert werden sollten. Inzwischen werden Methodeneffekte aber als integrale Bestandteile der Messungen angesehen, die genauer analysiert werden sollten (vgl. Eid et al. 2016; Eid et al. 2003).

25.3.2 Typische Entstehungsquellen von Methodeneffekten

Obwohl sich Methodeneffekte prinzipiell auch auf Itemebene (z. B. Ähnlichkeit von Formulierungen oder Verwendung derselben Ratingskala) auswirken können, werden wir in den folgenden Beispielen nur Methodeneffekte auf Test- bzw. Skalenebene behandeln, die neben Messungen mit derselben Methode auch Messungen durch denselben Beurteilertyp oder Messungen in derselben Situation umfassen (vgl. Podsakoff et al. 2003; Podsakoff et al. 2012).

Typische Entstehungsquellen von Methodeneffekten

- **Messinstrument (Method):** Wird dieselbe Art von Messinstrument (z. B. ein Fragebogen oder ein sprachfreier Test) zur Messung verschiedener Merkmale (z. B. logisches Denken, räumliches Verständnis) verwendet, so können die Zusammenhänge zwischen den Merkmalen möglicherweise über- oder unterschätzt werden, je nachdem, ob der Fragebogen oder der sprachfreie Test verwendet wird.
 - Der *messmethodenspezifische Bias* kann zu einer Verzerrung der korrelativen Beziehung zwischen den Merkmalen führen.
- **Beurteiler (Informant):** Verschiedene Typen von Beurteilern (z. B. Freunde, Eltern oder Lehrkräfte) schätzen eine Person bezüglich mehrerer Merkmale (z. B. Aggressivität, Risikobereitschaft) ein. Der Typus des Beurteilers kann einen systematischen Einfluss auf die Beziehung zwischen den Merkmalen haben, da Freunde möglicherweise ihre Einschätzung aufgrund gänzlich anderer Erfahrungen in unterschiedlichen Kontexten vornehmen als Eltern oder Lehrkräfte.
 - Der *beurteilerspezifische Bias* kann zu einer Verzerrung der korrelativen Beziehung zwischen den Merkmalen führen.
- **Kontext (Occasion):** In verschiedenen Situationen (z. B. normales Wetter oder ein schwülheißen Sommertag) werden Studierende bezüglich mehrerer Merkmale (z. B. Aufmerksamkeit, Gedächtnisleistung) untersucht. Die Umgebungsbedingungen können sich systematisch auf die Beziehung zwischen den Merkmalen auswirken, da an einem schwülheißen Sommertag möglicherweise andere Leistungen erbracht werden als bei Normalwetter.
 - Der *kontextspezifische Bias* kann zu einer Verzerrung der korrelativen Beziehung zwischen den Merkmalen führen.

25.3.3 Untereinander austauschbare vs. strukturell unterschiedliche Methoden

Ein wesentlicher Unterscheidungsgesichtspunkt von Methoden besteht darin, ob sie untereinander austauschbar oder strukturell unterschiedlich sind (vgl. ► Kap. 27; Eid et al. 2008). Werden z. B. Einschätzungen von verschiedenen Beurteilern erhoben, so können sich die Beurteiler sehr ähnlich sein (z. B. mehrere Mitschüler eines zu beurteilenden Schülers), sodass es keine Rolle spielt, welche Mitschüler die Einschätzungen vornehmen.

Untereinander austauschbare Methoden

Strukturell unterschiedliche Methoden

Die Beurteiler können sich aber auch systematisch voneinander unterscheiden (Eltern, Lehrkräfte und Freunde des zu beurteilenden Kindes), sodass es wichtig ist, welche Beurteiler ausgewählt werden. Die einzelnen Methoden können typischerweise nicht einfach durch andere, strukturell unterschiedliche Methoden ersetzt werden (z. B. Einschätzung der Mutter durch Einschätzung einer Lehrkraft), da strukturell unterschiedliche Methoden andere Informationen berücksichtigen und somit jeweils unterschiedliche und „einzigartige“ Perspektiven liefern.

25.3.4 Aufdeckung von Methodeneffekten

Der Ansatz von Campbell und Fiske war und ist der Ausgangspunkt für die Entwicklung einer Vielzahl verschiedener MTMM-Analysen, die explizit Methodeneffekte berücksichtigen und deren Bezeichnung von der Art der kontrollierten Methode abhängt (vgl. Eid und Diener 2006; Koch et al. 2018).

Je nach Fragestellung kann ein *messmethodenspezifischer*, ein *beurteilerspezifischer* oder ein *kontextspezifischer Bias* dadurch aufgedeckt werden, dass immer mehrere Methoden zur Erfassung eines Konstruktts verwendet werden. Nur so kann der Methodenanteil aus den Korrelationen herausgerechnet werden.

Art der MTMM-Analyse

- Multitrait-Multimethod-Analyse (Campbell und Fiske 1959): Zur Kontrolle messmethodenspezifischer Methodeneffekte werden mehrere Messinstrumente (*Methods*) eingesetzt.
- Multitrait-Multiinformant-Analyse (Biesanz und West 2004): Zur Kontrolle beurteilerspezifischer Methodeneffekte werden mehrere Beurteiler (*Informants*) eingesetzt.
- Multitrait-Multioccasion-Analyse (Biesanz und West 2004; Steyer et al. 1999): Zur Kontrolle kontextspezifischer Methodeneffekte werden mehrere Messzeitpunkte (*Occasions*) eingesetzt.

Mehrere Erfassungsmethoden verhindern Fehlschlüsse

Beispielsweise kann ein *messinstrumentspezifischer Bias* dadurch aufgedeckt werden, dass sowohl sprachgebundene als auch sprachfreie Tests eingesetzt werden. Ein *beurteilerspezifischer Bias* kann dadurch aufgedeckt werden, dass sowohl Selbsteinschätzungen als auch Fremdeinschätzungen erhoben werden, und ein *kontextspezifischer Bias* kann dadurch aufgedeckt werden, dass die Messungen zu verschiedenen Zeitpunkten (Situationen, Messgelegenheiten) durchgeführt werden.

Auf diese Weise soll verhindert werden, dass hohe Korrelationen zwischen Trait-Methoden-Einheiten fälschlicherweise im Sinne von hohen konvergenten Validitäten interpretiert werden, wenn sie maßgeblich auf die verwendete Erfassungsmethode zurückzuführen sind.

25.4 Das MTMM-Design

Mit der MTMM-Analyse wird die Konstruktvalidität der Messungen überprüft. Konstruktvalidität liegt nur dann vor, wenn einerseits Messungen desselben Konstruktts (Traits) mit verschiedenen Methoden zu hoher Merkmalskonvergenz führen (konvergente Validität); andererseits soll aber auch eine Merkmalsdiskrimination zwischen inhaltlich unterschiedlichen Traits sowohl innerhalb einer Methode als auch zwischen verschiedenen Methoden nachgewiesen werden können (diskriminante Validität).

Anhand der MTMM-Matrix, einer systematisch zusammengesetzten Korrelationsmatrix mit den Korrelationen aller Traits, die jeweils mit allen Methoden gemessen wurden, können die konvergente und die diskriminante Validität der Messungen bestimmt und Hinweise auf Methodeneffekte erhalten werden. Dafür sollten mindestens drei Traits jeweils durch mindestens drei Methoden gemessen werden.

MTMM-Matrix

25.4.1 Aufbau der MTMM-Matrix

Das allgemeine tabellarische Schema der MTMM-Matrix für drei Traits und drei Methoden ist in Abb. 25.1 dargestellt. In der Matrix sind die Korrelationskoeffizienten der Messungen von drei Traits (1, 2, 3) anhand von drei Methoden (A, B, C) enthalten. In der Hauptdiagonalen sind die Reliabilitäten eingetragen (Reliabilitätsdiagonale) und in den Nebendiagonalen die Koeffizienten der konvergenten Validität (Validitätsdiagonalen). Oberhalb und unterhalb der Validitätsdiagonalen befinden sich die Koeffizienten der diskriminanten Validität.

Konstruktvalide Messungen sollten einen möglichst geringen methodenspezifischen Anteil aufweisen. Deutliche Methodeneffekte würden sich in überhöhten Korrelationen zwischen den verschiedenen Traits zeigen, die mit derselben Methode erfasst wurden. Sind zusätzlich die Methoden miteinander korreliert, so würde sich dies in überhöhten Korrelationen zwischen Traits, die mit unterschiedlichen Methoden erfasst wurden, zeigen. Ein Beispiel wäre die Einschätzung des Verhaltens von Kindern sowohl durch die Mutter als auch durch den Vater. Es kann erwartet werden, dass die Beurteiltypen (Methoden) „Mutter“ und „Vater“ zumindest bei manchen Verhaltensweisen eine ähnliche Sichtweise in Bezug auf die Kinder haben, was zu überhöhten Korrelationen der Traits, die mit den beiden Methoden gemessen wurden, führen könnte.

Trait		Methode A			Methode B			Methode C		
		1	2	3	1	2	3	1	2	3
Methode A	1	<i>Rel(A1)</i>								
	2	<i>r_{A2A1}</i>	<i>Rel(A2)</i>							
	3	<i>r_{A3A1}</i>	<i>r_{A3A2}</i>	<i>Rel(A3)</i>						
Methode B	1	<i>r_{B1A1}</i>	<i>r_{B1A2}</i>	<i>r_{B1A3}</i>	<i>Rel(B1)</i>					
	2	<i>r_{B2A1}</i>	<i>r_{B2A2}</i>	<i>r_{B2A3}</i>	<i>r_{B2B1}</i>	<i>Rel(B2)</i>				
	3	<i>r_{B3A1}</i>	<i>r_{B3A2}</i>	<i>r_{B3A3}</i>	<i>r_{B3B1}</i>	<i>r_{B3B2}</i>	<i>Rel(B3)</i>			
Methode C	1	<i>r_{C1A1}</i>	<i>r_{C1A2}</i>	<i>r_{C1A3}</i>	<i>r_{C1B1}</i>	<i>r_{C1B2}</i>	<i>r_{C1B3}</i>	<i>Rel(C1)</i>		
	2	<i>r_{C2A1}</i>	<i>r_{C2A2}</i>	<i>r_{C2A3}</i>	<i>r_{C2B1}</i>	<i>r_{C2B2}</i>	<i>r_{C2B3}</i>	<i>r_{C2C1}</i>	<i>Rel(C2)</i>	
	3	<i>r_{C3A1}</i>	<i>r_{C3A2}</i>	<i>r_{C3A3}</i>	<i>r_{C3B1}</i>	<i>r_{C3B2}</i>	<i>r_{C3B3}</i>	<i>r_{C3C1}</i>	<i>r_{C3C2}</i>	<i>Rel(C3)</i>

Abb. 25.1 Tabellarisches Schema der MTMM-Matrix für drei Traits (1, 2, 3) und drei Methoden (A, B, C). Die durchgezogenen Striche markieren Monomethod-Blöcke, die gestrichelten Linien Heteromethod-Blöcke; *Rel* bezeichnet die Reliabilitätskoeffizienten in der Hauptdiagonalen („Reliabilitätsdiagonale“ innerhalb der Monomethod-Blöcke), die fettgedruckten Koeffizienten bezeichnen die konvergenten Validitäten in den Nebendiagonalen („Validitätsdiagonale“ innerhalb der Heteromethod-Blöcke)

25.4.2 Unterscheidung und Interpretation der Blöcke und Koeffizienten

In der MTMM-Matrix werden zwei Arten von Blöcken unterschieden, die als Monomethod- und Heteromethod-Blöcke bezeichnet werden. Außerdem werden vier verschiedene Arten von Koeffizienten unterschieden: die Monotrait- und die Heterotrait-Korrelationskoeffizienten, die jeweils unter der Monomethod- bzw. der Heteromethod-Bedingung erfasst werden.

■ ■ Blöcke

Monomethod-Blöcke

- Die *Monomethod-Blöcke*, zur Verdeutlichung in ▶ Abb. 25.1 mit durchgezogenen Strichen markiert, enthalten die Korrelationen zwischen den verschiedenen Traits, die jeweils mit der *gleichen* Methode erfasst wurden.

Heteromethod-Blöcke

- Die *Heteromethod-Blöcke*, in ▶ Abb. 25.1 gestrichelt markiert, enthalten die Korrelationen zwischen den untersuchten Traits, die jeweils mit *verschiedenen* Methoden erfasst wurden.

■ ■ Koeffizienten

Monotrait-Monomethod-Koeffizienten

- Die *Monotrait-Monomethod-Koeffizienten* in der Hauptdiagonalen der Matrix beinhalten die Reliabilitätskoeffizienten (▶ Kap. 14) der Messinstrumente, weshalb die Diagonale auch *Reliabilitätsdiagonale* genannt wird. So bezeichnet z. B. $Rel(A1)$ die Reliabilität von Trait 1, gemessen mit Methode A. Nach Campbell und Fiske (1959) sollten die Reliabilitätskoeffizienten möglichst hoch und nicht zu unterschiedlich sein. Diese Forderung kann jedoch nur selten eingehalten werden und wird von verschiedenen Autoren als unrealistisch angesehen.

Monotrait-Heteromethod-Koeffizienten

- Die *Monotrait-Heteromethod-Koeffizienten* in den Nebendiagonalen beinhalten die konvergenten Validitäten der Traits, weshalb die Nebendiagonalen auch *Validitätsdiagonalen* genannt werden. So bezeichnet z. B. r_{B1A1} die Korrelation (*konvergente Validität*) von Trait 1, gemessen mit den Methoden B und A; r_{C3B3} die konvergente Validität von Trait 3, gemessen mit den Methoden C und B.

Heterotrait-Monomethod-Koeffizienten

- Die *Heterotrait-Monomethod-Koeffizienten*, angeordnet in Dreiecksmatrizen der Monomethod-Blöcke unterhalb der Reliabilitätsdiagonalen, beinhalten die Korrelationen zwischen unterschiedlichen Traits, die jeweils mit derselben Methode gemessen wurden. So bezeichnet z. B. r_{A2A1} die Korrelation (*diskriminante Validität*) zwischen Trait 2 und Trait 1, gemessen mit Methode A.

Heterotrait-Heteromethod-Koeffizienten

- Die *Heterotrait-Heteromethod-Koeffizienten*, angeordnet in Dreiecksmatrizen der Heteromethod-Blöcke oberhalb und unterhalb der Validitätsdiagonalen, beinhalten die Korrelationen zwischen unterschiedlichen Traits, die jeweils mit unterschiedlichen Methoden gemessen wurden. So bezeichnet z. B. r_{B1A2} die Korrelation von Trait 1, gemessen mit Methode B, und Trait 2, gemessen mit Methode A; somit handelt es sich hierbei um Korrelationen zwischen verschiedenen Traits, die mit verschiedenen Methoden erfasst werden. Zusammen mit den Korrelationen zwischen unterschiedlichen Traits, die jeweils mit derselben Methode gemessen werden (Heterotrait-Monomethod-Koeffizienten, z. B. r_{A2A1}) sind sie Indikatoren der *diskriminanten Validität*. Die Heterotrait-Heteromethod-Koeffizienten bezeichnen aber die um den Einfluss der Methoden bereinigten diskriminanten Validitäten.

Zur Analyse der MTMM-Matrix haben sich zwei Methoden etabliert, die korrelationsbasierte Analyse, die in ▶ Abschn. 25.5 vorgestellt wird, und die konfirmatorische Faktorenanalyse (CFA), die sich in den letzten Jahren als statistisches Verfahren zur MTMM-Analyse durchgesetzt hat; sie werden in ▶ Abschn. 25.6 und 25.7 erläutert. Weitere Methoden finden sich z. B. bei Eid et al. (2006) sowie in Koch et al. (2018).

25.5 Korrelationsbasierte Analyse der MTMM-Matrix

Die korrelationsbasierte Analyse (Campbell und Fiske 1959) ist ein deskriptives Verfahren, das dazu eingesetzt werden kann, einen Überblick über die Struktur der Daten zu erhalten. Erwartet wird, dass die Korrelationen zwischen Messvariablen, die denselben Trait mit unterschiedlichen Methoden (konvergente Validität) erfassen, deutlich höher ausfallen als die Korrelationen zwischen Messvariablen, die unterschiedliche Traits erfassen (diskriminante Validität).

25.5.1 Nachweiskriterien für die konvergente und diskriminante Validität

Zum Nachweis der Konstruktvalidität nach den Campell-Fiske-Kriterien werden die Korrelationen in der Korrelationsmatrix durch paarweise Vergleiche dahingehend beurteilt, ob die Kriterien der konvergenten und der diskriminanten Validität erfüllt sind. Werden nicht alle Kriterien vollständig erfüllt, so spricht dies trotzdem nicht unbedingt gegen die Konstruktvalidität. Allerdings gibt es für die Beurteilung von Abweichungen keine verbindlichen Regeln, sodass es dem Beurteiler überlassen bleibt zu entscheiden, ob die Kriterien in hinreichendem Maße erfüllt wurden oder ob die Konstruktvalidität insgesamt infrage gestellt bzw. in bestimmten Teilspekten eingeschränkt werden muss.

Zum Nachweis der *konvergenten Validität* sollte folgendes Kriterium erfüllt sein:

1. Die Korrelationen von Messungen *eines Traits*, gemessen mit jeweils zwei *verschiedenen Methoden* (Monotrait-Heteromethod-Koeffizienten), in den Validitätsdiagonalen sollen statistisch signifikant von null verschieden und hoch sein. Ein absolutes Maß für die Höhe der Korrelationen wird von Campbell und Fiske (1959) nicht vorgegeben. Gelingt der Nachweis der konvergenten Validität nicht, so muss davon ausgegangen werden, dass mit den unterschiedlichen Methoden verschiedene Konstrukte gemessen werden oder dass zumindest eine der Methoden keine validen Messungen liefert.

Nachweis der konvergenten Validität

Zum Nachweis der *diskriminanten Validität* sollten drei Kriterien erfüllt sein:

1. *Verschiedene Traits*, die durch *dieselbe Methode* erfasst werden (Heterotrait-Monomethod-Koeffizienten, in Abb. 25.1 in den Dreiecksmatrizen unterhalb der Reliabilitätsdiagonalen), sollen miteinander geringer korrelieren als Messungen desselben Traits mit verschiedenen Methoden (Monotrait-Heteromethod-Koeffizienten in den Validitätsdiagonalen).
2. Korrelationen zwischen *verschiedenen Traits*, die durch *verschiedene Methoden* erfasst werden (Hetrotrait-Heteromethod-Koeffizienten, in Abb. 25.1 in den gestrichelten Dreiecksmatrizen über und unter den Validitätsdiagonalen), sollen niedriger sein als die konvergenten Validitätskoeffizienten in den Validitätsdiagonalen. Trifft dieses Kriterium nicht zu, so diskriminieren die Messungen nicht zwischen inhaltlich oder theoretisch verschiedenen Konstrukten. Ursache hierfür könnte z. B. ein gemeinsamer Faktor sein, der mehrere Traits in der MTMM-Matrix umfasst.
3. Die *Muster der Korrelationskoeffizienten* sollen sowohl innerhalb einer Methode (Dreiecksmatrizen unterhalb der Reliabilitätsdiagonalen) als auch zwischen den Methoden (Dreiecksmatrizen über und unter den Validitätsdiagonalen) in etwa gleich sein. Ein exaktes Kriterium für die Übereinstimmung wird nicht vorgegeben, je nach Autor werden unterschiedliche Auswertungsempfehlungen genannt. Am häufigsten wird überprüft, ob die Rangreihe der Korrelationen über alle Teilmatrizen hinweg konstant ist oder ob die Vorzeichen der Korrelationen in allen Heterotrait-Teilmatrizen (sowohl in den Monomethod-

Nachweis der diskriminanten Validität

als auch in den Heteromethod-Matrizen) übereinstimmen. Erhöhte Korrelationen verschiedener Traits innerhalb einer Methode können ebenso wie erhöhte Korrelationen verschiedener Traits zwischen unterschiedlichen Methoden auf Methodeneffekte hinweisen.

25.5.2 Empirisches Anwendungsbeispiel

Nachfolgend sollen die Ergebnisse einer MTMM-Analyse vorgestellt werden, bei der die Auswertung auf einer Multitrait-Multiinformant-Matrix beruht (Burns et al. 2013). Verwendet wird hier eine Teilstichprobe von $N = 801$ spanischen Grundschulkindern mit einem durchschnittlichen Alter von 8.3 Jahren. Die Kinder wurden jeweils von ihrer Mutter, ihrem Vater und einem Lehrer (bzw. einer Lehrerin) hinsichtlich drei Arten von Verhaltensstörungen unter Verwendung entsprechender Skalen beurteilt, und zwar Unaufmerksamkeit (UN), Hyperaktivität (HA) und Trotzverhalten (TV), die zu den Symptomen der ADHS (Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung) und der ODD (Oppositional Defiant Disorder, oppositionelles Trotzverhalten) zählen. In die MTMM-Analyse gingen somit neun Messvariablen (Skalensummenwerte) zur Erfassung von drei Arten von Verhaltensstörungen (Traits) anhand von drei Beurteilertypen (Methoden) ein (nähere Informationen finden sich bei Burns et al. 2013).

Die Korrelationen zwischen den neun manifesten Variablen sind in der MTMM-Matrix (► Abb. 25.2) aufgeführt. Die einzelnen Skalen weisen bei jedem der drei Beurteilertypen (Informants) hohe Reliabilitäten auf (alle ω -Koeffizienten in der Reliabilitätsdiagonalen $> .90$, zur Berechnung s. ► Kap. 15).

Nachweis der konvergenten Validität

■ ■ Nachweis der konvergenten Validität:

1. Im vorliegenden Beispiel sind alle neun Reliabilitätskoeffizienten mit Werten zwischen .29 und .82 signifikant von null verschieden und in ihrer Größe bedeutsam, sodass die konvergente Validität als nachgewiesen gelten kann. Besonders hoch fallen die konvergenten Validitäten zwischen den Methoden „Mutter“ und „Vater“ aus ($r \geq .72$).

		Trait	Methode A: Mutter			Methode B: Vater			Methode C: Lehrer		
			1-UN	2-HA	3-TV	1-UN	2-HA	3-TV	1-UN	2-HA	3-TV
Methode A: Mutter	1-UN	(.)94									
	2-HA	.63	(.95)								
	3-TV	.58	.61	(.97)							
Methode B: Vater	1-UN	.82			.55	.50	(.94)				
	2-HA	.56	.79		.53		.65	(.95)			
	3-TV	.45	.52	.72			.60	.66	(.97)		
Methode C: Lehrer	1-UN	.44			.35	.25	.42			.37	.20
	2-HA	.28	.47		.28		.29	.44		.25	.60
	3-TV	.25	.38	.32			.26	.33	.29		.51

■ Abb. 25.2 Tabellarisches Schema der empirischen Multitrait-Multiinformant-Matrix mit drei Traits („Unaufmerksamkeit“, „Hyperaktivität“, „Trotzverhalten“) und drei Methoden (Beurteilertypen: „Mutter“, „Vater“, „Lehrer“; vgl. Burns et al. 2013). Traits: 1-UN = Unaufmerksamkeit, 2-HA = Hyperaktivität, 3-TV = Trotzverhalten; Methoden (Beurteilertypen): A = Mutter, B = Vater, C = Lehrer. Die Reliabilitätskoeffizienten (McDonalds Omega) sind in der Hauptdiagonalen aufgeführt, die konvergenten Validitäten in den Nebendiagonalen

■■ Nachweis der diskriminanten Validität:

1. Die Korrelationen zwischen verschiedenen Traits, die durch die gleiche Methode gemessen werden (Heterotrait-Monomethod-Koeffizienten), sollten niedriger sein als die entsprechenden konvergenten Validitätskoeffizienten dieser Traits. Somit sollte z. B. die Korrelation zwischen HA und UN, gemessen mit der Methode „Mutter“ ($r_{A2A1} = .63$), niedriger sein als die entsprechenden vier Korrelationskoeffizienten der Validitätsdiagonalen, die Messungen der Methode „Mutter“ beinhalten. Die Korrelation zwischen HA und UN, gemessen jeweils mit der Methode „Mutter“, ist mit $r_{A2A1} = .63$ erwartungsgemäß niedriger als die Validitätskoeffizienten $r_{B1A1} = .82$ und $r_{B2A2} = .79$ (Methoden „Vater“ und „Mutter“), jedoch entgegen der Erwartung höher als $r_{C1A1} = .44$ und $r_{C2A2} = .47$ (Methoden „Lehrer“ und „Mutter“). Von diesen vier Vergleichen erfüllen somit nur zwei das erste Kriterium zum Nachweis der diskriminanten Validität. Insgesamt ist das erste Kriterium der diskriminanten Validität nur teilweise erfüllt, da von den 36 Vergleichen nur 12 den Erwartungen entsprechen.
2. Zusätzlich sollten die Koeffizienten der Heterotrait-Heteromethod-Blöcke (oberhalb und unterhalb der Validitätsdiagonalen) niedriger sein als ihre entsprechenden Validitätskoeffizienten. Es werden also die Korrelationen auf der Validitätsdiagonalen mit den Korrelationen derselben Zeile oder Spalte im gleichen Heteromethod-Block verglichen. Dies trifft auf fast alle 36 Vergleiche (12 pro Block) zu. Beispielsweise ist die Korrelation zwischen HA, gemessen mit der Methode „Vater“, und UN, gemessen mit der Methode „Mutter“, mit $r_{B2A1} = .56$ niedriger als die Korrelation zwischen UN, gemessen mit den Methoden „Vater“ und „Mutter“ ($r_{B1A1} = .82$). Dies trifft ebenfalls auf die Koeffizienten r_{B3A1} und r_{B1A2} sowie r_{B1A3} zu (.45, .55, .50 < .82). Insgesamt erfüllen die Koeffizienten damit das zweite Kriterium der diskriminanten Validität. Es gibt nur zwei Abweichungen: Lediglich die Koeffizienten $r_{C3A2} = .38$ (Korrelation zwischen TV, gemessen mit der Methode „Lehrer“, und HA, gemessen mit der Methode „Mutter“), sowie $r_{C3B2} = .33$ (Korrelation zwischen TV, gemessen mit der Methode „Lehrer“, und HA, gemessen mit der Methode „Vater“), sind jeweils höher als zwei korrespondierende Validitätskoeffizienten ($r_{C3A3} = .32$ bzw. $r_{C3B3} = .29$).
3. Die Muster der Merkmalsinterkorrelationen sind sowohl innerhalb als auch zwischen den Methoden (Beurteilern) in etwa gleich und erfüllen damit im Wesentlichen das dritte Kriterium der diskriminanten Validität. Die niedrigsten Korrelationen finden sich immer zwischen den Traits TV und UN, sowohl innerhalb der Monomethod-Blöcke als auch innerhalb der Heteromethod-Blöcke. Von insgesamt neun Dreieckmatrizen stimmt die Rangreihe der Koeffizienten innerhalb von fünf Dreieckmatrizen überein (HA/UN > TV/HA > TV/UN), während in vier Dreieckmatrizen TV/HA und HU/UN ihre Rangplätze tauschen, sodass die Rangreihe dort TV/HA > HA/UN > TV/UN lautet.

Wie bereits aus empirischen Studien bekannt ist, korrelieren auch hier die drei Konstrukte UN, HA und TV positiv und substantiell miteinander. Die höheren Korrelationen innerhalb der Monomethod-Blöcke im Vergleich zu den Heteromethod-Blöcken könnten – wie auch die teilweise Nichterfüllung des ersten Kriteriums der diskriminanten Validität (s. o.) – auf Methodeneffekte hinweisen, die aber anhand der korrelationsbasierten Analyse nicht quantifiziert werden können.

Nachweis der diskriminanten Validität

Insgesamt gibt die Analyse der MTMM-Matrix nach den Campbell-Fiske-Kriterien deutliche Hinweise auf die konvergente Validität der drei untersuchten Traits, während die Kriterien der diskriminanten Validität nur zum Teil erfüllt werden konnten. Die Heterotrait-Monomethod-Koeffizienten in den Dreiecksmatrizen unterhalb der Reliabilitätsdiagonalen, die durchweg höher sind als die Korrelationen zwischen den Traits, gemessen mit unterschiedlichen Methoden (Heterotrait-Heteromethod-

Koeffizienten), deuten darauf hin, dass hier Methodeneffekte vorliegen könnten, die zu einer Verfälschung der Ergebnisse führen.

25.5.3 Kritik an der korrelationsbasierten MTMM-Analyse

Die von Campbell und Fiske (1959) vorgeschlagene Auswertung auf Korrelationsebene erfolgt über einfache Häufigkeitsauszählungen bzw. viele Einzelvergleiche von Korrelationskoeffizienten und ist ein geeignetes Verfahren, um einen Überblick über die Datenstruktur zu gewinnen. Campbell und Fiske kommt damit das Verdienst zu, durch die Einführung ihres MTMM-Ansatzes das frühere nur auf dem Konvergenzprinzip aufbauende Validierungskonzept um den Aspekt der diskriminanten Validität und der Methodeneffekte erweitert zu haben.

Kritikpunkte

■ ■ Kritikpunkte

Mit der deskriptiven Auswertung auf der Korrelationsebene sind jedoch verschiedene Probleme verbunden:

1. Es werden einfache Häufigkeitsauszählungen bzw. viele Einzelvergleiche von Korrelationskoeffizienten vorgenommen. Eine Häufigkeitsauszählung von Korrelationen kann nur einen groben Überblick über die Struktur der Variablen geben. Es handelt sich somit um kein zufallskritisches Vorgehen, denn die Korrelationskoeffizienten werden lediglich ohne Berücksichtigung der zugehörigen Konfidenzintervalle als „größer“ oder „kleiner“ beurteilt.
2. Schwerwiegend ist auch das Problem, dass die Korrelationsmatrix selbst auf manifesten, messfehlerbehafteten Testwerten basiert. Vergleiche der manifesten Korrelationen sind insbesondere dann schwierig, wenn die Messvariablen in der MTMM-Matrix sehr unterschiedliche Reliabilitäten aufweisen. In diesem Fall könnten scheinbare Unterschiede der Korrelationskoeffizienten (oder die scheinbare Gleichheit solcher Koeffizienten) allein durch Messfehler verursacht und somit irreführend sein.
3. Die korrelationsbasierte MTMM-Analyse bietet zwar – wie in obigem Beispiel – Anhaltspunkte für das Vorhandensein von Methodeneffekten, die Größe dieser Effekte kann aber nicht separiert von konvergenten und diskriminanten Validitätskoeffizienten bestimmt werden, sodass die Methodeneffekte in den Korrelationskoeffizienten enthalten bleiben.
4. Es bleibt dem Anwender zu einem großen Teil selbst überlassen, ob er trotz Verletzung eines Kriteriums die konvergente und diskriminante Validität als nachgewiesen annehmen will oder nicht, da es keine exakten Entscheidungsregeln gibt. Die Auswertung ist somit subjektiv und sollte eher als Heuristik verstanden werden (vgl. Campbell und Fiske 1959).

25.6 Faktorenanalytische Ansätze: Klassische CFA-MTMM-Modelle

Wegen der aufgeführten Kritikpunkte an der korrelationsbasierten MTMM-Analyse, bei der viele wichtige Fragen offenbleiben, hat sich in den letzten Jahren die Konfirmatorische Faktorenanalyse (CFA, vgl. ► Kap. 24) als statistisches Verfahren zur Analyse von MTMM-Matrizen durchgesetzt. Das MTMM-Modell als faktorenanalytisches Modell geht auf Jöreskog (1971) zurück, der für jeden Trait und jede Methode jeweils einen latenten Faktor spezifiziert hat. Im Gegensatz zur korrelationsbasierten Analyse können mit der CFA Trait-, Methoden- und Messfehleranteile der Messungen voneinander separiert werden. Ein weiterer Vorteil besteht darin, dass die konvergente und die diskriminante Validität auch statistisch geprüft werden können. Die korrelationsbasierte Analyse behält trotzdem ihre

Berechtigung: Als deskriptives Verfahren kann sie auch weiterhin vor der faktorenanalytischen MTMM-Analyse eingesetzt werden, um einen ersten Überblick über die Struktur der Daten zu erhalten.

Mit verschiedenen Modellen der CFA können MTMM-Matrizen hinsichtlich konvergenter und diskriminanter Validität sowie Methodeneffekten untersucht werden (vgl. Eid und Diener 2006; Höfling et al. 2009; Kenny und Kashy 1992). CFA-Modelle haben den Vorteil, dass sie nicht nur eine Trennung von Trait-, Methoden- und Messfehleranteilen ermöglichen, sondern auch eine Überprüfung der Gültigkeit der zugrunde liegenden Modellannahmen erlauben. So kann z. B. die Eindimensionalität der einzelnen Messungen oder die Unkorreliertheit der verschiedenen Methodenfaktoren überprüft werden. Außerdem besteht die Möglichkeit, die latenten Trait-Faktoren mit Kriterien in Beziehung zu setzen, sodass zusätzlich zur Konstruktvalidität auch die Kriteriumsvalidität (vgl. ► Kap. 21) auf latenter Ebene überprüft werden kann.

Je nach den zugrunde liegenden Hypothesen können mit der CFA unterschiedliche Modelle spezifiziert werden. Zur Vorbereitung auf die neueren CFA-Modelle (► Abschn. 25.7) sollen hier aus historischen Gründen auch klassische CFA-MTMM-Modelle vorgestellt werden, obwohl ihre Annahmen und ursprünglichen Interpretationen inzwischen als überholt gelten.

Klassische CFA-MTMM-Modelle bestehen in der Regel aus mindestens drei Traits und drei Erhebungsmethoden, die durch mindestens neun ($3 \cdot 3$) Indikatoren (Trait-Methoden-Einheiten) gemessen werden (Y_{11} bis Y_{33}). Entsprechend der Annahmen der klassischen CFA-MTMM-Analyse soll jeder Indikator jeweils auf einem Trait-Faktor und auf einem Methodenfaktor laden, nicht jedoch auf den anderen Faktoren. Zwischen der Gruppe der Trait-Faktoren und der Gruppe der Methodenfaktoren sind keine Beziehungen erlaubt, da solche Modelle nicht identifiziert wären (s. dazu ► Kap. 24, ► Abschn. 24.2.6).

Vorteile von CFA-Modellen

Klassische CFA-MTMM-Modelle

Verschiedene klassische CFA-MTMM-Modelle

Abhängig davon, ob die Faktoren innerhalb einer Faktengruppe (Traits bzw. Methoden) miteinander korrelieren oder nicht, können verschiedene klassische CFA-Modelle unterschieden werden (vgl. Marsh und Grayson 1995; Widaman 1985):

- CTCM-Modell (Correlated-Trait-Correlated-Method-Modell):
CFA-Modell mit korrelierten Traits und korrelierten Methoden
- CTUM-Modell (Correlated-Trait-Uncorrelated-Method-Modell):
CFA-Modell mit korrelierten Traits und unkorrelierten Methoden
- UTCM-Modell (Uncorrelated-Trait-Correlated-Method-Modell):
CFA-Modell mit unkorrelierten Traits und korrelierten Methoden
- UTUM-Modell (Uncorrelated-Trait-Uncorrelated-Method-Modell):
CFA-Modell mit unkorrelierten Traits und unkorrelierten Methoden

25.6.1 Das CTCM-Modell

In dem klassischen CTCM-Modell korrelieren die Trait- und die Methodenfaktoren jeweils untereinander, was durch Doppelpfeile in ► Abb. 25.3 symbolisiert wird. Das CTCM-Modell weist in der Praxis jedoch häufig Identifikations- und Schätzprobleme (z. B. Konvergenzprobleme oder negative Varianzschätzungen) auf, die in Modellen mit unkorrelierten Methoden oder Traits in der Regel nicht auftreten. Zudem sind die Trait- und Methodenfaktoren in diesem Modell nur schwer zu interpretieren, sodass neuere Entwicklungen dieses klassischen Modells inzwischen ersetzt haben (vgl. ► Kap. 27; Eid und Diener 2006; Koch et al. 2018).

Identifikations- und Schätzprobleme

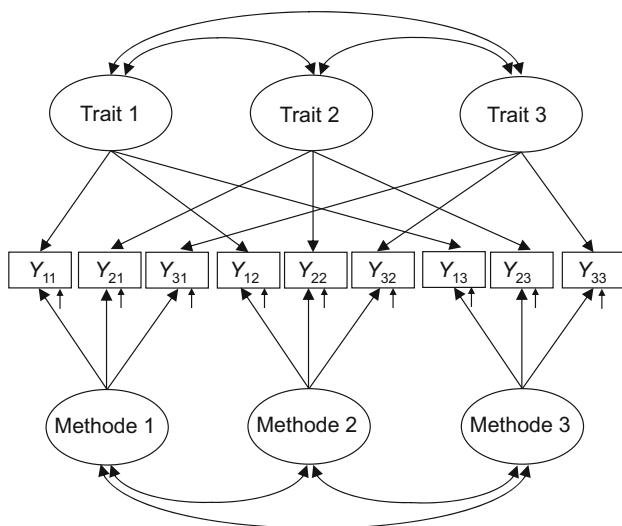


Abb. 25.3 Das CTCM-Modell zur Analyse einer MTMM-Matrix mit neun Indikatorvariablen, Y_{11} bis Y_{33} , drei Traits (*Trait 1*, *Trait 2*, *Trait 3*) und drei Methoden (*Methode 1*, *Methode 2*, *Methode 3*). Die Traits und die Methoden sind jeweils untereinander korreliert, was durch *gebogene Doppelpfeile* symbolisiert wird; die Messfehler sind durch *kleine Pfeile* nur angedeutet

Dennoch soll das CTCM-Modell nachfolgend ausführlich behandelt werden, weil es einerseits den Vorstellungen von Campbell und Fiske (1959) bezüglich der Trennung von Trait- und Methodenfaktoren am ehesten entspricht, andererseits aber auch die Grundlage für das CTC($M - 1$)-Modell bildet, das später ausführlicher behandelt wird. Durch Fixieren einer oder mehrerer Korrelationen zwischen Methoden oder Traits auf null kann das CTCM zudem leicht in andere klassische CFA-MTMM-Modelle (CTUM, UTUM, UTCM) überführt werden.

Vermeintlicher Vorteil des CTCM-Modells

Ein Vorteil des CTCM-Modells wurde ursprünglich darin gesehen, dass die Trait-Varianz und die Methodenvarianz der Indikatoren vermeintlich voneinander separiert und unabhängig von den Messfehlern geschätzt werden können. Konvergente Validität sollte anhand von hohen Faktorladungen auf den Trait-Faktoren, diskriminante Validität anhand von geringen Korrelationen zwischen den Traits und der Einfluss der Methoden anhand der Höhe der Faktorladungen auf den Methodenfaktoren nachgewiesen werden können. Somit wurde angenommen, dass sich jede Indikatorvariable entsprechend der Vorstellungen von Campbell und Fiske aus einem Trait-Anteil, einem Methodenanteil und einem unsystematischen Messfehler zusammensetzt. Es ist aber fraglich, ob Trait- und Methodenanteile durch dieses Modell tatsächlich voneinander getrennt werden können.

25.6.2 Empirisches Anwendungsbeispiel

Zum Vergleich mit dem korrelationsbasierten Ansatz (► Abschn. 25.5) soll die empirische Multitrait-Multiinformant-Matrix (► Abb. 25.2) nun faktorenanalytisch mit dem CTCM-Modell überprüft werden (vgl. ► Abb. 25.3). Dieses Modell wird hier gewählt, weil einerseits – der Theorie zufolge – die Traits miteinander korrelieren sollten, andererseits aber auch die Methoden, da systematische Beziehungen zwischen den Beurteiltypen „Mutter“, „Vater“ und „Lehrer“ angenommen werden können. Als Schätzmethode wird die robuste Maximum-Likelihood-Methode (MLR-Methode) verwendet, die von verschiedenen Programmen, u. a. dem Programm *Mplus* (Muthén und Muthén 2017), zur Verfügung gestellt wird. Diese Schätzmethode ermöglicht es, einerseits die Standardfehler und den Modelltest

hinsichtlich der Nichtnormalverteilung der Variablen zu korrigieren, andererseits fehlende Werte adäquat zu berücksichtigen (► Kap. 24).

■■ Ergebnisse

Überprüfung der Modellgüte Die Überprüfung der Güte des gesamten Modells zeigt (vgl. Schermelleh-Engel et al. 2003), dass eine gute Passung zwischen dem Modell und den Daten besteht. Der χ^2 -Wert, der möglichst klein und nicht signifikant sein sollte (vgl. auch ► Kap. 24), ist $\chi^2(12) = 27.40, p = .01$, und deutet mit $\chi^2/12 = 2.28$ zumindest auf einen zufriedenstellenden Modellfit hin. Hier scheint sich die große Stichprobe auszuwirken, da die Stichprobengröße in den χ^2 -Wert eingeht. Die deskriptiven Gütemaße sprechen sehr deutlich für einen guten Modellfit mit einem Root Mean Square Error of Approximation (RMSEA) von .04 (dieser Wert sollte $< .05$ liegen) und zwei weiteren Gütekriterien, dem Comparative Fit Index (CFI) und dem Tucker Lewis Index (TLI), die für einen guten Modellfit nahe bei eins liegen sollten: CFI = .99 und TLI = .98.

Modellgüte

Konsistenz und Methodenspezifität

Zur Beurteilung der Varianzanteile der manifesten Variablen, die auf den Trait und die Methode zurückgeführt werden können, kann eine Dekomposition der Varianzen in einzelne Komponenten vorgenommen werden (Eid und Schmidt 2014; vgl. auch Eid et al. 1994; Geiser et al. 2008).

Da es sich um kongenerische Messungen handelt (vgl. ► Kap. 13 und 24), setzt sich jede manifeste Variable Y_{jk} additiv aus einem Trait- und einem Methodenanteil sowie dem Messfehler zusammen, wobei die Trait- und Methodenfaktoren noch mit den entsprechenden Faktorladungen gewichtet werden:

$$Y_{jk} = \lambda_{Tjk} T_j + \lambda_{Mjk} M_k + \varepsilon_{jk} \quad (25.1)$$

Dabei ist $j = 1, 2, 3$ der Laufindex der Traits und $k = 1, 2, 3$ der Laufindex der Methoden; T = Trait, M = Methode und ε = Messfehler, λ_T = Ladung auf dem Traitfaktor, λ_M = Ladung auf dem Methodenfaktor.

Analog lässt sich nun die Varianz jeder manifesten Variablen Y_{jk} in zwei systematische Varianzanteile zerlegen, die Trait-Varianz und die Methodenvarianz, sowie in einen unsystematischen Fehlervarianzanteil:

$$\text{Var}(Y_{jk}) = \lambda_{Tjk}^2 \text{Var}(T_j) + \lambda_{Mjk}^2 \text{Var}(M_k) + \text{Var}(\varepsilon_{jk}) \quad (25.2)$$

Um die Anteile der Trait- und der Methodenvarianz jeder manifesten Variablen einfacher beurteilen zu können, werden diese an der beobachteten Varianz normiert (vgl. Eid und Schmidt 2014, S. 347; Eid et al. 1994; Newsom 2015). Es resultieren dann zwei Varianzkomponenten, die als *Konsistenz* (*Con*) bzw. als *Methodenspezifität* (*MSpe*) bezeichnet werden:

$$\begin{aligned} \text{Con}(Y_{jk}) &= \lambda_{Tjk}^2 \text{Var}(T_j) / \text{Var}(Y_{jk}) \\ \text{MSpe}(Y_{jk}) &= \lambda_{Mjk}^2 \text{Var}(M_k) / \text{Var}(Y_{jk}) \end{aligned} \quad (25.3)$$

Die Konsistenz gibt an, wie viel Varianz einer manifesten Variablen durch den jeweiligen Trait erklärt wird. Die Methodenspezifität gibt dagegen an, wie viel Varianz einer manifesten Variablen auf spezifische Effekte der jeweiligen Erfassungsmethode (im Beispiel: Beurteiler) zurückgeht.

Konsistenz und Methodenspezifität

Konsistenz (*Con*) und Methodenspezifität (*MSpe*)

Reliabilität als Summe zweier Varianzkomponenten

Die Summe dieser beiden systematischen Varianzkomponenten ergibt die Reliabilität einer beobachteten Variablen, d. h. die gesamte erklärte Varianz der betreffenden Variablen, relativiert an der Gesamtvarianz. Die Reliabilität setzt sich somit hier aus zwei Komponenten zusammen, aus der Konsistenz und aus der Methodenspezifität:

$$\begin{aligned} Rel(Y_{jk}) &= [\lambda_{Tjk}^2 Var(T_j) + \lambda_{Mjk}^2 Var(M_k)] / Var(Y_{jk}) \\ &= Con(Y_{jk}) + MSpe(Y_{jk}) \end{aligned} \quad (25.4)$$

Wurden die manifesten und latenten Variablen standardisiert, so kann die Reliabilität der manifesten standardisierten Variablen Z_{jk} einfach durch Addition der quadrierten standardisierten Faktorladungen $\lambda^{(s)2}$ der Trait- und der Methodenkomponente bestimmt werden:

$$Rel(Z_{jk}) = \lambda_{Tjk}^{(s)2} + \lambda_{Mjk}^{(s)2} = Con(Z_{jk}) + MSpe(Z_{jk}) \quad (25.5)$$

Faktorladungen und Zusammensetzung der Reliabilität In □ Tab. 25.1 sind für jeden Indikator die Faktorladungen auf den Trait- und Methodenfaktoren sowie die aus Konsistenz und Methodenspezifität zusammengesetzte Reliabilität übersichtlich zusammengestellt.

■ **Tabelle 25.1** Ergebnisse des CTCM-Modells mit Faktorladungen der neun manifesten Variablen Y_{11} bis Y_{33} auf drei Trait- und drei Methodenfaktoren sowie Aufteilung der Reliabilität in Konsistenz und Methodenspezifität

	Faktorladungen							Varianzkomponenten		
	Trait			Methode						
Indikator	UN	HA	TV	Mutter	Vater	Lehrer	Con	MSpe	Rel	
Y_{11}	.86			.48			.74	.23	.97	
Y_{21}		.53		.82			.28	.67	.95	
Y_{31}			.82	.50			.66	.25	.92	
Y_{12}	.71				.59		.51	.35	.85	
Y_{22}		.74			.67		.54	.45	.99	
Y_{32}			.59		.68		.35	.46	.81	
Y_{13}	.37					.63	.14	.39	.53	
Y_{23}		.23				.87	.05	.75	.80	
Y_{33}			.17			.77	.03	.60	.63	
	Korrelierte Traits			Korrelierte Methoden						
	1.0			1.0						
	.52	1.0		.72	1.0					
	.48	.47	1.0	.46	.41	1.0				

Anmerkung: Traits: UN = Unaufmerksamkeit, HA = Hyperaktivität, TV = Trotzverhalten; Methoden (Beurteiltypen): Mutter, Vater, Lehrer; Con = Konsistenz, MSpe = Methodenspezifität, Rel = Reliabilität. Die Faktorladungen wurden auf zwei Dezimalstellen gerundet, sodass sich die Varianzkomponenten Con, MSpe und Rel nicht exakt aus den Faktorladungen ergeben.

Die Reliabilität der Indikatorvariablen berechnet sich aus den quadrierten standardisierten Faktorladungen (Gl. 25.5). Beispielsweise wird die Reliabilität für Y_{11} mit einem Wert von $Rel(Y_{11}) = .97$ als Summe der Konsistenz (quadrierte Faktorladung auf den Trait Unaufmerksamkeit; $.86 \cdot .86 = .74$) und der Methodenspezifität (quadrierte Faktorladung auf die Methode „Mutter“, $.48 \cdot .48 = .23$), berechnet. Die Reliabilität setzt sich immer aus der Summe der Koeffizienten Konsistenz und Methodenspezifität zusammen (Genaueres s. im Folgenden).

Reliabilität berechnet über quadrierte standardisierte Faktorladungen

■■ Konvergente und diskriminante Validität

Wie die einzelnen Parameterschätzungen des CTCM-Modells zeigen (► Tab. 25.1), weisen alle mit den Methoden „Mutter“ und „Vater“ erhobenen Messungen substantielle (und signifikante) Faktorladungen auf allen drei Traits auf (Koeffizienten zwischen .53 und .86). Die höchsten Ladungen finden sich bei der Messung des latenten Traits Unaufmerksamkeit mit Werten von .71 und .86. Somit kann für diese Indikatorvariablen die *konvergente Validität* angenommen werden.

Konvergente Validität

Anders sieht es für die Messungen mit der Methode „Lehrer“ aus (s. ► Tab. 25.1), deren Faktorladungen auf den Traits relativ gering sind (Koeffizienten zwischen .17 und .37). Somit muss für die Messungen mit der Methode „Lehrer“ eine mangelnde konvergente Validität festgehalten werden.

Diskriminante Validität

Die *diskriminante Validität* zeigt sich darin, dass die Traits theoriekonform in mittlerer Höhe miteinander korrelieren (Korrelationskoeffizienten von .47, .48 und .52).

Methodeneffekte

Die *Methodeneffekte* zeigen sich durch substantielle und hohe Ladungen auf den Methodenfaktoren (.48 bis .87), die im Fall der Methode „Lehrer“ sogar erheblich höher sind, als die entsprechenden Faktorladungen auf den Trait-Faktoren. Die Methoden korrelieren miteinander, wobei die hohe Korrelation zwischen den Methoden „Mutter“ und „Vater“ ($r = .72$) zeigt, dass die Eltern in der Beurteilung der Verhaltensauffälligkeiten ihres Kindes in einem hohem Maß übereinstimmen, während die Übereinstimmung mit der Einschätzung des Lehrers geringer ausfällt ($r = .46$ bzw. $r = .41$).

Konsistenz und Methodenspezifität

Die *konvergente Validität* zeigt sich noch präziser im Vergleich der Varianzkomponenten Konsistenz (*Con*) und Methodenspezifität (*MSpe*). Diese beiden Komponenten der Reliabilität geben an, in welchem Ausmaß die Varianz einer Indikatorvariablen durch den jeweiligen Trait erklärt wird, der einen konsistenten Einfluss auf alle Variablen hat, und in welchem Ausmaß die Varianz zusätzlich durch den Einfluss der Methode bestimmt wird. Betrachtet man z. B. die Indikatorvariable Y_{11} (Unaufmerksamkeit, Methode „Mutter“), so ist die Konsistenz recht hoch mit $Con(Y_{11}) = .74$ ($= .86 \cdot .86$), die Methodenspezifität dagegen relativ gering ausgeprägt mit $MSpe(Y_{11}) = .23$ ($= .48 \cdot .48$) (s. ► Tab. 25.1).

Hohe Reliabilität kann unterschiedliche Ursachen haben

Ein anderes Bild ergibt sich hingegen z. B. bei der Indikatorvariablen Y_{21} (Hyperaktivität, Methode „Mutter“): Hier beträgt die Konsistenz nur $Con(Y_{21}) = .28$ ($= .53 \cdot .53$), die Methodenspezifität dagegen $MSpe(Y_{21}) = .67$ ($= .82 \cdot .82$). Beide Indikatoren sind hoch reliable Messungen, jedoch geht bei der einen Messung ein hoher Anteil der wahren Varianz auf den Trait zurück, bei der anderen Messung dagegen ein hoher Anteil der wahren Varianz auf die Methode. Somit unterscheiden sich die beiden Variablen bezüglich ihrer Konsistenz und Methodenspezifität. Ein hoher Methodenanteil ist aber vor allem bei den Indikatoren des Beurteiltyps „Lehrer“ zu verzeichnen, da die Varianzen dieser Indikatoren nur zu einem sehr geringen Anteil auf den jeweiligen Trait, dagegen zu einem recht hohen Anteil auf die Erfassungsmethode (Beurteiler) zurückzuführen sind.

Insgesamt zeigen die Ergebnisse der CTCM-Analyse in ► Tab. 25.1, dass im Wesentlichen sowohl konvergente als auch diskriminante Validität vorliegt, wobei eine Aufteilung in Trait- und Methodenanteile vorgenommen wurde. Allerdings ist die konvergente Validität nur für die Messungen mit den Methoden „Mutter“ und „Vater“ hoch (mit Ausnahme der Variablen Y_{21}), nicht jedoch für die Messungen mit der Methode „Lehrer“. Ein Grund für den hohen Methodenanteil in

den Einschätzungen des Beurteiltyps „Lehrer“ könnte darin gesehen werden, dass die Lehrer die Verhaltensweisen von Schülern nur relativ ungenau einschätzen können, weil möglicherweise gerade die auffälligen Verhaltensweisen eher zu Hause als in der Schule gezeigt werden. Hier zeigt sich, dass die Messungen der Traits methodenabhängig sind. Diskriminante Validität kann dagegen als nachgewiesen gelten, da die Korrelationen zwischen den Traits nicht zu hoch ausgeprägt sind.

25.6.3 Kritik am CTCM-Modell

■ ■ Schätz- und Identifikationsprobleme

Trotz der hohen Popularität dieses klassischen CFA-MTMM-Modells bestehen einige Probleme bei der Anwendung dieser Methode (vgl. ► Kap. 27; Eid 2000; Grayson und Marsh 1994; Kenny und Kashy 1992; Koch et al. 2018; Marsh 1989; Pohl und Steyer 2010). Dabei handelt es sich vor allem um Schätz- und Identifikationsprobleme, die sich u. a. in unplausiblen Parameterschätzungen, z. B. negativen Fehlervarianzen oder standardisierten Faktorladungen größer als eins, manifestieren können.

Auch im vorliegenden Beispiel sind die Parameterschätzungen des CTCM-Modells mit der CFA nicht stabil, da die Fehlervarianzen zum Teil nicht signifikant von null verschieden sind. Auch eine Gleichsetzung der Faktorladungen eines Methodenfaktors (Annahme der essentiellen τ -Äquivalenz) führt ebenso wie die Fixierung einer Korrelation zwischen zwei Methoden auf null nicht zu einem schlechteren Modellfit, sondern zu Schätz- und Konvergenzproblemen. Solche Schätzprobleme, die oftmals aus Identifikationsproblemen resultieren, sind vor allem dann wahrscheinlich, wenn Modelle mit wenigen Faktoren, z. B. drei Traits und drei Methoden, analysiert werden und wenn die untersuchte Stichprobe klein ist (vgl. Dumenci und Yates 2012).

■ ■ Interpretationsprobleme

Da im CTCM-Modell sowohl alle Trait-Faktoren untereinander als auch alle Methodenfaktoren untereinander korreliert sind, stellt sich die Frage der Interpretation dieser Faktoren. Worin unterscheiden sich Trait- von Methodenfaktoren? Implizit wird angenommen, dass ein Methodenfaktor ein Residualfaktor ist, der den Anteil an der systematischen Varianz eines Indikators erklärt, der nicht auf den Trait zurückgeführt werden kann. Formal trifft diese Unterscheidung aber nicht zu, da inhaltlich nicht zwischen Trait- und Methodenfaktoren unterschieden werden kann und damit unklar ist, wer von ihnen ein Residualfaktor ist. Der einzige Unterschied zwischen den beiden Arten von Faktoren besteht darin, welche Indikatoren auf welchem Faktor laden. Im empirischen Beispiel laden z. B. alle Indikatoren mit dem ersten Index $j = 1$ auf dem ersten Trait-Faktor und alle Indikatoren mit dem zweiten Index $k = 1$ auf dem ersten Methodenfaktor. Welcher der Faktoren als Trait- und welcher als Methodenfaktor bezeichnet wird, bleibt allerdings ungeklärt, da keine eindeutige Unterscheidung zwischen den beiden Typen von Faktoren möglich ist. Um hier Klarheit zu schaffen, werden die neueren CFA-MTMM-Modelle benötigt.

25.7 Faktorenanalytische Ansätze: Neuere CFA-MTMM-Modelle

Neuere CFA-MTMM-Modelle, z. B. das nachfolgend vorgestellte Correlated-Trait-Correlated-(Method-minus-1)-Modell, kurz CTC($M - 1$)-Modell (Eid 2000), basieren auf einer psychometrischen Theorie, definieren Trait- und Methodenfak-

toren designbasiert und erlauben multiple Indikatoren pro Trait-Methoden-Einheit anstatt eines singulären Indikators (Eid et al. 2016; Koch et al. 2018; ▶ Kap. 27).

25.7.1 Das CTC($M - 1$)-Modell

Eine Alternative zu den klassischen CFA-MTMM-Modellen stellt das CTC($M - 1$)-Modell dar (Eid 2000; Eid et al. 2008; s. auch ▶ Kap. 27), das auf dem True-Score-Konzept der Klassischen Testtheorie (KTT) basiert (▶ Kap. 13). Im CTC($M - 1$)-Modell geht man davon aus, dass Messungen nicht unabhängig von der verwendeten Methode sind und Traits somit nicht unabhängig von der verwendeten Methode erfasst werden können. Dies entspricht der Annahme von Campbell und Fiske (1959), dass sich jede Messung aus einer systematischen *Trait-Methoden-Einheit* und einem unsystematischen Fehleranteil zusammensetzt (Eid 2000; Koch et al. 2018).

■ ■ Referenzmethode

Im Gegensatz zum CTCM-Modell wird im CTC($M - 1$)-Modell eine Methode als Standardmethode (Referenzmethode) definiert und für diese Methode kein eigener Methodenfaktor ins Modell aufgenommen (Abb. 25.4). Für jede andere Methode wird ein eigener Faktor spezifiziert, wodurch eine Kontrastierung der Nichtreferenzmethoden gegen die Standardmethode ermöglicht wird (Eid 2000).

Die Wahl der Standardmethode basiert häufig auf theoretischen Annahmen, vorangegangenen empirischen Studien oder darauf, dass eine Methode sich besonders von übrigen Methoden abhebt (z. B. objektive Tests vs. Selbst- und Fremdeinschätzungen der Intelligenz; Geiser et al. 2008; Geiser et al. 2016). Die Referenzmethode definiert die latenten Trait-Faktoren, die damit jeweils eine klare Bedeutung erhalten.

■ ■ Bedeutung der Faktoren

Die Faktoren haben in diesem Modell eine klare Bedeutung: Die Trait-Faktoren können als gemeinsame True-Score-Variablen angesehen werden, die den beobachteten Variablen zugrunde liegen, und die im vorliegenden Anwendungsbeispiel

Referenzmethode ohne eigenen Methodenfaktor

Trait-Faktoren

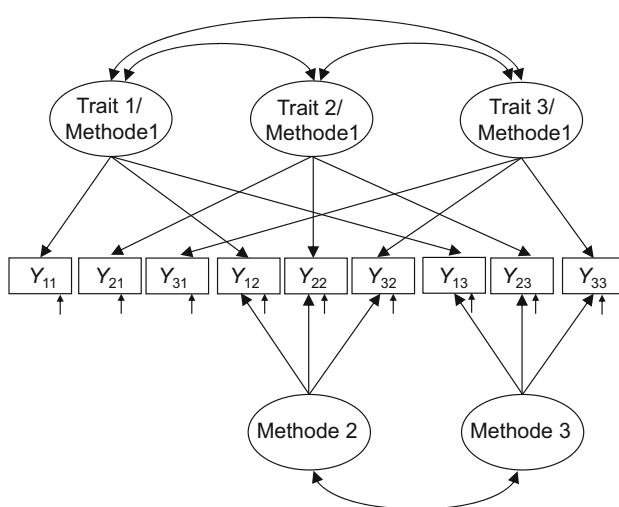


Abb. 25.4 Das CTC($M - 1$)-Modell zur Analyse einer MTMM-Matrix mit neun Indikatorvariablen, Y_{11} bis Y_{33} , drei Traits (*Trait 1*, *Trait 2*, *Trait 3*) und drei Methoden (*Methode 1*, *Methode 2*, *Methode 3*), wobei Methode 1 als Standardmethode gewählt wurde. Die drei Trait-Faktoren und die beiden Methodenfaktoren können jeweils untereinander korreliert sein, was durch *gebogene Doppelpfeile* symbolisiert wird; die Messfehler sind durch *kleine Pfeile* nur angedeutet

die Sichtweise der Mütter beinhalten. Sie werden explizit als Faktoren aufgefasst, die zwei Anteile an systematischer Varianz enthalten, und zwar jeweils einen spezifischen Trait-Anteil und einen Methodenanteil der Referenzmethode. Diese beiden Anteile können jedoch nicht voneinander getrennt werden. Die Trait-Faktoren beinhalten somit die latenten Trait-Werte von Personen, die mit der Referenzmethode gemessen wurden (im Anwendungsbeispiel z. B. Unaufmerksamkeit erfasst durch den Beurteilertyp „Mutter“).

Die beiden übrigen Methoden („Vater“, „Lehrer“) sind als Residualfaktoren hinsichtlich dieser Referenzmethode definiert und bilden jeweils diejenigen Aspekte der Nichtreferenzmethoden ab, die *nicht* mit der gewählten Referenzmethode geteilt werden. Weitere Ansätze, mit denen ebenfalls die Probleme des CTCM-Modells vermieden werden können, auf die aber hier nicht näher eingegangen wird, werden z. B. bei Koch et al. (2018) erläutert.

25.7.2 Empirisches Anwendungsbeispiel

Für die Anwendung des CTC($M = 1$)-Modells auf die Daten von Burns et al. (2013) wurde die Methode „Mutter“ als Referenzmethode gewählt, da die Einschätzungen von Müttern gegenwärtig die am häufigsten genutzte Informationsquelle in Studien zu ADHS-Symptomen darstellen. Somit enthält das Modell nur Methodenfaktoren für die Methoden „Vater“ und „Lehrer“ (vgl. □ Abb. 25.4; auf die Verwendung multipler Indikatoren pro Trait-Methoden-Einheit wurde hier aus Platzgründen verzichtet).

■■ Ergebnisse

Modellgüte

Überprüfung der Modellgüte Die Überprüfung der Güte des CTC($M = 1$)-Modells zeigt einen zum CTCM-Modell vergleichbaren Modellfit mit $\chi^2(17) = 39.61, p = .00$, RMSEA = .04, CFI = .99 und TLI = .98. Somit passt das Modell mit der Methode „Mutter“ als Referenzmethode hinreichend gut zu den Daten, und die geschätzten Parameter dürfen interpretiert werden.

Methodenspezifische Traitfaktoren

Konvergente und Diskriminante Validität Die Faktorladungen auf den drei methodenspezifischen Trait-Faktoren (□ Tab. 25.2) geben Auskunft darüber, in welchem Ausmaß die Indikatoren eines Traits, gemessen durch die Referenzmethode „Mutter“, mit den Messungen durch die Nichtreferenzmethoden „Vater“ und „Lehrer“ übereinstimmen. Diese Übereinstimmung zwischen den Methoden bezüglich eines Traits kann als *konvergente Validität* interpretiert werden.

Konvergente Validität

Wie die Ergebnisse zeigen, besteht eine hohe Übereinstimmung der Messungen des Beurteilertyps „Vater“ mit denen der Methode „Mutter“ (Faktorladungen der Variablen Y_{12} , Y_{22} und Y_{32} auf den drei Traits gemessen mit der Methode „Mutter“ UN/M, HA/M und TV/M zwischen .78 und .85). Auch die Einschätzungen des Beurteilertyps „Lehrer“ laden auf den Traitfaktoren positiv, allerdings sind die Faktorladungen der Variablen Y_{13} , Y_{23} und Y_{33} (Einschätzungen des Lehrers) mit Koeffizienten zwischen .36 und .45 deutlich geringer als die Faktorladungen des Beurteilertyps „Vater“. Somit kann zwar die konvergente Validität für alle Messungen nachgewiesen werden, aber die geringen Faktorladungen auf den Traits mit der Methode „Lehrer“ zeigen abweichende Einschätzungen zu den Beurteilungen durch die Eltern an.

Diskriminante Validität

Die *diskriminante Validität* zeigt sich an den Korrelationen zwischen den methodenspezifischen Trait-Faktoren. Diese Korrelationen sind recht hoch (.64 bis .71) und weisen darauf hin, dass innerhalb der Einschätzungen durch den Beurteilertyp „Mutter“ die diskriminante Validität recht gering ist.

Konsistenzkoeffizienten relativiert an den Reliabilitäten

Die konvergente Validität zeigt sich noch deutlicher, wenn man die Konsistenzkoeffizienten an den Reliabilitäten relativiert. Wie zu sehen ist, wird zwischen 72 %

Tabelle 25.2 Ergebnisse des CTC($M = 1$)-Modells mit der Methode „Mutter“ als Standardmethode, standardisierte Faktorladungen der neun manifesten Variablen Y_{11} bis Y_{33} auf drei Trait- und zwei Methodenfaktoren sowie Aufteilung der Reliabilität in Konsistenz und Methodenspezifität

	Faktorladungen						Varianzkomponenten		
	Trait/Methode Mutter			Methode					
Indikator	UN/M	HA/M	TV/M	Mutter	Vater	Lehrer	Con	MSpe	Rel
Y_{11}	.97			–			.94	–	.94
Y_{21}		.94		–			.88	–	.88
Y_{31}			.91	–			.84	–	.84
Y_{12}	.85				.40		.72	.16	.88
Y_{22}		.84			.40		.72	.16	.88
Y_{32}			.78		.49		.61	.24	.85
Y_{13}	.45					.57	.20	.33	.53
Y_{23}		.42				.80	.18	.63	.81
Y_{33}			.36			.70	.13	.48	.61
Korrierte Traits/Mutter			Korrierte Methoden						
1.0			–						
.69	1.0		–	1.0					
.64	.71	1.0	–	.09	1.0				

Anmerkung: Traits: UN = Unaufmerksamkeit, HA = Hyperaktivität, TV = Trotzverhalten; Methoden (Beurteiltypen): Mutter, Vater, Lehrer; Con = Konsistenz, MSpe = Methodenspezifität, Rel = Reliabilität. Die Faktorladungen wurden auf zwei Dezimalstellen gerundet, sodass sich die Varianzkomponenten Con, MSpe und Rel nicht exakt aus den Faktorladungen ergeben.

(.61/.85) und 82 % (.72/.88) der True-Score-Varianz der Einschätzungen durch den Beurteiltyp „Vater“ mit den Einschätzungen durch den Beurteiltyp „Mutter“ geteilt. Die Konvergenz der Einschätzungen des Beurteiltyps „Lehrer“ mit denen des Beurteiltyps „Mutter“ ist dagegen deutlich geringer, was sich an den niedrigeren Konsistenzkoeffizienten (.13 bis .20) und der jeweils höheren Methodenspezifität (.33 bis .63) zeigt. Die Einschätzungen durch den Beurteiler „Lehrer“ verglichen mit den Einschätzungen durch den Beurteiltyp „Mutter“ betragen lediglich zwischen 21 % (.13/.61) für das Trotzverhalten und 38 % (.20/.53) für die Unaufmerksamkeit. Diese Werte sprechen für eine substantielle, wenn auch geringe konvergente Validität.

■■ Methodeneffekte

Die Methodeneffekte zeigen sich durch substantielle und zum Teil hohe Ladungen auf den Methodenfaktoren (.40 bis .80) sowie den entsprechenden quadrierten standardisierten Faktorladungen, der *Methodenspezifität*. Für die Methode „Mutter“ gibt es hier keine Werte, da diese Methodeneffekte in den Traits enthalten sind.

Die Koeffizienten der Methodenspezifitäten der Methode „Vater“ ist gering mit Werten zwischen .16 und .24. Folglich bestehen zwischen den Einschätzungen der Mutter und des Vaters hohe Übereinstimmungen, was sich auch in den hohen Konsistenzkoeffizienten (.61 bis .72) widerspiegelt. Die Methodenspezifität der Einschätzungen des Lehrers sind dagegen deutlich höher mit Werten von .33,

Methodeneffekte der Mütter sind in den Traits enthalten

Hohe Übereinstimmungen zwischen den Eltern

Geringe Korrelation zwischen Nichtreferenzmethoden

.63 und .48. Hier zeigen sich Unterschiede zu den Einschätzungen der Mütter, die besonders stark bei der Hyperaktivität (.63) ausgeprägt sind.

Die sehr geringe Korrelation ($r = .09$, n. s.) zwischen den Nichtreferenzmethoden zeigt an, dass die Beurteiler „Vater“ und „Lehrer“ eine unterschiedliche Sicht auf die Kinder haben und keine spezifischen Gemeinsamkeiten über die gemeinsamen Anteile mit dem Beurteilertyp „Mutter“ hinaus aufweisen.

■■ Zusammenfassung der Ergebnisse

Insgesamt ist die Analyse des CTC($M = 1$)-Modells ein Beispiel für eine MTMM-Studie mit teils hoher (zwischen der Nichtreferenzmethode „Vater“ und der Referenzmethode „Mutter“) und teils geringer *konvergenter Validität* (zwischen der Nichtreferenzmethode „Lehrer“ und der Referenzmethode „Mutter“). Dies könnte darauf hindeuten, dass in der vorliegenden Studie nicht nur Methoden-, sondern möglicherweise auch Kontext- und/oder situative Effekte eine Rolle spielen. So könnte die hier untersuchte Symptomatik stark vom Kontext abhängig sein (kindliches Verhalten zu Hause vs. in der Schule), was die große Diskrepanz zwischen den Einschätzungen der Eltern und Lehrer erklären könnte. Die *diskriminante Validität* erscheint für die gegenwärtige Studie zufriedenstellend: Die drei Konstrukte korrelieren nicht extrem hoch und sind somit deutlich voneinander unterscheidbar.

25.7.3 Kritische Bewertung und Modellerweiterungen

Vorteile des CTC($M = 1$)-Modells

Das CTC($M = 1$)-Modell führt zu weniger Schätzproblemen als das CTCM-Modell (Geiser et al. 2015). Die beim CTCM-Modell auftretenden Interpretationsprobleme werden durch die bessere messtheoretische Fundierung des CTC($M = 1$)-Modells überwunden, indem z. B. die Trait-Faktoren als messfehlerfreie Einschätzungen der Traits durch die Mütter interpretiert werden. Durch die Verwendung einer Referenzmethode erhalten die Methodenfaktoren eine klare Bedeutung, da sie nun Residualfaktoren sind: Sie enthalten diejenigen Aspekte der Nichtreferenzmethoden, die nicht mit der gewählten Referenzmethode geteilt werden, also Über- oder Unterschätzungen des Traits im Vergleich zur Referenzmethode.

Das CTC($M = 1$)-Modell eignet sich besonders für strukturell unterschiedliche Methoden, z. B. im Anwendungsbeispiel die Einschätzung von Kindern durch Mutter, Vater und Lehrer. Strukturell unterschiedliche Methoden liefern jeweils unterschiedliche und „einzigartige“ Perspektiven, die gegeneinander kontrastiert werden können. Werden dagegen austauschbare Methoden verwendet, z. B. mehrere Freunde, die ein Kind beurteilen, so ist nicht klar, welcher Beurteiler als Referenz dienen soll, und Abweichungen wären nur schwer zu interpretieren.

Es gibt aber auch einige Kritikpunkte an diesem Modell. In der vorliegenden Studie wurde z. B. nur eine einzige Indikatorvariable pro Trait-Methoden-Einheit gewählt, was dem klassischen MTMM-Design von Campbell und Fiske (1959) entspricht. Es wird damit implizit angenommen, dass Methodeneffekte innerhalb jeder Methode eindimensional sind und sie sich somit gleichermaßen auf die Messung unterschiedlicher Traits auswirken. Die Annahme der Eindimensionalität ist aber immer dann verletzt, wenn Methodeneffekte trait-spezifisch sind (vgl. Eid et al. 2003). Im vorliegenden Beispiel kann die Verletzung der Eindimensionalitätsannahme u. a. zu einer Verzerrung der Reliabilitätsschätzungen geführt haben. So scheinen basierend auf den Modellergebnissen die Einschätzungen der Lehrer weniger reliabel zu sein als die Einschätzungen von Müttern und Vätern.

Eine neue Analyse mit drei Indikatoren pro Trait-Methoden-Einheit (die hier aus Umfangsgründen nicht erläutert werden kann) zeigt aber, dass die Reliabilität der Einschätzungen der Lehrer sehr hoch ist und sich nicht von der Reliabilität

Strukturell unterschiedliche Methoden

Ein Indikator pro Trait-Methoden-Einheit

Mehrere Indikatoren pro Trait-Methoden-Einheit

25.8 · Zusammenfassung

der Einschätzungen der Eltern unterscheidet. Somit sollten in der empirischen Anwendung möglichst mehrere Indikatoren pro Trait-Methoden-Einheit verwendet werden.

25.7.4 Weitere neue CFA-MTMM-Ansätze

Alternative CFA-MTMM-Ansätze, die ebenso wie das CTC($M - 1$)-Modell auf dem True-Score-Konzept der KTT beruhen, gehen von einer anderen Definition der Methodenfaktoren aus. Im *Latent-Difference-Modell* (Pohl et al. 2008; Steyer et al. 1997) werden Methodeneffekte als kausale Effekte angesehen. Daher werden die Methodenfaktoren in diesem Modell als True-Score-Differenzen und nicht wie im CTC($M - 1$)-Modell als True-Score-Residuen definiert.

Anwender sind nicht immer daran interessiert, Traits in Abhängigkeit von einer gemeinsamen Referenzmethode zu definieren. Im *Method-Effect-Modell* (Pohl und Steyer 2010) repräsentiert jede Trait-Variable den Durchschnitt der verschiedenen verwendeten Methoden, und Methodenfaktoren werden als Abweichungen der True-Score-Variablen vom Durchschnitt der True-Score-Variablen verstanden. Sowohl das *Latent-Difference-* als auch das *Method-Effect-Modell* haben jedoch die Einschränkung, dass diese Modelle nur dann sinnvoll anwendbar sind, wenn alle Methoden auf derselben Skala gemessen wurden (z. B. Selbst- und Fremdeinschätzungen mit demselben Fragebogen) oder die Messungen durch Reskalierung in eine gemeinsame Skala überführt werden können (Geiser et al. 2012).

Strukturell unterschiedliche Methoden benötigen andere Messdesigns als austauschbare Methoden (Eid et al. 2008; ► Kap. 27). Während für strukturell unterschiedliche Methoden (wie im hier verwendeten Beispiel) das CTC($M - 1$)-Modell angemessen ist, sollte für austauschbare Methoden ein Multilevel-CFA-Modell verwendet werden. Für austauschbare Methoden (z. B. Einschätzungen von Schülern) ist es unerheblich, welche Schüler ausgewählt werden, sie werden den Methoden deshalb zufällig zugeordnet. Liegt eine Kombination aus beiden Typen von Methoden vor, so bietet sich zur Schätzung der Methodeneffekte eine Integration des CTC($M - 1$)-Modells in das Multilevel-CFA-Modell an.

Latent-Difference-Modell

Method-Effect-Modell

25.8 Zusammenfassung

Nach Campbell und Fiske (1959) setzt sich jede Messung aus einer systematischen Trait-Methoden-Einheit und einem unsystematischen Fehleranteil zusammen, sodass nicht nur der gemessene Trait, sondern darüber hinaus die verwendete Methode als Bestandteil der Messung berücksichtigt werden muss. Konstruktvalidität liegt dem Konzept von Campbell und Fiske zufolge nur dann vor, wenn einerseits Messungen desselben Konstrukts mit verschiedenen Messmethoden zu einer hohen Übereinstimmung führen (konvergente Validität), andererseits eine Diskrimination zwischen inhaltlich unterschiedlichen Konstrukten sowohl innerhalb einer Messmethode als auch zwischen den Methoden nachgewiesen werden kann (diskriminante Validität).

Zum Nachweis der Konstruktvalidität anhand der korrelationsbasierten MTMM-Analyse werden die Korrelationskoeffizienten in der MTMM-Matrix durch systematische Vergleiche deskriptiv dahingehend beurteilt, ob die Kriterien der konvergenten und der diskriminanten Validität erfüllt sind. Die Methode beinhaltet eine Vielzahl von Problemen, die mit der CFA nicht auftreten.

Mit den Modellen der konfirmatorischen MTMM-Analyse ist es möglich, Trait-, Methoden- und unsystematische Messfehleranteile der gemessenen Variablen unabhängig voneinander zu schätzen und die Gültigkeit der zugrunde liegenden Annahmen inferenzstatistisch zu überprüfen. Mit dem CTCM-Modell

können die konvergente und die diskriminante Validität bestimmt werden, allerdings ist keine eindeutige Interpretation der Traits und Methoden möglich, da sowohl alle Trait-Faktoren untereinander als auch alle Methodenfaktoren untereinander korreliert sind. Bei diesem Modell treten häufig Interpretations- und Schätzprobleme auf, sodass es nur noch selten angewandt wird.

Mit der Verwendung des CTC($M-1$)-Modells können diese Probleme überwunden werden. In diesem Modell wird eine Referenzmethode festgelegt, die nicht mitmodelliert wird, sodass die Trait- und die Methodenfaktoren nun eine klare Bedeutung erhalten. Das CTC($M-1$)-Modell ermöglicht es, die konvergente und die diskriminante Validität bezogen auf die gewählte Referenzmethode zu interpretieren. Eine Erweiterung des Modells auf drei oder mehr Indikatoren pro Trait-Methoden-Einheit wird empfohlen, um Methodeneffekte als trait-spezifisch modellieren zu können.

25.9 EDV-Hinweise

Zur korrelationsbasierten MTMM-Analyse eignet sich jedes Statistikprogramm, z. B. SPSS oder R (R Development Core Team 2016), mit dem Korrelationen berechnet werden können.

Die verschiedenen Modelle der konfirmatorischen MTMM-Analyse können mit Verfahren zur Analyse von linearen Strukturgleichungsmodellen überprüft werden. Hierzu bieten sich u. a. die Programme Mplus (Muthén und Muthén 2017) oder das R-Paket lavaan (Rosseel 2012) an, mit denen CFA-Modelle entsprechend der Hypothesen spezifiziert und evaluiert werden können.

25.10 Kontrollfragen

Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was versteht man unter konvergenter Validität, was unter diskriminanter Validität im Rahmen der Multitrait-Multimethod-Analyse (MTMM-Analyse)?
2. Welche Arten von Koeffizienten befinden sich in der MTMM-Matrix?
3. Wie kann man die konvergente Validität nach den Kriterien von Campbell und Fiske (1959) nachweisen, wie die diskriminante Validität?
4. Was versteht man unter Methodeneffekten und welche ihrer Entstehungsquellen werden bei der MTMM-Analyse unterschieden?
5. Welches sind die Vorteile der konfirmatorischen MTMM-Analyse gegenüber der korrelationsbasierten MTMM-Analyse?
6. Welche wesentlichen Vorteile hat das CTC($M - 1$)-Modell gegenüber dem CTCM-Modell?
7. Was versteht man unter „Konsistenz“ und „Methodenspezifität“? In welchem Zusammenhang stehen diese beiden Koeffizienten?
8. Wie werden die Methodenfaktoren im CTC($M - 1$)-Modell interpretiert?

Literatur

- Biesanz, J. C. & West, S. G. (2004). Towards understanding assessments of the Big Five: multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, 72, 845–876.
- Burns, L. G., Walsh, J. A., Servera, M., Lorenzo-Seva, U., Cardo, E. & Rodríguez-Fornells, A. (2013). Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming DSM-V ADHD/ODD Scales. *Journal of Abnormal Child Psychology*, 41, 15–26.

- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.
- Dumenci, L. & Yates, P. D. (2012). Nonconvergence/improper solution problems with the correlated-trait correlated-method parameterization of a multitrait-multimethod matrix. *Educational and Psychological Measurement, 72*, 800–807.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241–261.
- Eid, M. & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
- Eid, M., Geiser, C. & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science, 25*, 275–280.
- Eid, M., Lischetzke, T., Nussbeck, F. & Trierweiler, L. (2003). Separating trait effects from traitspecific method effects in multitrait-multimethod analysis: A multiple indicator CTC($M-1$) model. *Psychological Methods, 8*, 38–60.
- Eid, M., Notz, P., Steyer, R. & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analyses. *Personality and Individual Differences, 16*, 63–76.
- 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., Nussbeck, F. W. & Lischetzke, T. (2006). Multitrait-Multimethod-Analyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 332–345). Göttingen: Hogrefe.
- Eid, M. & Schmidt, K. (2014). *Testtheorie*. Göttingen: Hogrefe.
- Geiser, C., Bishop, J. & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: Examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology: Quantitative Psychology and Measurement, 6*, 946. <https://doi.org/10.3389/fpsyg.2015.00946>
- Geiser, C., Eid, M. & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C($M-1$) model: A comment on Maydeu-Olivares & Coffman (2006). *Psychological Methods, 13*, 49–57.
- Geiser, C., Eid, M., West, S. G., Lischetzke, T. & Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling, 19*, 409–436.
- Geiser, C., Mandelman, S. D., Tan, M. & Grigorenko, E. L. (2016). Multitrait-multimethod assessment of giftedness: An application of the correlated traits-correlated (methods – 1) model. *Structural Equation Modeling, 23*, 76–90.
- Grayson, D. & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis: Multitrait-multimethod models. *Psychometrika, 59*, 121–134.
- Höfling, V., Schermelleh-Engel, K. & Moosbrugger, H. (2009). Analyzing multitrait-multimethod data: A comparison of three approaches. *Methodology, 5*, 99–111.
- Jöreskog, K. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.
- Kenny, D. A. & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.
- Koch, T., Eid, M., & Lochner, K. (2018). Multitrait-Multimethod-Analysis: The Psychometric Foundation of CFA-MTMM Models. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing. Volume II: A multidisciplinary reference on survey, scale and test development* (pp. 781–846). Hoboken, NJ: Wiley.
- Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.
- Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–187). Thousand Oaks, London: Sage.
- Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
- Muthén, L. K. & Muthén, B. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York: Routledge.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Psychological Methods, 88*, 879–903.
- Podsakoff, P. M., MacKenzie, S. B. & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.
- Pohl, S. & Steyer, R. (2010). Modelling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*, 45–72.
- Pohl, S., Steyer, R. & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society Series A, 171*, 41–63.

- R Development Core Team (2016). *R: A Language and Environment for Statistical Computing (version 3.2.4)*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/> [29.12.2019]
- Rosseel, I. (2012). *lavaan: An R Package for Structural Equation Modeling*. *Journal of Statistical Software*, 48, 1–36.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8, 23–74.
- Steyer, R., Eid, M. & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, 2, 21–33.
- Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389–408.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 10, 1–22.



Latent-State-Trait-Theorie (LST-Theorie)

Augustin Kelava, Karin Schermelleh-Engel und Axel Mayer

Inhaltsverzeichnis

26.1 Einleitung – 688

26.1.1 Unterscheidung zwischen State und Trait – 688

26.1.2 States und Traits in der LST-Theorie – 690

26.1.3 Klassische Testtheorie (KTT) als Grundlage – 691

26.2 LST-Theorie als Erweiterung der KTT – 692

26.2.1 Grundgedanke – 692

26.2.2 Der wahre Wert in der LST-Theorie und dessen Zerlegung – 692

26.2.3 Reliabilität in der LST-Theorie – 694

26.2.4 Methodeneffekte – 695

26.3 Modelltypen – 697

26.3.1 Multistate-Modell – 699

26.3.2 Multistate-Singltrait-Modell – 700

26.3.3 Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Faktoren – 701

26.4 Anwendungen der LST-Theorie – 703

26.4.1 Allgemeine Überlegungen und Voraussetzungen – 703

26.4.2 Empirisches Beispiel: Prüfungsangst – 704

26.4.3 Konfirmatorische Beurteilung der Modellgüte im Beispiel – 706

26.4.4 Varianzzerlegung im Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Faktoren – 707

26.4.5 Erweiterungen der LST-Theorie – 708

26.5 Zusammenfassung – 708

26.6 EDV-Hinweise – 709

26.7 Kontrollfragen – 709

Literatur – 709

i Die Latent-State-Trait-Theorie (LST-Theorie) stellt eine Erweiterung der Klassischen Testtheorie (KTT) dar. Sie berücksichtigt, dass sich Personen während einer Messung immer in einer bestimmten subjektiven Situation befinden, die in die Messung eingeht. Um die systematischen Anteile der wahren Werte unterscheiden zu können, müssen Messungen anhand von mindestens zwei Tests, Testhälften oder Items zu mindestens zwei Messgelegenheiten durchgeführt werden. Bei Gültigkeit der testbaren Modellannahmen erlaubt diese Vorgehensweise eine Varianzdekomposition. Die Gesamtvarianz einer Messung lässt sich zunächst in einen (a) wahren Anteil und (b) einen Messfehleranteil aufteilen. Die wahre Varianz lässt sich wiederum in zwei Bestandteile weiter aufteilen: (a.1) in eine personenspezifische stabile, d. h. zeitlich überdauernde Komponente (z. B. einen Trait als stabile Persönlichkeitsdisposition) und (a.2) in eine situationsspezifische, d. h. zur Messgelegenheit gehörige Komponente (die auch die Wechselwirkung von Person und Situation abbildet). Auf Basis dieser Varianzdekomposition lassen sich Kenngrößen quantifizieren (sog. „Koeffizienten“ der LST-Theorie), die in ihrer Summe die Reliabilität der Messung beschreiben. Grundsätzlich lassen sich aus der LST-Theorie verschiedene Modelle ableiten, deren Gültigkeit im Rahmen der konfirmatorischen Faktorenanalyse (CFA, ► Kap. 24) überprüft werden kann. In diesem Kapitel werden das Multistate-Modell, das Multistate-Singltrait-Modell und das Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren vorgestellt. Diese Modelle werden anhand eines empirischen Beispiels zur Prüfungsangst erläutert, es werden Modellüberprüfungen vorgenommen und die Schätzung der verschiedenen Koeffizienten der LST-Theorie demonstriert.

26.1 Einleitung

Ein Ziel der Einzelfalldiagnostik besteht oftmals darin, stabile Merkmale zu erfassen (z. B. Dispositionen eines Menschen, etwa im Bereich der Persönlichkeit die Merkmale Extraversion oder Gewissenhaftigkeit), deren Ausprägungen über Situationen hinweg konstant bleiben oder sich – zumindest über kurze Zeiträume – nicht substanzial ändern. Ein weiteres Ziel besteht ferner darin, kurzfristige Einflüsse von diesen Merkmalen zu separieren, die situationsspezifisch sind und sich damit der Messgelegenheit zuordnen lassen (z. B. Heiterkeit, Stimmung). Zeitlich überdauernde Merkmale, die über Situationen hinweg stabil sind, werden im Rahmen der LST-Theorie als „Traits“, bezeichnet, z. B. Trait-Anst; zeitlich instabile Merkmale hingegen werden als „States“ bezeichnet. Aber auch States, z. B. die State-Anst (► Exkurs 26.1), können ein zentraler Forschungsgegenstand sein.

26.1.1 Unterscheidung zwischen State und Trait

Definition

Unterscheidung zwischen State und Trait:

- **State** (Zustand): zeitlich instabiles, zustands- und situationsabhängiges Merkmal (z. B. aktuelle Stimmung oder momentanes Befinden)
- **Trait** (Eigenschaft, Disposition): zeitlich relativ stabiles, zustands- und situationsunabhängiges Merkmal (z. B. Extraversion oder Gewissenhaftigkeit)

Die Debatte über die Unterscheidung und über die Existenz von States und Traits reicht in der Differentiellen Psychologie weit zurück. Während das Trait-Konzept bereits in den Anfängen der Persönlichkeitsforschung entwickelt wurde (vgl. z. B. Allport 1937; Cattell 1946), wurde das State-Konzept später in die Persönlichkeitsforschung eingeführt (vgl. z. B. Bowers 1973). Die Trait-Forschung führte zur Entwicklung sehr bekannter Modelle der Persönlichkeitseigenschaften (vgl. z. B.

Exkurs 26.1**State-Angst und Trait-Angst**

Ein Beispiel für die explizite Unterscheidung von States und Traits ist die State-Trait-Theorie der Angst von Spielberger (1972). In dieser Theorie wird das Merkmal *State-Angst* als vorübergehender, situativer emotionaler Zustand aufgefasst, in dem sich eine Person zu einer gegebenen Messgelegenheit befindet. Die State-Angst fluktuiert über die Zeit hinweg, da sich die Person immer wieder in anderen (Lebens-)Situationen befindet. Die State-Angst kann prinzipiell aber auch die gleiche Ausprägung erneut aufweisen, wenn entsprechende Umweltbedingungen gegeben sind, oder aber andauern, falls dieselben Umweltbedingungen fortbestehen.

Das Merkmal *Trait-Angst* hingegen wird als ein zeitlich und über Situationen hinweg relativ überdauerndes Merkmal aufgefasst. Es beschreibt die spezifische Art und Weise, wie die Umwelt wahrgenommen wird bzw. wie auf Umweltgegebenheiten reagiert wird. Die Trait-Angst bedingt zum Teil die State-Angst: Personen, die eine hohe Trait-Angst aufweisen, werden in ganz unterschiedlichen Situationen (d. h. transsituativ) eher dazu neigen, mit erhöhter State-Angst auf Umweltgegebenheiten zu reagieren, als Personen, die nur eine geringe Trait-Angst aufweisen.

Eysenck 1947), die u. a. faktorenanalytische und biologische Grundlagen haben und auf der Annahme beruhen, dass interindividuelle Differenzen im Erleben und Verhalten zwischen Personen auf generalisierbare Unterschiede stabiler Merkmale zurückgeführt werden können.

Ob und wie sehr die Konsistenz interindividueller Merkmalsunterschiede, insbesondere mit Blick auf stabile Persönlichkeitseigenschaften, tatsächlich erfüllt ist, wurde in der Psychologie in der Vergangenheit kontrovers diskutiert (vgl. z. B. Schmitt 1990). In dieser sog. „Konsistenzkontroverse“ ging es um die prinzipielle Frage, ob die Annahme von überdauernden Persönlichkeitsmerkmalen oder Traits, anhand derer sich Verhalten erklären und vorhersagen lässt, überhaupt sinnvoll ist. Stark vereinfacht lässt sich die Kontroverse auf die Frage reduzieren, ob das Verhalten einer Person zu einer bestimmten Messgelegenheit (in Sinne eines Messzeitpunkts) durch deren Persönlichkeit determiniert ist (Dispositionismus, auch Personalismus) oder ob das Verhalten vielmehr von der konkreten Situation determiniert wird (Situationismus), in der sich die Person zur betreffenden Messgelegenheit befindet.

Lange Zeit stand für die in der Differentiellen Psychologie dominierende personalistische Position die *interindividuelle Variation* im Mittelpunkt des Interesses, was dazu führte, dass situative Einflüsse als störende Fehlergrößen interpretiert wurden, die folglich eliminiert oder zumindest statistisch kontrolliert werden sollten. Umgekehrt wurde im Rahmen des allgemeinpsychologischen Ansatzes versucht, die *intraindividuellen situativen Einflüsse*, z. B. durch experimentelle Manipulation, zu kontrollieren, wobei Unterschiede zwischen Personen, die derselben Versuchsbedingung angehörten, als störende Fehlergrößen interpretiert wurden (vgl. auch Steyer et al. 1992).

Während in den vergangenen zwei Jahrzehnten die Dispositionismus-Situationismus-Debatte als ausgefochten galt, indem man davon ausging, dass in jede psychodiagnostische Messung sowohl konsistente Merkmale der Person als auch inkonsistente Merkmale der Situation einfließen und dass diese Merkmale in Abhängigkeit voneinander interagieren (vgl. Fleeson 2004; Furr und Funder 2019; Geiser et al. 2015b), lässt sich heute von einem erneuten Aufkommen einer ähnlichen Diskussion berichten, wenngleich auch in abgewandelter und differenzierter Form (vgl. Roberts und Nickel 2017; Roberts et al. 2006).

Konsistenzkontroverse**Dispositionismus vs. Situationismus**

Die Unterscheidung zwischen States und Traits kann entweder erzielt werden durch unterschiedliche Operationalisierungen, d. h. durch unterschiedliche Itemformulierungen, oder aber durch ähnliche Itemformulierungen, die anhand von unterschiedlichen Instruktionen im Fragebogen eingeleitet und anhand von unterschiedlich bezeichneten Antwortkategorien beantwortet werden (► Beispiel 26.1).

Beispiel 26.1: Beispielitems

Zur Messung von State- und Trait-Angst in der deutschen Form des *State-Trait Anxiety Inventory* (STAII) werden unterschiedliche Instruktionen vorgegeben (vgl. Laux et al. 1981), die sich entweder auf den augenblicklichen Zustand der Person beziehen (State-Angrst) oder auf die allgemeine Tendenz, in verschiedenen Situationen in gleicher Weise ängstlich zu reagieren (Trait-Angrst). Testpersonen sollen entweder ankreuzen, wie sie sich im Moment fühlen (um die State-Angrst zu messen) oder aber, wie sie sich im Allgemeinen fühlen (um die Trait-Angrst zu messen). Die Items des Fragebogens unterscheiden sich jedoch inhaltlich nicht substantiell voneinander. Die *Ratingskalen* unterscheiden sich jedoch bezüglich der Ausprägung der Angstsymptome (State-Angrst) und der Häufigkeit ihres Auftretens (Trait-Angrst).

- Beispiel für ein State-Angrst-Item: „Ich bin besorgt, dass etwas schiefgehen könnte.“

Wählen Sie aus den vier Antworten diejenige aus, die angibt, wie Sie sich jetzt, d. h. *in diesem Moment fühlen*.

Ratingskala: 1 = überhaupt nicht, 2 = ein wenig, 3 = ziemlich, 4 = sehr

- Beispiel für ein Trait-Angrst-Item: „Ich mache mir Sorgen über ein mögliches Missgeschick.“

Wählen Sie aus den vier Antworten diejenige aus, die angibt, wie Sie sich *im Allgemeinen fühlen*.

Ratingskala: 1 = fast nie, 2 = manchmal, 3 = oft, 4 = fast immer

26.1.2 States und Traits in der LST-Theorie

Messung einer Person in einer Situation

In der LST-Theorie (Steyer 1987; Steyer et al. 1992; Steyer et al. 2015; Steyer et al. 1999) wird berücksichtigt, dass sich Personen während einer Messung immer in einer bestimmten Situation befinden. Dies bedeutet, dass ein State sowohl durch Persönlichkeitsmerkmale als auch durch die Situation, in der die Messung vorgenommen wurde, beeinflusst wird. Des Weiteren wird zusätzlich auch eine Interaktion zwischen Person und Situation zur Erklärung menschlichen Verhaltens und Erlebens als Teil des Modells aufgefasst.

Bezogen auf das Beispiel der Angst werden Veränderungen über mehrere Messgelegenheiten hinweg als Abweichung von einer stabilen Disposition, der Trait-Angrst, angesehen, die durch situative Einflüsse bedingt sind. Während früher davon ausgegangen wurde, dass sich ein Trait über die Zeit nicht verändert, hat sich inzwischen die Ansicht durchgesetzt, dass Traits durch Umwelteinflüsse, die Persönlichkeitsentwicklung und Reifung oder durch konkrete Interventionen veränderbar sind (vgl. Geiser et al. 2015c; Rieger et al. 2017; Steyer et al. 2015). In einer Metaanalyse konnten Roberts et al. (2017) zeigen, dass Persönlichkeitsmerkmale, z. B. emotionale Stabilität und Extraversion, durch klinische Interventionen veränderbar sind und dass diese Änderungen nach der Therapie bestehen bleiben.

Die von Steyer et al. (2015) vorgelegte Revision der LST-Theorie (LST-R) berücksichtigt – im Gegensatz zur ursprünglichen LST-Theorie –, dass eine Person zu einem spezifischen Zeitpunkt in ihrem Leben untersucht wird und sie sich über die Zeit aufgrund von vorausgegangenen situativen Einflüssen verändern kann (vgl. Geiser et al. 2015c; Koch et al. 2017). Formale Definitionen der States und Traits

Trait durch Intervention veränderbar

auf der Basis der Wahrscheinlichkeitstheorie finden sich bei Steyer et al. (2015). Es sei darauf hingewiesen, dass aus didaktischen Gründen eine verkürzte und vereinfachte Darstellung der LST-Theorie erfolgt.

26.1.3 Klassische Testtheorie (KTT) als Grundlage

Die LST-Theorie kann als eine Erweiterung der KTT aufgefasst werden (vgl. ► Kap. 13; Eid und Schmidt 2014; Steyer und Eid 2001; Zimmerman 1975). Um auf die formalen Grundlagen der LST-Theorie näher eingehen zu können, ist es zweckmäßig, die KTT nochmals als Ausgangspunkt heranzuziehen, da auf dieser Theorie die LST-Theorie aufgebaut ist.

Nach der KTT setzt sich jede beobachtete Messung y_{vi} einer Person v ($v = 1, \dots, n$) auf einer Variablen i ($i = 1, \dots, p$), z. B. auf der Antwortvariable eines Items, aus einem wahren Wert τ_{vi} und einem Fehlerwert ε_{vi} zusammen:

$$y_{vi} = \tau_{vi} + \varepsilon_{vi} \quad (26.1)$$

Diese Zerlegung des Messwertes y_{vi} in einen wahren Anteil und einen Fehleranteil gilt für alle Messungen, z. B. Antworten auf Items, zu Skalen oder Testhälften.

Der wahre Wert („True-Score“) τ_{vi} ist in der KTT definiert als der personenbedingte Erwartungswert der Variablen Y_{vi} (Guttman 1945; Lord und Novick 1968; Steyer und Eid 2001; Zimmerman 1975). Zur Verdeutlichung soll das folgende Gedankenexperiment helfen: Würde der Messwert y_{vi} einer Person v theoretisch unendlich oft anhand desselben Messinstruments (Item, Test) i erfasst, dann ergäbe sich eine intraindividuelle Verteilung der Messwerte y_{vi} für Person v auf Item i , deren Erwartungswert dem wahren Wert τ_{vi} , d. h. der wahren Ausprägung dieser Person auf Item i , entspricht (► Kap. 13):

$$\tau_{vi} := E(Y_{vi}) \quad (26.2)$$

Auf Variablenebene lässt sich jede beobachtete Variable Y_i in eine True-Score-Variable τ_i und eine Fehlervariable ε_i zerlegen:

$$Y_i = \tau_i + \varepsilon_i \quad (26.3)$$

Diese Beziehung ist in ► Abb. 26.1 grafisch dargestellt.

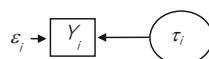
Im Rahmen der KTT kann sowohl für eine Itemvariable Y_i als auch für die über alle Items aufsummierte Testwertvariable Y die Messgenauigkeit bestimmt werden (vgl. ► Kap. 13). Üblicherweise wird in der KTT die Reliabilität einer Testwertvariable geschätzt (vgl. ► Kap. 14 und 15). Die Reliabilität in der KTT ist definiert als der Anteil der wahren Varianz $Var(T)$ an der Gesamtvarianz $Var(Y)$ der Testwertvariablen:

$$Rel(Y) := \frac{Var(T)}{Var(Y)} \quad (26.4)$$

In Gl. (26.4) bezeichnet T (großes griechisches Tau) die True-Score-Variable der Testwertvariable Y . Im Rahmen dieser formalen Konzeption der KTT ist es möglich, den Anteil der Messung zu bestimmen, der reliabel ist. Es ist aber nicht

Definition des wahren Wertes τ_{vi}

Reliabilität in der KTT



► Abb. 26.1 Zerlegung der beobachteten Variable Y_i in die Variable der wahren Werte τ_i (True-Score-Variable) und die Variable der Fehlerwerte ε_i

möglich, die Anteile der Messung zu separieren, die auf eine Disposition (Trait), eine Situation oder die Interaktion von Disposition und Situation zurückzuführen sind. Es ist auch nicht möglich, den Anteil an der Messung zu bestimmen, der auf die verwendete Methode zurückzuführen ist. Erst die Erweiterung der KTT zur LST-Theorie erlaubt eine solche Zerlegung.

26.2 LST-Theorie als Erweiterung der KTT

26.2.1 Grundgedanke

Die LST-Theorie (vgl. Steyer 1987; Steyer et al. 1999; Steyer et al. 2015) stellt eine Erweiterung der KTT dar. Ausgangspunkt der LST ist die Vorstellung, dass sich jede Person v bezüglich der Variablen i (Item, Testwert) zu einer Messgelegenheit t in einer persönlichen (Lebens-)Situation befindet. Da sich Menschen hinsichtlich ihrer erlebten Lebenssituationen unterscheiden, werden zu einer konkreten Messgelegenheit t interindividuelle Unterschiede dieser Situationen auftreten. Diese Vorstellung trägt dem Umstand Rechnung, dass die Messung nicht in einem situativen Vakuum stattfinden kann, sondern dass sich ein Messwert y_{vit} auf die „Person in einer Situation“ bezieht und folglich von der Situation beeinflusst wird. Im Gegensatz zur experimentellen Persönlichkeitsforschung wurde in der klassischen LST-Theorie (Steyer 1987; Steyer et al. 1992) nicht vorausgesetzt, dass die Situation, in der sich eine Person befindet, bekannt ist. Damit ist der wahre Wert einer Person in einer Situation zu einer gegebenen Messgelegenheit Gegenstand der Untersuchung (vgl. ► Beispiel 26.2). Die Zweifachindizierung in der KTT (τ_{vi}) wird in der LST-Theorie somit durch eine Dreifachindizierung (τ_{vit}) abgelöst, die die Unterscheidung der Messgelegenheiten erlaubt.

Messung der Person in einer Situation

Beispiel 26.2: Unterscheidung von Messgelegenheit und Situation

Um den Einfluss der Situation auf die Angst zu veranschaulichen, wollen wir folgendes Beispiel betrachten:

Kevin nimmt an einer Untersuchung teil, die vorsieht, dass zu *zwei Messgelegenheiten* die situative Angst *jeweils anhand zweier Unterkalen* eines Fragebogens erfasst wird. Bei der ersten Messgelegenheit ist Kevin sehr aufgeregt, weil er an diesem Tag noch seine Fahrprüfung ablegen soll. Bei der zweiten Messgelegenheit befindet er sich nicht mehr in einer für ihn persönlich beanspruchenden Situation. Demzufolge schätzt Kevin sich zur ersten Messgelegenheit in den beiden Unterkalen des Fragebogens deutlich ängstlicher ein (35, 33), als zur zweiten Messgelegenheit (14, 16).

Neben ihm nehmen auch andere, einander unbekannte Personen an dieser Untersuchung teil. Wie es den anderen Personen ergeht, wissen wir nicht. Fest steht aber, dass sich jede Person zu jeder der zwei Messgelegenheiten in einer *anderen Situation* befindet. Diese Situationsunterschiede stellen ihrerseits interindividuelle Differenzen dar.

26.2.2 Der wahre Wert in der LST-Theorie und dessen Zerlegung

Wahrer Wert τ_{vit}

Im Rahmen der LST-Theorie ist der wahre Wert τ_{vit} definiert als der zu erwartende Wert aus einer intraindividuellen Verteilung von möglichen Werten y_{vit} einer Person v in Antwortvariable i zu Messgelegenheit t . Der wahre Wert τ_{vit} ist konzipiert

26.2 · LST-Theorie als Erweiterung der KTT

als der zu erwartende (Durchschnitts-)Wert, wenn eine Person v in Antwortvariable i zu Messgelegenheit t unendlich oft wiederholt beobachtet werden könnte. Wie auch in der KTT tritt bei einer (konkreten) Messung y_{vit} ein Messfehler ε_{vit} auf:

$$\begin{aligned} y_{vit} &= \tau_{vit} + \varepsilon_{vit} \\ &= E(Y_{vit}) + \varepsilon_{vit} \end{aligned} \quad (26.5)$$

Im Unterschied zur KTT (vgl. Gl. 26.1) entspricht nun der wahre Wert dem Erwartungswert einer Verteilung von möglichen Messwerten einer Person in einer Situation zu einer Messgelegenheit¹. Der wahre Wert in der LST-Theorie ist der wahre Wert einer Person in einer Situation. Er wird deshalb als *latenter State-Wert* bezeichnet. An dieser Stelle sei darauf hingewiesen, dass der State-Begriff aus Gl. (26.5) eine Abweichung zum State-Begriff aus ► Abschn. 26.1.1 aufweist. Der State-Begriff hier ist vielmehr eine wahre Zustandsvariable, die die Summe aller stabilen und situativen Einflüsse (d. h., einen konkreten wahren Wert der „Person-in-der-Situation“) darstellt. Es ist nicht nur jener Teil des wahren Wertes, der dem situativen Einfluss ohne Persönlichkeitsmerkmal geschuldet ist (z. B. eine Zustandsangst, die allen Menschen bei gleichen äußeren Rahmenbedingungen gemein wäre).

In der LST-Theorie wird nun der latente State-Wert τ_{vit} in zwei weitere Werte zerlegt, und zwar in den latenten Trait-Wert ξ_{vit} und den latenten State-Residuum-Wert ζ_{vit} . Nach Steyer et al. (2015) berücksichtigt der Trait-Wert ξ_{vit} die Geschichte einer Person bis zur Messgelegenheit t . Das bedeutet beispielsweise, dass eine Person mit einer mittleren Trait-Angst zu einem früheren Zeitpunkt nach schlechten Erfahrungen (wenn sie z. B. öfter durch Prüfungen gefallen ist) eine höhere Trait-Angst entwickelt haben kann.

$$\tau_{vit} = \xi_{vit} + \zeta_{vit} \quad (26.6)$$

Im Folgenden betrachten wir nicht mehr den Wert y_{vit} einer konkreten Person, sondern die Zerlegung der Variable Y_{it} über alle Personen hinweg. Die sich daraus ergebende True-Score-Variable wird als *latente State-Variable* τ_{it} bezeichnet.

Die *latente Trait-Variable* ξ_{it} beinhaltet den zeitlich stabilen, d. h. situationsunabhängigen Einfluss der Merkmalsausprägungen der Personen auf die Messungen. Die Trait-Variable beschreibt somit die Variation der Dispositionen der Personen.

Die *latente State-Residuum-Variable* ζ_{it} repräsentiert die Einflüsse der Situation und der Interaktion von Person und Situation. Beide Einflüsse werden durch ζ_{it} repräsentiert. Man erhält ζ_{it} als Differenz von τ_{it} und ξ_{it} .

Die beobachtete Variable Y_{it} setzt sich somit wie folgt zusammen:

$$Y_{it} = \tau_{it} + \varepsilon_{it} = \xi_{it} + \zeta_{it} + \varepsilon_{it} \quad (26.7)$$

Diese Aufteilung der beobachteten Variablen Y_{it} wird in der ► Abb. 26.2 veranschaulicht.

Die Varianz der Antwortvariablen i zu Messgelegenheit t lässt sich nun in drei Teile zerlegen, in die Varianz der latenten Trait-Variablen, die Varianz des latenten State-Residuum und in die Fehlervarianz:

$$\begin{aligned} \text{Var}(Y_{it}) &= \text{Var}(\tau_{it}) + \text{Var}(\varepsilon_{it}) \\ &= \overbrace{\text{Var}(\xi_{it})} + \overbrace{\text{Var}(\zeta_{it})} + \text{Var}(\varepsilon_{it}) \end{aligned} \quad (26.8)$$

Latenter State-Wert

Zerlegung des latenten State-Wertes

Latente State-Variable τ_{it}

Latente Trait-Variable ξ_{it}

Latente State-Residuum-Variable ζ_{it}

Varianzzerlegung von Y_{it}

¹ In der revidierten LST-Theorie (Steyer et al. 2015) wird die True-Score-Variable als bedingte Erwartung $E(Y_{it} | U_t, S_t)$ definiert. Damit soll ausgedrückt werden, dass der erwartete Wert für eine Person U ($U = \text{unit}$) zur Messgelegenheit t (zeitpunktspezifische Person) in einer Situation S zur Messgelegenheit t betrachtet wird.

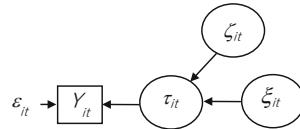


Abb. 26.2 Zerlegung der beobachteten Variablen Y_{it} ($i = \text{Indikator}$, $t = \text{Messgelegenheit}$) in die latente State-Variable τ_{it} und die Fehlervariable ε_{it} . Die State-Variable wird weiter zerlegt in die latente Trait-Variable ξ_{it} und das latente State-Residuum ζ_{it}

26.2.3 Reliabilität in der LST-Theorie

Reliabilität in der LST-Theorie

Im Rahmen der KTT ist es möglich, den Anteil der Messung zu bestimmen, der reliabel ist. Es ist aber nicht möglich, die Anteile der Messung zu separieren, die auf eine Disposition (Trait), eine Situation oder die Interaktion von Disposition und Situation zurückzuführen sind. Es ist auch nicht möglich, den Anteil an der Messung zu bestimmen, der auf die verwendete Methode zurückzuführen ist. Erst die Erweiterung der KTT zur LST-Theorie erlaubt eine solche Zerlegung.

Auf Grundlage der Varianzzerlegung kann in der LST-Theorie analog zur KTT ein Maß der *Reliabilität* definiert werden. Nachfolgend wird nicht zwischen verschiedenen Arten der Indikatorvariablen (Antwortvariablen) unterschieden, da es für die LST unerheblich ist, ob es sich um eine Indikatorvariable eine Testwertvariable oder eine Itemvariable handelt. Deshalb wird nachfolgend die Variable der wahren Werte mit τ und nicht mit T (vgl. ▶ Kap. 13) bezeichnet.

Analog zur KTT erfolgt in der LST-Theorie eine Aufteilung in wahre Varianz und Gesamtvarianz. Über die Erweiterung der KTT ist es möglich, die Reliabilität in die Maße *Konsistenz*, $Con(Y_{it})$, und *Spezifität*, $Spe(Y_{it})$, aufzuteilen (Steyer et al. 1999, 2015)².

$$Rel(Y_{it}) := \frac{Var(\tau_{it})}{Var(Y_{it})} = \frac{Var(\xi_{it}) + Var(\zeta_{it})}{Var(Y_{it})} \quad (26.9)$$

$$Con(Y_{it}) := \frac{Var(\xi_{it})}{Var(Y_{it})} \quad (26.10)$$

$$Spe(Y_{it}) := \frac{Var(\zeta_{it})}{Var(Y_{it})} \quad (26.11)$$

Wie man den Gln. (26.9) bis (26.11) entnehmen kann, gilt:

$$Rel(Y_{it}) := Con(Y_{it}) + Spe(Y_{it}) \quad (26.12)$$

Reliabilitätskoeffizient $Rel(Y_{it})$

Wie auch in der KTT gibt der *Reliabilitätskoeffizient* $Rel(Y_{it})$ an, wie hoch der Anteil der wahren Varianz an der Gesamtvarianz der Variablen Y_{it} ist. Die wahre Varianz entspricht der Varianz der latenten State-Variablen τ_{it} . Somit beschreibt der Reliabilitätskoeffizient die Messgenauigkeit der Variablen i zu Messgelegenheit t .

Der *Konsistenzkoeffizient* $Con(Y_{it})$ bezeichnet einen Teil des Reliabilitätskoeffizienten (vgl. Gl. 26.10). Er gibt den Anteil der Varianz der latenten Trait-Variablen ξ_{it} an der Gesamtvarianz der Variablen Y_{it} an, der auf die Personen zurückgeführt werden kann. Es handelt sich hierbei um konsistente, d. h. situationsunabhängige interindividuelle Unterschiede, die die Gesamtvarianz von Y_{it} bedingen. Je größer der Konsistenzkoeffizient ist, desto größer ist der Trait-Anteil an der Messung zu Messgelegenheit t . Der situative Einfluss ist dann entsprechend geringer.

² Der Begriff „Spezifität“, wie er im Rahmen der LST-Theorie gebraucht wird, ist nicht zu verwechseln mit dem Spezifitätsbegriff in der ROC-Analyse (▶ Kap. 9).

Der *Spezifitätskoeffizient Spe* (Y_{it}) (vgl. Gl. 26.11) ist das Gegenstück zum Konsistenzkoeffizienten. Er gibt den Anteil der systematischen Varianz der latenten State-Residuum-Variablen ζ_{it} an der Gesamtvarianz der Variablen Y_{it} an und beschreibt den Anteil an der Gesamtvarianz, der (a) auf die Situation sowie (b) auf die Interaktion von Person und Situation zurückzuführen ist. Ein hoher Spezifitätskoeffizient bedeutet, dass ein hoher situativer Einfluss auf die Messung vorliegt (d. h. ein Einfluss der Situation und der Interaktion von Person und Situation).

Bei gegebener Reliabilität sind die Konsistenz- und die Spezifitätsanteile der Varianz einer Antwortvariablen Y_{it} somit gegenläufig: Je größer der eine Koeffizient bzw. Anteil ist, desto kleiner ist der andere. Beide Anteile repräsentieren reliable Anteile der Gesamtvarianz, d. h. Anteile, die durch Person und Situation erkläbar sind.

Je nach Messintention werden die Messinstrumente (d. h. die Items und damit die Tests) so konstruiert, dass entweder der Konsistenzanteil (bei der Messung von Dispositionen) oder der Spezifitätsanteil (bei der Messung von situativen Einflüssen) überwiegt.

Spezifitätskoeffizient *Spe* (Y_{it})

Gegenläufigkeit von Konsistenz und Spezifität

26.2.4 Methodeneffekte

Traits und States werden im Rahmen der LST-Theorie anhand von mindestens zwei beobachtbaren Variablen zu mindestens zwei Messzeitpunkten gemessen. Da jede Variable mittels einer bestimmten Methode (z. B. Selbsteinschätzung, Fremdeinschätzung) gemessen wird, hat diese Methode ebenfalls einen Einfluss auf die Messung. Dies entspricht der Annahme von Campbell und Fiske (1959), dass Traits nicht unabhängig von der verwendeten Methode erfasst werden können, sondern dass sich jede Messung aus einer systematischen *Trait-Methoden-Einheit* und einem unsystematischen Fehleranteil zusammensetzt (vgl. auch ► Kap. 25 und 27). Somit werden Methodeneffekte inzwischen als integrale Bestandteile der Messungen angesehen, die genauer analysiert werden sollten (vgl. Eid et al. 2016; Eid et al. 2003; Pohl et al. 2008).

Um den Einfluss einer Methode bestimmen zu können, werden mehrere Methoden benötigt, also z. B. Selbst- und Fremdeinschätzungen, gemessen zu jeder Messgelegenheit (vgl. z. B. Campbell und Fiske 1959; Eid et al. 2003, 2008).

Wurde nur eine einzige Methode (z. B. Selbsteinschätzung anhand von Items eines Fragebogens) zur wiederholten Messung eines Konstrukt eingesetzt, so können entweder die Items selbst als Indikatoren verwendet werden oder durch Aufsummierung eines Teils der Items können aber auch (Item-)Parcels (z. B. Testhälften) gebildet und die Antwortvariablen der Testhälften als Indikatoren des Traits oder des States verwendet werden. Parcels werden z. B. dann verwendet, wenn kontinuierliche Variablen benötigt werden, die möglichst normalverteilt sein sollten. Weisen die Ratingskalen der Items nur wenige Kategorien auf oder sind die Antwortvariablen schiefverteilt, so würde die Bildung von Parcels eine Alternative darstellen. Unterscheiden sich die Parcels geringfügig inhaltlich oder in den Formulierungen der Items, so können diese als unterschiedliche Methoden angesehen und behandelt werden.

Werden unterschiedliche Methoden verwendet, sollten diese im LST-Modell adäquat repräsentiert werden. Hierfür stehen verschiedene Möglichkeiten zur Verfügung. In der Vergangenheit wurde häufig das LST-Modell mit orthogonalen Methodenfaktoren verwendet. In dem „orthogonalen Methodenfaktoransatz“ (Geiser und Lockhart 2012) werden alle Methoden als zusätzliche Faktoren ins Modell aufgenommen und diese als unabhängig voneinander (orthogonal) definiert (Steyer et al. 1992).

Bildung von (Item-)Parcels

LST-Modell mit orthogonalen Methodenfaktoren

Unklare Interpretation der Methodenfaktoren

LST-Modell mit indikatorsspezifischen Trait-Faktoren

LST-Modell mit ($M - 1$) Methodenfaktoren

Vorteile des LST-Modells mit orthogonalen Methodenfaktoren

1. Methoden und Traits werden als unterschiedliche Faktoren definiert.
2. Alle Methoden sind durch einen eigenen Methodenfaktor repräsentiert.
3. Die Methodenspezifität kann als Teil der Reliabilität berechnet werden.

Nach Eid (2000) sowie Geiser und Lockhart (2012) können bei diesem Ansatz aber verschiedene Probleme auftreten. Zunächst ist nicht klar, wie die Methodenfaktoren zu interpretieren sind, da es sich sowohl um spezifische Traits als auch um Residuen hinsichtlich des latenten Traits handeln könnte. Des Weiteren gibt es zumeist keine ausreichende theoretische Basis für die Annahme, dass die Methodenfaktoren unkorreliert sein müssen. In praktischen Anwendungen weist zudem häufig zumindest einer der Methodenfaktoren eine nicht signifikante latente Varianz oder nicht signifikante Faktorladungen auf, was schwierig zu interpretieren ist.

Als Alternative bietet sich ein LST-Modell mit indikatorsspezifischen Trait-Faktoren an (Eid 1996; Geiser und Lockhart 2012; Marsh und Grayson 1995; Steyer et al. 1999). In diesem Modell sind so viele Trait-Faktoren vorhanden, wie es unterschiedliche Methoden gibt. Würden z. B. zwei Methoden verwendet (Selbsteinschätzung, Fremdeinschätzung), so müssten zwei Trait-Faktoren ins Modell aufgenommen werden. Dabei laden die Variablen der Selbsteinschätzung, gemessen zu allen Messzeitpunkten, jeweils auf dem ersten Trait-Faktor und die Variablen der Fremdeinschätzung, gemessen zu allen Messzeitpunkten, jeweils auf dem zweiten Trait-Faktor. Die beiden Trait-Faktoren sind Trait-Methoden-Einheiten, die miteinander hoch korrelieren sollten, da beide dasselbe Konstrukt erfassen.

Vorteile des LST-Modells mit indikatorsspezifischen Trait-Faktoren

1. Alle Konstrukte können als Konzepte der LST-Theorie definiert werden.
2. Getrennte Methodenfaktoren sind nicht nötig, da die Methodeneffekte in den indikatorsspezifischen Trait-Faktoren enthalten sind.
3. Das Ausmaß der Methodenspezifität findet sich in der Höhe der Korrelationen zwischen den indikatorsspezifischen Trait-Faktoren.

Gegenüber dem LST-Modell mit orthogonalen Methodenfaktoren hat das LST-Modell mit indikatorsspezifischen Trait-Faktoren einige Vorteile (Geiser und Lockhart 2012). Das LST-Modell mit indikatorsspezifischen Trait-Faktoren ist besonders dann zu empfehlen, wenn sich die Methoden zur Messung eines Traits (oder States) deutlich unterscheiden (z. B. Selbst- und Fremdeinschätzung).

In diesem Fall würde sich auch das LST-Modell mit ($M - 1$) korrelierten Methodenfaktoren anbieten, das eine Methode weniger enthält, als Methoden vorhanden sind. Diesem Modell liegt dieselbe Logik zugrunde wie dem Correlated-Trait-Correlated-(Method-minus-1)-Modell, kurz CTC($M - 1$)-Modell (s. ▶ Kap. 25, ▶ Abschn. 25.7; Eid 2000; Eid et al. 2008; Geiser und Lockhart 2012), in dem eine Methode als Referenzmethode definiert und für diese Methode kein Methodenfaktor spezifiziert wird (im Sinne der Trait-Methoden-Einheit), während die übrigen Methodenfaktoren miteinander korrelieren dürfen. In den Traits sind somit die Effekte der Referenzmethode enthalten, während die Nichtreferenzmethoden Abweichungen von dieser Referenzmethode darstellen.

Ein wesentlicher Vorteil bei der Interpretation eines Traits besteht darin, dass dieser analog zum CTC($M - 1$)-Ansatz methodenspezifisch interpretiert wird und der Trait damit eine klare Bedeutung bekommt.

26.3 · Modelltypen

Ein vergleichbarer Ansatz ist ein LST-Modell mit Methoden als Differenz von State-Variablen (Pohl und Steyer 2010; Pohl et al. 2008). Auch in diesem Modell wird eine Referenzmethode ausgewählt, die keinen eigenen Faktor aufweist.

Vorteile des LST-Modells mit $M - 1$ Methodenfaktoren

1. Alle latenten Variablen können als Konzepte der LST-Theorie definiert werden.
2. Der Anteil der Referenzmethode ist in der Trait-Varianz enthalten.
3. Die Methodenfaktoren haben eine klare Bedeutung: Sie stellen die Abweichungen von der Referenzmethode dar.

Eine weitere Möglichkeit, Methodeneffekte in längsschnittlichen Modellen in Betracht zu ziehen, besteht darin, Kovarianzen zwischen Messfehlervariablen des jeweils gleichen Indikators zu unterschiedlichen Messgelegenheiten zuzulassen. Dieser Ansatz wird in der Literatur auch *Correlated-Uniqueness-Ansatz* (CU-Ansatz) genannt (Geiser und Lockhart 2012; Kenny 1976). Im Gegensatz zu den vorher genannten Ansätzen behandelt der CU-Ansatz Methodeneffekte nicht als eigenständige Faktoren, sondern als Teil des Messfehlers.

Zusammenfassend lässt sich festhalten, dass die Berücksichtigung von Methodenfaktoren auf verschiedene Weise erfolgen kann. Die genannten Herangehensweisen werden in der jüngeren Literatur stark diskutiert. Im weiteren Verlauf dieses Einführungskapitels in die LST-Theorie konzentrieren wir uns aus didaktischen Erwägungen auf das LST-Modell mit indikator spezifischen Trait-Faktoren, da für dieses Modell die grundlegenden Definitionen von States und Traits der LST-Theorie ausreichen und keine weitergehenden Definitionen von Methodenfaktoren erforderlich sind.

26.3 Modelltypen

Auf Grundlage der dargestellten Konzepte lassen sich innerhalb der LST-Theorie unterschiedliche Modelltypen spezifizieren. Um die situativen und dispositionellen Anteile der Messung trennen zu können, werden Messinstrumente zu mehreren Messgelegenheiten eingesetzt. Auch innerhalb jeder Messgelegenheit müssen mehrere Messungen anhand von Tests, Testhälften oder Items vorliegen, um dispositionelle und situationsbedingte Einflüsse von Messfehlereinflüssen getrennt bestimmen zu können.

Um die Modelle möglichst einfach darstellen zu können, sollen nachfolgend sehr strenge Annahmen hinsichtlich der Messäquivalenz der Indikatorvariablen Y_{it} bezüglich der latenten State- und Methodenvariablen sowie der Messäquivalenz der latenten State-Variablen bezüglich der latenten Trait-Variablen gemacht werden. Diese Annahmen implizieren, dass die Faktorladungen der beiden Indikatoren innerhalb jeder der drei Messgelegenheiten als gleich angenommen werden, indem sie alle auf den Wert eins fixiert werden. Im empirischen Beispiel wird zusätzlich zu diesen Annahmen noch die Annahme von parallelen Messungen innerhalb jeder Messgelegenheit bezüglich der latenten State- bzw. Trait-Variablen angenommen, sodass die Fehlervarianzen innerhalb einer Messgelegenheit gleichgesetzt werden, sich diese Parameter aber über die Messgelegenheiten hinweg unterscheiden dürfen. Es sei explizit darauf hingewiesen, dass die LST-Theorie diese starken Annahmen nicht treffen muss (vgl. Steyer et al. 2015). Sie werden hier lediglich aus didaktischen Gründen vorgenommen.

Strenge Annahmen

Messäquivalenz

Messäquivalenz

In einem Messmodell können im Wesentlichen drei unterschiedlich strenge Formen der Messäquivalenz unterschieden werden, die mittels CFA getestet werden können und die für die Reliabilitätsschätzung von Bedeutung sind (► Kap. 24). Das „sparsamste“ Modell mit den strengsten Annahmen und somit der geringsten Anzahl zu schätzender Parameter sollte verwendet werden. Nachfolgend wird τ als Symbol für alle latenten Variablen verwendet.

- **τ -Kongenerität:** Alle Faktorladungen und alle Fehlervarianzen werden frei geschätzt.
- **τ -Äquivalenz:** Alle Faktorladungen werden als gleich angenommen und alle Fehlervarianzen frei geschätzt.
- **τ -Parallelität:** Alle Faktorladungen werden als gleich angenommen und alle Fehlervarianzen werden ebenfalls als gleich angenommen.

Anmerkung: Da in den LST-Modellen, die in diesem Kapitel dargestellt werden, die Mittelwerte keine Rolle spielen und deshalb zentrierte Variablen mit einem Mittelwert von null verwendet werden, können keine Annahmen bezüglich der Interzepte (vgl. ► Kap. 24) formuliert werden und der Begriff „essentiell“ wird hier nicht verwendet.

Wie bei allen Längsschnittmodellen ist es auch hier nötig, die Messinvarianz sicherzustellen, um sinnvolle Vergleiche über die Zeit vornehmen zu können (vgl. ► Kap. 24). Nachfolgend soll die schwache Invarianzbedingung mit identischen Faktorladungen über die Zeit verwendet werden.

Messinvarianz

Messinvarianz

Die Frage, ob ein Messinstrument über mehrere Messzeitpunkte hinweg dasselbe Merkmal mit derselben Messgenauigkeit misst, wird anhand von aufeinander aufbauenden Stufen der Messinvarianz überprüft (► Kap. 24). Hier werden wiederum die Mittelwerte der Variablen nicht berücksichtigt:

- **Konfigurale Invarianz** setzt dieselbe Anzahl an Faktoren sowie dieselbe Zuordnung der Items zu den Faktoren über die Messzeitpunkte hinweg voraus. Die Gleichheit der Struktur stellt eine Mindestvoraussetzung für weitere Invarianztests dar.
- **Schwache Invarianz** setzt neben konfiguraler Invarianz voraus, dass die Faktorladungen jeder Indikatorvariablen auf einen Faktor über die Messzeitpunkte hinweg gleich sind. Sind die Faktorladungen invariant, haben die Faktoren zu allen Messzeitpunkten dieselbe Bedeutung.
- **Strikte Invarianz** setzt zusätzlich voraus, dass die Fehlervarianzen für jedes Item über die Messzeitpunkte hinweg gleich sind. Sind auch die Fehlervarianzen invariant, dann sind manifeste Varianzunterschiede zwischen den Messzeitpunkten nur auf Varianzunterschiede der Faktoren zurückzuführen.

Äquivalenzhypotesen

Die aus spezifischen „Äquivalenzhypotesen“ (vgl. Steyer et al. 1992, 2015; Yousfi und Steyer 2006) resultierenden Modelle erlauben eine Bestimmung der Reliabilität, der Konsistenz und der Spezifität der Messungen Y_{it} . Wenn diese Kennwerte der Messungen bekannt sind, kann man beurteilen, ob das entwickelte Messinstrument seinem Anwendungszweck gerecht wird. Um beispielsweise die situative Ängstlichkeit einer Person zu messen, wäre es notwendig, dass das Instrument eine hohe Reliabilität aufweist, die auf einem hohen Anteil an situationsbedingter Varianz (Spezifität) basiert. Um die Trait-Angst einer Person zu messen, wäre es

26.3 · Modelltypen

dagegen notwendig, dass die hohe Reliabilität des Messinstruments auf einem hohen Anteil an trait-spezifischer Varianz (Konsistenz) basiert. Darüber hinaus lassen sich mit Modellen der LST-Theorie weitere inhaltliche Fragestellungen untersuchen, z. B. zur Stabilität oder zur Variabilität von Traits.

Im Folgenden sollen drei ausgewählte Modelltypen vorgestellt werden: das Multistate-Modell, das Multistate-Singltrait-Modell und das Multistate-Multi-trait-Modell mit indikatorsspezifischen Trait-Faktoren (Geiser und Lockhart 2012; Steyer et al. 1999, 2015; Yousfi und Steyer 2006). Zur Vereinfachung wurde hier angenommen, dass sich der Trait über die Zeit nicht verändert.

26.3.1 Multistate-Modell

Das Multistate-Modell beinhaltet in unserem Beispiel (unter Verwendung von restriktiven Äquivalenzannahmen) auf Personenebene die Annahme, dass zwei Messwerte y_{vit} und y_{vjt} einer Person v bei zwei Messinstrumenten i und j (z. B. zwei Testhälften) zu Messgelegenheit t gleiche latente State-Werte ($\tau_{vit} = \tau_{vjt} = \tau_{vt}$) aufweisen:

$$y_{vit} = \tau_{vit} + \varepsilon_{vit} = \tau_{vt} + \varepsilon_{vit} \quad (26.13)$$

$$y_{vjt} = \tau_{vjt} + \varepsilon_{vjt} = \tau_{vt} + \varepsilon_{vjt} \quad (26.14)$$

Auf Variablenebene bedeutet dies, dass die Antwortvariablen Y_{it} und Y_{jt} zu Messgelegenheit t τ -äquivalente latente State-Variablen τ_{it} und τ_{jt} darstellen, woraus eine gemeinsame latente State-Variable τ_t definiert wird:

$$\tau_{it} = \tau_{jt} =: \tau_t \quad (26.15)$$

In Abb. 26.3 wird die Äquivalenzannahme gleicher State-Variablen ($\tau_{it} = \tau_{jt} = \tau_t$) zu Messgelegenheit t dadurch veranschaulicht, dass in dem Pfaddiagramm die Ladungen der beiden Indikatoren auf den latenten Variablen τ_1 und τ_2 jeweils auf den gleichen Wert eins fixiert sind.

Wie Abb. 26.3 entnommen werden kann, werden hierbei zwei Messinstrumente (bzw. zwei Testhälften oder zwei Items) zu zwei Messgelegenheiten verwendet. Unter der Annahme, dass die Variablen der wahren Werte der beiden Messungen tatsächlich zur Messgelegenheit t identisch sind, lässt sich die Reliabilität des Indikators i wie folgt darstellen:

$$Rel(Y_{it}) = \frac{Var(\tau_t)}{Var(Y_{it})} = \frac{Cov(\tau_t, \tau_t)}{Var(Y_{it})} = \frac{Cov(\tau_{it}, \tau_{jt})}{Var(Y_{it})} = \frac{Cov(Y_{it}, Y_{jt})}{Var(Y_{it})} \quad (26.16)$$

Dabei ist $Cov(\tau_{it}, \tau_{jt}) = Cov(Y_{it}, Y_{jt})$, da die Messfehlervariablen unteineander und mit den Variablen der wahren Werte unkorreliert sind (vgl. ► Kap. 24).

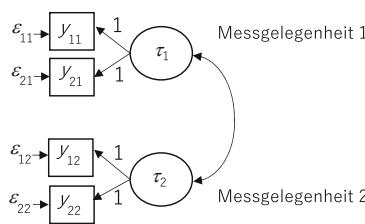


Abb. 26.3 Multistate-Modell mit jeweils zwei Indikatoren (z. B. Tests, Testhälften, Items), gemessen zu zwei Messgelegenheiten

Stabilität der latenten State-Variable

Wie man □ Abb. 26.3 darüber hinaus auch entnehmen kann, ist eine Modellierung der Stabilität der latenten State-Variablen anhand der Schätzung der Korrelation zwischen τ_1 und τ_2 möglich. Eine Trennung von Konsistenz und Spezifität ist im Multistate-Modell hingegen nicht möglich, da keine Modellierung der Trait-Variablen und des State-Residuums vorgenommen wurde.

26.3.2 Multistate-Singletrait-Modell**Keine Veränderung des Traits über die Zeit**

Eine solche Trennung von Konsistenz und Spezifität ist erst in einem Multistate-Singletrait-Modell möglich.

Unter der Annahme, dass keine Veränderung des Traits über die Zeit stattfindet und dass die Trait-Variablen beider Indikatoren parallele Messungen sind, kann man die Trait-Variablen (für zwei Indikatoren und zwei Messzeitpunkte) ξ_{11} , ξ_{21} , ξ_{12} und ξ_{22} durch eine einzige Trait-Variable ξ ersetzen. Formal bedeutet das also:

$$\xi_{it} = \xi, \quad \text{für alle } i, t. \quad (26.17)$$

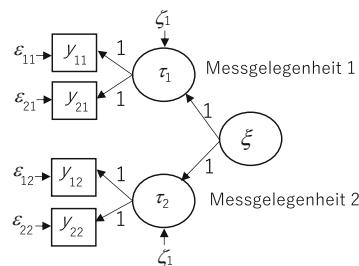
Für die Antwortvariable Y_{it} ergibt sich mit diesen zusätzlichen Annahmen die folgende Gleichung:

$$\begin{aligned} Y_{it} &= \tau_t + \varepsilon_{it} \\ &= \overbrace{\xi}^{\xi_t} + \overbrace{\zeta_t}^{\zeta_i} + \varepsilon_{it} \end{aligned} \quad (26.18)$$

Angenommen wird hier, dass die latente State-Residuum-Variable ζ_t und die Messfehlervariable ε_{it} untereinander und mit der latenten Trait-Variable ξ unkorreliert sind.

In □ Abb. 26.4 ist das Modell als Pfaddiagramm veranschaulicht, wie es z. B. im Rahmen von CFA (vgl. ▶ Kap. 24) erstellt und getestet werden könnte. Auch hier sind die Äquivalenzannahmen bezüglich der latenten Variablen durch Ladungen/Koeffizienten mit auf eins fixierten Werten repräsentiert.

In diesem Modell wird von der Stabilität der Trait-Variable ξ ausgegangen. Die Stabilität lässt sich konfirmatorisch aber nur testen, wenn mindestens vier Messgelegenheiten vorliegen. Dann könnte man den Trait zu den ersten beiden Messgelegenheiten mit dem Trait zu den letzten beiden Messgelegenheiten miteinander vergleichen (Steyer et al. 2015).



□ Abb. 26.4 Multistate-Singletrait-Modell

Reliabilität im Multistate-Singltrait-Modell

Unter der Gültigkeit der oben gemachten Annahmen lässt sich die *Reliabilität* des Indikators i (bzw. der Testhälfte, des Items) als Summe aus Konsistenz und Spezifität analog zum Multistate-Modell schätzen:

$$Rel(Y_{it}) = Con(Y_{it}) + Spe(Y_{it}) \quad (26.19)$$

Die *Konsistenz* wird wie folgt berechnet, wobei t und s unterschiedliche Messgelegenheiten bezeichnen und i und j unterschiedliche Indikatoren:

$$Con(Y_{it}) = \frac{Var(\xi)}{Var(Y_{it})} = \frac{Cov(\tau_t, \tau_s)}{Var(Y_{it})} = \frac{Cov(Y_{it}, Y_{js})}{Var(Y_{it})} \quad (26.20)$$

Die *Spezifität* ergibt sich wie folgt:

$$Spe(Y_{it}) = \frac{Var(\xi_t)}{Var(Y_{it})} = \frac{Cov(Y_{it}, Y_{jt}) - Cov(Y_{it}, Y_{js})}{Var(Y_{it})} \quad (26.21)$$

Die Reliabilität setzt sich somit additiv aus der Konsistenz und der Spezifität zusammen (vgl. auch Gl. 26.25):

$$\begin{aligned} Rel(Y_{it}) &= \frac{Var(\tau_t)}{Var(Y_{it})} = \frac{Var(\xi) + Var(\xi_t)}{Var(Y_{it})} \\ &= \frac{Cov(Y_{it}, Y_{js}) + [Cov(Y_{it}, Y_{jt}) - Cov(Y_{it}, Y_{js})]}{Var(Y_{it})} \\ &= \frac{Cov(Y_{it}, Y_{jt})}{Var(Y_{it})} \end{aligned} \quad (26.22)$$

Dies gilt unter der Annahme, dass die Ausprägungen der latenten Trait-Variablen über die Messgelegenheiten konstant und die State-Residuen unabhängig voneinander sind. Über die Messgelegenheiten und Antwortvariablen hinweg variieren lediglich die Messfehler und die State-Residuen. Unterschiedliche Messwerte einer Person v sind nur durch Messfehler und Situationseinflüsse sowie die Interaktion zwischen Person und Situation bedingt.

26.3.3 Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren

Ein weiterer interessanter Modelltyp ergibt sich, wenn man annimmt, dass jeder Indikator (z. B. Testhälfte) auf seinem eigenen indikatorsspezifischen Trait lädt, sodass im Modell so viele Trait-Variablen enthalten sind, wie es unterschiedliche Indikatoren (z. B. Testhälften) pro Messzeitpunkt gibt (Eid 1996; Geiser und Lockhart 2012; Steyer et al. 1999). Die Trait-Variablen sollten bei homogenen Indikatoren (Antwortvariablen) hoch miteinander korrelieren.

Der Index i der Trait-Variablen ξ_i verdeutlicht, dass die Trait-Variable nicht mehr einen generellen Trait abbildet, sondern dass der Trait ξ_i spezifisch ist für den Indikator i , gemessen zu mehreren Messgelegenheiten ($\xi_{it} = \xi_{is} = \xi_i$). Das bedeutet, dass es pro Indikator jeweils eine über die Messgelegenheiten hinweg stabile latente Trait-Variable ξ_i gibt:

$$Y_{it} = \xi_i + \zeta_t + \varepsilon_{it} \quad (26.23)$$

**Keine Veränderung
der indikatorsspezifischen Traits
über die Zeit**

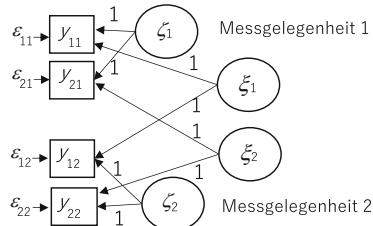


Abb. 26.5 Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Variablen ξ_1 und ξ_2 sowie State-Residuen ζ_1 und ζ_2

Unterschiedliche Trait-Variablen ξ_i und ξ_j zweier verschiedener Antwortvariablen Y_i und Y_j

Schätzung der Homogenität von Messinstrumenten

Gleichheit von State-Residuen innerhalb einer Messgelegenheit

Ein Modell mit indikatorspezifischen Trait-Faktoren hat mehrere Vorteile (Geiser und Lockhart 2012). So basieren alle latenten Variablen auf den Konzepten der LST- bzw. LST-R-Theorie und sind damit klar definiert. Des Weiteren werden keine Methodenfaktoren oder Fehlervarianzen ins Modell aufgenommen, da diese in der Kovarianz der Trait-Variablen enthalten sind. Die latenten Trait-Variablen ξ_i und ξ_j zweier verschiedener Antwortvariablen Y_i und Y_j können sich unterscheiden. Y_i und Y_j könnten z. B. zwei Fragebogenhälften zur Messung der Depressivität oder Ängstlichkeit darstellen.

Abb. 26.5 zeigt das Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Variablen anhand eines Pfaddiagramms.

Die Bestimmung der Korrelation zwischen den Trait-Variablen ξ_i und ξ_j erlaubt eine Schätzung der Homogenität von je zwei verschiedenen Messinstrumenten i und j über zwei verschiedene Messgelegenheiten t und s hinweg, z. B. eine Schätzung der Homogenität von zwei Fragebogen- oder Testhälften. Eine Populationskorrelation der indikatorspezifischen Trait-Variablen von 1.0 würde bedeuten, dass die beiden Messungen eindimensional sind und somit keine Methodeneffekte beinhalten. Eine geringe Korrelation würde dagegen bedeuten, dass starke Methodeneffekte vorliegen und damit deutliche Unterschiede zwischen den Messungen bestehen. Ebenso wie im CTC($M - 1$)-Modell (Eid 2000; s. auch ► Kap. 25) ist auch hier der Methodenanteil in der indikatorspezifischen Trait-Variablen enthalten.

Wie auch in den Modellen zuvor wird davon ausgegangen, dass die latenten State-Residuen ζ_{it} und ζ_{jt} innerhalb einer Messgelegenheit t identisch (d. h. $\zeta_{it} = \zeta_{jt} = \zeta_t$) und über die Messgelegenheiten hinweg unkorreliert sind. Darüber hinaus werden alle Messfehlervariablen voneinander als auch von den latenten Trait-Variablen sowie von den State-Residuen als unabhängig angenommen.

Die Antwortvariable Y_{it} lässt sich nun in die folgenden Komponenten zerlegen:

$$Y_{it} = \tau_{it} + \varepsilon_{it} \\ = \overbrace{\xi_i + \zeta_t} + \varepsilon_{it} \quad (26.24)$$

Die Reliabilität einer Antwortvariablen Y_{it} (z. B. Test- oder Fragebogenhälfte i zu Messgelegenheit t) erhält man ähnlich wie in den vorangegangenen Modellen (analog auch für die Messgelegenheit s):

$$\begin{aligned} Rel(Y_{it}) &= \frac{Var(\tau_{it})}{Var(Y_{it})} = \frac{Var(\xi_i) + Var(\zeta_t)}{Var(Y_{it})} \\ &= \frac{Cov(Y_{it}, Y_{is}) + Cov(Y_{it}, Y_{jt}) - Cov(Y_{jt}, Y_{is})}{Var(Y_{it})} \end{aligned} \quad (26.25)$$

Zur Bestimmung der *Konsistenz* wird die Varianz des indikatorsspezifischen Traits benötigt:

$$Con(Y_{it}) = \frac{Var(\xi_i)}{Var(Y_{it})} = \frac{Cov(\xi_{it}, \xi_{is})}{Var(Y_{it})} = \frac{Cov(\tau_{it}, \tau_{is})}{Var(Y_{it})} = \frac{Cov(Y_{it}, Y_{is})}{Var(Y_{it})} \quad (26.26)$$

Die *Spezifität* ergibt sich wiederum als:

$$\begin{aligned} Spe(Y_{it}) &= \frac{Var(\zeta_t)}{Var(Y_{it})} = \frac{Cov(Y_{it}, Y_{jt}) - Cov(\xi_i, \xi_j)}{Var(Y_{it})} \\ &= \frac{Cov(Y_{it}, Y_{jt}) - Cov(Y_{jt}, Y_{is})}{Var(Y_{it})} \end{aligned} \quad (26.27)$$

Wie zu sehen ist (vgl. □ Abb. 26.5), kann man unter der Gültigkeit der Modellannahmen im Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren die Konsistenz der Antwortvariablen Y_{it} schätzen, indem man die Kovarianz $Cov(Y_{it}, Y_{is})$ der interessierenden Antwortvariablen zu zwei verschiedenen Messgelegenheiten t und s an der Varianz $Var(Y_{it})$ der Antwortvariablen Y_{it} relativiert.

26.4 Anwendungen der LST-Theorie

26.4.1 Allgemeine Überlegungen und Voraussetzungen

Die Anwendung der LST-Theorie kann verschiedene Ziele haben. So kann man das Ziel verfolgen, Maße der Reliabilität, Konsistenz und Situationsspezifität zu schätzen, um die Messeigenschaften der Antwortvariablen eines Messinstruments (eines Items, einer Testhälfte oder eines Tests) zu bestimmen. Man kann aber ebenso das Ziel verfolgen, latente state-, trait- und messgelegenheitsspezifische Unterschiede durch weitere Variablen zu erklären. So ist es z. B. in der therapeutischen Praxis wichtig zu wissen, ob die Veränderungen der Depressivitätsmaße auf Veränderungen der Trait-Variablen oder auf situative Einflüsse zurückzuführen sind.

In jedem Fall lassen sich diese Ziele nur dann erreichen, wenn man Messungen zu mindestens zwei (besser: mindestens drei) Messgelegenheiten anhand von mindestens zwei (besser drei oder mehr) Tests oder Testhälften durchführt. Wie wir in ▶ Abschn. 26.3 gesehen haben, müssen darüber hinaus Modellannahmen hinsichtlich der Messäquivalenz von Variablen angestellt werden (s. dortige Ausführungen).

Hat man z. B. das Konstruktionsziel, ein Messinstrument zu entwickeln, das einen relativ stabilen Trait als Merkmal misst, dann geht damit die Vorstellung einher, dass in der Varianz der Antwortvariablen Y_{it} möglichst wenige situative Einflüsse nachweisbar sein sollen. Wird die Kontrolle potentiell situativer Einflüsse auf die Messung bei der Entwicklung des Messinstruments vernachlässigt, so kann die Anwendung des Messinstruments leicht zu Fehlinterpretationen führen. Die Konstruktvalidität wäre etwa dann gemindert, wenn situative Einflüsse irrtümlich als Trait-Einflüsse interpretiert würden. Das Ziel, „im Wesentlichen einen Trait zu messen“, setzt also voraus, dass das Messinstrument eine hohe Reliabilität und einen hohen Konsistenzkoeffizienten aufweist. Von Interesse ist es zudem, die Konsistenz in einen Trait- und einen Methodenanteil zu unterteilen.

Die Separierbarkeit der beiden Konzepte „Konsistenz“ und „Spezifität“ ermöglicht somit eine Präzisierung des Begriffs der *Messgenauigkeit*. Die Konsistenz ist als Messgenauigkeit hinsichtlich der Traits zu begreifen und die Reliabilität als Messgenauigkeit hinsichtlich der allgemeinen State-Variablen. Die Spezifität bezeichnet den Anteil an der Reliabilität, der auf die situativen Einflüsse zurückgeht.

**Mehrere Messungen und
Messzeitpunkte nötig**

Konsistenz

**Präzisierung des Konzepts
der Messgenauigkeit**

Wie in ▶ Abschn. 26.1.1 bereits erwähnt, ist die State-Anst-Skala des STAI (Lau et al. 1981) ein Beispiel für eine gelungene Konstruktion zur Messung der State-Anst. Diese Skala weist explizit einen hohen Spezifitätsanteil auf, um situationsspezifische Einflüsse zu erfassen.

26.4.2 Empirisches Beispiel: Prüfungsangst

Test Anxiety Inventory (TAI-G)

Im Folgenden wollen wir anhand eines konkreten empirischen Beispiels zur Prüfungsangst zeigen, wie die Messeigenschaften der Skala „Interferenz“ der deutschen Version des „Test Anxiety Inventory“ (TAI-G; Hodapp 1991, 1996) überprüft und die Reliabilitätskoeffizienten auf der Basis der LST-Theorie geschätzt werden können. Die Analyse basiert auf einer Stichprobe von $N = 302$ Studierenden, die den TAI-G zu drei Messgelegenheiten im Abstand von jeweils zwei Wochen ausgefüllt haben (Keith et al. 2003). Überprüft werden soll hier, ob die Subskala „Interferenz“ des Prüfungsangstfragebogens eher einen Trait oder eher einen State misst.

Die Skala „Interferenz“ umfasst sechs Items, die sich auf Kognitionen beziehen, die aufgabenbezogene Handlungen behindern können. Beispielitems sind „Mir schießen plötzlich Gedanken durch den Kopf, die mich blockieren“ oder „Ich vergesse Dinge, weil ich einfach zu viel mit mir selbst beschäftigt bin“. Aus den sechs Items wurden durch Aufsummierung von jeweils drei Items zwei möglichst parallele Testhälften gebildet, die auch als „(Item-)Parcels“ bezeichnet werden (▶ Abschn. 26.2.4, vgl. Bandalos 2002, 2008). Der Vorteil der Bildung von Testhälften besteht u. a. darin, dass die Summenvariablen im Vergleich zu Items einerseits eine größere Anzahl an Abstufungen aufweisen und andererseits eher die Annahme der Normalverteilung erfüllen. Allerdings sollte die Bildung von Parcels gut begründet werden, da auch Nachteile mit dieser Methode verbunden sind (vgl. Little et al. 2013; Rhemtulla 2016).

Angenommen wird hier, dass sich einerseits der State pro Messgelegenheit gleichermaßen auf die Indikatoren auswirkt und sich andererseits der Trait über die Zeit nicht verändert. Für die LST-Modelle bedeutet das, dass sowohl die Faktorladungen auf die State-Variablen (Annahme der τ -Äquivalenz) als auch die Faktorladungen auf die Trait-Variable (Annahme der ξ -Äquivalenz) jeweils gleichgesetzt und somit auf eins fixiert werden.

Schließlich werden noch die Messfehlervarianzen pro Messzeitpunkt gleichgesetzt (vgl. ▶ Kap. 24). Zu jeder der drei Messgelegenheiten lagen somit dieselben zwei Testhälften zur Messung des Konstrukt „Interferenz“ vor, insgesamt also sechs Antwortvariablen (Indikatoren).

Im vorliegenden Beispiel werden die folgenden drei konkurrierenden Modelle analysiert (vgl. □ Abb. 26.6):

1. Multistate-Modell
2. Multistate-Singletrait-Modell
3. Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Faktoren

Wenn ein guter Modellfit vorliegt, kann davon ausgegangen werden, dass das jeweilige Modell zu den Daten passt und die Reliabilitätskoeffizienten des LST-Modells geschätzt werden können. Anhand der Modelle soll überprüft werden, ob mit den Messungen neben dem gemeinsamen Trait „Interferenz“ auch situationsspezifische Anteile erfasst werden.

Im *Modell 1*, dem Multistate-Modell (□ Abb. 26.6a), laden beide Indikatoren zu jedem Messzeitpunkt auf einer State-Variablen, die wiederum miteinander korreliert sind. Damit wird die Annahme geprüft, dass die Fragebogenhälften (Parcels) pro Messzeitpunkt ausschließlich einen gemeinsamen State messen. Angenommen

Parcels

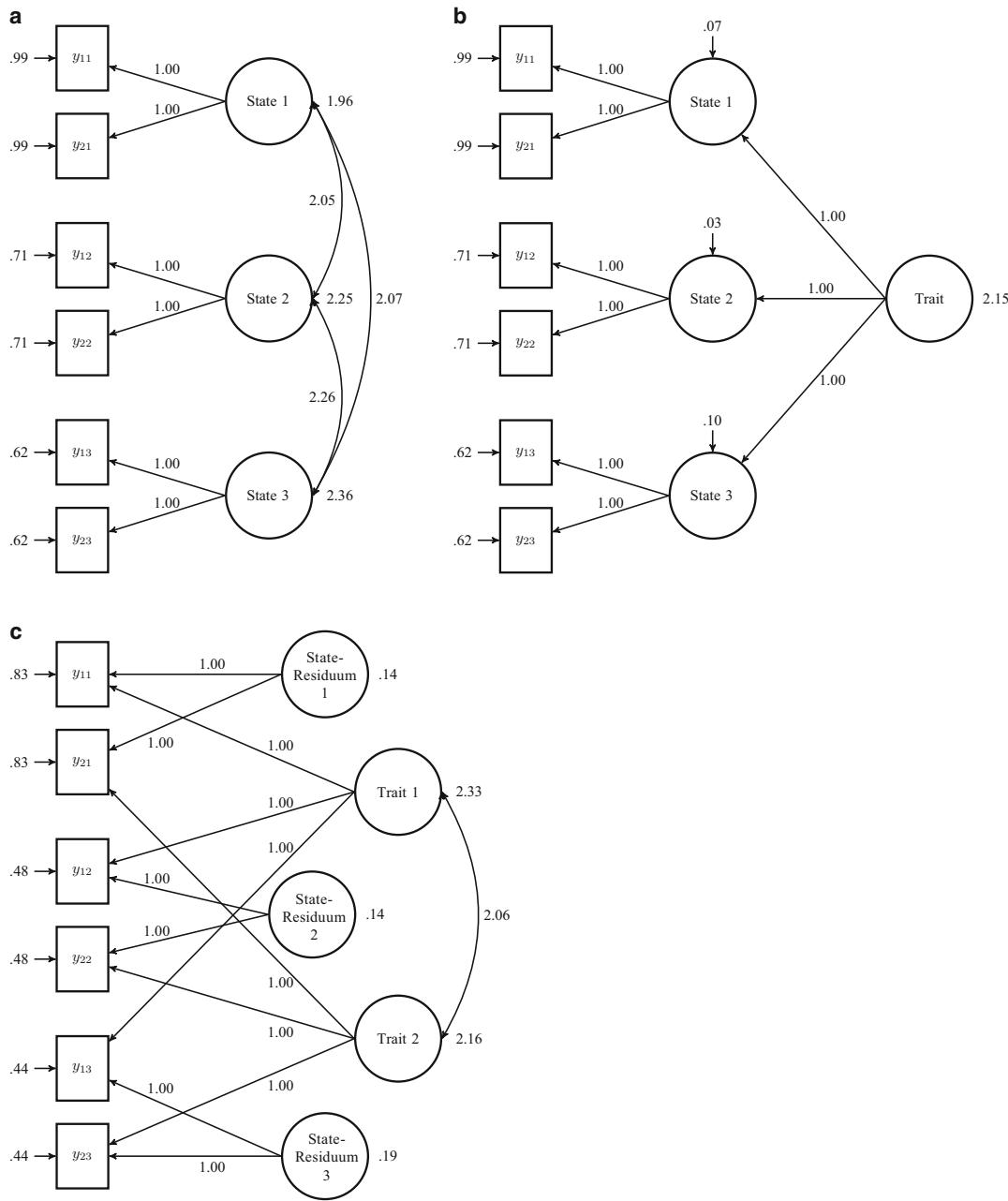
Messäquivalenz

Testhälften der Skala „Interferenz“

Analyse von drei Modellen

Multistate-Modell

26.4 · Anwendungen der LST-Theorie



■ Abb. 26.6 Unstandardisierte Parameterschätzungen von drei LST-Modellen: **a** Multistate-Modell (Modell 1), **b** Multistate-Singletrait-Modell (Modell 2) und **c** Multistate-Multitrait-Modell mit indikatorspezifischen Trait-Faktoren (Modell 3)

wird, dass alle Messungen identische Faktorladungen und Fehlervarianzen aufweisen (essentielle τ -Parallelität).

Im *Modell 2*, dem Multistate-Singletrait-Modell (■ Abb. 26.6b), wird zusätzlich zu den messzeitpunktspezifischen States ein gemeinsamer Trait ins Modell aufgenommen. Der durch den Trait nicht erklärte Anteil an den State-Variablen ist in den State-Residuen enthalten, die frei geschätzt werden. Damit kann die Annahme geprüft werden, dass die Indikatoren einen gemeinsamen Trait und zusätzlich auch situationsspezifische Anteile messen. Angenommen wird, dass pro Messzeitpunkt die beiden Faktorladungen erster Ordnung auf die State-Faktoren und die beiden Fehlervarianzen jeweils identisch sind (τ -Parallelität), ebenso wie die beiden Faktorladungen zweiter Ordnung auf den Trait-Faktor (ξ -Äquivalenz).

Multistate-Singletrait-Modell

Multistate-Multitrait-Modell mit methodenspezifischen Trait-Faktoren

Im *Modell 3*, dem Multistate-Multitrait-Modell mit methodenspezifischen Trait-Variablen (► Abb. 26.6c), wird für jede Messung ein indikatorspezifischer Trait angenommen. Somit sind Trait- und Methodeneffekte im methodenspezifischen Trait-Faktor enthalten, wobei jeder Indikator einen eigenen Trait misst. Die Kovarianz zwischen den methodenspezifischen Trait-Faktoren repräsentiert die gemeinsame Trait-Varianz, die unabhängig von den Methoden ist. Ist die Korrelation 1.0, so sind keine spezifischen Methodenanteile in den Messungen enthalten. Angenommen wird hier, dass die drei Faktorladungen des ersten Parcels auf dem indikatorspezifischen Trait-Faktor 1 und die drei Faktorladungen des zweiten Parcels auf dem indikatorspezifischen Trait-Faktor 2 jeweils identisch sind (ξ -Äquivalenz), ebenso wie die Faktorladungen der beiden Parcels auf die State-Residuen zu den drei Messzeitpunkten und die beiden Fehlervarianzen zu demselben Messzeitpunkt (ζ -Parallelität).

26.4.3 Konfirmatorische Beurteilung der Modellgüte im Beispiel

Gütekriterien

Die Parameterschätzung erfolgte jeweils unter Verwendung der Maximum-Likelihood-Methode des Programms *Mplus*, Version 8.0 (Muthén und Muthén 2017). Hinweis: Alternativ kann die kostenfreie R-project-Software und das dortige laavaan package (Rosseel 2012) verwendet werden, für das wir online Syntax bereitstellen (► Abschn. 26.6). Zur konfirmatorischen Beurteilung der Modellgüte wurden verschiedene Gütekriterien verwendet. Der χ^2 -Test ist das einzige Verfahren zur *inferenzstatistischen* Beurteilung der Modellgüte (vgl. Schermelleh-Engel et al. 2003) und sollte deshalb immer zusammen mit dem *p*-Wert berichtet werden. Zusätzlich wurden deskriptive Gütekriterien herangezogen, der *Root Mean Square Error of Approximation* (RMSEA) zusammen mit seinem Konfidenzintervall, der *Comparative Fit Index* (CFI), der *Tucker Lewis Index* (TLI) und der *Index Standardized Root Mean Square Residual* (SRMR). Für einen guten Modellfit sollten folgende Werte vorliegen: *p*-Wert > .01, RMSEA ≤ .05, CFI ≥ .97, TLI ≥ .97 und SRMR ≤ .05 (► Kap. 24).

Die beiden Modelle 1 und 2 unterscheiden sich nur darin, dass im Multistate-Modell die State-Variablen miteinander korrelieren, während sie im LST-Modell auf einem gemeinsamen Trait mit identischen Faktorladungen laden, sodass in diesem Modell die Trait-Anteile von den situationsbedingten Anteilen getrennt werden können. Die Modelltests zeigen, dass beide Modelle keinen zufriedenstellenden Modellfit aufweisen (► Tab. 26.1), da der χ^2 -Test jeweils signifikant ist ($p < .01$), wobei das Multistate-Modell einen χ^2 -Wert von 62.46 (12 *df*) aufweist und das Multistate-Singletrait-Modell einen χ^2 -Wert von 68.69 (14 *df*). Die beiden Modelle sind somit nicht mit den Daten vereinbar und müssen verworfen werden.

► Tabelle 26.1 Ergebnisse der Modellanpassung

Modell	χ^2 -Wert	<i>df</i>	<i>p</i>	RMSEA	90 %-KI	CFI	TLI	SRMR
1. Multistate-Modell	62.463	12	.000	.118	[.090, .148]	.967	.959	.034
2. Multistate-Singletrait-Modell	68.694	14	.000	.114	[.088, .141]	.965	.962	.056
3. Multistate-Multitrait-Modell mit indikator-spezifischen Trait-Faktoren	15.378	12	.221	.031	[.000, .070]	.998	.997	.047

CFI = Comparative Fit Index; *df* = Freiheitsgrade; *p* = *p*-Wert; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; TLI = Tucker Lewis-Index; 90 %-KI = 90 %-Konfidenzintervall

werden, die Parameterschätzungen dürfen nicht interpretiert werden. Der schlechte Modellfit könnte daran liegen, dass keine Methodeneffekte berücksichtigt wurden oder dass die Restriktionen bezüglich der Messäquivalenz zu restriktiv sind. Die Methodeneffekte werden im Modell 3 berücksichtigt.

Modell 3, das Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren, das im Unterschied zu Modell 2 zwei Trait-Faktoren (je einen für jeden Indikator) enthält, weist hingegen einen guten Modellfit auf. Dies zeigt sich am nicht signifikanten χ^2 -Wert ($\chi^2(12) = 15.38, p = .22$), am RMSEA, der kleiner als .05 ist, mit einem Konfidenzintervall, das auf der linken Seite den Wert null umfasst (RMSEA = .03, KI: .00, .07), an den Gütemaßen CFI und TLI, die jeweils fast den optimalen Wert von 1.0 erreichen (CFI = .998, TLI = .997) und am SRMR, der kleiner als .05 ist (SRMR = .047). Somit kann das Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren angenommen werden.

Die Methodenanteile in Modell 3 sind in den Trait-Faktoren enthalten und können nicht separiert werden. Dafür wären Modelle nötig, die weiter in ► Abschn. 26.2.4 erwähnt sind.

26.4.4 Varianzzerlegung im Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren

Auf Grundlage der im Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren (Modell 3) enthaltenen Varianzanteile ($Var(Y_{it}) = Var(\xi_i) + Var(\zeta_i) + Var(\varepsilon_{it})$) lassen sich nun die Reliabilitätskoeffizienten für die beiden Testhälften zu den drei Messgelegenheiten schätzen (► Tab. 26.2). Eine Schätzung der Reliabilität des Tests als Summe der beiden Testhälften können nach der Formel des Omega-Koeffizienten vorgenommen werden (vgl. ► Kap. 15).

Wie die Ergebnisse für Modell 3 zeigen, sind die Reliabilitätskoeffizienten der beiden Testhälften über die Messzeitpunkte immer etwas unterschiedlich, indem die zweite Testhälfte niedrigere Reliabilitätskoeffizienten aufweist als die erste Testhälfte.

Die Konsistenz ergibt sich für die Antwortvariable jeder Testhälfte aus dem Verhältnis der über die Zeit konsistenten Varianzanteile, hier der indikatorsspezifischen Trait-Varianz, bezogen auf die Gesamtvarianz der jeweiligen Testhälfte. Zur ersten Messgelegenheit beträgt die Konsistenz der ersten Testhälfte $(2.33)/(2.33 + .83 + .14) = .71$, zur zweiten Messgelegenheit ist sie $(2.33)/(2.33 + .48 + .14) = .79$, und zur dritten Messgelegenheit beläuft sie sich ebenfalls auf $(2.33)/(2.33 + .44 + .19) = .79$.

Modell 3: Guter Modellfit

Reliabilität der Testwertvariable über Omega bestimmbar

Hohe Konsistenz zu den drei Messgelegenheiten

■ **Tabelle 26.2** LST-Koeffizienten der beiden Testhälften zur Messung des Konstrukt „Interferenz“ zu den drei Messgelegenheiten auf Basis des LST-Modells mit indikatorsspezifischen Trait-Faktoren (Modell 3)

	MG1		MG2		MG3	
	Y_{11}	Y_{21}	Y_{12}	Y_{22}	Y_{13}	Y_{23}
<i>Con</i>	.706	.690	.788	.775	.787	.774
<i>Spe</i>	.044	.046	.049	.052	.063	.067
<i>Rel</i>	.750	.736	.837	.826	.851	.841
95 %-KI(<i>Rel</i>)	[.696, .797]	[.680, .785]	[.794, .873]	[.781, .863]	[.812, .883]	[.800, .875]

MG = Messgelegenheit; Con = Konsistenz; Spe = Situationsspezifität; Rel = Reliabilität; 95 %-KI(*Rel*) = 95 %-Konfidenzintervall der Reliabilitätsschätzung; Rel = Con + Spe

Geringe Situationsspezifität zu den drei Messgelegenheiten

95 %-Konfidenzintervall der Reliabilitätsschätzungen

Die Spezifität ergibt sich aus dem Verhältnis der State-Residuum-Varianz zur Gesamtvarianz einer Testhälfte. Zur ersten Messgelegenheit beträgt die Spezifität der ersten Testhälfte $.14/3.31 = .04$, zur zweiten Messgelegenheit ist sie $.14/.296 = .05$, und zur dritten Messgelegenheit liegt sie bei $.19/2.97 = .06$. Die situationsspezifischen Einflüsse sind in diesem Beispiel somit sehr gering.

Das 95 %-Konfidenzintervall gibt den Wertebereich an, in dem der wahre Wert der Reliabilität mit einer Wahrscheinlichkeit von 95 % liegt. Hier wurde das asymmetrische Konfidenzintervall verwendet, das für Reliabilitätskoeffizienten, die auf den Wertebereich von null bis eins beschränkt sind, geeigneter ist als das symmetrische Konfidenzintervall (Raykov und Marcoulides 2011, S. 166; s. auch ► Kap. 15). Antwortvariablen der Testhälften sind natürlich weniger reliabel als der gesamte Test, sodass die Werte der unteren Grenze hier als zufriedenstellend interpretiert werden können mit Werten knapp unter .70 bis zu Werten knapp über .80. Das Konfidenzintervall sollte immer angegeben werden, um die Präzision der Reliabilitätsschätzung beurteilen zu können.

Anhand der LST-Modelle konnte somit gezeigt werden, dass die Antwortvariablen der Testhälften tatsächlich im Wesentlichen den Trait messen und dass nur geringe situative Einflüsse bestehen. Durch die Erweiterung der KTT zur LST-Theorie war es möglich, getrennte Schätzungen der Trait-Anteile sowie der situationsspezifischen Anteile an der wahren Varianz vorzunehmen. Auch wenn der Einfluss der Situation (und der Person \times Situation-Interaktion) im vorliegenden Beispiel relativ gering ist, so wurde doch deutlich, dass eine Messung nicht in einem situativen Vakuum stattfindet. Die Ergebnisse zeigen, dass Messungen abhängig sind vom Indikator und dass unterschiedliche Trait-Faktoren für die Indikatoren oder ggf. andere Möglichkeiten zur Modellierung von Methodeneffekten (► Abschn. 26.2.4) mitberücksichtigt werden sollten.

26.4.5 Erweiterungen der LST-Theorie

In den vergangenen Jahren wurden mehrere Erweiterungen der LST-Modelle vorgeschlagen, darunter LST-Modelle mit autoregressiven Effekten (u. a. Cole et al. 2005; Eid et al. 2012; Eid et al. 2017; Geiser et al. 2018), mit indikatorspezifischen Effekten (Eid et al. 1999; Geiser und Lockhart 2012) sowie mit festen oder zufälligen Situationen (Geiser et al. 2015b). Des Weiteren wurden hierarchische LST-Modelle (Schermelleh-Engel et al. 2004), latente Klassen-LST-Modelle (Eid und Langeheine 1999, 2007) und Mischverteilungs-LST-Modelle (Courvoisier et al. 2007) vorgeschlagen.

Diese Methoden beruhen jeweils auf unterschiedlichen Annahmen und führen damit zu unterschiedlichen Modellen und unterschiedlichen Interpretationen der latenten Variablen. Auf die Vielzahl dieser Modelle kann in einem Einführungs Kapitel nicht eingegangen werden, jedoch finden sich ausführliche Erläuterungen und weitere Literaturangaben z. B. bei Eid et al. (2008), Geiser und Lockhart (2012), Geiser et al. (2015a), Hintz et al. (2019), Koch et al. (2018) sowie Newsom (2015).

26.5 Zusammenfassung

Der State- und der Trait-Begriff sind in der Differentiellen Psychologie und in der psychologischen Diagnostik etablierte Konzepte. Der State-Begriff beschreibt einen Zustand, in dem sich eine Person in einer Situation befindet, während der Trait-Begriff eine mehr oder weniger zeitlich überdauernde Merkmalsausprägung beschreibt.

Ausgehend von der KTT führte dieses Kapitel zunächst in die formale Repräsentation der LST-Theorie ein, die eine Erweiterung der KTT darstellt. Dabei ist die

latente State-Variable τ_{it} die Variable der wahren Werte der Personen in Item i , gemessen in der Situation t , und entspricht der Variable der wahren Werte in der KTT. Mit ξ_{it} wurde die *latente Trait-Variable* eingeführt, die die Variable der Erwartungswerte der Personen in der Situation t darstellt. Die latente *State-Residuum-Variable* ζ_{it} entspricht formal der Differenz zwischen der latenten State-Variablen und der Trait-Variablen. Inhaltlich repräsentiert die State-Residuum-Varianz den Anteil an der wahren Varianz, der nicht durch die Person, sondern durch die Situation und die Interaktion von Person und Situation bedingt ist.

Nach der formalen Definition der Konzepte der Latent-State-Trait-Theorie erfolgte die Darstellung von drei typischen Modellen der LST-Theorie und ihren inhaltlichen Eigenschaften: das *Multistate-Modell*, das *Multistate-Singletrait-Modell* und das *Multistate-Multitrait-Modell mit indikatorsspezifischen Trait-Faktoren*. Abschließend wurden Anwendungen der LST-Theorie anhand eines empirischen Beispiels zur Prüfungsangst vorgestellt und die Schätzung der Reliabilitätskoeffizienten demonstriert.

26.6 EDV-Hinweise

Rechentechnische Hinweise anhand eines Datenbeispiels zur Durchführung einer LST-Analyse anhand von *Mplus* und *lavaan* in R finden sich im Bereich EDV-Hinweise unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

26.7 Kontrollfragen

- ?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <http://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).
1. Was versteht man unter einem Trait, was unter einem State?
 2. Welche Erweiterung erfolgt in der LST-Theorie gegenüber der KTT?
 3. In welche Koeffizienten wird der Reliabilitätskoeffizient in der LST-Theorie weiter zerlegt, wenn das Multistate-Singletrait-Modell zugrunde gelegt wird?
 4. In welcher Beziehung stehen der Konsistenzkoeffizient und der Spezifitätskoeffizient zueinander?
 5. Woran erkennt man in einem Multistate-Singletrait-Modell, ob die einzelnen Indikatoren eher einen Trait oder eher einen State messen?

Literatur

- Allport, G. W. (1937). *Personality, a psychological interpretation*. New York: Holt & Co.
- Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, 9, 78–102.
- Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, 15, 211–240.
- Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review*, 80, 307–336.
- Cattell, R. (1946). *The description and measurement of personality*. New York: World Book.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cole, D. A., Martin, N. M. & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, 10, 3–20.
- Courvoisier, D. S., Eid, M. & Nussbeck, F. W. (2007). Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods*, 12, 80–104.

- Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research – Online*, 1, 65–85.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Eid, M., Courvoisier, D. S. & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 384–406). New York, NY: Guilford Press.
- Eid, M., Geiser, C. & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science*, 25, 275–280.
- Eid, M., Holtmann, J., Santangelo, P. & Ebner-Priemer, U. (2017). On the definition of latent state-trait models with autoregressive effects: Insights from LST-R theory. *European Journal of Psychological Assessment*, 33, 285–295.
- Eid, M. & Langeheine, R. (1999). The measurement of consistency and occasion specificity with latent class models: A new model and its application to the measurement of affect. *Psychological Methods*, 4, 100–116.
- Eid, M. & Langeheine, R. (2007). Detecting population heterogeneity in stability and change of subjective well-being by mixture distribution models. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Handbook of methods in positive psychology*, (pp. 501–607). Oxford, NY: Oxford University Press.
- Eid, M., Lischetzke, T., Nußbeck, F. & Trierweiler, L. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod analysis: A multiple indicator CTC($M-1$) model. *Psychological Methods*, 8, 38–60.
- 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. & Schmidt, K. (2014). *Testtheorie*. Göttingen: Hogrefe.
- Eid, M., Schneider, C. & Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality*, 13, 283–306.
- Eysenck, H. (1947). *Dimensions of personality*. London: Routledge & Keagan Paul.
- Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and opportunity of within-person variability. *Current Directions in Psychological Science*, 13, 83–87.
- Furr, R. M. & Funder, C. F. (2019). Persons, situations, and person-situation interactions. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (4th ed.). New York: Guilford.
- Geiser, C., Hintz, F., Burns, G. L. & Servera, M. (2018). Latent variable modeling of person-situation data. In D. Funder, J. Rauthman, & R. Sherman (Eds.), *The Oxford handbook of psychological situations*. Oxford, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190263348.013.15>
- Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A. & Koch, T. (2015a). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavior Research Methods*, 47, 172–203.
- Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., Servera, M. & Shiffman, S. (2015b). Analyzing person, situation and person \times situation interaction effects: Latent state-trait models for the combination of random and fixed situations. *Psychological Methods*, 20, 165–192.
- Geiser, C. & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, 17, 255–283.
- Geiser, C., Lockhart, G., Keller, B. T., Eid, M., Koch, T. & Cole, D. A. (2015c). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavior Research Methods*, 47, 172–203.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Hintz, F., Geiser, C. & Shiffman, S. (2019). A latent state-trait model for analyzing states, traits, situations, method effects, and their interactions. *Journal of Personality*, 87, 434–454.
- Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten. *Zeitschrift für Pädagogische Psychologie*, 5, 121–130.
- Hodapp, V. (1996). The TAI-G: A multidimensional approach to the assessment of test anxiety. In C. Schwarzer & M. Zeidner (Eds.), *Stress, anxiety, and coping in academic settings* (pp. 95–130). Tübingen: Francke.
- Keith, N., Hodapp, V., Schermelleh-Engel, K. & Moosbrugger, H. (2003). Cross-sectional and longitudinal confirmatory factor models for the German test anxiety inventory: A construct validation. *Anxiety, Stress & Coping*, 16, 251–270.
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247–252.
- Koch, T., Eid, M. & Lochner, K. (2018). Multitrait-multimethod-analysis: The psychometric foundation of CFA-MTMM models. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing. Volume II: A multidisciplinary reference on survey, scale and test development* (pp. 781–846). Hoboken, NJ: Wiley.

- Koch, T., Schultze, M., Holtmann, J., Geiser, C. & Eid, M. (2017). A multimethod latent state-trait model for structurally different and interchangeable methods. *Psychometrika*, 82, 17–47.
- Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C. (1981). *STAI. Das State-Trait-Angst-Inventar: Theoretische Grundlagen und Handweisung*. Weinheim: Beltz Testgesellschaft.
- Little, T. D., Rhemtulla, M., Gibson, K. & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Marsh, H. W. & Grayson, D. A. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousand Oaks, CA: Sage.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Newsom, J. T. (2015). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. New York, NY: Routledge.
- Pohl, S. & Steyer, R. (2010). Modelling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*, 45, 1–28.
- Pohl, S., Steyer, R. & Kraus, K. (2008). Modeling method effects as individual causal effects. *Journal of the Royal Statistical Society, Series B*, 171, 41–63.
- Raykov, T. & Marcoulides, G. A. (2011). *Psychometric theory*. New York, NY: Routledge.
- Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, 21, 348–368.
- Rieger, S., Göllner, R., Spengler, M., Trautwein, U., Nagengast, B. & Roberts, B. W. (2017). Social cognitive constructs are just as stable as the Big Five between grades 5 and 8. *AERA Open*, 3, 1–9.
- Roberts, B. W. & Nickel, L. (2017). A critical evaluation of the neo-socioanalytic model of personality. In J. Specht (Ed.), *Personality development across the life span* (pp. 157–177). London: Elsevier.
- Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R. & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, 143, 117–141.
- 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.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36. Retrieved from <http://www.jstatsoft.org/v48/i02/> [29.12.2019]
- Schermelleh-Engel, K., Keith, N., Moosbrugger, H. & Hodapp, V. (2004). Decomposing person and occasion-specific effects: An extension of latent state-trait theory to hierarchical LST models. *Psychological Methods*, 9, 198–219.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research – Online*, 8 (2), 23–74.
- Schmitt, M. (1990). *Konsistenz als Persönlichkeitseigenschaft? Moderatorvariablen in der Persönlichkeits- und Einstellungsorschung*. Berlin: Springer.
- Spielberger, C. (1972). *Anxiety: Current trends in research*. London: Academic Press.
- Steyer, R. (1987). Konsistenz und Spezifität: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 8, 245–258.
- Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin, Heidelberg: Springer.
- Steyer, R., Ferring, D. & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, 8, 79–98.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389–408.
- Yousfi, S. & Steyer, R. (2006). Latent-State-Trait-Theorie. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 346–357). Göttingen: Hogrefe.
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, 40, 395–412.



Konvergente und diskriminante Validität über die Zeit: Integration von Multitrait- Multimethod-Modellen (MTMM-Modellen) und der Latent-State-Trait-Theorie (LST-Theorie)

Fridtjof W. Nussbeck, Michael Eid, Christian Geiser, Delphine S. Courvoisier und David A. Cole

Inhaltsverzeichnis

- 27.1 Einleitung – 715**
 - 27.1.1 MTMM-Modelle – 715
 - 27.1.2 Methodeneffekte – 716
 - 27.1.3 LST-Modelle – 718
 - 27.1.4 Beschränkungen der LST- und MTMM-Modelle – 719
 - 27.1.5 Verbindung beider Ansätze – 720
- 27.2 Längsschnittliche MTMM-Modelle – 721**
 - 27.2.1 Multioccasion-MTMM-Modell – 721
 - 27.2.2 Multiconstruct-LST-Modell – 722
 - 27.2.3 Multimethod-LST-Modell – 724
 - 27.2.4 Vergleich der drei längsschnittlichen MTMM-Modelle – 727
- 27.3 Multiconstruct-LST- und Multimethod-LST-Modell
in der empirischen Anwendung – 730**
 - 27.3.1 Ergebnisse mit dem Multiconstruct-LST-Modell – 730
 - 27.3.2 Ergebnisse mit dem Multimethod-LST-Modell – 733
 - 27.3.3 Fazit der Anwendungen der beiden multimethodalen
LST-Modelle – 735
- 27.4 Praktische Hinweise zur Analyse longitudinaler multimodaler
Modelle – 735**

27.5 Zusammenfassung – 736

27.6 EDV-Hinweise – 736

27.7 Kontrollfragen – 737

Literatur – 737

i Im diagnostisch-therapeutischen Kontext könnten folgende Fragen von hoher Relevanz sein: Wie stark ähneln sich Schüler- und Lehrereinschätzungen in Bezug auf z. B. Ängstlichkeit und Depressivität? Generalisieren Lehrer im Sinne eines Halo-Effekts (Thorndike 1920), kommen sie also zu nahezu identischen Einschätzungen von Ängstlichkeit und Depressivität? Wie sehr hängen die Einschätzungen der Lehrer von stabilen oder situativen Faktoren ab? Können Lehrer die Schwankungen der Ängstlichkeit ihrer Schüler nachvollziehen? Diese Fragen können weder mit LST- noch mit MTMM-Modellen umfassend beantwortet werden, sondern müssen in Kombinationen der beiden Ansätze untersucht werden. Im folgenden Abschnitt werden drei Modelle vorgestellt, die eine Beantwortung dieser Fragen ermöglichen.

27.1 Einleitung

Psychologische Merkmale und somit auch psychologische Messwerte unterliegen einer Vielzahl von Einflüssen. Beispielsweise hängt die Ausprägung der Ängstlichkeit von Schulkindern nicht nur von ihrer dispositionellen Ängstlichkeit als über Situationen hinweg stabile Eigenschaft ab, sondern auch von (momentan wirkenden) situativen Einflüssen, z. B. einem gerade wütenden Sturm oder dem Albtraum der letzten Nacht. Außerdem fallen die Messungen je nach Messmethode, z. B. ob die Schüler sich selbst einschätzen oder ob sie von ihren Lehrern oder Eltern eingeschätzt werden, unterschiedlich aus. In vorangegangenen Kapiteln wurden bereits statistische Modelle beschrieben, die verschiedene Einflussfaktoren auf Messungen trennbar und in ihrer Größe messbar machen können. Schermelleh-Engel, Geiser und Burns beschreiben in ► Kap. 25 Modelle, die den Einfluss verschiedener Messmethoden auf die Messergebnisse untersuchen. Die Messmethoden beinhalten ganz unterschiedliche inhaltliche Aspekte. Bezüglich der Messergebnisse verschiedener Fragebogen können mit den vorgestellten Multitrait-Multimethod-Modellen (MTMM-Modellen) z. B. die Einflüsse unterschiedlicher Beurteiler (*Rater*), unterschiedlicher Facetten eines Konstrukts oder verschiedener Messgelegenheiten analysiert werden. Liegen wiederholte Messungen vor, bieten sich Modelle der Latent-State-Trait-Theorie (LST-Theorie, vgl. ► Kap. 26; Steyer 1987, 1989; Steyer et al. 1992; Steyer et al. 2015) an. LST-Modelle teilen viele strukturelle Bestandteile mit den MTMM-Modellen mit latenten Variablen, entstammen jedoch einer eigenständigen Forschungstradition.

Psychologische Messungen unterliegen Einflüssen der Messmethode und der Situation

27.1.1 MTMM-Modelle

MTMM-Modelle können herangezogen werden, um die konvergente und die diskriminante Validität von psychologischen Messungen, beispielsweise von Fremdeinschätzungen (Lehrerratings) und Selbsteinschätzungen (Schülerratings) der Ängstlichkeit und Depressivität von Kindern zu bestimmen. Mithilfe von MTMM-Modellen lassen sich zunächst die Messfehler von den wahren Werten trennen. Die wahren Werte können dann in die Bestandteile, die auf den Einfluss des Konstrukts (z. B. Ängstlichkeit) und der Messmethode (Lehrer- oder Schülerrating) zurückzuführen sind, zerlegt werden. Je nach ausgewähltem Modell (s. Eid et al. 2006; Eid et al. 2008) erhält man einen gemeinsamen Faktor als Trait und zusätzlich Abweichungsvariablen für die Selbstratings der Schüler sowie für die Fremdratings der Lehrer als Methodenfaktoren. Beispiele hierfür sind Correlated-Trait-Uncorrelated-Method- (CTUM-) oder Correlated-Trait-Correlated-Method-Modelle (CTCM-Modelle, ► Abschn. 25.6). In diesen Modellen werden die Trait-Variablen wie ein Faktor in der Faktorenanalyse bestimmt. Alle Urteile mit allen Messmethoden tragen zur Schätzung des Faktorwertes (Ausprägung des Traits) bei. Die gemeinsamen systematischen Abweichungen der Urteile

Konvergente und diskriminante Validität

innerhalb der Methoden (also der gemeinsame Varianzanteil der Urteile, der nicht durch den Faktor erklärt werden kann) bilden die Methodenvariable.

Im Sinne des *Correlated-Trait-Correlated-(Method-minus-1)-Modells*, kurz CTC($M - 1$)-Modell (► Abschn. 25.7; Eid 2000), kann eine Methode als Standard gewählt werden, die allein zur Schätzung der Trait-Variablen herangezogen wird; die anderen Methoden werden gegen diese *Standardmethode* kontrastiert. Dienen beispielsweise die Ratings der Schüler als Standardmethode, so entspricht die Trait-Ausprägung dem True-Score-Wert der Schülerratings (also dem wahren Wert im Sinne der Klassischen Testtheorie, KTT, ► Kap. 13). In der True-Score-Variable der Schüler sind somit sowohl Einflüsse des tatsächlich zu messenden Merkmals als auch der Methode (Selbsteinschätzung) enthalten. Die Trait-Variable, die mit der Standardmethode gemessen wird, wird explizit als Variable aufgefasst, die diese beiden Komponenten enthält. Aus diesem Grund muss die Wahl der Standardmethode aus theoretischen Überlegungen erfolgen, wobei z. B. die Methode als Standard gewählt werden kann, von der die beste Einschätzung des zu messenden Merkmals zu erwarten ist. Die Trait-Variable dient als Prädiktor in einer latenten Regression zur Vorhersage der True-Score-Variablen der Lehrer (Nichtstandardmethoden). Die Abweichungen der True-Score-Werte der Lehrer werden dann in den Methodenfaktoren abgebildet (s. Eid et al. 2003; Eid et al. 2006, 2008; ► Kap. 25). Die Methodenfaktoren sind in diesem Modell Residuen einer latenten Regression; sie bilden Abweichung der Nichtstandardmethoden von der Vorhersage durch die Standardmethode ab.

27.1.2 Methodeneffekte

Im Rahmen von MTMM-Modellen gibt es verschiedene Möglichkeiten, Methodeneffekte zu konzeptualisieren. In den vorangegangenen Kapiteln (► Kap. 24, 25 und 26) wurden bereits mehrere auf Strukturgleichungsmodellen aufbauende Ansätze vorgestellt. Zur Modellierung ist es notwendig zu unterscheiden, welche inhaltlichen Fragen mit den Modellen beantwortet werden sollen und welche Datenstruktur vorliegt. Es geht darum, die Frage zu erörtern, wie der manifeste Messwert eines Konstrukt Y_{ikmt} ¹ (i = Indikator, in der Regel Items bzw. Itempäckchen², k = Konstrukt, m = Messmethode/Beurteiler/Rater/Informant und t = Messgelegenheit/Zeitpunkt/Occasion) mit bedeutungsvollen latenten Variablen erklärt werden kann.

Dabei ist es wichtig, die Art der eingesetzten Methoden zu beachten (Eid 2006; Eid et al. 2006, 2008). Messmethoden können untereinander austauschbar, strukturell unterschiedlich oder gleichwertig sein:

a. *Die Methoden sind untereinander austauschbar:*

- Dieser Fall tritt z. B. ein, wenn mehrere Messmethoden (Beurteiler) eingesetzt werden, sich keine der Methoden von den anderen abhebt und die Messmethoden somit statistisch einer Zufallsauswahl entsprechen. Dies wäre der Fall, wenn Schüler einer Klasse per Zufall ausgewählt werden, um die Qualität des Unterrichts einzuschätzen. Alle Schüler entstammen der gleichen Population (Schüler einer Klasse) und liefern somit gleichwertige Einschätzungen der Lehrqualität.
- Ein weiteres Beispiel sind Messzeitpunkte, z. B. wenn Schüler in regelmäßigen Abständen in Bezug auf ihre Ängstlichkeit befragt würden. Dabei darf es keine Kriterien für die Auswahl der Messgelegenheiten geben, die mit

¹ Der Index m beschreibt im Unterschied zu den vorangegangenen Kapiteln nicht die Zahl der Items (l, \dots, m), sondern eine ausgewählte Methode m .

² Um metrische Indikatoren zu erhalten, können mehrere Items, die ein Konstrukt messen, gemittelt werden. Die neu entstandene Variable nennt man Itempäckchen (Item-Parcel). Dieses Vorgehen ist nicht gänzlich unumstritten (s. dazu Little et al. 2002).

der Ausprägung des Konstrukts zusammenhängen, sodass sich die Schüler zu jeder Messgelegenheit in einer zufälligen Situation befinden, deren Einfluss auf das Konstrukt unbekannt ist. Dies wäre nicht mehr der Fall, wenn sie vor und nach einem Selbstbewusstseinstraining befragt würden, da davon auszugehen ist, dass die Ängstlichkeitseinschätzungen nach dem Training geringer ausfallen sollten. Für den Fall austauschbarer Methoden (Occurrences) ist es sinnvoll, den Trait als einen gemeinsamen Faktor zu definieren (z. B. im CTUM- oder LST-Modell; Eid et al. 2008; □ Abb. 27.1a). Die Abweichungen der austauschbaren Methoden (einer Messgelegenheit: O_{+kmt} von Occasion) vom gemeinsamen Faktor sind oft von substantiellem Interesse. Erklärende Konstrukte können evtl. herangezogen werden, um die Abweichung der Methoden vom gemeinsamen Faktor zu erklären.

- b. *Die Methoden sind nicht austauschbar; sie unterscheiden sich strukturell:* Werden z. B. Schüler und Lehrer zur Ängstlichkeit der Schüler befragt, stehen den Beurteilern unterschiedliche Informationen zur Verfügung, da die Schüler eine Selbsteinschätzung abgeben, während die Lehrer aus der Fremdperspektive beurteilen müssen (vgl. □ Abb. 27.1b). Im Schülerrating (Selbsteinschätzung) können die Schüler sowohl ihr eigenes Verhalten als auch ihr Erleben als Grundlage ihrer Einschätzungen nutzen. Im Lehrerrating (Fremdeinschätzung) müssen die Lehrer hingegen auf Verhaltensbeobachtungen oder Äußerungen der Schüler und Mutmaßungen über das Erleben der Schüler zurückgreifen. So gibt es Schüler, denen man ihre Angst „ansieht“, und wieder andere, die sie recht gut zu verbergen wissen. Manche Schüler sind vorsichtiger und werden alleine deshalb als ängstlicher eingeschätzt. Deshalb ist es sinnvoll, die Schülerratings als Standardmethode im CTC($M - 1$)-Modell einzusetzen und die Lehrerratings gegen diese zu kontrastieren. Ungünstig wäre es hingegen, einen Trait als das den verschiedenen Ratings gemeinsame Merkmal zu definieren: Denkt beispielsweise ein Kind, dass es das traurigste Kind der Welt ist, der Lehrer hingegen, dass es ein sehr glückliches Kind sei, so ist es nicht sinnvoll, den Mittelwert der beiden Ratings als „wahre Stimmung“ zu definieren. Viel aussagekräftiger ist es, beide Methoden zu kontrastieren, um untersuchen zu können, warum sich die Schüler- von den Lehrerratings unterscheiden.
- c. *Die Methoden sind gleichwertige Repräsentationen eines Traits*, z. B. Testhälften oder parallele Tests (□ Abb. 27.1c). In MTMM-Datensätzen kann es vorkommen, dass Indikatoren trotz sehr hoher Korrelationen mit den Indikatoren desselben Traits einen sehr spezifischen systematischen Varianzanteil binden. Dies liegt daran, dass sich die Inhalte der Items in den verschiedenen Indikatoren nicht vollständig replizieren, sondern neben den gemeinsamen sprachlichen Inhalten auch stets eigene Aspekte beinhalten.

Vor allem in longitudinalen Modellen kommt es oft vor, dass Indikatoren über Messgelegenheiten hinweg stärker miteinander korrelieren als mit den anderen Indikatoren des gleichen Konstrukts zur gleichen Messgelegenheit (Eid 1996; Marsh und Grayson 1994; Steyer et al. 1992). Dieser *Autokorrelationseffekt* entspricht der spezifischen, zeitlich stabilen und reliablen Varianz eines einzelnen Indikators. Die Autokorrelation kann man auch als Methodeneffekt des Indikators (oder indikator-spezifischen Effekt) interpretieren.

Oft ist dieser Effekt (der Unterschied zwischen den Indikatoren) aber nicht von substantiellem Interesse, sondern eher nebensächlich, wenn es darum geht, die Varianzquellen Methode, Messgelegenheit und Messfehler vom Trait zu trennen. Aus diesem Grund spezifiziert man in vielen Anwendungen separate Trait-Faktoren für beide Testhälften, obwohl beide zum gleichen Konstrukt gehören (vgl. ► Kap. 26). In den meisten Fällen korrelieren diese Faktoren sehr hoch positiv miteinander, was darauf hinweist, dass die indikator-spezifischen Effekte eher gering ausfallen. Bei den hier vorgestellten Modellen ist jedoch davon auszugehen, dass ein Autokorrelationseffekt eintritt.

Strukturell unterschiedliche Methoden

Gleichwertige Methoden

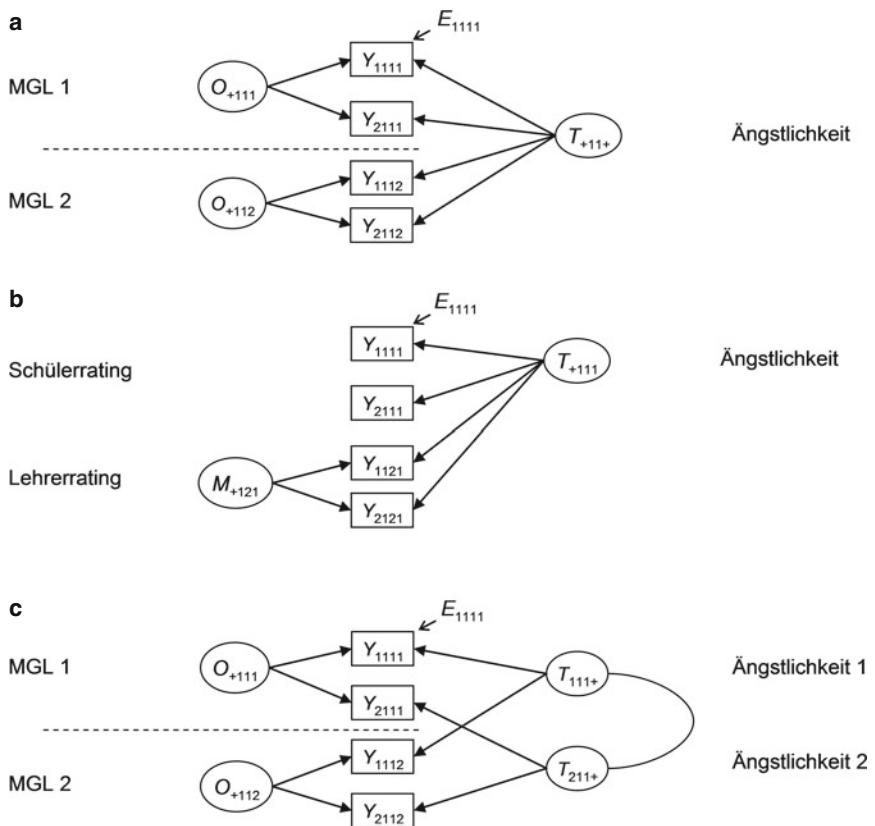


Abb. 27.1 Verschiedene Arten von Methodenfaktoren: **a** austauschbare Messgelegenheiten/Occasions (O_{+kmj}), **b** Methodenfaktor/Beurteiler als Residualfaktor [M_{+kmi} ; CTC($M - 1$)-Modell] und **c** Methodeneffekte, die durch separate Trait-Variablen abgebildet sind (T_{ikm+} und $T_{i'km+}$). MGL: Messgelegenheit; „+“ zeigt an, dass diese Variable sich auf mehrere Ziffern des betreffenden Indexes bezieht (Platzhalter). Fehlervariablen sind nur für den jeweils ersten Indikator (Item bzw. Itempäckchen) abgebildet. Die Trait-Variablen in den Modellen **a** und **c** sind als gemeinsame Faktoren definiert, sie sind zwischen den Methoden abgebildet. Die Trait-Variablen im Modell **b** entspricht dem Faktor für das Schülerrating, sie ist deswegen auf der Höhe der Indikatoren des Schülerratings abgebildet

Methodeneffekte sind trait-spezifisch

Allen Konzeptualisierungen der verschiedenen Arten von Methoden ist gemeinsam, dass Methodeneffekte *trait-spezifisch* sind. Wir gehen also davon aus, dass der Effekt einer Methode bei verschiedenen Merkmalen unterschiedlich ausfällt. Daraus folgt, dass für jede Kombination von Trait und Methode ein Methodenfaktor geschätzt werden muss³. Fremdratings können beispielsweise bei Einschätzungen der Depressivität starke Verzerrungen, bei der Einschätzung der Ängstlichkeit hingegen nur geringe Verzerrungen aufweisen; ebenso können die Messwerte der Ängstlichkeit zu bestimmten Messgelegenheiten erhöht sein, die Messwerte der Depressivität hingegen nicht.

27.1.3 LST-Modelle

LST-Modelle können genutzt werden, um zeitstabile Einflüsse von messgelegenheitsspezifischen Einflüssen auf Merkmalsausprägungen zu trennen (vgl. ► Kap. 26). Hier werden ebenfalls zunächst wahre Werte und Messfehler getrennt; sodann werden die wahren Werte in zeitstabile Anteile (die Trait-Einflüsse)

3 Für die Standardmethode im CTC($M - 1$)-Modell wird jedoch kein separater Methodenfaktor spezifiziert.

und messgelegenheitsspezifische Anteile (Messgelegenheitseinflüsse) zerlegt. Die situativen Einflüsse schwanken dabei um den stabilen Trait-Wert.

Situative Schwankungen können viele Ursachen haben. Einerseits könnten die Schüler (aus obigem Beispiel) bei der Einschätzung ihrer Ängstlichkeit stark durch momentane Einflüsse gesteuert werden: Dies könnte ein vorherrschendes schwieres Gewitter sein oder der gerade vergangene Aufenthalt im Schullandheim. Sie können aber auch durch nicht oder nur schwer messbare Faktoren beeinflusst sein: Dies könnte ein Albtraum in der vergangenen Nacht, eine Magenverstimmung oder ein Streit mit einem der Geschwister sein. Die messgelegenheitsspezifischen Faktoren umfassen alle möglichen momentanen Einflüsse und integrieren sie in einen sog. „inneren Zustand“.

Der wesentliche Unterschied zwischen MTMM- und LST-Modellen ist, dass in LST-Modellen typischerweise ein bestimmter Trait unter expliziter Berücksichtigung von zeitlichen Schwankungen über einen längeren Zeitraum wiederholt gemessen wird, während in MTMM-Modellen mehrere Traits ohne Berücksichtigung von zeitlichen Schwankungen mit verschiedenen Methoden gemessen werden. LST-Modelle werden vornehmlich in Längsschnittanalysen zur Bestimmung der Reliabilität, Konsistenz und Messgelegenheitsspezifität eingesetzt (vgl. ► Kap. 26; Yousfi und Steyer 2006), können aber auch zur Überprüfung inhaltlicher Hypothesen eingesetzt werden (s. dazu Courvoisier et al. 2007). MTMM-Modelle werden vornehmlich in Querschnittsanalysen eingesetzt, um Konstrukte hinsichtlich ihrer konvergenten und diskriminanten Validität zu untersuchen (u. a. Eid et al. 2006, 2008).

Unterschied zwischen MTMM- und LST-Modellen

27.1.4 Beschränkungen der LST- und MTMM-Modelle

Die bislang vorgestellten Modelle erlauben eine Zerlegung der Varianz der beobachteten multimethodal erhobenen Daten in verschiedene Bestandteile. Im LST-Modell werden die Bestandteile geschätzt, die von stabilen Dispositionen (Traits), messgelegenheitsspezifischen Einflüssen und Messfehlern abhängen. In MTMM-Modellen werden die Bestandteile geschätzt, die von stabilen Dispositionen (Traits), unterschiedlichen Messmethoden (Methoden) und Messfehlern abhängen. Im ersten Fall können Hypothesen getestet werden, wie sehr Messungen über die Zeit um einen stabilen Trait-Wert schwanken und ob ein Trait tatsächlich zeitlich stabil ist; im zweiten Fall kann die konvergente und die diskriminante Validität psychologischer Messungen überprüft werden. In konventionellen LST-Modellen kann jedoch die Hypothese, dass Lehrer- und Schülerratings übereinstimmen, nicht überprüft werden. In MTMM-Modellen kann nicht überprüft werden, ob die konvergente und diskriminante Validität der Messungen zeitlich stabil sind.

Gerade in den empirischen Sozialwissenschaften können wir jedoch nicht uneingeschränkt davon ausgehen, dass die Einflüsse unterschiedlicher Methoden auf ein Messergebnis über die Zeit stabil bleiben oder sich homogen verändern. Dies bedeutet, dass das Ergebnis einer querschnittlichen MTMM-Studie nicht ohne Zusatzannahmen auf spätere Zeitpunkte übertragen werden kann. Wenn zu einer Messgelegenheit eine hohe Konvergenz von Lehrer- und Schülerratings festgestellt werden konnte, können wir nicht automatisch davon ausgehen, dass sich ein halbes Jahr später keine Veränderungen der Übereinstimmung der beiden Ratings zeigen werden.

Die Entwicklung der konvergenten und diskriminanten Validität über die Zeit hinweg kann nur mithilfe *längsschnittlicher MTMM-Modelle* untersucht werden. Ebenso kann die konvergente und diskriminante Validität auf der Ebene von Traits und messgelegenheitsspezifischen Einflüssen untersucht werden. Dabei geht es um die stabilen Anteile der konvergenten Validität, die auf das überdauernde Merkmal

Längsschnittliche MTMM-Modelle

Längsschnittliche Analyse der konvergenten und diskriminanten Validität

Beispiel aus der Persönlichkeitspsychologie

Beispiel aus der Entwicklungspsychologie

Variable Zustände

zurückzuführen sind und um variable Anteile der konvergenten Validität, die dadurch zustande kommen, dass sich Effekte der Situation und der Interaktion von Situation und Person auf alle Methoden gleichmäßig auswirken.

27.1.5 Verbindung beider Ansätze

Bislang sind wenige Modelle formuliert worden, die eine Verbindung der beiden Ansätze ermöglichen. Aber gerade eine *längsschnittliche Analyse der konvergenten und diskriminanten Validität* kann von sehr großem Nutzen sein, da Querschnittsmodelle immer nur eine Momentaufnahme sein können. Es ist plausibel anzunehmen, dass situative Schwankungen oder situative Merkmale zu einer erheblichen Erhöhung oder Verringerung der Koeffizienten der konvergenten und diskriminanten Validität führen.

So könnte die konvergente Validität zwischen zwei Beurteilern und die diskriminante Validität zwischen den Konstrukten „Extraversion“ und „Verträglichkeit“ bei einer einzigen Messgelegenheit verzerrt sein, z. B. bei einer studentischen Stichprobe, bei der kurz vor der MTMM-Untersuchung die Big-Five-Persönlichkeitsfaktoren (Costa und McCrae 1998) Gegenstand in der Vorlesung zur Persönlichkeitspsychologie waren. Nach der Veranstaltung haben die studentischen Selbst- und Fremdbeurteiler die Konstrukte „Extraversion“ und „Verträglichkeit“ besonders gut verstanden und evtl. mit Kommilitonen darüber diskutiert. Sie haben sich vielleicht sogar im Hinblick auf diese beiden Konstrukte untereinander verglichen. In diesem Moment wird die Studie im Vergleich zu „normalen“ Bedingungen verzerrte konvergente und diskriminante Validitäten aufweisen, da die Studierenden die Konstrukte gerade kognitiv besser verfügbar haben und feiner zwischen ihnen unterscheiden können, als dies üblicherweise der Fall wäre.

Ein anderes Beispiel kommt aus der Entwicklungspsychologie: Kinder reifen sehr schnell, lernen viele neue Verhaltensweisen und entwickeln in kurzer Zeit neue Fähigkeiten. Cole und Martin (2005) berichten, dass die Selbsteinschätzung der Depressivität bei Kindern zunächst stark von situativen Faktoren abhängt. Bei älteren Kindern im Übergang zur Pubertät werden die Einschätzungen jedoch stabiler und das Konstrukt der Depressivität bekommt stärker den Charakter eines Traits. Eltern hingegen schätzen die Depressivität ihrer Kinder von Beginn an eher wie einen stabilen Trait ein, sodass die Konvergenz der Ratings im Laufe der Entwicklung zunehmen sollte.

Mit der Kombination von MTMM- und LST-Modellen können Analysen der konvergenten und diskriminanten Validität über die Zeit mit systematischen Veränderungen auch für *variable Zustände* vorgenommen werden. Variable Zustände wie Stimmungen zeichnen sich dadurch aus, dass sie über die Zeit schwanken. Sie variieren jedoch nicht beliebig, sondern um ein „mittleres Niveau“, das für jede Person unterschiedlich sein kann. Für Stimmungsforscher ist es daher interessant, die konvergente und diskriminante Validität auf der Ebene der Traits und auf der Ebene der Messgelegenheiten zu untersuchen. Die Frage, ob sich dieselben situativen Effekte auf mehrere Methoden homogen auswirken, ist dabei von besonderem Interesse.

Im diagnostisch-therapeutischen Kontext könnten folgende Fragen von hoher Relevanz sein: Wie stark konvergieren Selbst- und Lehrereinschätzungen in Bezug auf die Ängstlichkeit und Depressivität? Generalisieren Lehrer im Sinne eines Halo-Effekts (Thorndike 1920), kommen sie also zu nahezu identischen Einschätzungen von Ängstlichkeit und Depressivität? Wie sehr hängen die Einschätzungen der Lehrer von stabilen oder situativen Faktoren ab? Können Lehrer die Schwankungen der Ängstlichkeit ihrer Schüler nachvollziehen?

Diese Fragen können weder mit LST- noch mit MTMM-Modellen umfassend beantwortet werden, sondern müssen in Kombinationen der beiden Ansätze unter-

27.2 · Längsschnittliche MTMM-Modelle

sucht werden. In ► Abschn. 27.2 werden die folgenden drei Modelle vorgestellt, die eine Beantwortung dieser Fragen ermöglichen:

1. Multioccasion-MTMM-Modell (z. B. Burns und Haynes 2006)
2. Multiconstruct-LST-Modell (Schermelleh-Engel et al. 2004)
3. Multimethod-LST-Modell (Courvoisier 2006; Courvoisier et al. 2008)

Die drei Ansätze unterscheiden sich darin, wie sie den Verlauf über die Zeit und die Einflüsse unterschiedlicher Methoden in die Modellierung aufnehmen. Da das LST-Modell strukturell einem CTUM-Modell entspricht (Marsh und Grayson 1995), werden im folgenden Abschnitt auch Messgelegenheiten als Methoden aufgefasst.

27.2 Längsschnittliche MTMM-Modelle

Längsschnittuntersuchungen, in denen mehrere Konstrukte mit mehreren Methoden wiederholt gemessen werden, sind aufwendig erhobene Datensätze, die gewissen Anforderungen genügen müssen. Die Konstrukte sollten zu allen Messgelegenheiten mit den gleichen Methoden erhoben werden. Darüber hinaus sollten identische Indikatoren zur Messung der einzelnen Trait-Methoden-Einheiten erhoben werden. Aufgrund dieser recht hohen Anforderungen an die Daten und die Komplexität der zu analysierenden Strukturgleichungsmodelle ist es kaum verwunderlich, dass bislang kaum Längsschnitt-MTMM-Modelle vorgestellt wurden.

Anforderungen an die Daten

27.2.1 Multioccasion-MTMM-Modell

Burns und Haynes (2006) stellen eine längsschnittliche Erweiterung des CTCM-Modells (Marsh und Grayson 1995) vor, in dem zu jeder der Messgelegenheiten (Occasions) ein MTMM-Modell geschätzt wird. In ihrem Multioccasion-MTMM-Modell kann die *Stabilität von States und Methodeneffekten* überprüft werden. Da zu jeder Messgelegenheit ein MTMM-Modell geschätzt wird, sind die „Trait-Variablen“ dieser Modelle State-Variablen im Sinne der LST-Theorie. Im Folgenden werden wir deshalb von States sprechen. Die Korrelationen derselben State-Variablen zu unterschiedlichen Messzeitpunkten gibt die Stabilität der Rangplätze der Testpersonen auf dem Merkmal wieder. Korrelationen zwischen den Methodenvariablen zu verschiedenen Messgelegenheiten zeigen an, ob sich die Einflüsse der Methoden über die Zeit verändern. Überträgt man den Ansatz des CTC($M - 1$)-Modells auf den Vorschlag von Burns und Haynes (2006), erhält man ein *Multioccasion-Correlated-States-Correlated-(Method-minus-1)-Modell*, kurz Multioccasion-CSC($M - 1$)-Modell (s. auch Geiser et al. 2008).

Multioccasion-CSC($M - 1$)-Modell

Das Multioccasion-CSC($M - 1$)-Modell in □ Abb. 27.2 kann dazu genutzt werden, die Stabilität von Schülern im Hinblick auf ihre Depressivitäts- und Ängstlichkeitszustände zu analysieren. Die bivariate Korrelation zwischen den Depressivitätsvariablen (oder Ängstlichkeitsvariablen) zur ersten und zweiten Messgelegenheit spiegelt diese Stabilität wider. Die Methodenfaktoren stellen die Abweichungen der Lehrer- von den Schülerratings dar. Korrelationen zwischen den Methodenfaktoren zeigen somit an, ob Lehrer, die die Depressivität einer Schülerin oder eines Schülers zur ersten Messgelegenheit überschätzen (unterschätzen), dies auch zur zweiten Messgelegenheit tun. Die trait-spezifischen Methodenfaktoren der Lehrer (M_{+kmi} ; „+“ zeigt an, dass sich der Methodenfaktor auf beide Indikatoren auswirkt) dürfen auch über die Konstrukte hinweg korrelieren (in □ Abb. 27.2 nicht dargestellt). Diese Korrelationen geben den Grad der Generalisierbarkeit von Methodeneffekten an. Je höher diese Korrelation ausfällt, desto homogener ist der Einfluss der Methode auf die Verzerrung der Lehrerratings. Sie

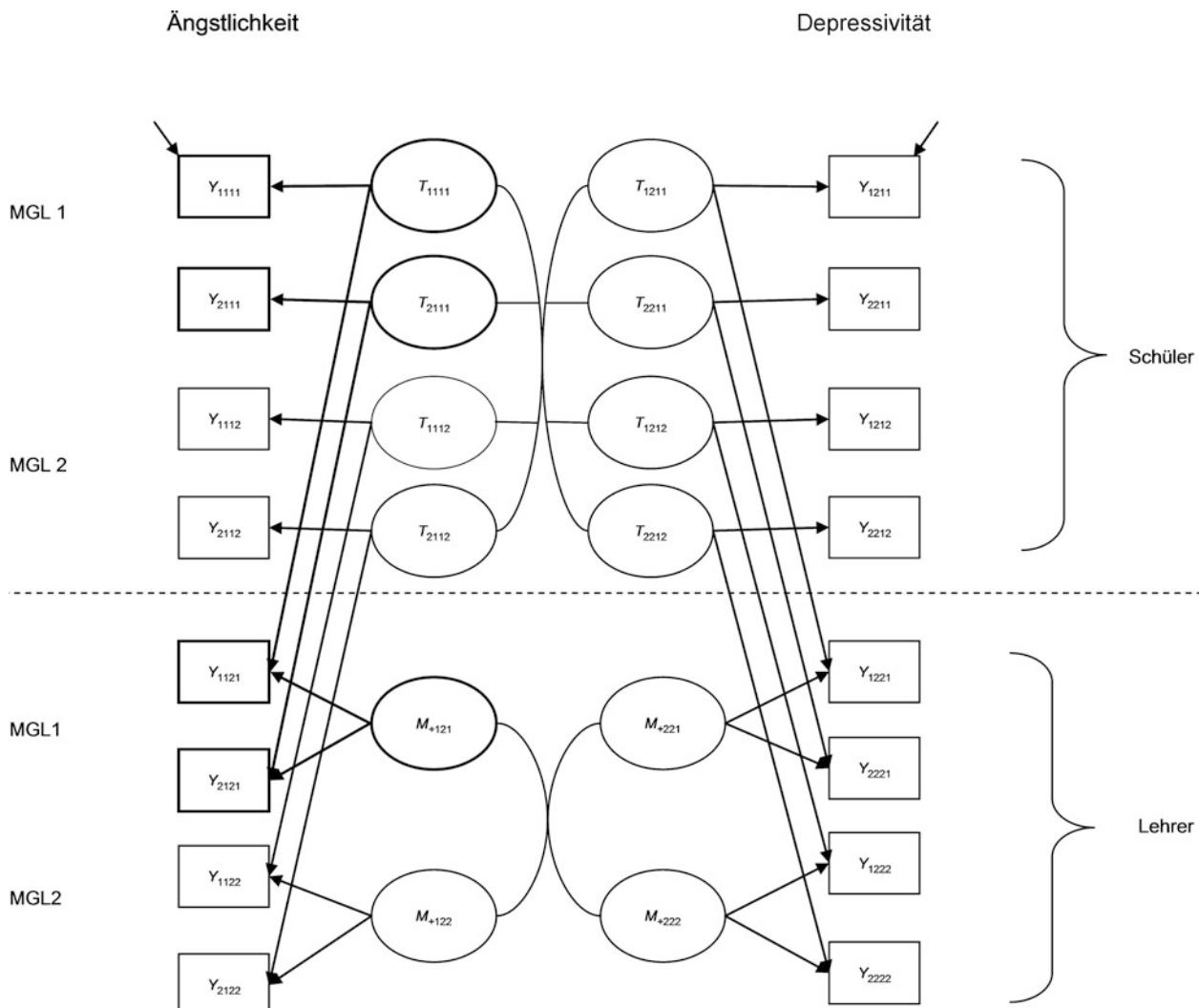


Abb. 27.2 Multioccasion-CSC($M - 1$)-Modell für zwei Traits, gemessen mit zwei Methoden zu zwei Messgelegenheiten. T_{ikmt} : Trait-Faktor; M_{+kmt} : Methodenfaktor; MGL : Messgelegenheit. Fehlervariablen sind nur für die ersten Indikatoren abgebildet

gibt an, ob Lehrer die Depressivität und die Ängstlichkeit der Kinder in gleichem Maße über- bzw. unterschätzen. Die Methodenfaktoren können auch über Traits und Messgelegenheiten hinweg korrelieren. Diese Korrelationen geben z. B. an, ob Lehrer, die zu einer Messgelegenheit die Depressivität überschätzen, die Ängstlichkeit zu einer späteren Messgelegenheit ebenfalls überschätzen.

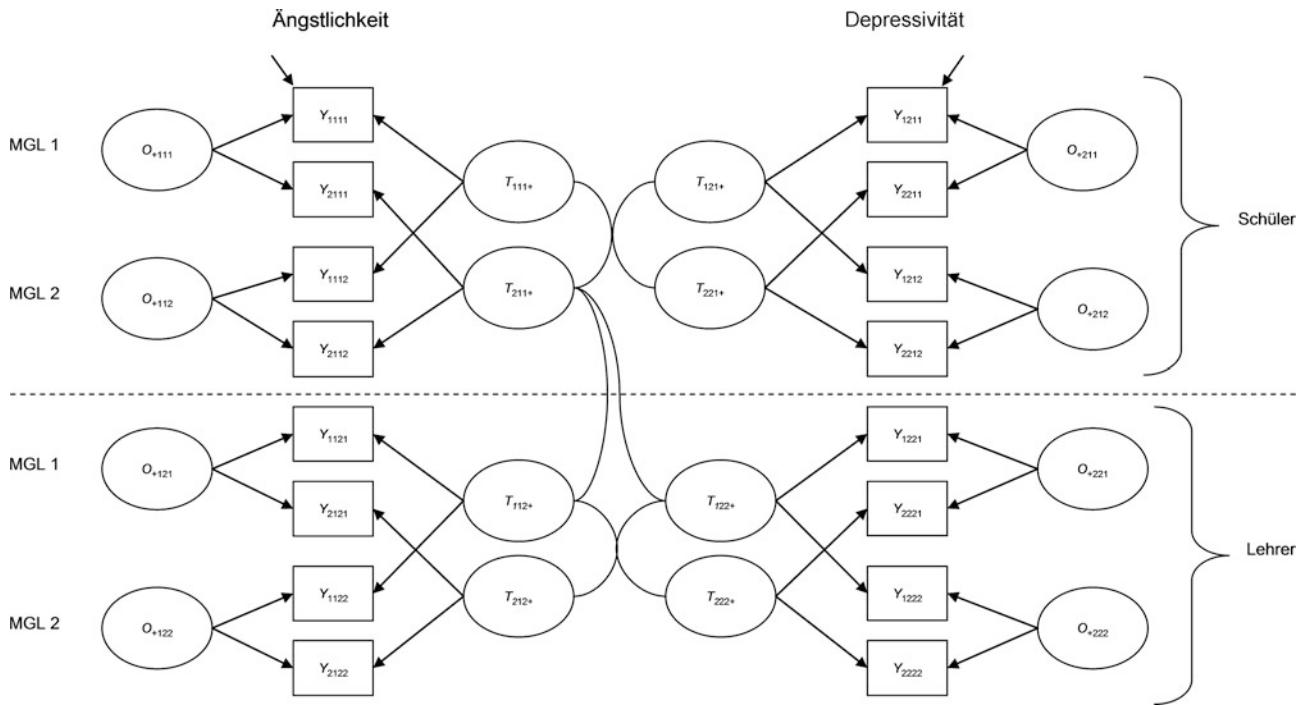
27.2.2 Multiconstruct-LST-Modell

Multiconstruct-LST-Modell

Das *Multiconstruct-LST-Modell* (Dumenci und Windle 1998; Eid et al. 1994; Majcen et al. 1988; Schermelleh-Engel et al. 2004; Schmitt 2000; Steyer 1989; Steyer et al. 1990) wird zur Erfassung mehrerer Konstrukte im zeitlichen Verlauf eingesetzt. Im Gegensatz zu dem von Burns und Haynes (2006) vorgeschlagenen Modell werden hier nicht zu jeder Messgelegenheit MTMM-Modelle geschätzt, sondern für jede Trait-Methoden-Einheit ein LST-Modell (Abb. 27.3).

Wie in den klassischen LST-Modellen gibt es eine indikatorspezifische Trait-Variable (T_{ikm+}) für jede Kombination von Indikator, Trait und Methoden, d. h., es gibt beispielsweise zwei indikatorspezifische Trait-Variablen für die Einschätzung

27.2 · Längsschnittliche MTMM-Modelle



■ Abb. 27.3 Multicreate-LST-Modell. T_{ikm+} : Trait-Variablen; O_{+kmt} : Messgelegenheitsvariable; MGL : Messgelegenheit. Die zulässigen Korrelationen zwischen den Messgelegenheitsvariablen sind aus Gründen der Lesbarkeit nicht alle abgebildet. Korrelationen zwischen Trait-Variablen sind nur exemplarisch eingezeichnet. Prinzipiell korrelieren alle Trait-Variablen miteinander. Fehlervariablen sind nur für die ersten Indikatoren dargestellt

der Ängstlichkeit durch die Schüler und zwei indikatorsspezifische Trait-Variablen für die Einschätzung durch die Lehrer. Die Korrelationen dieser Trait-Variablen eines Konstrukts, gemessen mit unterschiedlichen Methoden, geben in diesem Modell die konvergente Validität an, so z. B. $\text{Corr}(T_{1k1+}, T_{1k2+})$, die Korrelation der ersten Trait-Variablen ($i = 1$) für ein beliebiges Konstrukt (k), gemessen anhand des Schülerratings ($m = 1$) und des Lehrerratings ($m = 2$). Die Korrelationen verschiedener Trait-Variablen für unterschiedliche Konstrukte (Ängstlichkeit und Depressivität) geben die diskriminante Validität an, so z. B. $\text{Corr}(T_{111+}, T_{121+})$, die Korrelation der ersten Trait-Variablen ($i = 1$) der beiden Konstrukte ($t = 1$ und $t = 2$), gemessen anhand des Schülerratings ($m = 1$). Korrelationen zwischen einer Trait-Variablen und der Trait-Variablen eines anderen Konstrukts, das mit einer anderen Methode gemessen wurde, spiegeln ebenfalls die diskriminante Validität wider, wobei in diese zusätzlich die Unterschiede der Methoden einfließen, so z. B. $\text{Corr}(T_{111+}, T_{122+})$, die Korrelation der ersten Trait-Variablen ($i = 1$) der beiden Konstrukte ($t = 1$ und $t = 2$), gemessen anhand des Schülerratings ($m = 1$) und des Lehrerratings ($m = 2$).

Die Korrelation der messgelegenheitsspezifischen Variablen einer Methode und einer anderen Methode desselben Konstrukts zu einer Messgelegenheit spiegelt dabei wider, ob die beiden Methoden zu dieser Messgelegenheit in gleicher Weise vom jeweiligen stabilen Trait abweichen, so z. B. $\text{Corr}(O_{+111}, O_{+121})$, die Korrelation der situationsspezifischen Abweichungen vom stabilen Trait eines Konstrukts ($k = 1$), gemessen anhand des Schülerratings ($m = 1$) und des Lehrerratings ($m = 2$) zur gleichen Messgelegenheit ($k = 1$). Dies entspricht einem Effekt der Messgelegenheit, der sich auf die Messungen beider Methoden auswirkt und gleichzeitig ein weiteres Kennzeichen für die konvergente Validität darstellt. Die Korrelation zweier messgelegenheitsspezifischer Variablen einer Methode, die zu zwei Konstrukten gehören, zeigt den Einfluss der Messgelegenheit auf beide Konstrukte an, so z. B. $\text{Corr}(O_{+111}, O_{+211})$, die Korrelation der situationsspezifischen

Anwendung auf das Beispiel

Abweichungen vom stabilen Trait zweier Konstrukte ($k = 1$ und $k = 2$), gemessen anhand des Schülerratings ($m = 1$) zur gleichen Messgelegenheit ($k = 1$); sie ist ein Maß für die diskriminante Validität der Konstrukte auf messgelegenheitspezifischer Ebene. Die Korrelationen zwischen den messgelegenheitsspezifischen Variablen einer Methode mit den messgelegenheitsspezifischen Variablen der anderen Methode sind auch über die Zeit hinweg erlaubt, jedoch nicht einfach zu interpretieren. Sie geben keinen Hinweis auf die konvergente oder diskriminante Validität und werden deshalb hier nicht weiter erläutert (s. dazu Courvoisier 2006).

Wendet man das Multiconstruct-LST-Modell erneut auf das Beispiel der Ängstlichkeits- und Depressivitätsbeurteilungen bei Schulkindern an, so müssten vier LST-Modelle geschätzt werden:

1. Depressivität im Schülerrating
2. Depressivität im Lehrerrating
3. Ängstlichkeit im Schülerrating
4. Ängstlichkeit im Lehrerrating

Findet sich eine hohe Korrelation der Schüler- und Lehrerratings für die Einschätzungen der Depressivität, ist dies ein Beleg für die Übereinstimmung der Methoden und somit für die konvergente Validität. Die Korrelationen der Trait-Variablen für Ängstlichkeit und Depressivität des Schülerratings (Lehrerratings) gibt die diskriminante Validität an. Die Korrelation der messgelegenheitsspezifischen Variablen der Schüler und der Lehrer eines Konstrukt zu einer Messgelegenheit spiegelt wider, ob die beiden Methoden in gleicher Weise vom jeweiligen stabilen Trait abweichen (konvergente Validität). Schätzt sich der Schüler momentan erhöht auf der Skala der Depressivität ein, so ist dies auch tendenziell für den Lehrer der Fall (bei positiver Korrelation).

Im Gegensatz zum Multioccasion-MTMM-Modell können in diesem Modell die stabilen Varianzkomponenten von den situativ bedingten Varianzkomponenten und den Residualkomponenten getrennt werden. Eine Erweiterung des Multiconstruct-LST-Modells zum hierarchischen LST-Modell findet sich bei Schermelleh-Engel et al. (2004).

27.2.3 Multimethod-LST-Modell

Verknüpfung von LST- und MTMM-Modellen

Das Multioccasion-MTMM- und das Multiconstruct-LST-Modell ermöglichen die Analyse unterschiedlicher Aspekte der Daten. Beim ersten Modell steht der MTMM-Charakter, beim zweiten der LST-Charakter stärker im Vordergrund. Je nach wissenschaftlicher Fragestellung kann eines dieser Modelle ausgewählt werden und entsprechend aussagekräftige Resultate liefern. Liegt das Augenmerk jedoch auf der Analyse der konvergenten und diskriminanten Validität über die Zeit und möchte man die unterschiedlichen Varianzkomponenten identifizieren, die durch den Trait, die Methode, die Messgelegenheit und den Messfehler bedingt sind, so muss die Verknüpfung von LST- und MTMM-Modellen noch stärker erfolgen als in den vorangegangenen Modellen. Courvoisier (2006) und Courvoisier et al. (2008) stellen das Multimethod-LST-Modell vor, bei dem die longitudinalen und multimethodalen Bestandteile der einzelnen Modelle nicht nebeneinander gestellt, sondern ineinander verschränkt werden (Abb. 27.4).

Das hier vorgestellte Modell verbindet die Eigenschaften des LST-Modells mit denen des CTC($M - 1$)-Modells. Das Modell lässt sich in sinnvolle Submodelle unterteilen. In der oberen Hälfte des abgebildeten Modells befinden sich zwei klassische LST-Modelle für die Standardmethode (die Schüler).

Die indikatorspezifischen Trait-Variablen (T_{ik1+}) entsprechen den stabilen Komponenten der Standardmethode, also ihren stabilen Trait- und Methodeneinflüssen in Bezug auf einen Indikator. Wie im klassischen CTC($M - 1$)-Modell ist

Obere Hälfte des abgebildeten Modells

27.2 · Längsschnittliche MTMM-Modelle

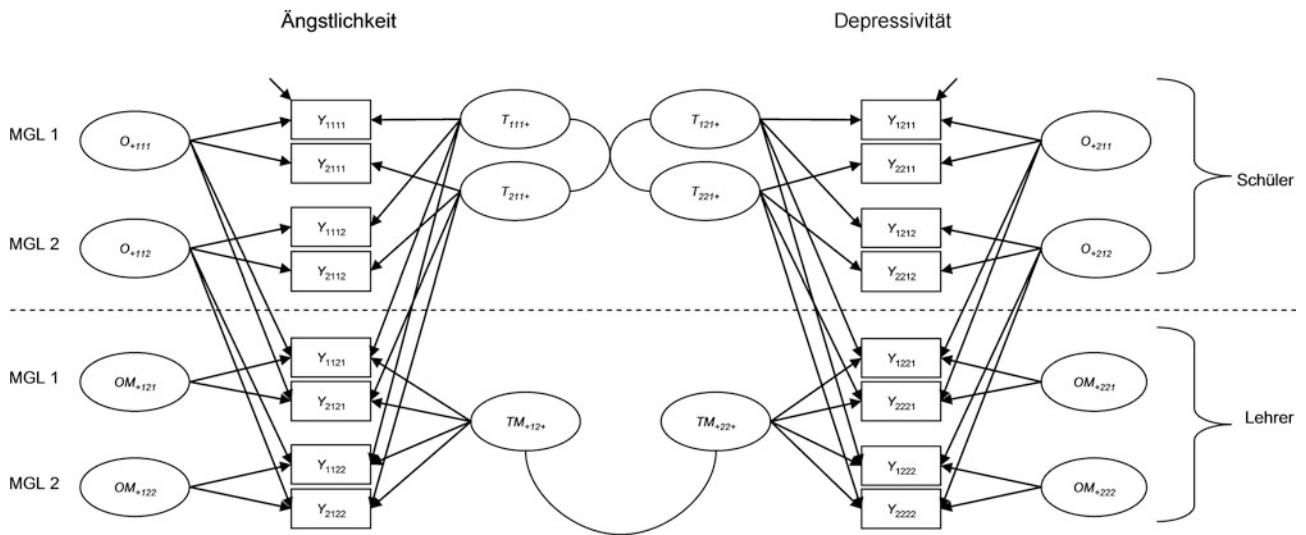


Abb. 27.4 Multimethod-LST-Modell. T_{ikm+} : Trait-Variable; TM_{+km+} : Methodenvariable; O_{+kmt} : messgelegenheitsspezifische Variable; OM_{+kmt} : messgelegenheitsspezifische Abweichungsvariable; MGL : Messgelegenheit. Die zulässigen Korrelationen zwischen sämtlichen messgelegenheitsspezifischen Variablen sind aus Gründen der Lesbarkeit nicht abgebildet. Fehlervariablen sind nur für die obersten Indikatoren dargestellt

es wichtig, die Standardmethode anhand theoretischer Überlegungen zu wählen, da sämtliche Methodeneinflüsse und deren Interaktionen mit dem Trait in die Trait-Variable einfließen.

Die messgelegenheitsspezifischen Variablen (O_{+kt}) entsprechen den situationsspezifischen Abweichungen der Standardmethode (Schüler) von den stabilen Trait-Variablen.

Die Methodenfaktoren der Nichtstandardmethoden (TM_{+km+} , Lehrer) entsprechen dem zeitlich stabilen Bias der Nichtstandardmethoden gegenüber dem Standard, d. h. der stabilen positiven oder negativen Abweichung vom erwarteten Rating.

Die messgelegenheitsspezifischen Abweichungsvariablen (OM_{+kmt}) entsprechen den situationspezifischen Abweichungen der Nichtstandardmethoden (Lehrer) von den durch die ersten drei Variablen (T_{ik1+} , O_{+k1t} und TM_{+km+}) vorhergesagten Werten. Hierdurch kann geprüft werden, ob das Lehrerrating stärker oder schwächer abweicht, als es aufgrund des Schülerratings (stabilen Anteil und situationspezifischer Anteil) und der stabilen Abweichung des Lehrers zu erwarten ist.

Die Korrelationen zwischen den Variablen der Standardmethode haben praktisch dieselben Bedeutungen wie im Multiconstruct-LST-Modell. Korrelationen zwischen Trait-Variablen spiegeln die konvergente oder diskriminante Validität, Korrelationen der messgelegenheitsspezifischen Variablen zu einer Messgelegenheit den trait-übergreifenden Einfluss der Messgelegenheit wider.

Die Korrelationen in der unteren Hälfte des abgebildeten Modells sind aus der Perspektive des CTC($M - 1$)-Modells zu interpretieren. Die Korrelationen zwischen den Methodenfaktoren zeigen den generellen Methodenbias, so z. B. $\text{Corr}(TM_{+12+}, TM_{+22+})$. Korrelationen zwischen den messgelegenheitsspezifischen Abweichungsfaktoren zu einer Messgelegenheit (z. B. $\text{Corr}(OM_{+121}, OM_{+221})$, in Abb. 27.4 nicht eingezeichnet) geben den merkmalsübergreifenden Einfluss der Messgelegenheit, der sich ausschließlich auf die Nichtstandardmethode auswirkt, wieder (die gemeinsame Auswirkung auf die Standard- und Nichtstandardmethode steckt bereits in dem messgelegenheitsspezifischen Faktor O_{+k1t} der Standardmethode).

Bezogen auf das Beispiel der Ängstlichkeit und der Depressivität von Schulkindern bietet es sich an, das Schülerrating als Standardmethode auszuwählen, um die Lehrerratings vorherzusagen. Die Interpretation der Variablen für die Schüler

Untere Hälfte des abgebildeten Modells

Anwendung auf das Beispiel

entspricht der Interpretation derselben Variablen im klassischen LST-Modell. Ihre manifesten Variablen werden in einen stabilen Anteil (T) und einen messgelegenheitsspezifischen Anteil (O) zerlegt. Die Ratings der Lehrer werden ebenfalls durch diese Variablen vorhergesagt. Darüber hinaus wird die stabile Abweichung der Lehrer in der Methodenvariable abgebildet (TM). Dies ist die stabile Abweichung (Über- oder Unterschätzung) der Lehrerratings von den Erwartungen aufgrund der stabilen Schülerratings, sie ist unabhängig von situativen Einflüssen auf die Lehrerratings. Situationsspezifische Abweichungen der Lehrerratings werden durch die messgelegenheitsspezifische Abweichungsvariable abgebildet (OM). Diese Variable zeigt Veränderungen des Bias der Lehrer von ihren stabilen Abweichungen an. Sie kann als interner Zustand der Lehrer aufgefasst werden, der nicht mit dem internen Zustand der Schüler geteilt wird. Beispielsweise wären gemeinsame Einflüsse auf Schüler- und Lehrerratings durch die bevorstehenden Ferien in der messgelegenheitsspezifischen Variablen (O) erfasst; Einflüsse, beispielsweise die zusätzliche Arbeit, die fälligen Zeugnisse auszustellen, beeinflussen nur die Lehrer und werden in der messgelegenheitsspezifischen Abweichungsvariablen (OM) abgebildet.

Die Konvergenz der Lehrer- und Schülerratings ergibt sich aus dem Anteil der Varianz der Lehrerratings, der durch den Trait *und* die messgelegenheitsspezifische Variable der Schüler vorhergesagt werden kann. Sie besteht also aus stabilen und variablen Anteilen. Ob die Lehrer die Schwankungen der Schüler in Bezug auf die Depressivität (Ängstlichkeit) richtig nachvollziehen können, zeigt sich in der Größe des Einflusses der messgelegenheitsspezifischen Variablen der Schüler (O_{+k1t}) auf die beobachteten Variablen der Lehrer (dieser Einfluss wäre bei optimaler Übereinstimmung genau so hoch wie der Einfluss dieser Variablen auf die Schülerratings). Die Unterschiede zwischen den beiden Methoden werden durch zwei Faktoren abgebildet. Zunächst wirkt sich der stabile Bias der Lehrer aus, d. h., ob ein Lehrer die Depressivität (Ängstlichkeit) eines Schülers konsistent höher oder niedriger einschätzt als erwartet. Da sich aber auch der Lehrer zu jeder Messgelegenheit in einer bestimmten Situation befindet, wird noch eine messgelegenheitsspezifische Abweichungsvariable in Betracht gezogen, die den momentanen Bias der Lehrer darstellt (OM_{+kmt}). Die diskriminante Validität der beiden Konstrukte „Ängstlichkeit“ und „Depressivität“ zeigt sich in der Höhe der Korrelation der beiden Trait-Variablen, so z. B. in $\text{Corr}(T_{111+}, T_{121+})$.

Generalisierbarkeit eines messgelegenheitsspezifischen Einflusses

Die Generalisierbarkeit eines messgelegenheitsspezifischen Einflusses kann über die Korrelation der messgelegenheitsspezifischen Variablen zu einer Messgelegenheit ermittelt werden ($\text{Corr}(O_{+111}, O_{+211})$, homogener Einfluss der Messgelegenheit auf die Schülerratings). Ob der stabile Anteil des Bias der Lehrer über beide Konstrukte generalisiert, kann an der Korrelation $\text{Corr}(TM_{+12+}, TM_{+22+})$ der beiden Methodenvariablen abgelesen werden. Inwiefern der instabile (momentane) Bias der Lehrer für die Ängstlichkeit und die Depressivität identisch ist, kann mit der Korrelation $\text{Corr}(OM_{+121}, OM_{+221})$ der beiden messgelegenheitsspezifischen Abweichungsvariablen zu einer Messgelegenheit überprüft werden.

Im Multimethod-LST-Modell sind noch weitere Korrelationen latenter Variablen zulässig, die wir hier aber nicht besprechen werden, da sie für die Bestimmung der konvergenten und der diskriminanten Validität im Längsschnitt von nachgeordneter Priorität sind. Nicht erlaubt sind Korrelationen von Variablen, die in derselben Modellgleichung für eine bestimmte beobachtete Variable vorkommen. So sind die Trait-Variablen (T), die messgelegenheitsspezifischen Variablen (O), die Methodenvariablen (TM) und die messgelegenheitsspezifischen Abweichungsvariablen (OM) unkorreliert, wenn sie zur Erklärung einer bestimmten beobachteten Variablen herangezogen werden. Folglich sind alle latenten Variablen einer Trait-Methoden-Einheit zu einer Messgelegenheit unkorreliert (das sind fast alle latenten

27.2 · Längsschnittliche MTMM-Modelle

Variablen, die zu einem Konstrukt gehören)⁴. Hingegen können alle latenten Variablen einer Trait-Methoden-Einheit mit allen latenten Variablen einer anderen Trait-Methoden-Einheit korrelieren. Zu beachten ist, dass die oben aufgeführten Korrelationen am leichtesten und sinnvoll zu interpretieren sind. Andere Korrelationen sind testtheoretisch erlaubt, jedoch schwierig zu interpretieren.

27.2.4 Vergleich der drei längsschnittlichen MTMM-Modelle

In der folgenden Box sind die Modellgleichungen und Varianzdekompositionen der drei longitudinalen MTMM-Modellen übersichtlich zusammengestellt.

Modellgleichungen und Varianzdekompositionen der längsschnittlichen MTMM-Modelle (für Abweichungsvariablen, s. z. B. Bollen 1989)

Für alle Modelle gilt:

- α_{ik1t} Interzept des Indikators
- λ_{Tikmt} Ladungsparameter auf dem Traitfaktor
- λ_{Oikmt} Ladungsparameter auf dem messgelegenheitsspezifischen Faktor
- λ_{Mik2t} Ladungsparameter auf dem Methodenfaktor (λ_{TMik2t} im Multi-method-LST-Modell)
- λ_{OMikmt} Ladungsparameter auf dem messgelegenheitsspezifischen Methodenfaktor
- i Indikator;
- k Trait;
- m Methode;
- t Messgelegenheit

Multioccasion-MTMM-Modell

Standardmethode zu allen Messgelegenheiten:

$$\begin{aligned} Y_{ik1t} &= \alpha_{ik1t} + \lambda_{Tik1t} S_{ik1t} + E_{ik1t} \\ \Rightarrow \text{Var}(Y_{ik1t}) &= \lambda_{Tik1t}^2 \text{Var}(S_{ik1t}) + \text{Var}(E_{ik1t}) \end{aligned}$$

Nichtstandardmethode zu allen Messgelegenheiten:

$$\begin{aligned} Y_{ik2t} &= \alpha_{ik2t} + \lambda_{Tik2t} S_{ik1t} + \lambda_{Mik2t} M_{+k2t} + E_{ik2t} \\ \Rightarrow \text{Var}(Y_{ik2t}) &= \lambda_{Tik2t}^2 \text{Var}(S_{ik1t}) + \lambda_{Mik2t}^2 \text{Var}(M_{+k2t}) + \text{Var}(E_{ik2t}) \end{aligned}$$

Multiconstruct-LST-Modell

Für alle Methoden zu allen Messgelegenheiten:

$$\begin{aligned} Y_{ikmt} &= \alpha_{ikmt} + \lambda_{Tikmt} T_{ikmt} + \lambda_{Oikmt} O_{ikmt} + E_{ikmt} \\ \Rightarrow \text{Var}(Y_{ikmt}) &= \lambda_{Tikmt}^2 \text{Var}(T_{ikmt}) + \lambda_{Oikmt}^2 \text{Var}(O_{ikmt}) + \text{Var}(E_{ikmt}) \end{aligned}$$

⁴ Lediglich die messgelegenheitsspezifischen Variablen dürfen mit den messgelegenheitspezifischen Abweichungsvariablen zu einem anderen Messzeitpunkt korrelieren, so z.B. $\text{Corr}(O_{+111}, OM_{+112})$. Diese Korrelation ist jedoch theoretisch nicht einfach zu interpretieren. In vielen Anwendungen bietet es sich an, diese Korrelationen nicht zuzulassen.

Multimethod-LST-Modell

Standardmethode zu allen Messgelegenheiten:

$$Y_{ik1t} = \alpha_{ik1t} + \lambda_{Tik1t} T_{ik1t} + \lambda_{Oik1t} O_{ik1t} + E_{ik1t}$$

$$\Rightarrow \text{Var}(Y_{ik1t}) = \lambda_{Tik1t}^2 \text{Var}(T_{ik1t}) + \lambda_{Oik1t}^2 \text{Var}(O_{ik1t}) + \text{Var}(E_{ik1t})$$

Nichtstandardmethode zu allen Messgelegenheiten:

$$Y_{ik2t} = \alpha_{ik2t} + \lambda_{Tik2t} T_{ik1t} + \lambda_{Oik2t} O_{ik1t} + \lambda_{TMik2t} TM_{+k2t}$$

$$+ \lambda_{OMik2t} OM_{+k2t} + E_{ik2t}$$

$$\Rightarrow \text{Var}(Y_{ik2t}) = \lambda_{Tik2t}^2 \text{Var}(T_{ik1t}) + \lambda_{Oik2t}^2 \text{Var}(O_{+k1t})$$

$$+ \lambda_{TMik2t}^2 \text{Var}(TM_{+k2t}) + \lambda_{OMik2t}^2 \text{Var}(OM_{+k2t})$$

$$+ \text{Var}(E_{ik2t})$$

Vergleich der Modellgleichungen und der spezifischen Varianzkomponenten

Die drei hier vorgestellten längsschnittlichen MTMM-Modelle unterscheiden sich im Wesentlichen in Bezug auf ihre Komplexität. Das *Multioccasion-MTMM-Modell* und das *Multiconstruct-LST-Modell* ermöglichen die Varianzzerlegung in drei Komponenten: den Anteil des Traits, den Anteil einer Methode oder der Messgelegenheit sowie den Anteil des Fehlerterms. Das *Multimethod-LST-Modell* zerlegt die Varianz in fünf Bestandteile, und zwar den Anteil des Traits, der Nichtstandardmethode, des messgelegenheitsspezifischen Einflusses auf die Standardmethode, des messgelegenheitsspezifischen Einflusses auf die Nichtstandardmethode und des Fehlerterms.

Der State im *Multioccasion-MTMM-Modell* enthält sowohl stabile Bestandteile als auch messgelegenheitsspezifische Bestandteile. Die Methodenvariable enthält die momentane Abweichung der Nichtstandardmethode, die sich theoretisch aus stabilen und situativen Komponenten zusammensetzen könnte.

Die Trait-Variable im *Multiconstruct-LST-Modell* umfasst die zeitlich stabilen Anteile des Ratings mit der jeweiligen Methode. Sie beinhaltet damit sowohl Anteile, die auf die wahre Ausprägung zurückzuführen sind, als auch Anteile des Methodenbiases. Die messgelegenheitsspezifischen Variablen setzen sich ebenfalls aus Einflüssen der Messgelegenheit und Methodeneinflüssen zusammen und spiegeln die momentanen Abweichungen vom stabilen Anteil wider.

Lediglich das komplexeste Modell, das *Multimethod-LST-Modell*, ermöglicht es, die stabilen Einflüsse von Trait und Nichtstandardmethode von den variablen Einflüssen der Messgelegenheiten und messgelegenheitsspezifischen Methodeneffekte zu trennen. In diesem Modell ist die feinste Zerlegung der Varianz einer beobachteten Variablen möglich.

In □ Tab. 27.1 sind die Verfahren zur Bestimmung der Stabilität, der Messgelegenheitsspezifität, der konvergenten Validität sowie der Methodenspezifität in den drei longitudinalen MTMM-Modellen übersichtlich zusammengestellt.

Beim Einsatz des Multioccasion-MTMM-Modells entspricht die Anzahl der Traitfaktoren dem Produkt von Traits und Messgelegenheiten pro Methode. So entstehen schon bei nur drei Messgelegenheiten eine Vielzahl von Korrelationen, die die zeitliche Stabilität des Konstrukt, aber auch die diskriminante Validität zwischen den Traits kennzeichnen. Ein Maß für die mittlere Konvergenz der Methoden oder die mittlere diskriminante Validität ist nicht einfach zu berechnen (s. Bortz und Schuster 2010, zur allgemeinen Verrechnung von Korrelationen). Aus diesem Grund werden im folgenden Beispiel nur das Multiconstruct-LST-Modell und das Multimethod-LST-Modell eingesetzt.

Vorteile des Multimethod-LST-Modells

Vergleich der Verfahren zur Bestimmung der Stabilität, der Messgelegenheitsspezifität, der konvergenten Validität sowie der Methodenspezifität

27.2 · Längsschnittliche MTMM-Modelle

Multioccasion-MTMM-Modell		Multiconstruct-LST-Modell	Multimethod-LST-Modell
Stabilität	Korrelation der State-Variablen über die Zeit: $\text{Corr}(S_{ik1t}, S_{ik1t'}) , t \neq t'$	$\frac{\lambda_{Tik1t}^2 \text{Var}(T_{ik1t})}{\text{Var}(Y_{ik1t})}$	Für die Standardmethode: $\frac{\lambda_{Tik1t}^2 \text{Var}(T_{ik1t})}{\text{Var}(Y_{ik1t})}$
Messgelegenheitspezifischer Einfluss	-	$\frac{\lambda_{Oikmt}^2 \text{Var}(O_{+kmt})}{\text{Var}(Y_{ikmt})}$	Für die Nichtstandardmethode: $\frac{\lambda_{Tikmt}^2 \text{Var}(T_{ikmt}) + \lambda_{Oikmt}^2 \text{Var}(TM_{+kmt})}{\text{Var}(Y_{ikmt})}$
Konvergente Validität	Für Nichtstandardmethoden: $\frac{\lambda_{Tikmt}^2 \text{Var}(S_{ik1t})}{\text{Var}(Y_{ikmt})}$	-	Für die Standardmethode: $\frac{\lambda_{Oikmt}^2 \text{Var}(O_{+kt1})}{\text{Var}(Y_{ikmt})}$
Diskriminante Validität	Korrelation der State-Variablen zweier Konstrukte zu einer Messgelegenheit: $\text{Corr}(S_{ik1t}, S_{ik1t'}) , k \neq k'$	Korrelationskoeffizienten der Trait-Faktoren: $\text{Corr}(T_{ikmt}, T_{ik'mt}) , k \neq k'$	Korrelationskoeffizienten der Trait-Faktoren: $\text{Corr}(T_{ik1t}, T_{ik'1t}) , k \neq k'$
Methodenspezifität	Für die Nichtstandardmethode: $\frac{\lambda_{Mikmt}^2 \text{Var}(M_{+kmt})}{\text{Var}(Y_{ikmt})}$	-	Für die Standardmethode: $\frac{\lambda_{TMikmt}^2 \text{Var}(TM_{+kmt}) + \lambda_{Oikmt}^2 \text{Var}(OM_{+kmt})}{\text{Var}(Y_{ikmt})}$
Generalisierbarkeit der Methodeneffekte	Korrelation der Methodeneffekte: $\text{Corr}(M_{+kmt}, M_{+k'mt}) , k \neq k'$	-	Davon stabil: $\frac{\lambda_{TMikmt}^2 \text{Var}(TM_{+kmt})}{\text{Var}(Y_{ikmt})}$
Reliabilität	$\text{Rel}(Y_{ikmt}) = \frac{\text{Rel}(Y_{ikmt})}{\lambda_{Tikmt}^2 \text{Var}(S_{ikmt}) + \lambda_{Mikmt}^2 \text{Var}(M_{+kmt})}$	$\text{Rel}(Y_{ikmt}) = \frac{\lambda_{Tikmt}^2 \text{Var}(T_{ikmt}) + \lambda_{Oikmt}^2 \text{Var}(O_{ikmt})}{\text{Var}(Y_{ikmt})}$	Davon messgelegenheitspezifisch: $\frac{\lambda_{Oikmt}^2 \text{Var}(OM_{+kmt})}{\text{Var}(Y_{ikmt})}$
Abhängig von der Situation:		Stabil: $\text{Corr}(TM_{+kmt}, TM_{+k'mt}) , k \neq k'$ Abhängig von der Situation: $\text{Corr}(TM_{+kmt} + OM_{+kmt}, TM_{+k'mt} + OM_{+k'mt}) , k \neq k'$	Stabil: $\text{Corr}(TM_{+kmt}, TM_{+k'mt}) , k \neq k'$ Abhängig von der Situation: $\text{Corr}(TM_{+kmt} + OM_{+kmt}, TM_{+k'mt} + OM_{+k'mt}) , k \neq k'$
		Für die Standardmethode: $\text{Rel}(Y_{ik1t}) = \frac{\lambda_{Tik1t}^2 \text{Var}(T_{ik1t}) + \lambda_{Oik1t}^2 \text{Var}(O_{+kt1})}{\text{Var}(Y_{ik1t})}$	Für die Nichtstandardmethode: $\text{Rel}(Y_{ik2t}) = \frac{\lambda_{Tik2t}^2 \text{Var}(T_{ik2t}) + \lambda_{Oik2t}^2 \text{Var}(O_{+kt2}) + \lambda_{OMik2t}^2 \text{Var}(OM_{+kt2})}{\text{Var}(Y_{ik2t})}$

Anmerkung: Leere Zellen zeigen an, dass es keinen direkten Koeffizienten gibt. Oft lassen sich allerdings Korrelationen oder Vergleiche von Koeffizienten im Sinne dieser Koeffizienten interpretieren.

27.3 Multiconstruct-LST- und Multimethod-LST-Modell in der empirischen Anwendung

Cole und Kollegen (Cole und Martin 2005; Cole et al. 1996; Cole et al. 1997) befragten u. a. 375 Schüler und deren Lehrer einer amerikanischen Grundschule („elementary school“) zu vier Messgelegenheiten im Hinblick auf ihre Depressivität und Ängstlichkeit. Die Depressivität der Schüler wurde mit dem Child Depression Inventory (Kovacs 1981, 1982) und ihre Ängstlichkeit mit der Revised Children’s Manifest Anxiety Scale (Reynolds und Richmond 1978) erfasst. Die Ratings der Lehrer wurden mit dem Teacher Report Index of Depression (Cole und Jordan 1995) und dem Teacher Report Index of Anxiety (Cole und Jordan 1995; Lefkowitz und Tesiny 1980) erfasst.

Das Multiconstruct-LST-Modell (Dumenci und Windle 1998; Eid et al. 1994; Majcen et al. 1998; Schermelleh-Engel et al. 2004; Schmitt 2000; Steyer et al. 1989, 1990) und das Multimethod-LST-Modell (Courvoisier 2006; Courvoisier et al. 2008) wurden in einer Reanalyse mit dem robusten Maximum-Likelihood-Schätzer (MLR, *Maximum Likelihood Robust Estimator*) und der *Complex-Option* von Mplus (Muthén und Muthén 2017) geschätzt⁵.

27.3.1 Ergebnisse mit dem Multiconstruct-LST-Modell

Das Multiconstruct-LST-Modell zeigte in der ersten Schätzung einen nur mäßigen Datenfit ($\chi^2 = 842.2$, $df = 396$, $p = .00$, Comparative Fit Index [CFI] = .964, Root Mean Square Error of Approximation [RMSEA] = .055). Ein möglicher Grund für die Fehlanpassung kann darin liegen, dass die Lehrer von der zweiten zur dritten Messgelegenheit mit dem Schuljahr wechselten. Dadurch ist die Annahme eines stabilen Traits für die Lehrer zu stark. Teilt man die Submodelle für die Schüler und Lehrer in je zwei neue LST-Modelle (entsprechend dem Schuljahr) und setzt man zusätzlich alle Trait-Ladungen gleich⁶, verbessert sich die Modellanpassung erheblich und weist einen guten Modelfit auf ($\chi^2 = 362.4$, $df = 312$, $p = .03$, CFI = .996, RMSEA = .021; Teilmodell für die Ängstlichkeit in Abb. 27.5).

Stabilität, Variabilität und Reliabilität

In Tab. 27.2 präsentieren wir nur einen Auszug der Ergebnisse der Stabilität, Variabilität (Messgelegenheitsspezifität) und Reliabilität für die Ängstlichkeit der Schüler, da an diesem Auszug bereits die wesentlichen Merkmale des Multiconstruct-LST-Modells für die Varianzzusammensetzung deutlich gemacht werden können. Die vollständigen Ergebnisse (einschließlich der Korrelationstabellen) sind in detaillierter Form bei Courvoisier et al. (2008) beschrieben.

Zunächst können wir feststellen, dass alle Indikatoren hoch reliabel gemessen wurden. Sowohl die Schüler als auch die Lehrer zeigten eine recht stabile Einschätzung der Ängstlichkeit, da ein großer Teil der aufgeklärten Varianz auf die Trait-Variablen zurückgeführt werden kann (52 bis 87 %). Auffallend ist weiterhin, dass mit zunehmender Dauer der Untersuchung die Stabilität der Ratings zunimmt. Bei den Schülern sinkt die Variabilität von der ersten bis zur dritten Messung und steigt dann wieder an. Bei den Lehrern ist die jeweils zweite Messgelegenheit eines Schuljahres durch eine geringere Variabilität gekennzeichnet (Messgelegenheiten 2 und 4), wenngleich die Lehrer im zweiten Schuljahr der Untersuchung etwas variablere Einschätzungen zeigen.

⁵ Diese Spezifikation berücksichtigt die Homogenität der Schüler in den einzelnen Schulklassen (Schachtelung von Ratings, Multilevelstruktur) und korrigiert die Standardfehler der Parameterschätzungen und den χ^2 -Wert entsprechend.

⁶ Durch diese Restriktion wird sichergestellt, dass in beiden Modellen die gleiche Beziehung zwischen Indikatoren und Traits besteht, also das identische Konstrukt gemessen wird.

27.3 · Multiconstruct-LST- und Multimethod-LST-Modell in der empirischen Anwendung

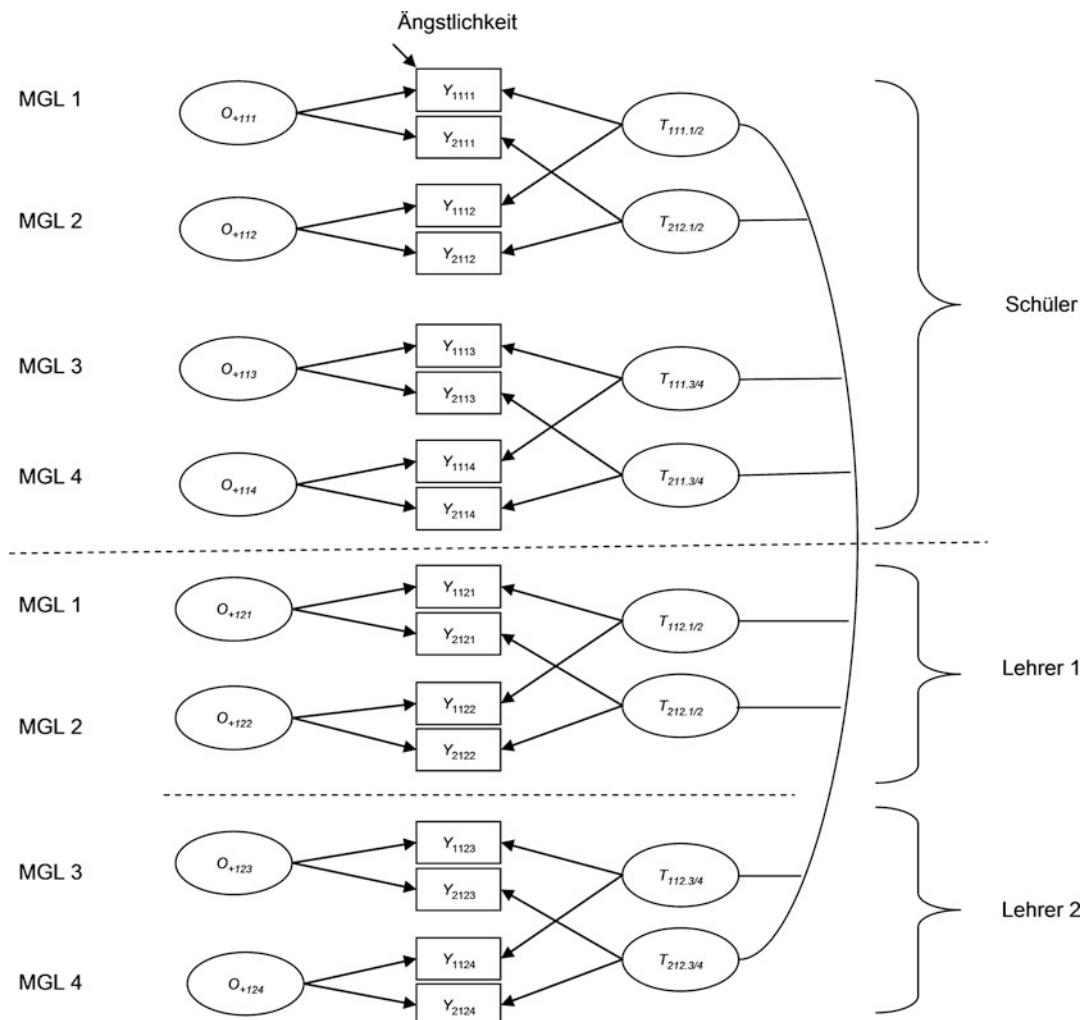


Abb. 27.5 Darstellung einer Trait-Einheit des Multiconstruct-LST-Modells für vier Messgelegenheiten. T_{ikm+} : Trait-Variable; O_{+kmt} : Messgelegenheitsvariable; T_{ikmt} : $t = 1/2$ kennzeichnet die Trait-Variablen des ersten Schuljahres (MGL 1 und 2), $3/4$ die Trait-Variablen des zweiten Schuljahres (MGL 3 und 4); MGL: Messgelegenheit. Die Korrelationen zwischen den Messgelegenheitsfaktoren sind aus Gründen der Lesbarkeit nicht abgebildet. Die Fehlervariable ist nur für den ersten Indikator angegeben

Wie sehr hängen nun die Schüler- und Lehrerratings zusammen? Eine Analyse der Korrelationen der indikator spezifischen Trait-Variablen zeigt, dass diese nur zu .19 bis .29 zwischen den Schülern und den Lehrern korrelieren. Lehrer und Schüler stimmen offensichtlich in ihren Ratings nicht sehr hoch überein. Darüber hinaus stimmen die Lehrer der unterschiedlichen Schuljahre in ihren Einschätzungen auch nur zu einem sehr geringen Teil überein (alle Korrelationen sind kleiner als .17). Die testhälftenspezifischen Traits der Schüler und der beiden Lehrer sind jedoch sehr hoch miteinander korreliert, was auf die Homogenität der Testhälften hinweist (alle Korrelationen sind höher als .86 für die Schüler und .94 für die Lehrerratings innerhalb eines Schuljahrs).

Die in der Ergebnisliste nicht wiedergegebenen Korrelationen zwischen den Traits „Ängstlichkeit“ und „Depressivität“ der Schüler sind alle größer als .68, ein deutliches Zeichen für die Verwandtschaft und mangelnde diskriminante Validität beider Konstrukte. Bei den Lehrern wird die mangelnde Trennbarkeit der beiden Konstrukte noch stärker sichtbar. Die Korrelationen der Trait-Variablen der beiden Konstrukte in einem Schuljahr sind alle größer als .83.

Konvergente und diskriminante Validität

Tabelle 27.2 Stabilitäten, Messgelegenheitsspezifitäten und Reliabilitäten der Indikatoren im Multiconstruct-LST-Modell für die Ängstlichkeit			
Testhälften	Stabilität	Messgelegenheitsspezifität	Reliabilität
<i>Ängstlichkeit im Schülerrating</i>			
Y_{1111}	.59	.31	.90
Y_{2111}	.56	.32	.88
Y_{1112}	.86	.07	.93
Y_{2112}	.82	.08	.90
Y_{1113}	.83	.07	.90
Y_{2113}	.87	.06	.93
Y_{1114}	.69	.21	.90
Y_{2114}	.73	.19	.92
<i>Ängstlichkeit im Lehrerrating</i>			
Y_{1121}	.60	.31	.91
Y_{2121}	.58	.32	.90
Y_{1122}	.87	.04	.91
Y_{2122}	.85	.04	.89
Y_{1123}	.55	.37	.92
Y_{2123}	.52	.38	.90
Y_{1124}	.62	.29	.91
Y_{2124}	.61	.31	.92

Wie gut können die Lehrer Schwankungen der Ängstlichkeit und der Depressivität der Schüler nachvollziehen? Wirken sich die Situationen in gleichem Ausmaß auf die Einschätzungen der beiden Konstrukte für die Schüler und die Lehrer aus? Zur Beantwortung dieser Fragen können die Korrelationen der messgelegenheitspezifischen Variablen herangezogen werden. Die Schwankungen der Einschätzungen der Schüler und der Lehrer sind sehr unterschiedlich. Schätzen die Schüler ihre momentane Ängstlichkeit höher ein als gewöhnlich, so schätzen die Lehrer die Schüler tendenziell sogar weniger ängstlich ein als gewöhnlich (zwei der vier möglichen Korrelationen der Messgelegenheitsfaktoren sind negativ [−.09 und −.25], eine ist 0, lediglich eine Korrelation ist positiv mit .19). Ein ähnliches Muster zeigt sich für die messgelegenheitspezifischen Variablen der Depressivität.

Die messgelegenheitsspezifischen Variablen der Schüler für Ängstlichkeit und Depressivität sind zwischen .41 und .74 korreliert, was für einen gleichartigen Einfluss der Messgelegenheit auf beide Konstrukte spricht. Für die Lehrer zeigen sich sehr hohe Korrelationen der messgelegenheitsspezifischen Variablen für die beiden Konstrukte zur jeweils ersten Messgelegenheit (.80 und .84). Die situativen Einflüsse generalisieren somit sehr stark über die beiden Konstrukte hinweg. Zur zweiten Messgelegenheit fallen diese Korrelationen auf .40 und .58. Dies spricht dafür, dass die Lehrer (zumindest in ihrer Wahrnehmung) besser zwischen den beiden Konstrukten differenzieren, wenn sie die Schüler am Ende des Schuljahres besser kennen.

27.3.2 Ergebnisse mit dem Multimethod-LST-Modell

Das Multimethod-LST-Modell mit den Selbsteinschätzungen der Schüler als Standardmethode (Teilmodell für die Ängstlichkeit in Abb. 27.6) passt ebenfalls gut auf die Daten ($\chi^2 = 338.4$, $df = 288$, $p = .02$, CFI = .996, RMSEA = .022).

Die Reliabilitäten und Varianzkomponenten für Ängstlichkeit sind in Tab. 27.3 wiedergegeben. Im Gegensatz zum vorherigen Modell können hier die gemeinsamen Varianzkomponenten direkt bestimmt werden. Wieder sind die Ratings der Schüler relativ stabil über die Zeit. Auch die Ratings der Lehrer zeigen sich fast ebenso stabil. Analysiert man die Stabilität etwas genauer, so zeigt sich, dass sie fast ausschließlich auf die Stabilität des Methodeneffekts zurückgeführt werden kann und nicht an einer erhofften Übereinstimmung mit den Schülern liegt (die maximale Stabilität, die auf den Trait der Schüler zurückzuführen ist, liegt bei .06). Die Ratings der Lehrer sind in keiner Weise von den situativen Schwankungen der Schüler beeinflusst, sondern die Schwankungen der Lehrer hängen ausschließlich von den internen Zuständen der Lehrer ab (alle auf die situativen Variablen der Schüler zurückzuführenden Varianzkomponenten sind kleiner als .01).

Ergebnisse für Ängstlichkeit

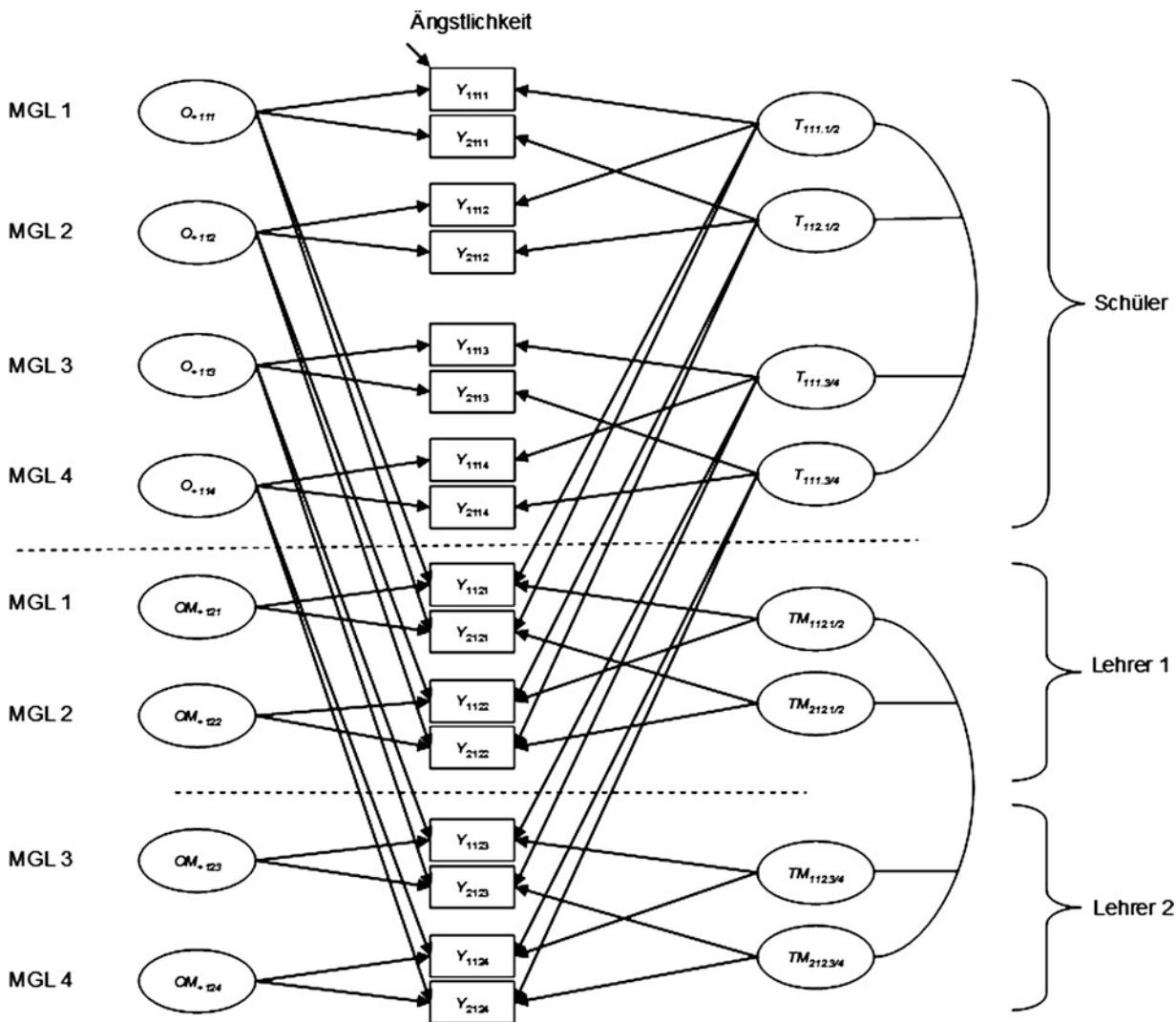


Abb. 27.6 Multimethod-LST-Submodell zur Ängstlichkeit für das erste Schuljahr. $T_{ikm.1/2}$: Trait-Variable; $TM_{ikm.1/2}$: Methodenvariable; O_{+kmt} : messgelegenheitsspezifische Variable; OM_{+kmt} : messgelegenheitsspezifische Abweichungsvariable; MGL : Messgelegenheit. Korrelationen sind aus Gründen der Lesbarkeit nicht abgebildet. Die Fehlervariable ist nur für den ersten Indikator angegeben

Tabelle 27.3 Stabilitäten, Variabilitäten und Reliabilitäten der Indikatoren im Multimethod-LST-Modell für die Ängstlichkeit

Testhälfte	Stabilität	Stabilität des Traits	Stabilität der Methode	Messgelegenheits-spezifität	Mit den Schülern geteilte Messgelegenheitsspezifität	Messgelegenheits-spezifität der Lehrer	Reliabilität
<i>Ratings der Schüler</i>							
Y_{1111}	.59			.28			.87
Y_{2111}	.55			.35			.91
Y_{1112}	.86			.06			.92
Y_{2112}	.80			.12			.92
Y_{1113}	.84			.03			.87
Y_{2113}	.86			.14			1.00
Y_{1114}	.68			.24			.92
Y_{2114}	.74			.17			.91
<i>Ratings der Lehrer</i>							
Y_{1121}	.76	.06	.70	.16	.00	.16	.92
Y_{2121}	.74	.03	.71	.17	.00	.17	.91
Y_{1122}	.69	.05	.64	.21	.01	.20	.89
Y_{2122}	.67	.04	.63	.21	.00	.21	.88
Y_{1123}	.62	.01	.61	.30	.00	.30	.92
Y_{2123}	.60	.02	.58	.31	.00	.31	.91
Y_{1124}	.58	.04	.54	.33	.00	.33	.91
Y_{2124}	.56	.05	.52	.35	.00	.35	.91

Ergebnisse für Depressivität

Die Ergebnisse der Depressivitäts- weichen kaum von den Ergebnissen der Ängstlichkeitsratings ab. Das Multimethod-LST-Modell verdeutlicht somit unverkennbar die starke Divergenz der Schüler- und Lehrerratings. Die Korrelationen zwischen den Trait-Faktoren und zwischen den Methodenvariablen zeigen ein nahezu identisches Bild wie im Multiconstruct-LST-Modell. Es findet sich eine hohe Korrelation der Trait-Variablen der Schüler, die auf mangelnde diskriminante Validität hindeutet, und eine hohe Korrelation der Methodenvariablen der Lehrer, die auf einen starken Generalisierungseffekt der Methode hindeutet. Das heißt: Über-schätzt ein Lehrer die Ängstlichkeit der Schüler, so überschätzt er tendenziell auch deren Depressivität. Zur besseren Anpassung des Modells wurden für jede Test-hälfte separat Methodenvariablen geschätzt, wobei sich die Einflüsse der Methode als recht homogen auf beide Testhälften erwiesen.

Die messgelegenheitsspezifischen Effekte der Schüler wirken sich auf beide Konstrukte auf einem ähnlich hohen Niveau wie im vorhergehenden Modell aus. Die messgelegenheitsspezifischen Abweichungsvariablen der Lehrer generalisieren ebenfalls über die Konstrukte hinweg. Die messgelegenheitsspezifischen Variablen der Schüler können nicht herangezogen werden, um die Abweichungen der Lehrer bei der Einschätzung des anderen Konstrukt zu erklären; d. h., Schüler, die sich selbst ängstlicher als sonst erleben, werden von den Lehrern nicht depressiver als sonst eingeschätzt (die Korrelationen zwischen den messgelegenheitsspezifischen Variablen der Schüler für die Ängstlichkeit mit den messgelegenheitsspezifischen Abweichungsvariablen der Lehrer für die Depressivität und vice versa liegen alle nahe 0).

27.3.3 Fazit der Anwendungen der beiden multimethodalen LST-Modelle

Die Anwendungen der beiden multimethodalen LST-Modelle illustrieren den Nutzen dieser Modelle und ermöglichen interessante Einblicke in den Zusammenhang von Schüler- und Lehrerratings: Lehrerratings können *nicht* dazu herangezogen werden, die von den Schülern berichtete Ängstlichkeit und Depressivität angemessen widerzuspiegeln. Vielmehr zeigen Lehrer sowohl im Querschnitt als auch im Verlauf eines Schuljahres einen stabilen Bias in Bezug auf ihre Einschätzungen. Sie sind nicht nur nicht in der Lage, die stabile Ängstlichkeit oder Depressivität der Schüler einzuschätzen, sie können auch die jeweiligen situativen Schwankungen der Ängstlichkeit oder der Depressivität der Schüler nicht nachvollziehen. Gesteh man den Schülern der vierten und fünften Klasse in US-amerikanischen Schulen die Fähigkeit zu, ihre eigene Ängstlichkeit und Depressivität angemessen einschätzen zu können, können sich Psychologen diesen Ergebnissen zufolge keinesfalls auf die Einschätzungen der Lehrer verlassen. Die landläufige Annahme, dass Lehrer nach einem Schuljahr in der Lage sein sollten, die Gefühle der Schüler ihrer Klasse einzuschätzen zu können, wird zumindest mit dieser Studie widerlegt.

Querschnittlicher und längsschnittlicher Bias im Lehrerrating

27.4 Praktische Hinweise zur Analyse longitudinaler multimodaler Modelle

Im folgenden sollen praktische Hinweise zur Analyse der komplexen Modelle geben werden. Da in diesen Modellen eine Vielzahl von Parametern geschätzt werden, kann erst ab relativ hohen Stichprobengrößen mit verlässlichen Ergebnissen gerechnet werden. Die Bestimmung der benötigten minimalen Stichprobengrößen ist Gegenstand gegenwärtiger Forschung.

Es empfiehlt sich, longitudinale multimethodale Modelle schrittweise in den empirischen Anwendungen aufzubauen. Wir wollen den Aufbau dieser Modelle in drei Schritten skizzieren und auf mögliche Probleme eingehen:

■■ Schritt 1: Ein Konstrukt und eine Methode

Zunächst empfiehlt es sich, für jedes Konstrukt und jede Methode getrennte Analysen vorzunehmen. Für jede dieser Trait-Methoden-Einheiten kann ein Modell geschätzt werden, in dem für jede Messgelegenheit nur eine State-Variablen einge führt wird. Die Korrelationen dieser (Single-Method-)State-Variablen sollten signifikant sein, damit im nächsten Schritt überprüft werden kann, ob ein LST-Modell diese Daten angemessen repräsentiert (s. auch Geiser et al. 2010). Korrelieren die State-Variablen nicht miteinander, d.h., es handelt sich um ein vollständig zeitlich instabiles Merkmal, so kann auch kein LST-Modell angepasst werden. Eine Möglichkeit, konvergente und diskriminante Validität im zeitlichen Verlauf festzustellen, bietet dann das Multioccasion-MTMM-Modell. Es sei ausdrücklich darauf hingewiesen, dass Methoden gut übereinstimmen können, auch wenn keine zeitliche Stabilität vorliegt. Psychologische Konstrukte haben nicht immer einen stabilen zeitlichen Charakter (Stimmungen, Hormonlevel, Ärgererleben etc.).

Korrelationen der State-Variablen sollten signifikant sein

■■ Schritt 2: LST-Modell

Im zweiten Schritt kann für jede Trait-Methoden-Einheit ein (Mono-Method-)LST-Modell geschätzt werden. Mit diesem Modell kann die zeitliche Stabilität eines Konstruktts anhand einer einzigen Methode überprüft werden. Mögliche Gründe für eine schlechte Anpassung des Modells an die Daten können hier neben mangelnder Stabilität noch weitere Faktoren sein:

Gründe für mangelnden Modellfit

- Der sog. „Sokrates-Effekt“: Testpersonen kennen nicht alle psychologischen Konstrukte und verfügen folglich nicht über eine kognitive Repräsentation der

Sokrates-Effekt

Mangelnde messgelegenheitsspezifische Varianz

Trait-Ausprägung. Durch die Befragung können Testpersonen eine Vorstellung darüber entwickeln, welches Konstrukt gemessen werden sollte, sodass sie in weiteren Befragungen eventuell konformer im Hinblick auf das von ihnen vermutete Konstrukt antworten. Bei Modellen, in denen die Annahme der Messinvarianz getroffen wurde, kann dies zu Problemen führen. Liegt ein Sokrates-Effekt vor, können die Ladungen der Indikatoren des ersten Messzeitpunkts von der Annahme der Messinvarianz ausgenommen werden (d. h. frei geschätzt werden).

- Eine oder mehrere messgelegenheitsspezifische Variablen haben keine Varianz: Dieses Phänomen tritt bei sehr stabilen Merkmalen (wie der Intelligenz) auf, die keinen oder nur sehr geringen situativen Schwankungen unterliegen; es kann aber auch bei weniger stabilen Konstrukten auftreten. Durch Eliminierung der betreffenden Messgelegenheitsvariablen kann das LST-Modell an die Daten angepasst werden. Im extremen Fall, wenn alle Messgelegenheitsvariablen eliminiert werden, bleibt ein reines Latent-Trait-Modell bestehen.
- Die verschiedenen Indikatoren eines Traits sind nicht perfekt homogen: Dieses Problem wurde bereits in ► Abschn. 27.2 behandelt.
- Das Konstrukt unterliegt einem Veränderungsprozess und ist nicht stabil über die Zeit: In diesem Fall müssen andere longitudinale Modelle wie etwa das Latent-Curve-Modell oder ein Modell mit autoregressiven Strukturen (s. dazu Bollen und Curran 2006) herangezogen werden.

■ ■ Schritt 3: Longitudinales multimethodales Modell

Je nach inhaltlicher Fragestellung kann in einem dritten Schritt eines der vorgestellten MTMM-LST-Modelle analysiert werden.

27.5 Zusammenfassung

Ziel dieses Kapitels war es, zu verdeutlichen, dass Merkmalsausprägungen von Individuen über die Zeit schwanken können, und dass somit auch die konvergente und diskriminante Validität verschiedener Methoden und Konstrukte zeitlichen Veränderungen unterworfen sind. Die Analyse konvergenter und diskriminanter Validität ist Basis jeder diagnostischen Entscheidung. Nur bei gesicherter Qualität der eingesetzten Verfahren können zuverlässig Indikationen für mögliche Interventionen getroffen werden. Besonders bei Kindern, die sich in einem Entwicklungsprozess befinden, aber auch bei Erwachsenen ist es notwendig, die zeitliche Stabilität der gefundenen Testscores zu untersuchen. Nur bei gegebener Stabilität der Messungen kann von einem stabilen Trait ausgegangen werden. Darüber hinaus ist es wichtig, zu analysieren, wie sich die konvergente Validität verschiedener Messmethoden über die Zeit entwickelt.

Drei longitudinale multimethodale Modelle für mehrere Traits wurden vorgestellt, die es erlauben, die Konvergenz verschiedener Methoden und die diskriminante Validität von Traits und States zu untersuchen. Die empirischen Anwendungen zeigen deutlich, dass implizite Annahmen zur Übereinstimmung verschiedener Methoden prinzipiell überprüft werden müssen.

27.6 EDV-Hinweise

Die hier vorgestellten longitudinalen multimethodalen Modelle können mit gängigen EDV-Programmen zur Analyse von Strukturgleichungsmodellen vorgenommen werden (s. hierzu ► Abschn. 25.9).

27.7 Kontrollfragen

?) Die Antworten auf die folgenden Fragen finden Sie im Lerncenter zu diesem Kapitel unter ► <https://www.lehrbuch-psychologie.springer.com> (Projekt Testtheorie und Fragebogenkonstruktion).

1. Was wird unter austauschbaren, strukturell unterschiedlichen und gleichwertigen Methoden im Zusammenhang mit MTMM-Modellen verstanden? Was sind die wesentlichen Unterschiede zwischen diesen Methoden?
2. Worin liegt der Vorteil trait-spezifischer Methodeneffekte in MTMM-Modellen?
3. Worin besteht der Unterschied zwischen LST- und MTMM-Modellen?
4. Bei welchen wissenschaftlichen Fragestellungen sollte das Multioccasion-MTMM-, das Multiconstruct-LST- oder das Multimethod-LST-Modell eingesetzt werden?

Literatur

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bollen, K. A. & Curran, P. J. (2006). *Latent curve models: A structural equations perspective*. Hoboken, NJ: Wiley.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.
- Burns, G. L. & Haynes, S. N. (2006). Clinical Psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid and E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 401–418). Washington, DC: American Psychological Association.
- Cole, D. A. & Jordan, A. E. (1995). Competence and Memory: Integrating psychosocial and cognitive correlates of child depression. *Child Development*, 66, 459–473.
- Cole, D. A. & Martin N. C. (2005). The longitudinal structure of the children's depression inventory: Testing a latent trait-state model. *Psychological Assessment*, 17, 144–155.
- Cole, D. A., Martin, J. M., Powers, B. & Truglio, R. (1996). Modeling causal relations between academic and social competence and depression: A multitrait-multimethod longitudinal study of children. *Journal of Abnormal Psychology*, 105, 258–270.
- Cole, D. A., Truglio, R. & Peeke, L. (1997). Relation between symptoms of anxiety and depression in children: A multitrait-multimethod-multigroup assessment. *Journal of Consulting and Clinical Psychology*, 65, 110–119.
- Costa, P. T. & McCrae, R. R. (1998). Trait theories of personality. In D. F. Barone, M. Hersen & V. B. Van Hasselt (Eds.), *Advances Personality* (pp. 103–121). New York: Plenum Press.
- Courvoisier, D. S. (2006). *Unfolding the constituents of psychological scores: Development and application of mixture and multitrait-multimethod LST models*. Unpublished doctoral dissertation, University of Geneva, Switzerland.
- Courvoisier, D. S., Eid, M. & Nussbeck, F. W. (2007). Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods*, 12, 80–104.
- Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C. & Cole, D. A. (2008). Analyzing the convergent and discriminant validity of states and traits: Development and applications of multimethod latent state-trait models. *Psychological Assessment*, 20, 270–280.
- Dumenci, L. & Windle, M. (1998). A multitrait-multioccasion generalization of the latent trait-state model: Description and application. *Structural Equation Modeling*, 5, 391–410.
- Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition on the basis of stochastic measurement theory. *Methods of Psychological Research*, 1, 65–85.
- Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 223–230). Washington, DC: American Psychological Association.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Eid, M., Lischetzke, T. & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid and E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association.
- Eid, M., Lischetzke, T., Nussbeck, F. W. & Trierweiler, L. I. (2003). Separating trait effects from method-specific effects in multitrait-multimethod models: A multiple indicator CTC(M-1) model. *Psychological Methods*, 8, 38–60.
- Eid, M., Notz, P., Steyer, R. & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analysis. *Personality and Individual Differences*, 16, 63–76.

- 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.
- Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S. & Cole, D. A. (2008). Analyzing the convergent and discriminant validity of change: Structural equation modeling of multitrait-multimethod-multiplication data. *Psychological Assessment, 20*, 270–280.
- Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S. & Cole, D. A. (2010). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. *Developmental Psychology, 46*, 29.
- Kovacs, M. (1981). Rating scales to assess depression in school-aged children. *Acta Paedopsychiatrica, 46*, 305–315.
- Kovacs, M. (1982). *The children's depression inventory: A self-rating depression scale for school-aged youngsters*. Unpublished test manual.
- Lefkowitz, M. & Tesiny, E. (1980). Assessment of childhood depression. *Journal of Consulting Clinical Psychologists, 48*, 43–50.
- Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.
- Majcen, A.-M., Steyer, R. & Schwenkmezger, P. (1988). Konsistenz und Spezifität bei Eigenschafts- und Zustandsangst. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 105–120.
- Marsh, H. W. & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling, 1*, 116–145.
- Marsh, H. W. & Grayson, D. (1995) Latent variable models of multitrait-multimethod data. In R. E. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 117–187). Thousand Oaks, London: Sage.
- Muthén, L. K. & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Reynolds, C. R. & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6*, 271–280.
- Schermelleh-Engel, K., Keith, N., Moosbrugger, H. & Hodapp, V. (2004). Decomposing person and occasion-specific effects: An extension of the latent state-trait (LST) Theory to hierarchical LST models. *Psychological Methods, 9*, 198–219.
- Schmitt, M. J. (2000). Mother-daughter attachment and family cohesion: Single and multi-trait latent state-trait models of current and retrospective perceptions. *European Journal of Psychological Assessment, 16*, 115–124.
- Steyer, R. (1987). Konsistenz und Spezifität: Definition zweier zentraler Begriffe der Differentiellen Psychologie und einfaches Modell zu ihrer Identifikation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 8*, 245–258.
- Steyer, R. (1989). Models of classical test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, and testability. *Methodika, 3*, 25–60.
- Steyer, R., Majcen, A.-M., Schwenkmezger, P. & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research, 1*, 281–299.
- Steyer, R., Ferring, D. & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 79–98.
- Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology, 11*, 71–98.
- Steyer, R., Schwenkmezger, P. & Auer, A. (1990). The emotional and cognitive components of trait anxiety: A latent state trait anxiety model. *Personality and Individual Differences, 11*, 125–134.
- Thorndike, E. L. (1920). A constant error in psychological rating. *Journal of Applied Psychology, 4*, 25–29.
- Youssi, S. & Steyer, R. (2006). Latent-State-Trait-Theorie. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 346–357). Göttingen: Hogrefe.

Serviceteil

[Übersicht der griechischen Buchstaben – 740](#)

[Verteilungsfunktion der Standardnormalverteilung \(z-Tabelle\) – 741](#)

[Glossar – 744](#)

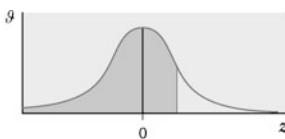
[Stichwortverzeichnis – 761](#)

Übersicht der griechischen Buchstaben

Großbuchstaben	Kleinbuchstaben	Großbuchstaben	Kleinbuchstaben	Großbuchstaben	Kleinbuchstaben	Name
A	α	A	α	A	α	Alpha
B	β	B	β	B	β	Beta
Γ	γ	Γ	γ	Γ	γ	Gamma
Δ	δ	Δ	δ	Δ	δ	Delta
E	ε	E	ε	E	ε	Epsilon
Z	ζ	Z	ζ	Z	ζ	Zeta
H	η	H	η	H	η	Eta
Θ	θ, ϑ	Θ	θ, ϑ	Θ	θ, ϑ	Theta
I	ι	I	ι	I	ι	Iota
K	κ	K	κ	K	κ	Kappa
Λ	λ	Λ	λ	Λ	λ	Lambda
M	μ	M	μ	M	μ	My
N	ν	N	ν	N	ν	Ny
Ξ	ξ	Ξ	ξ	Ξ	ξ	Xi
O	\circ	O	\circ	O	\circ	Omi-kron
Π	π	Π	π	Π	π	Pi
P	ρ, ϱ	P	ρ, ϱ	P	ρ, ϱ	Rho
Σ	σ	Σ	σ	Σ	σ	Sigma
T	τ	T	τ	T	τ	Tau
Υ	υ	Υ	υ	Υ	υ	Ypsilon
Φ	ϕ, φ	Φ	ϕ, φ	Φ	ϕ, φ	Phi
X	χ	X	χ	X	χ	Chi
Ψ	ψ	Ψ	ψ	Ψ	ψ	Psi
Ω	ω	Ω	ω	Ω	ω	Omega

Verteilungsfunktion der Standardnormalverteilung (z-Tabelle)

Quelle: Glass, G.V. & Stanley, J.C. (1970). *Statistical methods in education and psychology* (S. 513–519). Englewood Cliffs, NJ: Prentice-Hall.



z	Fläche	Ordinate	z	Fläche	Ordinate	z	Fläche	Ordinate
-3,00	0,0013	0,0044	-2,45	0,0071	0,0198	-1,90	0,0287	0,0656
-2,99	0,0014	0,0046	-2,44	0,0073	0,0203	-1,89	0,0294	0,0669
-2,98	0,0014	0,0047	-2,43	0,0075	0,0208	-1,88	0,0301	0,0681
-2,97	0,0015	0,0048	-2,42	0,0078	0,0213	-1,87	0,0307	0,0694
-2,96	0,0015	0,0050	-2,41	0,0080	0,0219	-1,86	0,0314	0,0707
-2,95	0,0016	0,0051	-2,40	0,0082	0,0224	-1,85	0,0322	0,0721
-2,94	0,0016	0,0053	-2,39	0,0084	0,0229	-1,84	0,0329	0,0734
-2,93	0,0017	0,0055	-2,38	0,0087	0,0235	-1,83	0,0336	0,0748
-2,92	0,0018	0,0056	-2,37	0,0089	0,0241	-1,82	0,0344	0,0761
-2,91	0,0018	0,0058	-2,36	0,0091	0,0246	-1,81	0,0351	0,0775
-2,90	0,0019	0,0060	-2,35	0,0094	0,0252	-1,80	0,0359	0,0790
-2,89	0,0019	0,0061	-2,34	0,0096	0,0258	-1,79	0,036	0,0804
-2,88	0,0020	0,0063	-2,33	0,0099	0,0264	-1,78	0,0375	0,0818
-2,87	0,0021	0,0065	-2,32	0,0102	0,0270	-1,77	0,0384	0,0833
-2,86	0,0021	0,0067	-2,31	0,0104	0,0277	-1,76	0,0392	0,0848
-2,85	0,0022	0,0069	-2,30	0,0107	0,0283	-1,75	0,0401	0,0863
-2,84	0,0023	0,0071	-2,29	0,0110	0,0290	-1,74	0,0409	0,0878
-2,83	0,0023	0,0073	-2,28	0,0113	0,0297	-1,73	0,0418	0,0893
-2,82	0,0024	0,0075	-2,27	0,0116	0,0303	-1,72	0,0427	0,0909
-2,81	0,0025	0,0077	-2,26	0,0119	0,0310	-1,71	0,0436	0,0925
-2,80	0,0026	0,0079	-2,25	0,0122	0,0317	-1,70	0,0446	0,0940
-2,79	0,0026	0,0081	-2,24	0,0125	0,0325	-1,69	0,0455	0,0957
-2,78	0,0027	0,0084	-2,23	0,0129	0,0332	-1,68	0,0465	0,0973
-2,77	0,0028	0,0086	-2,22	0,0132	0,0339	-1,67	0,0475	0,0989
-2,76	0,0029	0,0088	-2,21	0,0136	0,0347	-1,66	0,0485	0,1006
-2,75	0,0030	0,0091	-2,20	0,0139	0,0355	-1,65	0,0495	0,1023
-2,74	0,0031	0,0093	-2,19	0,0143	0,0363	-1,64	0,0505	0,1040
-2,73	0,0032	0,0096	-2,18	0,0146	0,0371	-1,63	0,0516	0,1057
-2,72	0,0033	0,0099	-2,17	0,0150	0,0379	-1,62	0,0526	0,1074
-2,71	0,0034	0,0101	-2,16	0,0154	0,0387	-1,61	0,0537	0,1092
-2,70	0,0035	0,0104	-2,15	0,0158	0,0396	-1,60	0,0548	0,1109
-2,69	0,0036	0,0107	-2,14	0,0162	0,0404	-1,59	0,0559	0,1127
-2,68	0,0037	0,0110	-2,13	0,0166	0,0413	-1,58	0,0571	0,1145
-2,67	0,0038	0,0113	-2,12	0,0170	0,0422	-1,57	0,0582	0,1163
-2,66	0,0039	0,0116	-2,11	0,0174	0,0431	-1,56	0,0594	0,1182
-2,65	0,0040	0,0119	-2,10	0,0179	0,0440	-1,55	0,0606	0,1200
-2,64	0,0041	0,0122	-2,09	0,0183	0,0449	-1,54	0,0618	0,1219
-2,63	0,0043	0,0126	-2,08	0,0188	0,0459	-1,53	0,0630	0,1238
-2,62	0,0044	0,0129	-2,07	0,0192	0,0468	-1,52	0,0643	0,1257
-2,61	0,0045	0,0132	-2,06	0,0197	0,0478	-1,51	0,0655	0,1276
-2,60	0,0047	0,0136	-2,05	0,0202	0,0488	-1,50	0,0668	0,1295
-2,59	0,0048	0,0139	-2,04	0,0207	0,0498	-1,49	0,0681	0,1315
-2,58	0,0049	0,0143	-2,03	0,0212	0,0508	-1,48	0,0694	0,1334
-2,57	0,0051	0,0147	-2,02	0,0217	0,0519	-1,47	0,0708	0,1354
-2,56	0,0052	0,0151	-2,01	0,0222	0,0529	-1,46	0,0721	0,1374
-2,55	0,0054	0,0154	-2,00	0,0228	0,0540	-1,45	0,0735	0,1394
-2,54	0,0055	0,0158	-1,99	0,0233	0,0551	-1,44	0,0749	0,1415
-2,53	0,0057	0,0163	-1,98	0,0239	0,0562	-1,43	0,0764	0,1435
-2,52	0,0059	0,0167	-1,97	0,0244	0,0573	-1,42	0,0778	0,1456
-2,51	0,0060	0,0171	-1,96	0,0250	0,0584	-1,41	0,0793	0,1476
-2,50	0,0062	0,0175	-1,95	0,0256	0,0596	-1,40	0,0808	0,1497
-2,49	0,0064	0,0180	-1,94	0,0262	0,0608	-1,39	0,0823	0,1518
-2,48	0,0066	0,0184	-1,93	0,0268	0,0620	-1,38	0,0838	0,1539
-2,47	0,0068	0,0189	-1,92	0,0274	0,0632	-1,37	0,0853	0,1561
-2,46	0,0069	0,0194	-1,91	0,0281	0,0644	-1,36	0,0869	0,1582

(Fortsetzung)

<i>z</i>	Fläche	Ordinate	<i>z</i>	Fläche	Ordinate	<i>z</i>	Fläche	Ordinate
-1,35	0,0885	0,1604	-0,65	0,2578	0,3230	0,05	0,5199	0,3984
-1,34	0,0901	0,1626	-0,64	0,2611	0,3251	0,06	0,5239	0,3982
-1,33	0,0918	0,1647	-0,63	0,2643	0,3271	0,07	0,5279	0,3980
-1,32	0,0934	0,1669	-0,62	0,2676	0,3292	0,08	0,5319	0,3977
-1,31	0,0951	0,1691	-0,61	0,2709	0,3312	0,09	0,5359	0,3973
-1,30	0,0968	0,1714	-0,60	0,2749	0,3332	0,10	0,5398	0,3970
-1,29	0,0985	0,1736	-0,59	0,2776	0,3352	0,11	0,5438	0,3965
-1,28	0,1003	0,1758	-0,58	0,2810	0,3372	0,12	0,5478	0,3961
-1,27	0,1020	0,1781	-0,57	0,2843	0,3391	0,13	0,5517	0,3956
-1,26	0,1038	0,1804	-0,56	0,2877	0,3410	0,14	0,5557	0,3951
-1,25	0,1056	0,1826	-0,55	0,2912	0,3429	0,15	0,5596	0,3945
-1,24	0,1075	0,1849	-0,54	0,2946	0,3448	0,16	0,5636	0,3939
-1,23	0,1093	0,1872	-0,53	0,2981	0,3467	0,17	0,5675	0,3932
-1,22	0,1112	0,1895	-0,52	0,3015	0,3485	0,18	0,5714	0,3925
-1,21	0,1131	0,1919	-0,51	0,3050	0,3503	0,19	0,5753	0,3918
-1,20	0,1151	0,1942	-0,50	0,3085	0,3521	0,20	0,5793	0,3910
-1,19	0,1170	0,1965	-0,49	0,3121	0,3538	0,21	0,5832	0,3902
-1,18	0,1190	0,1989	-0,48	0,3156	0,3555	0,22	0,5871	0,3894
-1,17	0,1210	0,2012	-0,47	0,3192	0,3572	0,23	0,5910	0,3885
-1,16	0,1230	0,2036	-0,46	0,3228	0,3589	0,24	0,5948	0,3876
-1,15	0,1251	0,2059	-0,45	0,3264	0,3605	0,25	0,5987	0,3867
-1,14	0,1271	0,2083	-0,44	0,3300	0,3621	0,26	0,6026	0,3857
-1,13	0,1292	0,2107	-0,43	0,3336	0,3637	0,27	0,6064	0,3847
-1,12	0,1314	0,2131	-0,42	0,3372	0,3653	0,28	0,6103	0,3836
-1,11	0,1335	0,2155	-0,41	0,3409	0,3668	0,29	0,6141	0,3825
-1,10	0,1357	0,2179	-0,40	0,3446	0,3683	0,30	0,6179	0,3814
-1,09	0,1379	0,2203	-0,39	0,3483	0,3697	0,31	0,6217	0,3802
-1,08	0,1401	0,2227	-0,38	0,3520	0,3712	0,32	0,6255	0,3790
-1,07	0,1423	0,2251	-0,37	0,3557	0,3725	0,33	0,6293	0,3778
-1,06	0,1446	0,2275	-0,36	0,3594	0,3739	0,34	0,6331	0,3765
-1,05	0,1469	0,2299	-0,35	0,3632	0,3752	0,35	0,6368	0,3752
-1,04	0,1492	0,2323	-0,34	0,3669	0,3765	0,36	0,6406	0,3739
-1,03	0,1515	0,2347	-0,33	0,3707	0,3778	0,37	0,6443	0,3725
-1,02	0,1539	0,2371	-0,32	0,3745	0,3790	0,38	0,6480	0,3712
-1,01	0,1562	0,2396	-0,31	0,3783	0,3802	0,39	0,6517	0,3697
-1,00	0,1587	0,2420	-0,30	0,3821	0,3814	0,40	0,6554	0,3683
-0,99	0,1611	0,2444	-0,29	0,3859	0,3825	0,41	0,6591	0,3668
-0,98	0,1635	0,2468	-0,28	0,3897	0,3836	0,42	0,6628	0,3653
-0,97	0,1660	0,2492	-0,27	0,3936	0,3847	0,43	0,6664	0,3637
-0,96	0,1685	0,2516	-0,26	0,3974	0,3857	0,44	0,6700	0,3621
-0,95	0,1711	0,2541	-0,25	0,4013	0,3867	0,45	0,6736	0,3605
-0,94	0,1736	0,2565	-0,24	0,4052	0,3876	0,46	0,6772	0,3589
-0,93	0,1762	0,2589	-0,23	0,4090	0,3885	0,47	0,6808	0,3572
-0,92	0,1788	0,2613	-0,22	0,4129	0,3894	0,48	0,6844	0,3555
-0,91	0,1814	0,2637	-0,21	0,4168	0,3902	0,49	0,6879	0,3538
-0,90	0,1841	0,2661	-0,20	0,4207	0,3910	0,50	0,6915	0,3521
-0,89	0,1867	0,2685	-0,19	0,4247	0,3918	0,51	0,6950	0,3503
-0,88	0,1894	0,2709	-0,18	0,4286	0,3925	0,52	0,6985	0,3485
-0,87	0,1922	0,2732	-0,17	0,4325	0,3932	0,53	0,7019	0,3467
-0,86	0,1949	0,2756	-0,16	0,4364	0,3939	0,54	0,7054	0,3448
-0,85	0,1977	0,2780	-0,15	0,4404	0,3945	0,55	0,7088	0,3429
-0,84	0,2005	0,2803	-0,14	0,4443	0,3951	0,56	0,7123	0,3410
-0,83	0,2033	0,2827	-0,13	0,4483	0,3956	0,57	0,7157	0,3391
-0,82	0,2061	0,2850	-0,12	0,4522	0,3961	0,58	0,7190	0,3372
-0,81	0,2090	0,2874	-0,11	0,4562	0,3965	0,59	0,7224	0,3352
-0,80	0,2119	0,2897	-0,10	0,4602	0,3970	0,60	0,7257	0,3332
-0,79	0,2148	0,2920	-0,09	0,4641	0,3973	0,61	0,7291	0,3312
-0,77	0,2206	0,2966	-0,08	0,4681	0,3977	0,62	0,7324	0,3292
-0,78	0,2217	0,2943	-0,07	0,4721	0,3980	0,63	0,7357	0,3271
-0,76	0,2236	0,2989	-0,06	0,4761	0,3982	0,64	0,7389	0,3251
-0,75	0,2266	0,3011	-0,05	0,4801	0,3984	0,65	0,7422	0,3230
-0,74	0,2296	0,3034	-0,04	0,4840	0,3986	0,66	0,7454	0,3209
-0,73	0,2327	0,3056	-0,03	0,4880	0,3988	0,67	0,7486	0,3187
-0,72	0,2358	0,3079	-0,02	0,4920	0,3989	0,68	0,7517	0,3166
-0,71	0,2389	0,3101	-0,01	0,4960	0,3989	0,69	0,7549	0,3144
-0,70	0,2420	0,3123	0,00	0,5000	0,3989	0,70	0,7580	0,3123
-0,69	0,2451	0,3144	0,01	0,5040	0,3989	0,71	0,7611	0,3101
-0,68	0,2483	0,3166	0,02	0,5080	0,3989	0,72	0,7642	0,3079
-0,67	0,2514	0,3187	0,03	0,5120	0,3988	0,73	0,7673	0,3056
-0,66	0,2546	0,3209	0,04	0,5160	0,3986	0,74	0,7704	0,3034

Verteilungsfunktion der Standardnormalverteilung (z-Tabelle)

(Fortsetzung)

z	Fläche	Ordinate	z	Fläche	Ordinate	z	Fläche	Ordinate
0,75	0,7734	0,3011	1,50	0,9332	0,1295	2,25	0,9878	0,0317
0,76	0,7764	0,2989	1,51	0,9345	0,1276	2,26	0,9881	0,0310
0,77	0,7794	0,2966	1,52	0,9357	0,1257	2,27	0,9884	0,0303
0,79	0,7852	0,2920	1,53	0,9370	0,1238	2,28	0,9887	0,0297
0,78	0,7823	0,2943	1,54	0,9382	0,1219	2,29	0,9890	0,0290
0,80	0,7881	0,2897	1,55	0,9394	0,1200	2,30	0,9893	0,0283
0,81	0,7910	0,2874	1,56	0,9406	0,1182	2,31	0,9896	0,0277
0,82	0,7939	0,2850	1,57	0,9418	0,1163	2,32	0,9898	0,0270
0,83	0,7967	0,2827	1,58	0,9429	0,1145	2,33	0,9901	0,0264
0,84	0,7995	0,2803	1,59	0,9441	0,1127	2,34	0,9904	0,0258
0,85	0,8023	0,2780	1,60	0,9452	0,1109	2,35	0,9906	0,0246
0,86	0,8051	0,2756	1,61	0,9463	0,1092	2,36	0,9909	0,0246
0,87	0,8078	0,2732	1,62	0,9474	0,1074	2,37	0,9911	0,0241
0,88	0,8106	0,2709	1,63	0,9484	0,1057	2,38	0,9913	0,0235
0,89	0,8133	0,2685	1,64	0,9495	0,1040	2,39	0,9916	0,0229
0,90	0,8159	0,2661	1,65	0,9505	0,1023	2,40	0,9918	0,0224
0,91	0,8186	0,2637	1,66	0,9515	0,1006	2,41	0,9920	0,0219
0,92	0,8212	0,2613	1,67	0,9525	0,0989	2,42	0,9922	0,0213
0,93	0,8238	0,2589	1,68	0,9535	0,0973	2,43	0,9925	0,0208
0,94	0,8264	0,2565	1,69	0,9545	0,0957	2,44	0,9927	0,0203
0,95	0,8289	0,2541	1,70	0,9554	0,0940	2,45	0,9929	0,0198
0,96	0,8315	0,2516	1,71	0,9564	0,0925	2,46	0,9931	0,0194
0,97	0,8340	0,2492	1,72	0,9573	0,0909	2,47	0,9932	0,0189
0,98	0,8365	0,2468	1,73	0,9582	0,0893	2,48	0,9934	0,0184
0,99	0,8389	0,2444	1,74	0,9591	0,0878	2,49	0,9936	0,0180
1,00	0,8413	0,2420	1,75	0,9599	0,0863	2,50	0,9938	0,0175
1,01	0,8438	0,2396	1,76	0,9608	0,0848	2,51	0,9940	0,0171
1,02	0,8461	0,2371	1,77	0,9616	0,0833	2,52	0,9941	0,0167
1,03	0,8485	0,2347	1,78	0,9625	0,0818	2,53	0,9943	0,0163
1,04	0,8508	0,2323	1,79	0,9633	0,0804	2,54	0,9945	0,0158
1,05	0,8531	0,2299	1,80	0,9641	0,0790	2,55	0,9946	0,0154
1,06	0,8554	0,2275	1,81	0,9649	0,0775	2,56	0,9948	0,0151
1,07	0,8577	0,2251	1,82	0,9656	0,0761	2,57	0,9949	0,0147
1,08	0,8599	0,2227	1,83	0,9664	0,0748	2,58	0,9951	0,0143
1,09	0,8621	0,2203	1,84	0,9671	0,0734	2,59	0,9952	0,0139
1,10	0,8643	0,2179	1,85	0,9678	0,0721	2,60	0,9953	0,0136
1,11	0,8665	0,2155	1,86	0,9686	0,0707	2,61	0,9955	0,0132
1,12	0,8686	0,2131	1,87	0,9693	0,0694	2,62	0,9956	0,0129
1,13	0,8708	0,2107	1,88	0,9699	0,0681	2,63	0,9957	0,0126
1,14	0,8729	0,2083	1,89	0,9706	0,0669	2,64	0,9959	0,0122
1,15	0,8749	0,2059	1,90	0,9713	0,0656	2,65	0,9960	0,0119
1,16	0,8770	0,2036	1,91	0,9719	0,0644	2,66	0,9961	0,0116
1,17	0,8790	0,2012	1,92	0,9726	0,0632	2,67	0,9962	0,0113
1,18	0,8810	0,1989	1,93	0,9732	0,0620	2,68	0,9963	0,0110
1,19	0,8830	0,1965	1,94	0,9738	0,0608	2,69	0,9964	0,0107
1,20	0,8849	0,1942	1,95	0,9744	0,0596	2,70	0,9965	0,0104
1,21	0,8869	0,1919	1,96	0,9750	0,0584	2,71	0,9966	0,0101
1,22	0,8888	0,1895	1,97	0,9756	0,0573	2,72	0,9967	0,0099
1,23	0,8907	0,1872	1,98	0,9761	0,0562	2,73	0,9968	0,0096
1,24	0,8925	0,1849	1,99	0,9767	0,0551	2,74	0,9969	0,0093
1,25	0,8944	0,1826	2,00	0,9772	0,0540	2,75	0,9970	0,0091
1,26	0,8962	0,1804	2,01	0,9778	0,0529	2,76	0,9971	0,0088
1,27	0,8980	0,1781	2,02	0,9783	0,0519	2,77	0,9972	0,0086
1,28	0,8997	0,1758	2,03	0,9788	0,0508	2,78	0,9973	0,0084
1,29	0,9015	0,1736	2,04	0,9793	0,0498	2,79	0,9974	0,0081
1,30	0,9032	0,1714	2,05	0,9798	0,0488	2,80	0,9974	0,0079
1,31	0,9049	0,1691	2,06	0,9803	0,0478	2,81	0,9975	0,0077
1,32	0,9066	0,1669	2,07	0,9808	0,0468	2,82	0,9976	0,0075
1,33	0,9082	0,1647	2,08	0,9812	0,0459	2,83	0,9977	0,0073
1,34	0,9099	0,1626	2,09	0,9817	0,0449	2,84	0,9977	0,0071
1,35	0,9115	0,1604	2,10	0,9821	0,0440	2,85	0,9978	0,0069
1,36	0,9131	0,1582	2,11	0,9826	0,0431	2,86	0,9979	0,0067
1,37	0,9147	0,1561	2,12	0,9830	0,0422	2,87	0,9979	0,0065
1,38	0,9162	0,1539	2,13	0,9834	0,0413	2,88	0,9980	0,0063
1,39	0,9177	0,1518	2,14	0,9838	0,0404	2,89	0,9981	0,0061
1,40	0,9192	0,1497	2,15	0,9842	0,0396	2,90	0,9981	0,0060
1,41	0,9207	0,1476	2,16	0,9846	0,0387	2,91	0,9982	0,0058
1,42	0,9222	0,1456	2,17	0,9850	0,0379	2,92	0,9982	0,0056
1,43	0,9236	0,1435	2,18	0,9854	0,0371	2,93	0,9983	0,0055
1,44	0,9251	0,1415	2,19	0,9857	0,0363	2,94	0,9984	0,0053
1,45	0,9265	0,1394	2,20	0,9861	0,0355	2,95	0,9984	0,0051
1,46	0,9279	0,1374	2,21	0,9864	0,0347	2,96	0,9985	0,0050
1,47	0,9292	0,1354	2,22	0,9868	0,0339	2,97	0,9985	0,0048
1,48	0,9306	0,1334	2,23	0,9871	0,0332	2,98	0,9986	0,0047
1,49	0,9319	0,1315	2,24	0,9875	0,0325	2,99	0,9986	0,0046
					3,00	0,9987	0,0044	

Glossar

Adaptiver Algorithmus Ein adaptiver Algorithmus ist ein Regelsystem, mit dem beim adaptiven Testen die Itemauswahl zu Beginn und während des Tests geregelt sowie Kriterien der Testbeendigung spezifiziert werden.

Adaptives Testen Ein spezielles Vorgehen bei der Messung individueller Ausprägungen von Personenmerkmalen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items an der Leistungsfähigkeit der untersuchten Testpersonen orientiert, die während der Testung berechnet wird.

Adjustiertes Bayesian Information Criterion (aBIC) Das aBIC ist eine Abwandlung des Bayesian Information Criterion (BIC), bei dem der Einfluss der Stichprobe kontrolliert wird. Es zeigt verlässlichere Eigenschaften bei Modellvergleichen im Kontext von Mischverteilungsmodellen.

Akaike Information Criterion (AIC) Unter dem AIC (auch Akaike-Informationskriterium) versteht man ein Maß für die Anpassungsgüte des geschätzten Modells an die vorliegenden empirischen Daten (Stichprobe) unter Berücksichtigung der Komplexität des Modells. Daraus hervorgegangen sind das Bayesian Information Criterion (BIC), das adjustierte Bayesian Information Criterion (aBIC) und Consistent Akaike Information Criterion (CAIC).

Akquieszenz Mit Akquieszenz bezeichnet man die Antworttendenz, auf Aussagen (Statements) unabhängig vom Inhalt mit Zustimmung zu reagieren.

Austauschbare Methoden Austauschbare Methoden in MTMM-Modellen sind solche Methoden, die einer Zufallsauswahl aus einer Menge gleichberechtigter (gleichadäquater) Methoden entsprechen. Beispielsweise wären verschiedene Messgelegenheiten austauschbar, wenn sich keine der Messgelegenheiten von den anderen Messgelegenheiten strukturell unterscheidet.

Auswahllaufgaben Aufgabentyp, bei dem die Testpersonen vor die Anforderung gestellt werden, aus mehreren vorgegebenen Antwortalternativen die richtige bzw. für sie zutreffende Antwort zu identifizieren.

Auswertungsobjektivität (Gütekriterium) Ein Test gilt als auswertungsobjektiv, wenn das Testergebnis unabhängig davon ist, wer den Test auswertet.

Autokorrelationseffekt In längsschnittlichen (longitudinalen) Strukturgleichungsmodellen korrelieren Indikatoren oft stärker über die Messgelegenheiten hinweg als mit anderen Indikatoren derselben Messgelegenheit, die dasselbe Konstrukt messen. Der Autokorrelationskoeffizient quantifiziert die Stärke der Korrelation der Indikatoren über die Zeit.

Axiom Axiome sind theoretische Grundannahmen, die als geltend angesehen werden und auf denen das Theoriegebäude aufgebaut wird.

Bayesian Information Criterion (BIC) Unter dem BIC (auch Bayes-Informationskriterium) wird ein dem AIC ähnliches Kriterium der Anpassungsgüte des Modells an die Daten verstanden, das im Unterschied zum AIC die Verletzung des Gebotes der Sparsamkeit (s. Parsimonie) von Modellparametern stärker bestraft.

Bedingte Antwortmusterwahrscheinlichkeit $P(a_v | g)$ Bei der dichotomen LCA: Wahrscheinlichkeit eines Antwortmusters a_v unter der Bedingung, dass die Person v zur Klasse g gehört.

Bedingte Itembejahungswahrscheinlichkeit $P(y_{vi} = 1 | g)$ Bei der dichotomen LCA: Wahrscheinlichkeit, mit der ein Item i bejaht wird, wenn die entsprechende Person v zur Klasse g gehört.

Bedingte Kategorienwahrscheinlichkeit $P(y_{vi} = k | g)$ Bei der polytomen LCA: Wahrscheinlichkeit, mit der ein Item i mit der Antwortkategorie k beantwortet wird, wenn die entsprechende Person v zur Klasse g gehört.

Bedingte Klassenzuordnungswahrscheinlichkeit $P(g | a_v)$ Bei der dichotomen LCA: Wahrscheinlichkeit, mit der eine Person v mit dem Antwortmuster a_v zur Klasse g gehört.

Beurteilungsaufgaben Aufgabentyp, bei dem der individuelle Zustimmungs- oder Ablehnungsgrad zu einer vorgelegten Aussage (Statement) erfasst wird.

BIC s. Bayesian Information Criterion

Bifaktormodell Faktorenanalytisches Modell, bei dem alle Items auf einem Generalfaktor und Teile der Items jeweils auf einem spezifischen Faktor laden.

Birnbaum-Modell Zweiparameter-logistisches Modell (auch zweiparametrisches logistisches Modell, 2PL-Modell) mit Itemschwierigkeitsparameter β_i und Diskriminationsparameter λ_i

CAIC Consistent Akaike Information Criterion (CAIC)

Consistent Akaike Information Criterion (CAIC) Das CAIC ist eine Abwandlung des AIC, bei dem der Stichprobenumfang Berücksichtigung findet.

Cronbachs Alpha (α) Reliabilitätsmaß, dessen Berechnung essentielle τ -Äquivalenz von eindimensionalen Items voraussetzt.

Cut-off-Wert (oder Cut-off-Score) Der Cut-off-Wert ist ein Schwellenwert der Merkmalsausprägung. Bei Überschreitung des Schwellenwertes (z. B. IQ = 130) erfolgt eine Klassifikation in eine bestimmte Gruppe (z. B. „Hochbegabte“), bei Unterschreitung hingegen nicht.

Debriefing Das Debriefing beschreibt die Qualitätssicherungsmaßnahme, nach der Testung den Testleiter nach Besonderheiten während der Testung zu befragen.

Deterministische Modelle Deterministische Modelle nehmen an, dass die Wahrscheinlichkeit, ein Item zu lösen oder ihm zuzustimmen, nur 0 oder 1 betragen kann, wobei die Wahrscheinlichkeit ab einer bestimmten Schwelle auf der Merkmalsdimension η von 0 auf 1 „springt“. Die Itemcharakteristische Funktion (IC-Funktion) entspricht einer Sprungfunktion.

Diagnostik- und Testkuratorium (DTK) Neuere Bezeichnung für das Testkuratorium (TK)

DIN 33430 Die DIN 33430 ist eine verbindliche Norm von Qualitätsstandards für die berufsbezogene Eignungsbeurteilung bezüglich der verwendeten Tests und der diagnostischen Ablaufschritte.

Disjunkttheit von Antwortalternativen Disjunkttheit von Antwortalternativen liegt vor, wenn die Antwortalternativen logisch nicht gleichzeitig gültig sein können.

Diskriminante Validität Im Rahmen der Konstruktvalidierung gilt die diskriminante Validität als nachgewiesen, wenn Messungen verschiedener Konstrukte mit derselben Methode nicht oder nur gering miteinander korrelieren.

Diskriminationsindex Unter dem Diskriminationsindex versteht man einen Kennwert zur Identifizierung „nicht trennscharfer“ Items bei der LCA.

Distraktoren Als Distraktoren bezeichnet man plausibel erscheinende, aber nicht zutreffende Antwortalternativen bei Auswahlaufgaben.

Dreiparameter-logistisches Modell (auch dreiparametrisches logistisches Modell, 3PL-Modell, Rate-Modell von Birnbaum) Im 3PL-Modell wird neben dem Schwierigkeits- und dem Diskriminationsparameter des 2PL-Modells noch die Ratewahrscheinlichkeit als Parameter ρ_i in das Modell aufgenommen (Birnbaum-Modell).

Durchführungsobjektivität (Gütekriterium) Ein Test ist dann durchführungsobjektiv, wenn das Testergebnis unabhängig davon ist, von welcher Testleitung der Test durchgeführt wird.

Eichstichprobe Stichprobe, die zur Normierung eines Tests eingesetzt wird. Die Eichstichprobe besteht idealerweise aus einer hinreichend großen, repräsentativen Zufallsstichprobe der Zielpopulation, für die der Test beim späteren Einsatz Gültigkeit haben soll.

Eichung (Gütekriterium) s. Normierung

Eigenwert Der Eigenwert eines Faktors gibt an, wie viel Varianz aller Itemvariablen durch diesen Faktor erklärt wird.

Einparameter-logistisches Modell (auch einparametrisches logistisches Modell, 1PL-Modell, Rasch-Modell) Das 1PL-Modell der IRT beschreibt den Zusammenhang zwischen dem beobachtbaren dichotomen Antwortverhalten und dem dahinterstehenden latenten Merkmal auf Grundlage einer logistischen Wahrscheinlichkeitsfunktion mit einem Itemparameter, nämlich dem Schwierigkeitsparameter β_i .

Erschöpfende (suffiziente) Statistiken Die Zeilen- und Spaltensummenscores einer (0/1)-Datenmatrix werden als suffizient bezeichnet, wenn die Wahrscheinlichkeit der Daten nicht davon abhängt, welche Personen welche Items gelöst haben, sondern lediglich davon, wie viele Personen ein Item gelöst haben (Schwierigkeit des Items) bzw. wie viele Items eine Person lösen konnte (Fähigkeit der Person). Die Zeilen- und Spaltensummenscores reichen dann jeweils aus, um die Personen- und Itemparameter zu schätzen.

Essentielle τ -Äquivalenz In der KTT Bezeichnung für eindimensionale Items, wo bei die Messmodelle der Items unterschiedliche Leichtigkeitsparameter α und unterschiedliche Fehlervarianzen aufweisen dürfen; die Diskriminationsparameter λ müssen hingegen identisch sein.

Essentielle τ -Parallelität In der KTT Bezeichnung für eindimensionale Items, wo bei die Messmodelle der Items unterschiedliche Leichtigkeitsparameter α aufweisen dürfen; die Diskriminationsparameter λ und die Fehlervarianzen müssen hingegen identisch sein.

Exhaustivität von Antwortalternativen Exhaustivität von Antwortalternativen liegt vor, wenn alle möglichen Antworten auf den vorgegebenen Antwortalternativen abgebildet werden können.

Exploratorische Faktorenanalyse (EFA) Die EFA ist ein statistisches Verfahren, das auf Annahmen beruht. Es kommt typischerweise dann zur Anwendung, wenn keine Hypothesen über die Anzahl der zugrunde liegenden Faktoren und über die Zuordnung der beobachteten Variablen zu den Faktoren vorliegen. Es ist ein sog. struktursuchendes und dimensionalitätsreduzierendes Verfahren.

Exposure Control Strategie zur Vermeidung der öffentlichen Bekanntheit von Items durch unerwünscht häufige Vorgabe der Items oder der Itemgruppen. Beim adaptiven Testen kann Exposure Control leichter erzielt werden.

Fairness (Gütekriterium) Ein Test erfüllt das Gütekriterium der Fairness, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen.

Faking good/bad Antwortverhalten, mit dem die Testperson fälschlicherweise eine zu gute/schlechte Merkmalsausprägung vortäuscht.

Faktorladung Die Gewichtungszahl λ_{jk} einer beobachteten Variablen j auf dem latenten Faktor k heißt Faktorladung und beschreibt die Stärke des Zusammenhangs zwischen Faktor und Variable (meist Item). Sie kann bei orthogonal rotierten Faktoren als Korrelation interpretiert werden.

Faktorwert (Faktorscore) Der Faktorwert η_{kv} gibt an, wie stark ein Faktor η_k bei der v -ten Person ausgeprägt ist. Faktorscores werden in der KTT als Personenparameter verwendet.

Fehlervarianz $Var(\varepsilon)$ Die Varianz der Fehlerwerte $Var(\varepsilon)$ der Personen stellt in der KTT den unerklärten Anteil der Testwertevarianz $Var(Y)$ dar.

Freies Antwortformat Bei Aufgaben mit einem freien Antwortformat sind keine Antwortalternativen vorgegeben. Die Antwort wird von der Person selbst formuliert bzw. produziert.

Geschwindigkeitstests s. Speedtests

Gleichwertige Methoden Im Rahmen von MTMM-Modellen sind gleichwertige Methoden solche Methoden, die das zu erfassende Trait gleichwertig repräsentieren. Beispielsweise sind parallele Tests oder Testhälften gleichwertige Methoden. Im Unterschied zu austauschbaren Methoden ist die Erklärung der Methodeneffekte für gleichwertige Methoden nachrangig.

Gütekriterien s. Testgütekriterien, s. aber auch Informationskriterien

Halbtest Aufteilung eines Tests in zwei Testhälften, z. B. zur Reliabilitätsbestimmung, s. auch Itempaare

Hauptachsenanalyse Methode der EFA, mit der versucht wird, das Beziehungs muster zwischen den manifesten Variablen mit möglichst wenigen dahinterliegenden latenten Faktoren zu erklären.

Hauptkomponentenanalyse (PCA) Die PCA (Principal Component Analysis) ist ein mathematisches Verfahren zur Bildung von Linearkombinationen von Items mit dem Ziel, möglichst viel Varianz der Items durch eine Abfolge von – hinsichtlich ihrer Varianzstärke gereichten – Hauptkomponenten zu erklären.

Hierarchisch geschachtelte Modelle Verschiedene CFA-Modelle werden als hierarchisch geschachtelt bezeichnetet, wenn sie dieselbe Modellstruktur aufweisen und durch Parameterrestriktionen bzw. -freisetzungene ineinander übergeführt werden können, s. auch Nested Models.

Homogenität Homogenität von Items liegt vor, wenn die verschiedenen Items eines (Sub-)Tests dasselbe Merkmal messen, s. auch Itemhomogenität.

IC-Funktion Itemcharakteristische Funktion

Informationskriterien Maße zur deskriptiven, relativen Beurteilung der Güte eines Modells. Häufig verwendete Informationskriterien sind das Akaike Information Criterion (AIC), das Bayesian Information Criterion (BIC) und das Consistent Akaike Information Criterion (CAIC).

Inkrementelle Validität Inkrementelle Validität bezeichnet das Ausmaß, in dem die Vorhersage eines externen Kriteriums verbessert werden kann, wenn zusätzlich Testaufgaben oder (Sub-)Tests (und allgemeiner: Informationen) zu den bereits eingesetzten Verfahren hinzugenommen werden.

Interne Konsistenz (Konsistenzanalyse) Methode der Reliabilitätsschätzung. Die Kovarianzen zwischen den Items eines Tests werden als wahre Varianz angesehen und zur Bestimmung der Reliabilität verwendet. Siehe auch Cronbachs Alpha (α).

Interpretationsobjektivität (Gütekriterium) Ein Test ist dann interpretationsobjektiv, wenn bezüglich der Interpretation der Testwerte eindeutige Richtlinien (z. B. Normentabellen) vorliegen.

Invertierte Items Invertierte Items sind „umgepolte“ Items, bei denen nicht die Bejahung, sondern die Verneinung symptomatisch für eine hohe Merkmalsausprägung ist, s. auch Item-Wording.

Itemcharakteristische Funktion (IC-Funktion) Die IC-Funktion beschreibt den Zusammenhang zwischen dem manifesten Antwortverhalten der Testpersonen auf die Items und dem dahinterliegenden latenten Persönlichkeitsmerkmal. Die IRT ist vor allem für dichotome Itemvariablen konzipiert und geht von einem logistischen Zusammenhang aus, die KTT hingegen von einem linearen Zusammenhang mit kontinuierlichen Itemvariablen.

Itemhomogenität Verschiedene Items sind bezüglich einer latenten Dimension η dann homogen, wenn das Antwortverhalten auf die Items nur von diesem Merkmal (der latenten Dimension) und keinem anderen systematisch beeinflusst wird und die Items dem zuvor spezifizierten funktionalen Zusammenhang (d. h. dem vorgegebenen logistischen Modell) folgen.

Iteminformation Die Iteminformation I_i gibt in der IRT an, wie groß der Informationsgehalt eines Items i bezüglich der Merkmalsausprägung η einer Testperson v ist. Die Iteminformation eines Items i ist maximal, wenn die Itemschwierigkeit mit der jeweiligen Merkmalsausprägung der Testperson v auf der Joint-Scale übereinstimmt. Die Iteminformationen können zur Testinformation aufaddiert werden, mit

deren Hilfe Konfidenzintervalle für die wahre Merkmalsausprägung der Testpersonen gebildet werden können.

Itempaare (auch Itemzwillinge) Bei essentieller τ -Parallelität können aus einer Menge eindimensionaler Testitems zwei Halbtests gebildet werden, wobei die Items von Itempaaren mit gleichen Leichtigkeits- und Diskriminationsparametern den jeweiligen Halbtesthälften zugeordnet werden. Die resultierende Halbtestreliabilität kann dann mit der Spearman-Brown-Formel der Testverlängerung zur Reliabilität des Gesamttests aufgewertet werden.

Itemparcels Zusammenfassung mehrere Items zu Päckchen, z. B. zu Halbtests, s. auch Parcels

Itempool Eine Menge von Items, für die mit einem geeigneten Testmodell (z. B. Rasch-Modell) Itemhomogenität (s. auch Messäquivalenz) festgestellt wurde; beim adaptiven Testen werden die informationsstärksten Items aus dem Itempool zur Vorgabe ausgewählt.

Item-Response-Theorie (IRT) Die IRT (auch probabilistische Testtheorie) beschreibt den Zusammenhang zwischen beobachtbarem Antwortverhalten und dem dahinterstehenden Persönlichkeitsmerkmal (Personenparameter) auf Grundlage eines wahrscheinlichkeitstheoretischen Modells. Dabei wird die Wahrscheinlichkeit für das beobachtbare (gezeigte) Antwortverhalten als von der latenten Merkmalsausprägung abhängig modelliert. Siehe auch Itemcharakteristische Funktion (IC-Funktion).

Itemschwierigkeit/Schwierigkeitsindex Die Itemschwierigkeit wird in der deskriptivstatistischen Itemanalyse durch den Schwierigkeitsindex ausgedrückt. Er beschreibt das mit 100 multiplizierte Verhältnis der tatsächlich erreichten Itempunktsumme aller Testpersonen zur maximal möglichen Itempunktsumme. Je größer der Schwierigkeitsindex ist, desto leichter ist das Item.

Itemschwierigkeitsparameter Schwierigkeitsparameter β_i (IRT), Leichtigkeitsparameter α_i (KTT)

Itemselektion Die Itemselektion beschreibt den Prozess, Items hinsichtlich ihrer Eignung zur Erfassung des interessierenden Merkmals auszuwählen. Neben der Betrachtung deskriptivstatistisch gewonnener Kennwerte (z. B. Itemschwierigkeit, Itemtrennschärfe und Itemvarianz) fließen auch inhaltliche und modelltheoretische Überlegungen in den Selektionsprozess ein.

Itemtrennschärfe Die Trennschärfe eines Items gibt in der deskriptivstatistischen Itemanalyse an, wie stark die mit dem jeweiligen Item erzielte Differenzierung zwischen den Testpersonen mit der Differenzierung durch den Gesamttest übereinstimmt.

Itemvarianz Die Varianz eines Items ist ein Maß für die Differenzierungsfähigkeit des Items. Die Itemvarianz gibt an, wie unterschiedlich die Testpersonen auf das Item antworten.

Item-Wording Variation der Formulierung eines Items (Statements) durch Veränderung der Wortwahl, z. B. in positiv gepolter Form oder in „invertierter“ negativ gepolter Form zu Aufdeckung von Akquieszenz oder von Methodeneffekten.

Joint-Scale Gemeinsame Skala von Personenfähigkeit und Itemschwierigkeit in der IRT.

Klassische Testtheorie (KTT) Die KTT (auch Messfehlertheorie) beschreibt den Zusammenhang zwischen dem beobachtbaren Antwortverhalten und dem dahinterstehenden wahren Testwert τ_v bzw. der latenten Merkmalsausprägung η_v (Personenparameter) auf Grundlage der Annahme, dass sich der Messwert y_{vi} einer Person v in einem Testitem i immer aus zwei Komponenten zusammensetzt. Diese sind ein wahrer Wert τ_{vi} und ein Messfehlerwert ε_{vi} . Der Zusammenhang zwischen den Messwerten und den wahren Werten bzw. latenten Merkmalsausprägungen wird in der KTT als linear angenommen.

Kognitives Vortesten Beim kognitiven Vortesten legt die Testleitung in Erprobung befindliche Items vor und bittet die Testpersonen, alle Überlegungen, die zur Beantwortung der Frage führen, zu formulieren. Diese Äußerungen werden meist auf Video aufgenommen.

Kommunalität Die Kommunalität h_i^2 einer Variablen i gibt an, in welchem Ausmaß die Varianz der Variablen durch die extrahierten q Faktoren erklärt wird.

Konfidenzintervall Das Konfidenzintervall kennzeichnet denjenigen Bereich um einen empirisch ermittelten individuellen Testwert Y_v , in dem sich 95 % (99 %) aller möglichen wahren Testwerte τ_v befinden, die den Testwert Y_v erzeugt haben können.

Konfirmatorische Faktorenanalyse (CFA) Die konfirmatorische Faktorenanalyse (CFA) ist ein Verfahren, mit dem Hypothesen über die Zuordnung von beobachteten Variablen zu dahinterliegenden (latenten) Faktoren über die Anzahl der Faktoren sowie über die Korrelationen zwischen den Faktoren theoriegeleitet überprüft werden können. Die CFA zählt zu den Verfahren der Strukturgleichungsmodelle.

Konsistenz Die Konsistenz einer Messvariablen beschreibt in der LST-Theorie das Ausmaß der durch einen Trait erklärten Varianz relativiert an der Gesamtvarianz der Messvariablen; siehe aber auch Interne Konsistenz.

Konsistenzeffekte Konsistenzeffekte treten dann auf, wenn Testpersonen versuchen, solche Antworten zu geben, die ihnen bezüglich ihrer Antworten auf vorangegangene Items als „stimmig“ erscheinen.

Konstrukt Bezeichnung für ein nicht direkt beobachtbares, aber operationalisierbares latentes Persönlichkeitsmerkmal.

Konstruktäquivalenz Die Konstruktäquivalenz ist die empirisch bestätigte Äquivalenz eines psychologischen Konstrukts über Sprachen und Kulturen hinweg.

Konstruktvalidität Konstruktvalidität liegt vor, wenn ein Test tatsächlich das Konstrukt erfasst, das er erfassen soll, s. auch konfirmatorischen Faktorenanalyse (CFA).

Konvergente Validität Im Rahmen der Konstruktvalidierung gilt die konvergente Validität als nachgewiesen, wenn Messungen eines Konstrukts (oder verwandter Konstrukte), das mit verschiedenen Messmethoden erfasst wird, hoch miteinander korrelieren.

Kriteriumsorientierte Testwertinterpretation Bei der kriteriumsorientierten Testwertinterpretation erfolgt die Interpretation des Testwertes nicht in Bezug zur Testwertverteilung einer Bezugsgruppe (s. Normorientierte Testwertinterpretation), sondern in Bezug auf ein spezifisches inhaltliches Kriterium. Es wird vorab

festgelegt, welches Testergebnis mindestens vorliegen muss, um das Kriterium zu erreichen.

Kriteriumsvalidität Kriteriumsvalidität liegt vor, wenn von einem Testergebnis auf ein für diagnostische Entscheidungen praktisch relevantes Kriterium außerhalb der Testsituation geschlossen werden kann. Kriteriumsvalidität kann durch empirische Zusammenhänge zwischen dem Testwert und möglichen Außenkriterien belegt werden.

Latent-Class-Analyse (LCA) Probabilistisches Verfahren zur Kategorisierung von Personen (Objekten) in qualitative latente Klassen.

Latent-Class-Modelle Bezeichnung für IRT-Modelle, die davon ausgehen, dass das latente Persönlichkeitsmerkmal zur Charakterisierung von Personenunterschieden aus qualitativen kategorialen latenten Klassen besteht.

Latent-State-Trait-Theorie (LST-Theorie) Die LST-Theorie ist eine formale Erweiterung der KTT, die neben der Aufteilung der Messvariablen Y_{it} einer Messung i zu Messgelegenheit t in eine Messfehlervariable ε_{it} und in eine Variable der wahren Werte τ_{it} auch eine Trennung von situationalen und dispositionellen Einflüssen erlaubt. Dazu wird die Variable der wahren Werte τ_{it} einer Messung Y_{it} zusätzlich in eine Trait-Variable η_{it} und in eine State-Residuum-Variable ζ_{it} zerlegt: $Y_{it} = \tau_{it} + \varepsilon_{it} = \eta_{it} + \zeta_{it} + \varepsilon_{it}$.

Latent-Trait-Modelle Bezeichnung für IRT-Modelle, die davon ausgehen, dass es sich bei dem latenten Persönlichkeitsmerkmal zur Charakterisierung von Personenunterschieden um eine quantitative kontinuierliche latente Dimension handelt.

Latente Dimension Nicht direkt beobachtbare Variable (auch Faktor, Konstrukt, Trait) zur Erfassung von Merkmalsausprägungen in Leistungs-, Einstellungs- oder Persönlichkeitsmerkmalen, von denen das manifeste Verhalten als abhängig angesehen wird.

Latentes State-Residuum Das State-Residuum ist der Teil eines States, der ausschließlich die Situation und die Interaktion zwischen Person und Situation repräsentiert.

Leichtigkeitsparameter In den Messmodellen der KTT wird der Leichtigkeitsparameter eines Items mit α_i (Interzept der linearen IC-Funktion) bezeichnet. Je höher α , desto einfacher ist das Item zu lösen/bejahren (vgl. Schwierigkeitsparameter der IRT).

Leistungstests Tests zur Erfassung der individuellen kognitiven Leistungsfähigkeit in Problemlösesituationen. Beispiele: Intelligenztests, Konzentrationstests etc.

Likelihood/IRT In der IRT ist die Likelihood das Anpassungskriterium bei der Parameterschätzung. Sie ist dort definiert als die Wahrscheinlichkeit aller beobachteten Daten in Abhängigkeit der gewählten Modellparameter und unter Annahme der Modellgültigkeit. Bei der Parameterschätzung werden die Parameter iterativ so lange verändert, bis die Likelihood maximal ist.

Likelihood/LCA In der LCA ist die Likelihood das Anpassungskriterium bei der Parameterschätzung. Es ist dort definiert als das Produkt der unbedingten Antwortmusterwahrscheinlichkeiten $P(a_v)$ über alle Antwortmuster in der Stichprobe (N_a) hinweg.

Likelihood-Ratio-Test (LRT) Möglichkeit zur inferenzstatistischen Absicherung der Güte von IRT-Modellen. Der LRT wird zur inferenzstatistischen Absicherung des Unterschieds zweier geschachtelter Modelle (Nested Models) verwendet.

Linear-logistische Modelle Linear-logistische Modelle zerlegen die Schwierigkeitsparameter der Items in für die Bearbeitung des Items erforderliche Basisoperationen. Jeder der Schwierigkeitsparameter wird als Linearkombination einer möglichst geringen Anzahl von Basisparametern ausgedrückt.

Lizenzprüfung nach DIN 33430 Nachweis einschlägiger Kenntnisse für den diagnostischen Prozess von Auftragnehmern (Lizenz A), bzw. Mitwirkenden an Verhaltensbeobachtungen (Lizenz MV) und von Mitwirkenden an Eignungsinterviews (Lizenz ME) gemäß den Anforderungen der DIN 33430.

Lokale stochastische Unabhängigkeit Bedingung, die erfüllt sein muss, um die Korrelation zwischen zwei Testitems auf eine dahinterliegende latente Persönlichkeitsvariable zurückführen zu können. Die lokale stochastische Unabhängigkeit liegt dann vor, wenn die Korrelation zwischen den Items verschwindet, wenn man sie auf den einzelnen („lokalen“) Stufen des latenten Persönlichkeitsmerkmals untersucht.

LST-Theorie s. Latent-State-Trait-Theorie

Manifeste Variablen Variablen zur Erfassung des beobachtbaren Antwortverhaltens mit verschiedenen Items, die Indikatoren für die latenten Dimensionen darstellen.

McDonalds Omega Reliabilitätsmaß, dessen Berechnung τ -Kongenerität von eindimensionalen Items voraussetzt.

Messäquivalenz In der KTT Oberbegriff für verschiedene strenge Formen von Parallelität eindimensionaler Testitems: τ -Kongenerität, essentielle τ -Äquivalenz, essentielle τ -Parallelität.

Messeffizienz Die Effizienz eines Tests berechnet sich als Quotienten aus Messpräzision und Testlänge, wobei Letztere häufig durch die Anzahl der präsentierten Items quantifiziert wird.

Messmodell Im Rahmen von Strukturgleichungsmodellen werden die Teilmodelle, in denen die Zuordnungen der beobachteten Variablen zu den latenten Variablen (Faktoren) erfolgt, als Messmodelle bezeichnet. In der KTT erfordern verschiedene Messmodelle unterschiedliche Reliabilitätsmaße.

Messpräzision Grad der Übereinstimmung von wahren Merkmalsausprägungen und den Testwerten. Auf Skalenebene oft durch die mittlere quadratische Abweichung von wahrer und geschätzter Merkmalsausprägung bestimmt.

Methodeneffekte Ein Sammelbegriff für verschiedene systematische Varianzquellen bei der MTMM-Analyse, die sich über den Trait hinausgehend auf die Validität der Messung auswirken können. Hierbei handelt es sich vor allem um Charakteristika der eingesetzten Messinstrumente, der Beurteiler oder der Situationen, in der eine Messung erfolgt.

Methodenspezifitätskoeffizient Der Methodenspezifitätskoeffizient gibt den Anteil an beobachteter Varianz wieder, der auf den Einfluss eines Methodeneffekts

zurückzuführen ist. Je höher der Methodenspezifitätskoeffizient ausfällt, desto stärker ist der Einfluss der Messmethode auf die Messung.

Mischverteilungs-Rasch-Modelle (Mixed-Rasch Models) Kombination aus Rasch-Modell und LCA. Innerhalb jeder latenten Klasse wird versucht, jeweils ein eigenes Rasch-Modell anzupassen. Zwischen den latenten Klassen unterscheiden sich die Parameter des Rasch-Modells.

Mixed-Rasch Models Mischverteilungs-Rasch-Modelle

Modelldifferenztest Werden mit der CFA hierarchisch geschachtelte Modelle spezifiziert und gegeneinander getestet, so kann der Unterschied im Modellfit statistisch über die Differenz der χ^2 -Werte beider Modelle überprüft werden, die wiederum χ^2 -verteilt ist.

Modellfit Der Modellfit bezeichnet in der Statistik ganz allgemein die Güte der Passung zwischen Modell und Daten. Je ungünstiger der zur Beurteilung der Passung gewählte Index (z. B. χ^2 -Wert, BIC etc.) ausfällt, desto schlechter ist die Passung.

Multidimensionales adaptives Testen Eine spezielle Form des adaptiven Testens, bei der mehrere latente Dimensionen als ursächlich für das beobachtete Antwortverhalten angesehen werden; aus den Antworten wird simultan auf mehrere latente Merkmale geschlossen.

Multiple Regression Mittels einer multiplen Regression werden die Ausprägungen einer manifesten Kriteriumsvariablen bestmöglich auf die Ausprägungen mehrerer manifesten Prädiktorvariablen zurückgeführt.

MTMM-Analyse Die Multitrait-Multimethod-Analyse ist ein Verfahren zum Nachweis der Konstruktvalidität unter Berücksichtigung einer systematischen Kombination von mehreren Traits und mehreren Messmethoden.

Nested Models Hierunter versteht man hierarchisch geschachtelte Modelle, die durch Parameterrestriktionen ineinander überführbar sind.

Niveautests Powertests

Nomologisches Netz Ein nomologisches Netz stellt ein Beziehungsgeflecht zwischen (latenten) Konstrukten und beobachtbaren Testvariablen dar. Die beiden Ebenen werden mit theoretischen Annahmen bzw. empirischen Evidenzen beschrieben und durch Korrespondenzregeln miteinander verbunden.

Norm(en)aktualisierung Unter Norm(en)aktualisierung versteht man eine erneute Testeichung, sobald die empirische Überprüfung der Gültigkeit von Normen ergeben hat, dass sich die Merkmalsverteilung in der Bezugsgruppe seit der vorherigen Testeichung bedeutsam verändert hat.

Normalisierung Bei der Normalisierung wird eine nicht normalverteilte Testwertvariable zur besseren Interpretierbarkeit so transformiert, dass die Variable danach normalverteilt ist. Die Normalisierung ist von der Normierung zu unterscheiden, die bei der Testeichung vorgenommen wird.

Norm(en)differenzierung Unter Norm(en)differenzierung versteht man die Bildung von separaten Normen für einzelne Subpopulationen aus der Eichstichprobe hinsichtlich eines mit dem Untersuchungsmerkmal korrelierten Hintergrundfaktors (z. B. separate Normen für Männer und Frauen).

Normentabelle s. Normierung

Normierung, auch Testeichung (Gütekriterium) Die Normierung dient dazu, Vergleichswerte zur normorientierten Testwertinterpretation zu gewinnen. Dazu werden Testergebnisse von Personen einer Eichstichprobe in Norm(en)tabellen zusammengestellt.

Normorientierte Testwertinterpretation Bei der normorientierten Testwertinterpretation wird der Testwert (d. h. die individuelle Merkmalsausprägung einer Testperson) mit den Normwerten einer Bezugsgruppe (Eichstichprobe) verglichen, um die relative Position der Testperson innerhalb der Bezugsgruppe zu beurteilen.

Normwert Ein Normwert (z. B. Prozentrang, z_v -Wert) ermöglicht es, den Testwert Y_v einer Testperson hinsichtlich seiner Position in der Testwertverteilung einer bestimmten Bezugsgruppe zu interpretieren.

Nützlichkeit (Gütekriterium) Ein Test ist dann nützlich, wenn die auf seiner Grundlage getroffenen Entscheidungen (Maßnahmen) mehr Nutzen als Schaden erwarten lassen.

Objektivität (Gütekriterium) Ein Test ist dann objektiv, wenn das Testergebnis unabhängig davon ist, wer den Test durchführt, auswertet und interpretiert.

Omega-Koeffizient McDonalds Omega (ω) oder Bollens Omega (ω^*)

Ordnungsaufgabe Aufgabentyp, bei dem die einzelnen Bestandteile der Aufgabe so umgeordnet oder einander zugeordnet werden, dass idealerweise eine logisch passende Ordnung entsteht.

Parallele Tests Messäquivalenz

Paralleltest-Reliabilität Methode der Reliabilitätsschätzung. Die Reliabilität eines Tests, von dem zwei parallele Formen existieren, wird über die Korrelation der Testwerte der beiden parallelen Testformen geschätzt.

Parcels s. Itemparcels

Parsimonitätsprinzip Wissenschaftliches Prinzip, demzufolge „sparsamere“ Modelle mit wenigen Parametern bei gleicher Qualität gegenüber aufwendigeren Modellen bevorzugt werden sollten.

Personenparameter Der Personenparameter kennzeichnet in der IRT die Merkmalsausprägung η_v einer Person v auf der latenten Variable η . In der KTT können Faktorscores als Personenparameter verwendet werden.

Persönlichkeitsmerkmale Persönlichkeitsmerkmale sind mehr oder weniger zeitlich stabile psychische und physische Eigenschaften von Testpersonen (z. B. Extraversion, Körpergröße).

Persönlichkeitstests Persönlichkeitstests dienen der Erfassung von individuell typischem Verhalten als Indikator für die Ausprägung von Persönlichkeitsmerkmalen (Verhaltens- oder Erlebensdispositionen).

Perzentil Das Perzentil bezeichnet jenen Testwert Y_v , der einem bestimmten Prozentrang in der Normierungsstichprobe entspricht. Beispielsweise wird derjenige

Testwert, der von 30 % der Testpersonen unterschritten bzw. höchstens erreicht wird, als 30. Perzentil bezeichnet.

Powertests, auch Niveautests Powertests sind Leistungstests mit eher schwierigen Aufgaben, wobei erhoben wird, welches Schwierigkeitsniveau der Aufgaben die Testperson ohne Zeitbegrenzung bewältigen kann.

Probabilistische Modelle Im Unterschied zu deterministischen Modellen gehen probabilistische Modelle davon aus, dass bei dichotomen Items die Wahrscheinlichkeit, ein Item zu lösen bzw. ihm zuzustimmen, in Abhängigkeit von der latenten Merkmalsausprägung nicht von 0 auf 1 springt, sondern jeden Wert zwischen 0 und 1 annehmen kann. In der IRT wird die Antwortwahrscheinlichkeit durch eine monoton steigende, meist logistische IC-Funktion modelliert.

Projektive Tests Bei projektiven Tests kommt mehrdeutiges Stimulusmaterial (meist Bilder) zum Einsatz. Es wird angenommen, dass Testpersonen unbewusste oder verdrängte Bewusstseinsinhalte in das Bildmaterial hineinprojizieren und dadurch Persönlichkeitsmerkmale ermittelt werden können. Die erforderlichen Gütekriterien werden durch projektive Tests häufig nicht erfüllt.

Prozentrang Ein Prozentrang gibt an, wie viel Prozent der Bezugsgruppe bzw. Normierungsstichprobe einen Testwert erzielt haben, der niedriger oder maximal ebenso hoch ist wie der Testwert Y_v der Testperson v .

Quartil Als erstes, zweites bzw. drittes Quartil (Q1, Q2, Q3) werden diejenigen Testwerte Y_v bezeichnet, die von 25 %, 50 % bzw. 75 % der Testpersonen unterschritten bzw. höchstens erreicht werden (vgl. Perzentil).

Rasch-Modelle Rasch-Modelle stellen eine Klasse von spezifisch objektiven Modellen in der IRT dar. Einparameter-logistisches Modell (1PL-Modell).

Ratekorrektur Die Ratekorrektur zieht bei der Testwertbestimmung jene Anzahl an „richtigen“ Lösungen ab, die nur durch zufälliges Raten der richtigen Antworten entstanden ist.

Rate-Modell von Birnbaum Dreiparameter-logistisches Modell (3PL-Modell)

Ratingsskala Beurteilungsskala mit mehr als zwei (zumeist 3–7) Antwortabstufen.

Receiver-Operating-Characteristics-Analyse ROC-Analyse

Reliabilität (Gütekriterium) Reliabilität bezeichnet die Messgenauigkeit eines Tests. Ein Testverfahren ist perfekt reliabel, wenn die damit erhaltenen Testwerte frei von zufälligen Messfehlern sind. Je größer die Einflüsse der Messfehler sind, desto weniger reliabel ist das Testverfahren.

Reliabilitätskoeffizient/KTT Konkrete Bezeichnung für die Messgenauigkeit eines Tests (Reliabilität). In der KTT wird der Reliabilitätskoeffizient (Rel) als das Verhältnis zwischen True-Score-Varianz $Var(\tau)$ und Testwertevarianz $Var(Y)$ definiert.

Repräsentative Aufgabenstichprobe Eine repräsentative Aufgabenstichprobe stimmt hinsichtlich der Schwierigkeitsverteilung mit der Grundgesamtheit aller merkmalsrelevanten Aufgaben überein und erlaubt somit eine kriteriumsorientierte Testwertinterpretation in Bezug auf die Aufgabeninhalte.

Repräsentativität Eine Stichprobe ist dann repräsentativ, wenn sie hinsichtlich ihrer Zusammensetzung die jeweilige Zielpopulation möglichst genau abbildet.

Retest-Reliabilität Methode der Reliabilitätsschätzung. Ein Test wird zu zwei Messzeitpunkten der gleichen Stichprobe vorgegeben. Die Korrelation der zu beiden Messzeitpunkten gemessenen essentiell τ -parallelen Testwertvariablen dient als Maß der Reliabilität des Tests.

Retrospektive Befragung In der Testentwicklungsphase wird die Testperson „rückblickend“ über Probleme bei der Beantwortung der einzelnen Items befragt.

ROC-Analyse Die ROC-Analyse (Receiver-Operating-Characteristics-Analyse) ermöglicht für eine binäre Klassifikation (z. B. gefährdet vs. nicht gefährdet) den zur Fallunterscheidung verwendeten Schwellenwert optimal in der Weise festzulegen, dass die Trefferquote und die Quote korrekter Ablehnungen maximiert werden.

Schwellenwert (Cut-off-Score) Im Rahmen der kriteriumsorientierten Testwertinterpretation bezeichnet ein Schwellenwert jenen Testwert, ab dem das Kriterium als erreicht angenommen wird. Schwellenwerte können z. B. mittels ROC-Analyse empirisch bestimmt werden.

Schwierigkeitsparameter/IRT Der Schwierigkeitsparameter β_i ist in der IRT ein Itemparameter, der durch jene Merkmalsausprägung η definiert ist, bei der die Lösungswahrscheinlichkeit des Items 50 % beträgt. Je höher β , desto schwieriger ist das Item; vgl. Leichtigkeitsparameter α_i in der KTT.

Sensitivität/ROC-Analyse Die Sensitivität (Trefferquote) in der ROC-Analyse ist das Verhältnis von „richtig positiv“ (RP) klassifizierten Merkmalsträgern zu der Summe von „falsch negativ“ (FN) und „richtig positiv“ (RP) klassifizierten Merkmalsträgern. Sie bezeichnet damit die Wahrscheinlichkeit, dass ein Fall, der ein Kriterium erfüllt, auch entsprechend als positiv klassifiziert wird.

Sicherung Unter Sicherung versteht man die Pflicht zur Regelung der Verfügbarkeit, Aufbewahrungsdauer und Verwendung von Testdaten (inklusive des Testprotokolls und aller schriftlichen Belege) und Schutz der Identität von Testpersonen.

Skalierung (Gütekriterium) Ein Test erfüllt das Gütekriterium Skalierung, wenn die laut Verrechnungsregel resultierenden Testwerte die empirischen Merkmalsrelationen adäquat abbilden.

Soziale Erwünschtheit, auch soziale Desirabilität Die Soziale Erwünschtheit beinhaltet die Antworttendenz einer Testperson, sich selbst so darzustellen, wie es soziale Normen ihrer Wahrnehmung nach erfordern.

Spearman-Brown-Formel der Testverlängerung Reliabilitätsmaß, dessen Berechnung essentielle τ -Parallelität von eindimensionalen Items voraussetzt.

Speedtest, auch Geschwindigkeitstest Speedtests sind Leistungstests mit meist einfachen Aufgaben, wobei erhoben wird, wie viele der Aufgaben unter Zeitdruck gelöst werden können.

Spezifische Objektivität/IRT Spezifische Objektivität liegt vor, wenn alle IC-Funktionen die gleiche Form aufweisen, d.h. lediglich entlang der η -Achse parallel verschoben sind. Ist dies der Fall, kann der Schwierigkeitsunterschied zweier Items ($\beta_j - \beta_i$) unabhängig davon festgestellt werden, ob Personen mit niedrigen oder hohen Merkmalsausprägungen η untersucht wurden. Umgekehrt kann auch der Fä-

higkeitsunterschied zweier Personen ($\eta_w - \eta_v$) unabhängig von den verwendeten Items festgestellt werden.

Spezifität/LST-Theorie Die Spezifität einer Messvariablen beschreibt in der LST-Theorie das Ausmaß der durch die Situation und die Person-Situation-Interaktion erklärten Varianz relativiert an der Gesamtvarianz der Messvariablen.

Spezifität/ROC-Analyse Die Spezifität (Quote korrekter Ablehnungen) in der ROC-Analyse ist das Verhältnis von „richtig negativ“ (RN) klassifizierten Merkmalsträgern zu der Summe von „falsch positiv“ (FP) und „richtig negativ“ (RN) klassifizierten Merkmalsträgern. Sie bezeichnet damit die Wahrscheinlichkeit, dass ein Fall, der ein Kriterium nicht erfüllt, auch entsprechend als negativ klassifiziert wird.

Split-Half-Reliabilität (Testhalbierungs-Reliabilität) Methode der Reliabilitäts schätzung unter bestimmten Voraussetzungen (Messäquivalenz). Aus den Items eines Tests werden zwei parallele Testhälften gebildet (s. Itempaare). Aus der Korrelation der Testwerte der Halbtests wird mittels Spearman-Brown-Formel der Testverlängerung die Reliabilität des Gesamttests geschätzt.

Standardabweichung $SD(Y)$ Die Standardabweichung ist ein Streuungsmaß der Testwertvariablen Y um den Mittelwert \bar{Y} an. Die Standardabweichung wird als Wurzel aus der Testwertvarianz $Var(Y)$ gewonnen. Ist die Testwertvariable normalverteilt, so befinden sich im Bereich $\bar{Y} \pm 1SD(Y)$ ca. 68 % der Testwerte, im Bereich $\bar{Y} \pm 2SD(Y)$ ca. 95 % der Testwerte.

Standardmessfehler $SD(\varepsilon)$ Der Standardmessfehler $SD(\varepsilon)$ eines Tests resultiert aus der Unreliabilität des Tests und errechnet sich als Wurzel aus der Fehlervarianz der Testwertvariablen. Dabei gilt: $SD(\varepsilon) = SD(Y) \cdot \sqrt{1 - Rel}$. Der Standardmessfehler ist bei höherer Reliabilität kleiner und bei niedrigerer Reliabilität größer.

Standardnormen Als Standardnormen werden die z -Norm sowie weitere durch Lineartransformationen gewonnene Normen (z. B. IQ- oder T-Norm) bezeichnet.

State Ein State ist ein zeitlich begrenzter biologischer, emotionaler und kognitiver Zustand, in dem sich eine Person befindet. Er kennzeichnet sich durch personenbedingte (d. h. traitbedingte), situativ bedingte und durch die Interaktion zwischen Person und Situation bedingte Einflüsse.

Stichprobenunabhängigkeit Stichprobenunabhängigkeit bedeutet, dass in Rasch-Modellen die Itemparameter unabhängig von den Personen und die Personenparameter unabhängig von den Items geschätzt werden können.

Strukturell unterschiedliche Methoden Als strukturell unterschiedlich werden Methoden dann bezeichnet, wenn sie nicht austauschbar sind, weil sie sich qualitativ von einander unterscheiden und keine Zufallsauswahl darstellen. Strukturell unterschiedliche Methoden sind z. B. Selbst- und Fremdbeurteilungen.

Suffiziente Statistik s. erschöpfende (suffiziente) Statistik

τ -Kongenerität In der KTT ist die τ -Kongenerität eine Bezeichnung für eindimensionale Items, wobei die Messmodelle der Items unterschiedliche Leichtigkeitsparameter α , unterschiedliche Diskriminationsparameter λ sowie unterschiedliche Fehlervarianzen aufweisen dürfen (s. auch Messäquivalenz).

TBS-TK Das TBS-TK ist ein veröffentlichtes Testbeurteilungssystem des Testkuratoriums (TK) zur standardisierten Erstellung und Publikation von Testrezensionen anhand eines vorgegebenen Kriterienkatalogs; s. auch Testkuratorium.

Tendenz zur Mitte Als Tendenz zur Mitte wird eine Antworttendenz bezeichnet, bei der extreme Antworten eher vermieden und mittlere Antwortkategorien eher bevorzugt werden.

Testadaptation Testadaptation bezeichnet den Prozess einer qualitativ hochwertigen Übertragung (Übersetzung unter Berücksichtigung von Konstruktäquivalenz) und empirischen Evaluation psychologischer Tests aus anderen Sprachen und in andere Sprachen unter Beachtung kultureller Unterschiede.

Testeichung Die Testeichung dient dazu, Normwerte zur normorientierten Testwertinterpretation zu gewinnen. Dazu wird der Test an Personen einer Normierungsstichprobe durchgeführt, die hinsichtlich einer definierten Bezugsgruppe repräsentativ ist.

Testgütekriterien/Gütekriterien Testgütekriterien stellen ein System zur Qualitätsbeurteilung psychologischer Tests dar. Üblicherweise werden folgende zehn Kriterien unterschieden: Objektivität, Reliabilität, Validität, Skalierung, Normierung (Eichung), Testökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und Fairness.

Testitem Zu beantwortende/beurteilende Aufgabenstellung (Frage, Statement etc.) eines Tests.

Testkuratorium (TK)/Diagnostik- und Testkuratorium (DTK) Das Testkuratorium (TK) ist ein Gremium der Föderation Deutscher Psychologievereinigungen (Deutsche Gesellschaft für Psychologie [DGPs] e. V. und Berufsverband Deutscher Psychologinnen und Psychologen [BDP] e. V.), dessen Aufgabe es ist, die Öffentlichkeit vor unzureichenden diagnostischen Verfahren und vor der unqualifizierten Anwendung diagnostischer Verfahren zu schützen. Seit Sommer 2011 lautet die Bezeichnung „Diagnostik- und Testkuratorium (DTK)“.

Testnormen s. Normierung

Testökonomie (Gütekriterium) Ein Test erfüllt das Gütekriterium Ökonomie, wenn er – gemessen am diagnostischen Erkenntnisgewinn – relativ wenig Ressourcen wie Zeit, Geld o. Ä. beansprucht.

Teststandards Teststandards sind vereinheitlichte Leitlinien, in denen sich allgemein anerkannte Zielsetzungen zur Entwicklung, Adaptation, Anwendung und Qualitätsbeurteilung/Validierung psychologischer und pädagogischer Tests widerspiegeln.

Testwert Der Testwert (= Rohwert) Y_v ist das individuelle numerische Testresultat und wird aus den registrierten Antworten einer Testperson durch Anwendung definierter Regeln gebildet (vgl. aber Personenparameter).

Testwertestreuung SD (Y) Die Testwertestreuung der Testwertverteilung sagt aus, wie breit die empirisch gewonnenen Testwerte einer Stichprobe um den Mittelwert der Testwerte verteilt sind. Die Streuung der Testwerte wird meist als Standardabweichung $SD (Y)$ angegeben; man gewinnt sie als Wurzel aus der Testwertevarianz $Var (Y)$.

Testwertevarianz $Var(Y)$ Die Testwertevarianz $Var(Y)$ ist die Varianz der beobachteten Testwerte. In der KTT setzt sie sich aus der wahren Varianz $Var(T)$ und der Fehlervarianz $Var(E)$ zusammen.

Trait Ein Trait ist ein zeitlich stabiles Merkmal (Disposition), das personeninhärent und transsituativ überdauernd ist.

Trait-Methoden-Einheit In der MTMM-Analyse wird angenommen, dass in jeder Messung Einflüsse des zu messenden Konstrukts und der verwendeten Messmethode zu finden sind. Messungen eines Traits repräsentieren somit eine Trait-Methoden-Einheit.

Treffsicherheit Index zur Beurteilung der Güte eines LCA-Modells. Definiert als die durchschnittliche Höhe der maximalen bedingten Klassenzuordnungswahrscheinlichkeit $P^{\max}(g|a_v)$ über alle in der Stichprobe vorkommenden Antwortmuster (N_a) hinweg.

Trennschärfe Itemtrennschärfe

True-Score τ_v Der True-Score bzw. wahre Wert τ_v ist die wahre Ausprägung der Testperson v in dem von einem Test gemessenen Merkmal. Da Messungen in der Regel fehlerbehaftet sind, stimmen Testwert Y_v und wahrer Wert τ_v nicht völlig überein. Ein Konfidenzintervall für τ_v kann mithilfe des Standardmessfehlers bestimmt werden.

Unbedingte Antwortmusterwahrscheinlichkeit $P(a_v)$ Bei der dichotomen LCA: Wahrscheinlichkeit eines Antwortmusters a_v in der Stichprobe.

Unbedingte Itembejahungswahrscheinlichkeit $P(y_{vi} = 1)$ Bei der dichotomen LCA: Wahrscheinlichkeit, mit der ein Item i bejaht wird.

Unbedingte Kategorienwahrscheinlichkeit $P(y_{vi} = k)$ Bei der polytomen LCA: Wahrscheinlichkeit, mit der ein Item i mit der Antwortkategorie k beantwortet wird.

Unbedingte Klassenzuordnungswahrscheinlichkeit $P(g)$ Bei der dichotomen LCA: Wahrscheinlichkeit, mit der eine beliebige Person v zur Klasse g gehört (auch relative Klassengröße π_g).

Unverfälschbarkeit (Gütekriterium) Unverfälschbarkeit eines Tests liegt vor, wenn das Verfahren derart konstruiert ist, dass die zu testende Person durch vorgetäusches Verhalten (s. Faking good/bad) die konkreten Ausprägungen ihrer Testwerte nicht steuern bzw. verzerrnen kann.

Validität (Gütekriterium) Ein Test gilt dann als valide („gültig“), wenn er das Merkmal, das er messen soll, auch wirklich misst – und nicht irgendein anderes. Validität bezeichnet darüber hinaus die Gültigkeit einer Menge zutreffender Schlussfolgerungen, die aus einem Testergebnis gezogen werden können.

Visuelle Analogskala Eine visuelle Analogskala ist eine kontinuierliche Skala ohne konkrete Skalenstufen; meist ist sie eine Linie, auf der lediglich die Anfangs- und Endpunkte als extreme Zustände markiert sind (z. B. keine Schmerzen vs. unerträgliche Schmerzen). Die Testperson kann durch eine Markierung auf der Linie seine Merkmalsausprägung (aktueller Schmerz) angeben.

Wahre Varianz Die wahre Varianz $Var(T)$ ist die Varianz der wahren Werte τ_v in einem Test. Sie ist meistens kleiner als die Testwertevervarianz $Var(Y)$. Aus dem Verhältnis beider Varianzanteile resultiert in der KTT die Reliabilität.

Youden-Index In der ROC-Analyse wird der Youden-Index als Wert definiert, der sich aus der Berechnung $Sensitivität + Spezifität - 1$ ergibt. Der Youden-Index dient der Schwellenwertbestimmung. Der Schwellenwert ist dann optimal, wenn der Youden-Index maximal groß ist. Dann gelingt die Trennung der zu klassifizierenden Fälle am besten.

Zielpopulation Die im Rahmen der Testeichung zu definierende Zielpopulation ist diejenige Bezugsgruppe, für welche die zu erstellenden Testnormen gelten sollen und aus der entsprechend die Eichstichprobe zu ziehen ist.

Zumutbarkeit (Gütekriterium) Zumutbarkeit liegt vor, wenn ein Test absolut sowie relativ zu dem aus seiner Anwendung resultierenden Nutzen die zu testende Person in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet.

Zweiparameter-logistisches Modell (auch zweiparametrisches logistisches Modell, 2PL-Modell, Birnbaum-Modell) Im Unterschied zum 1PL-Modell wird beim 2PL-Modell ein zusätzlicher Itemparameter λ_i ins Modell aufgenommen, der die Diskriminierungsfähigkeit des Items (ähnlich der Itemtrennschärfe) repräsentiert.

z_v -Normwert Der z_v -Normwert gibt an, wie stark der Testwert Y_v einer Testperson v vom Mittelwert \bar{Y} der Verteilung der Normierungsstichprobe (Bezugsgruppe) in Einheiten der Standardabweichung $SD(Y)$ abweicht.

Stichwortverzeichnis

A

Abbruchkriterium 515, 590
 Aberrant response patterns 398
 Ablaufsteuerung 132
 Ablehnungstendenz 83
 Abruf 80
 Acceleration Model 435
 Adaptation 203
 Adaptives Testen 24, 53, 131, 133, 264, 270, 381, 396, 406, 413, 502, 516
 – computerisiertes 132, 331, 413
 – multidimensionales 520
 Ad-hoc-Stichprobe 191
 Akquieszenz 83, 564
 Aktivitätsauswahl 133
 Algorithmus, adaptiver 503
 Alpha, Cronbachs (α) 29, 260, 290, 301, 314, 332, 338, 340, 342, 345
 – Berechnung 314
 – Eindimensionalität 320
 – Fehlerkovarianz 319
 – Formel 341
 – Probleme 318
 – Voraussetzungen 317
 – Voraussetzungstestung 318
 Alternativhypothese 540
 American Educational Research Association (AERA) 199, 219
 American Psychological Association (APA) 199, 219
 Analogska 105
 – kontinuierliche 106
 – visuelle 106
 Analyse
 – korrelationsbasierte 669
 – längsschnittlicher konvergenter und diskriminanter Validität 720
 – latente semantische 130
 Analysepotential 135
 Analysestichprobe 61
 Andrich's reliability 494
 Angoff-Verfahren 236
 Animation 128
 Ankereffekt 85
 Ankeritem 412
 Antwortabgabe 80
 Antwortbewertung 129
 Antworthandlung 129
 Antwortmuster 550
 – auffälliges 398
 – empirisch beobachtetes 551
 – maximal mögliches 551
 Antwortmusterwahrscheinlichkeit 454, 485
 – bedingte 552
 – unbedingte 552
 Antwortrichtung 111
 Antwortskala
 – bipolare 106
 – unipolare 106
 Antwortstil 81
 Antworttendenz 81, 570
 Antwortvariable
 – dichotome 372
 – kontinuierliche 443
 Antwortwahl 80
 A-posteriori-Verteilung 467, 468

B

A-priori-Aufgabenmerkmal 418, 420
 A-priori-Schätzung 509
 A-priori-Verteilung 467
 Äquivalenzhypothese 698
 Arithmetischer Mittelwert 160
 Assessment 121
 – ambulantes 134
 – computerbasiertes 121
 – summatives 136
 Assessmentdesign 225
 Assessment-Triade 225
 Assessmentzyklus 124
 Assimilationseffekt 85
 Audiomaterial 128
 Aufgabeninhalt 74
 Aufgabenmerkmal 420
 Aufgabentyp 93
 Aufgabenvorlage 133
 Auftraggeber 199
 Auftragnehmer 199
 Augenscheininvalidität 30, 32, 531
 Auswahlaufgabe 99
 Auswahldiagnostik 540
 Autokorrelation 475, 717
 Autokorrelationseffekt 717
 Automated Test Assembly 131
 Axiom 254, 279, 532

 Basisparameter, linear kombinierter 405
 Bayesian-Frequentist Debate 467
 Bayes-Inferenz 467
 Bayes-Modal-Schätzung, marginale 470
 Bayes-Schätzer 393, 510
 Bayes-Statistik 450, 462, 466
 Bayes-Theorem 462, 554
 Bayes'scher Personenparameterschätzer 487
 Bayes'sches Schätzverfahren 392, 466, 610
 – nicht simulationsbasiertes 470
 – simulationsbasiertes 474
 Bearbeitungsstil, unangemessener 398
 Bearbeitungszeit 136
 Befragung, retrospektive 59
 Behavior coding 59
 Beurteiler
 – Austauschbarkeit 665
 – strukturelle Unterschiedlichkeit 665
 Beurteilungsaufgabe 105
 Beurteilungsskala
 – asymmetrische 111
 – diskrete 105
 – kontinuierliche 105
 Bewertungsmodell 129
 Bias
 – beurteilerspezifischer 665
 – kontextspezifischer 665
 – messmethodenspezifischer 665
 Bifaktormodell 260, 350, 355, 357, 442, 639, 643
 Big-Data-Assessment 122
 Bildungsmonitoring 222
 Biquartimin 600

Birnbaum-Modell (2PL-Modell) 266, 267, 399, 414, 550
 – Modellgleichung 400
 – zweidimensionales 439
 Blueprint 130
 Bollens Omega 338, 343, 347, 348, 360
 – Formel 341
 Bookmark-Methode 236
 Branching 132
 Bundesdatenschutzgesetz 243
 Burn-in-Periode 482

C

Category response function 433
 Caution-Index 398
 CFA-MTMM-Modell 672, 678
 Chi-Quadrat-Differenztest (χ^2 -Differenztest) 652
 Chi-Quadrat-Test (χ^2 -Test) 258, 288, 398, 559, 648
 Chi-Quadrat-Wert (χ^2 -Wert) 648
 CML-Schätzverfahren 390, 391, 456
 Cognitive pretesting 59
 Comparative Fit Index (CFI) 649
 Comprehension 80
 Computer-Aided Instruction 137
 Computerisiertes adaptives Testen 502, 520
 Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) 503
 Conditional Likelihood-Ratio-Test 397
 Conditional Maximum Likelihood (CML) 390, 456
 Constraint-Management-Methode 514
 Constructed Response 125
 Constructive Alignment 225, 241
 Content-Management 515
 Contrasting-Groups-Methode 236
 Correlated-Trait-Correlated-(Method-minus-1)-Modell (CTC(M - 1)-Modell) 678, 679, 696, 716
 Correlated-Trait-Correlated-Method-Modell (CTCM-Modell) 673, 715
 Correlated-Trait-Uncorrelated-Method-Modell (CTUM-Modell) 673, 715
 Correlated-Uniqueness-Ansatz (CU-Ansatz) 697
 Cronbachs Alpha (α) 29, 260, 290, 301, 314, 332, 338, 340, 342, 345
 – Berechnung 314
 – Eindimensionalität 320
 – Fehlervarianz 319
 – Formel 341
 – Probleme 318
 – Voraussetzungen 317
 – Voraussetzungstestung 318
 Culture-Fair-Test 26
 Cumulative Score Category Response Function (CSCRF) 433
 Curation-Lifecycle-Modell 243
 Cut-off-Wert (Cut-off-Score) 199, 210, 235, 607, 649

D

Daten, personenbezogene 243
 Datenarchiv 243
 Datenmanagement 242
 Datenmanagementplan 242
 Datenmatrix 145
 Datenschutz 243
 Datensicherheit 243
 Debriefing 59
 Delivery 133
 Demografische Angaben 56
 Denken, lautes 59
 DESI-Studie 421

Determinationskoeffizient 282
 Diagnostik- und Testkuratorium (DTK) der Föderation Deutscher Psychologievereinigungen 199
 Dichotome Aufgabe 102
 Differential Item Functioning 397
 Differenz, kritische 297
 Differenzmodell 267
 DIFFTEST 610
 Diffusion Model 443
 Dimensionalität 233
 DIN 33430 190, 198
 Direct Quartimin 600
 Disjunkttheit 101
 Diskriminationsindex 562
 Diskriminationsparameter 255, 262, 265, 284, 293, 312, 399, 511
 Dispositionismus 689
 Distraktor 97, 99
 Dokumentation 36
 Domäne 221
 Domänenanalyse 227
 Domänenmodellierung 228
 Dominanzansatz 111
 Durchführungsfairness 26
 Durchführungsobjektivität 18

E

Echtdatensimulation 521
 Eichstichprobe 62
 Eigenvektor 587
 Eigenwert 582, 587
 Eignungsbeurteilung 199
 Eindeutigkeit 76
 Eindimensionalität 255, 311, 339, 340, 634
 Ein-Facetten-Design 300
 Einfachstruktur 578, 596, 635
 Einzeltestung 54
 Eisbrecheritem 509
 Elbow-Kriterium 591
 Enemies 131
 Entscheidung
 – absolute 301
 – relative 301
 Entscheidungsstudie 299
 Ergänzungsaufgabe 95
 Erinnerungseffekt 327
 Erwartungswert 625
 Erwünschtheit, soziale 82
 Ethikkommission 245
 Ethikrichtlinie 220, 244
 Evaluationsstudie 60
 Evidenz
 – diskriminante 540
 – konvergente 540
 Evidenzakkumulation 134
 Evidenzidentifikation 134
 Exhaustivität 101
 Expectation-Maximization-Algorithmus (EM-Algorithmus) 464
 Expected Score 429
 Expected-a-posteriori-Schätzer (EAP-Schätzer) 393, 488, 495
 Experience-Sampling-Methode 134
 Explanatory Item Response Model 422
 Exploratorische Faktorenanalyse (EFA) 34, 552
 – Ablaufschritte 578
 – Faktoreninterpretation 605
 – Itemauswahl 606

Stichwortverzeichnis

- Modellannahmen 580
- Modellevaluation 604
- Exploratory Structural Equation Modeling 608
- Exposure Control 515
- Externes Kriterium 180
- Exzess 162

- F**
- Facette 299
- Factor-Score 298
- Fairness 25, 239
- Fairnessanalyse 239
- Faking bad 44, 571
- Faking good 44, 571
- Faktorenanalyse
 - exploratorische (EFA) 34, 552, 578, 580, 604–606
 - konfirmatorische (CFA) 35, 617, 619, 634, 643, 648, 650, 651, 672
- Faktorextraktion 585
- Faktoreindeterminiertheit 595
- Faktorenrotation 595
- Faktorladung 34, 284, 293, 579, 623
 - standardisierte 623
- Faktorladungsmatrix 580
- Faktormodell 578
 - höherer Ordnung 360, 636
 - mit korrelierten Faktoren 360
- Faktorscore 606
- Faktorskalierung 625
- Faktorwert 298, 606
- Falsifikationsprinzip 540
- Feedback 223, 241
- Fehlerkovarianz 634
- Fehlervariable 254
- Fehlervarianz 312
 - absolute 301
 - relative 301
- Fehlervarianzquelle 299
- Filtering 132
- Filterregel 132
- Fixierungsrestriktion 565
- Flynn-Effekt 193
- Forschungsethik 244
- Frage
 - direkte 73
 - indirekte 73
 - offene 95
- Frageform
 - abstrakte 74
 - biografiebezogene 73
 - depersonalisierte 74
 - emotionalisierende 74
 - emotionsneutrale 74
 - hypothetische 73
 - konkrete 74
 - personalisierte 74
- Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II) 26, 44, 129, 518
- Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2) 19, 44, 103
- Freiheitsgrade (df) 289, 589
- Fremdeinschätzung 54
- Fremdtäuschung 82
- Fundamentaltheorem 578

- G**
- Gegenwahrscheinlichkeit 374
- Geltungsbereich 50
- Generalfaktor 350, 351, 637, 639
- Generalisierbarkeitskoeffizient 301
- Generalisierbarkeitsstudie 299
- Generalisierbarkeitstheorie 299
- Generalized Partial-Credit-Modell (GPCM) 266, 267, 426
 - Parametrisierung 427
- Geomin-Rotation 601
- Gesamttestwert 635
- Geschwindigkeitstest 44
- Gesetz, empirisches 532
- Gibbs-Sampler 477
- Gleichheitsrestriktion 565
- Graded-Response-Modell (GRM) 266, 267, 432
- Grenzwert 540
- Grundannahme 536
- Grundfunktion, symmetrische 459
- Gruppentestung 54
- Gütekriterium 17, 530, 648
 - allgemeines 17
 - spezielles testtheoriebasiertes 17

- H**
- Halbtest-Korrelation 324
- Handeln, evidenzbasiertes 221
- Harris-Kaiser-Rotation 599
- Hauptachsenanalyse 588
- Hauptgütekriterium 17
- Hauptkomponente 586
- Hauptkomponentenanalyse 586
- Hesse-Matrix 390, 453
- Heteromethod-Block 668
- Heterotrait-Heteromethod-Koeffizient 668
- Heterotrait-Monomethod-Koeffizient 668
- High-Stakes-Testung 224
- Hitrate 561
- Homogenität 702
- Hotspotaufgabe 125
- Hypothese, konkurrierende 537
- Hypothetisch-deduktiver Ansatz 532

- I**
- Idealnorm 173
- Idealpunktansatz 111
- Implizite Assoziationstests 49
- Impression Management 82
- Indikator 539
- Indikatorvariable
 - kategoriale 647
 - kontinuierliche 644, 645
- Inferenz 536
- Inhaltsvalidität 31, 32, 200, 531
- Instruktion 55
- Integrationsverfahren, numerisches 463
- Intelligentes Tutorielles System 137
- Intelligenzforschung 222
- Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R) 212
- Interaktionsgrad 127
- International Test Commission (ITC) 199
- International Test Commission Guidelines on Test Use (ITC-G-TU) 205
- Interpretationsobjektivität 21
- Interquartilabstand 161
- Interzept 284, 294, 312, 622
- Invarianz

- konfigurale 698
- schwache 698
- strikte 698
- Invarianzbedingung 698
- IRT-Modell 202, 266, 364, 412, 414, 426
- eindimensionales 265, 372
- mehrdimensionales 267
- multidimensionales (mIRT) 438
- Ising-Modell 444
- Item 42, 69
 - invertiertes 150
 - kalibriertes 413
 - multidimensionales 43
 - Rasch-homogenes 377
 - unidimensionales 43
- Item Category Response Function (ICRF) 427
- Itemanalyse 153, 379
- Itemanzahl 50, 330
- Itemauswahl 503, 510
 - adaptive 509
 - eingeschränkt adaptive 513
 - voll adaptive 510
- Itembank 130
- Itembearbeitung 79
- Itemcharakteristik 255
- Itemcharakteristische Funktion (IC-Funktion) 202, 261, 373
 - logistische 373, 415
 - Steigung 374, 394
 - Wendepunkt 374
- Itemformulierung 75
- Itemgenerierung 57, 71, 201
 - automatische 133
- Itemheterogenität 330
- Itemhomogenität 156, 330, 381, 569
- Itemkalibrierung 504
- Itemkategorienparameter 427
- Itemkovarianz 286
- Item-Parcel 364, 695
- Itempolung 84
- Itempool 59, 450, 504, 507
 - optimaler 508
- Itemreihenfolge 56, 85
- Itemreliabilität 281, 293, 312, 623
- Item-Response-Modell, erklärendes 422
- Item-Response-Theorie (IRT) 186, 252, 260, 504
 - generalisierte lineare 270
 - Grundüberlegung 371
 - Reliabilität 364
- Itemschwierigkeit 70, 146, 155, 623
- Itemselektion 60, 155, 396
- Itemsensitivität 400
- Itemtrennschärfe 153, 156, 623
 - part-whole-korrigierte 154
 - unkorrigierte 154
- Itemuniversum 32
- Itemvariable 308
- Itemvarianz 151, 152, 155, 285
- Itemwert 145
- Item-Wording 84
- Itemzwillinge 29

J

JML-Schätzverfahren 387, 453
 Joint Maximum Likelihood (JML) 387, 453
 Joint Scale 262, 374, 376, 413, 505

Judgment 80

K

Kaiser-Guttman-Kriterium 590
 Kalibrierungsstudie 507
 Kategorienantwortwahrscheinlichkeit 435

- kumulative 435, 441

 Kategoriencharakteristik 404
 Kategorienparameter, invariante 430
 Kategorienwahlmodellierung 426
 Kategorienwahrscheinlichkeit 426, 430

- bedingte 567
- unbedingte 428

 Klassenmodell, latentes 361
 Klassenzuordnungswahrscheinlichkeit, bedingte 552
 Klassische Testtheorie (KTT) 252, 254, 309, 620, 691

- empirisches Beispiel 291
- Grundannahmen 277
- Grundgleichung 279

 Kodierschlüssel 134
 Kommunalität 583
 Kommunalitätenproblem 588
 Kompetenz 240
 Kompetenzdiagnostik 268
 Kompetenzniveau 235, 237, 417
 Kompetenzniveaudefinition 419
 Kompetenzniveauschwelle 417
 Kompetenzskala 235
 Kompetenztest 542
 Komplexität 125
 Komplexitätsmaß 596
 Konfidenzintervall 291, 295, 338, 394, 395

- asymmetrisches 291, 348, 349, 708

 Konfidenzintervallsschätzung 331
 Konfirmatorische Faktorenanalyse (CFA) 35, 634, 672

- Einsatzgebiete 617
- Messmodell 619
- Modellevaluation 648
- Modellparameterbeurteilung 650
- Modellstrukturmodifikation 651
- Parameterschätzung 643

 Konkordanzkoeffizient 20
 Konsistenz 703

- interne 201, 319

 Konsistenzeffekt 85
 Konsistenzkoeffizient 694
 Konsistenzkontroverse 689
 Konstrukt, latentes 578
 Konstruktäquivalenz 203
 Konstruktdefinition 230
 Konstruktionsphase 57
 Konstruktionsstrategie

- faktorenanalytische 72
- intuitive 71
- kriteriumsorientierte 71
- rationale 71

 Konstruktrepräsentation 124
 Konstruktunterrepräsentation 537
 Konstruktvalidierung 35
 Konstruktvalidität 31, 33, 200, 531, 532, 663, 669
 Kontrasteffekt 85
 Kontrollskala 83
 Konvergenz 557
 Konvergenzdiagnostik 481
 Korrelation 270

Stichwortverzeichnis

- polychorische 484
- tetrachorische 484
- Korrelationsmatrix 607
 - empirische 581
 - modellimplizierte 581
 - reduzierte 588
- Korrespondenzregel 532
- Kovarianzmatrix 607
 - modellimplizierte 624
 - Kovarianzzerlegung 286
 - Kredibilitätsellipsoid 520
 - Kredibilitätsintervall 468
 - Kriterium, externes 180
 - Kriteriumsvalidität 31, 32, 200, 531
- KTT-Modell 256
 - eindimensionales 259
 - multidimensionales 260
- Kurzaufsatzaufgabe 94

- L**
- Längsschnittmodell 268
- Large-Scale-Assessment 222, 412
- Latent-Class-Analyse (LCA) 35, 550
 - exploratorische Anwendung 562
 - konfirmatorische Anwendung 564
 - Modellgleichung 552
 - Modellvergleichstest 566
 - Parameterschätzung 556
- Latent-Class-Modell 261
- Latent-Curve-Modell 736
- Latent-Difference-Modell 683
- Latent-State-Trait-Theorie (LST-Theorie) 35, 260, 690, 692
 - Reliabilität 694
- Latent-Trait-Modell 35, 261, 372
 - polytomous 403
- Law School Admission Test (LSAT) 372
- Layout 57
- Leichtigkeitsparameter 255, 257, 284, 294, 312, 622
- Leistungsdiagramm 44
- Leistungssimulation 44
- Leistungstest 44
- Lernen
 - formelles 223
 - informelles 223
- Lerngegenstand 240
- Lernverhaltensverbesserung 242
- Lernziel 185
- Lernzielerreichung 542
- Likelihood 388, 557, 560
 - der Datenmatrix 382
 - standardisierte 557
- Likelihood-Funktion 387, 451
 - bedingte 459
 - Maximum 390
- Likelihood-Quotienten-Test, bedingter 264, 397, 566
- Likert-Skala 105
- Linear Integer Programming 131
- Linearkombination 588
- Linear-logistisches Modell 405
 - mit relaxierten Annahmen (LLRA) 406
- Linear-logistisches Testmodell (LLTM) 405, 422
- Logarithmierung 162
- Logfile 135
- Logit 380
- Logit-Schreibweise 380
- Logit-Skala 377, 413
- Logit-Wert 380, 413
- Log-Likelihood-Funktion 452, 589
- Lokale stochastische Unabhängigkeit 261, 269, 381, 384, 454, 553
- Lösungswahrscheinlichkeit 372, 374
- Low-Stakes-Testung 224
- LST-Modell 695, 718, 719
- Lückentext 95
- Lügenskala 83

- M**
- Marginal Maximum Likelihood (MML) 391, 461
- Markieraufgabe 125
- Markov-Chain-Monte-Carlo-Verfahren (MCMC-Verfahren) 474
- Markov-Kern 475
- Markov-Kette
 - aperiodische 475
 - ergodische 475
 - irreduzible 475
- Materialbearbeitungstest 98
- Matrix-Sampling 412
- Maximum-a-posteriori-Schätzer (MAP-Schätzer) 393, 487, 495
- Maximum-Likelihood-Faktorenanalyse (ML-EFA) 589
- Maximum-Likelihood-Methode (ML-Methode) 288, 390, 450, 557, 644, 706
- McDonalds Omega 260, 290, 293, 332, 338, 342, 346
 - Formel 341, 343
- Median 160, 174
- Medienverwendung 128
- Mehrdimensionalität 340
- Mehrebenenmodell 361
- Mehrfachwahlaufgabe 104
- Merkmalsausprägung 294
- Merkmalsbreite 50
- Merkmalsdefinition 41
 - operationale 42
- Merkmalsindikator 42
- Merkmalskonfundierung 78
- Merkmalsstabilität 327, 328
- Merkmalsvariabilität 330
- Messäquivalenz 283, 288, 312, 629, 633, 697
 - τ -Äquivalenz 698
 - essentielle τ -Äquivalenz 256, 259, 284, 290, 338, 342, 632
 - essentielle τ -Parallelität 256, 258, 285, 290, 324, 632, 705
 - τ -Kongenerität 256, 284, 290, 338, 342, 629, 698
 - τ -Parallelität 698
- Messäquivalenzmodell 290
- Messfehler 278
 - systematischer 278, 665
 - unsystematischer 278
- Messfehlertheorie 254, 277
- Messfehlervariable 279
- Messgenauigkeit 703
- Messinvarianz 259
 - strikte 324
- Messinvarianztestung 653
- Messmodell 309, 617
- Messmodellgleichung 283, 619
- Messpräzision 517
- Methode
 - austauschbare 716
 - gleichwertige 717
 - strukturell unterschiedliche 717
- Method-Effect-Modell 683
- Methodeneffekt 319, 340, 664, 665, 695, 696, 702, 716

- trait-spezifischer 718
- Methodenfaktor 673
- orthogonaler 695
- Methodenspezifität 299, 696
- Metropolis-Hastings-Algorithmus (MH-Algorithmus) 476
- mIRT-Modell 438
 - kompensatorisches 439
 - Mischverteilungsmodell 405, 444, 567
 - Mittelwertevektor
 - empirischer 625
 - modellimplizierter 625
 - Mitwirkender 199
 - Mixed-Rasch-Modell 404, 569, 570
 - Mixture-IRT-Modell 444
 - ML-Personenparameterschätzer 494
 - ML-Schätzer 451, 452, 509, 706
 - ML-Schätzung 450
 - Genauigkeit 452
 - gewichtete 486
 - Prinzip 450
 - ML-Scoring 485
 - MML-Schätzverfahren 401, 461
 - Modalwert 160
 - Modell
 - eindimensionales 339
 - geschachteltes 652
 - höherer Ordnung 642
 - korrelierter Faktoren 360, 635, 642
 - linear-logistisches 405
 - mehrdimensionales 350, 634
 - motivationales 79
 - nicht geschachteltes 653
 - psychometrisches 504
 - Modelldifferenztest 594
 - Modellevaluation 648
 - Modellfit 258, 341, 605, 648
 - Modellgüte 258, 552, 706
 - Modellgütekriterium 589
 - Modellidentifikation 626
 - Modellierung
 - schrittweise 436
 - von Antwortzeiten 443
 - Modellkonformität 256, 258, 264, 265, 396
 - globale 398
 - Modellkontrolle, empirische 396
 - Modellparameter 619
 - nicht separierbarer 267
 - separierbarer 265
 - Modellparameterseparierbarkeit 265
 - Modellpassung 396
 - Modellrestriktion 627
 - Modellstrukturmodifikation 651
 - Modelltest 258, 264, 288
 - grafischer 264, 396
 - Modellvergleich 652
 - Modellvergleichstest 566
 - Modifikationsindex 611, 651
 - Moduseffekt 122
 - Monomethod-Block 668
 - Monotonie 373
 - Monotrait-Heteromethod-Koeffizient 668
 - Monotrait-Monomethod-Koeffizient 668
 - Monte-Carlo-Prinzip 474
 - Motivation zur Testbearbeitung 519
 - MTMM-Modell 715, 719
 - längsschnittliches 719, 721
 - Modellgleichungen 727
 - Varianzdekomposition 727
 - Multiconstruct-LST-Modell 722, 727, 730
 - Multidimensional Difficulty (MDIFF) 439
 - Multidimensional Discrimination (MDISC) 439
 - Multidimensionales Generalized Partial-Credit-Modell (mGPCM) 440
 - Multidimensionales Graded-Rating-Scale-Modell (mGRSM) 442
 - Multidimensionales Graded-Response-Modell (mGRM) 441
 - Multilevel-CFA-Modell 683
 - Multimethod-LST-Modell 724, 728, 733
 - Multioccasion-Correlated-States-Correlated-(Method-minus-1)-Modell (Multioccasion-CSC(M – 1)-Modell) 721
 - Multioccasion-MTMM-Modell 721, 727
 - Multiple-Choice-Aufgabe 104
 - Multiplikationstheorem für unabhängige Ereignisse 382
 - Multistage-Test 131
 - Multistate-Modell 699, 704
 - Multistate-Multitrait-Modell 707
 - mit indikatorsspezifischen Trait-Faktoren 701, 706
 - Multistate-Singletrait-Modell 700, 705
 - Multitrait-Multiinformant-Analyse 666
 - Multitrait-Multimethod-Analyse (MTMM-Analyse) 35, 260, 663, 666, 672
 - Multitrait-Multimethod-Matrix (MTMM-Matrix) 667, 669
 - konfirmatorische Faktorenanalyse 672
 - Multitrait-Multimethod-Modell (MTMM-Modell) 715
 - Multitrait-Multioccasion-Analyse 666
 - Mustermatrix 599

N

- National Council on Measurement in Education (NCME) 199, 219
- Navigation 131
- Nebengütekriterium 17
- Nested model 566
- Netzwerkmodell 444
- Newton-Raphson-Algorithmus 455
- Nichtstandardmethode 716
- Niveautest 45, 148
- Nominal-Response-Modell 434
- Nomologisches Netz 532
- Nonnormed Fit Index (NNFI) 649
- Norm, verteilungsunabhängige 175
- Normaktualisierung 193
- Normalisierung 164, 178
 - der Testwerte 164
- Normalverteilung 163, 589
- Normalverteilungsannahme 594
- Normdifferenzierung 188
- Normengültigkeit 193
- Normentabelle 62
- Normierung 22, 164, 379
- Normierungsstichprobe 190, 192
- Normierungstechnik 22
- Normtabelle 174
- Normwert 173
- Nullhypothese, globale 618
- Nützlichkeit 24

O

- Objective Function 131
- Objektivität 17, 309
 - spezifische 255, 259, 261, 262, 380
- Oblimin-Rotation 599
- Oblique Rotation 598
- Odds 413

Stichwortverzeichnis

offene Frage 95
 Ökonomie 23
 Omega 707
 – hierarchisch 338, 339, 352, 353, 356, 360
 – spezifisch 338, 339, 352, 353, 356, 357, 360
 – total 338, 339, 352, 353, 356, 360
 Omega-Koeffizient 351, 353, 360
 – Beurteilung 364
 Omega-Subskala
 – hierarchisch 338, 339, 353, 354
 – spezifisch 338, 339, 353, 354
 – total 338, 339, 353, 354
 One-Item-one-Screen 132
 Operationalisierung 58
 Optimalitätskriterium 510
 Optimizing 79
 Optimizing-Satisficing-Modell 79
 Ordnungsaufgabe 97
 Ordnungsrelation 565
 Orthogonale Rotation 597
 Overadjustment 189

P

Paper-Pencil-Test 53
 Paradaten 136
 Parallelanalyse 592
 Paralleltest 328
 Paralleltest-Reliabilität 29, 322
 – Probleme 328
 Parameter
 – inzidenteller 456
 – struktureller 456
 Parameternormierung 379, 427
 Parameterrestriktion 564
 Parameterschätzung 387, 401, 470, 643
 Parameterseparierbarkeit 390, 428
 Parcel 364, 695, 704
 Parsimonieprinzip 283, 560, 652
 Partial-Credit-Modell (PCM) 265, 266, 404, 428
 – lineares 406
 Perfektionismus 344, 355
 Person in einer Situation 692
 Personenbezogene Daten 243
 Personenparameter 263, 267, 377
 Personenparameterschätzung 484, 509
 Personenselektion 398
 Personenseparierbarkeit 264
 Personenwert, latenter 258, 298
 Person-Fit-Index 398
 Persönlichkeitsfragebogen 47
 Persönlichkeitstest 47, 149
 – objektiver 47
 Perzentil 174
 Pfaddiagramm 619
 Phi-Koeffizient 372
 Pick any out of n 104
 Pilotstudie 60
 PISA-Studie 86, 122, 203, 222, 406, 413, 490, 583
 Plausible Values (PVs) 489
 Polarität der Antwortskala 106
 Post-hoc-Analyse 418
 Postkorbaufgabe 98
 Potential-Scale-Reduction-Faktor 482
 Powertest 45
 Präsentation 133

Primacy Effect 86
 Priming-Effekt 85
 Principal Component Analysis (PCA) 586
 Principal Factor Analysis (PFA) 588
 Processing Function 433
 Programme for International Student Assessment (PISA) 86, 122, 203, 222, 406, 413, 490, 583
 Promax-Rotation 599
 Prompting 134
 Prozentrang 174
 Prozentrangnorm 174
 Prozess
 – bedingter 436
 – kognitiver 433
 Prozessdaten 135
 Prüfungsangst 704
 Prüfungsformvariation 240
 Pseudo-Rateparameter 511

Q

Q-Diffusion Model 443
 Quadraturformel 463
 Qualitätsbeurteilung 198, 210
 Quartile 174
 Quartimax-Rotation 598
 Quotenstichprobe 191

R

Randomisierung 85
 Range 161
 Rasch-Homogenität 265, 373
 Rasch-Modell (1PL-Modell) 262, 264–268, 373, 414, 428, 550, 567
 – ordinale 404
 – sequentielles 437
 Ratekorrektur 149
 Rate-Modell nach Birnbaum (3PL-Modell) 401, 414
 Rateparameter 401
 Rating-Scale-Modell (RSM) 265, 266, 403, 430
 – lineares 406
 Ratingskala 105, 106
 – kombinierte 108
 – numerische 107
 – optische 108
 – verbale 107
 Realnorm 173
 Receiver-Operating-Characteristics-Analyse (ROC-Analyse) 182, 184
 Receiver-Operating-Characteristics-Kurve (ROC-Kurve) 183
 Recency Effect 86
 Regressionsmethode 298
 Reihenfolgeeffekt 86
 Reliabilität 27, 201, 234, 254, 257, 264, 277, 281, 293, 307, 691, 694, 702, 707, 730
 – Daumenregel 330
 – Definition 282, 308
 – Einflussfaktoren 330
 – Konsistenz 675, 694, 707
 – marginale 494, 495
 – Messeigenschaften der Itemvariablen 312
 – Messmodelle 309
 – Methodenspezifität 675
 – Schätzmethode 308
 – Spearman-Brown-Formel der Testverlängerung 260
 – Spezifität 694, 708
 – Vergleichbarkeit unterschiedlicher Maße 329
 Reliabilitätsdiagonale 668

- Reliabilitätskoeffizient 28, 282, 289, 339, 694
 – Formel 342
 Reliabilitätsmaß 331
 Reliabilitätsschätzung
 – klassische 337
 – modellbasierte 338
 – ordinalskalierter Variablen 363
 – Probleme 362
 – Vorteile 361
 Repräsentativität 190
 – spezifische 190
 Residualfaktor 351
 Residualmatrix 604
 Response Reporting 80
 Response Selection 80
 Response Set 81, 570
 Response Style 81
 Response-Bias 81
 Restriktion 562
 Retest-Intervall 328
 Retest-Reliabilität 28, 322
 – Probleme 327
 Retrieval 80
 Revision der Latent-State-Trait-Theorie (LST-R-Theorie) 690
 Robuste Maximum-Likelihood-Methode (RML-Methode) 289
 Rohwert 172
 Root Mean Square Error of Approximation (RMSEA) 649
 Rotation
 – oblique 598
 – orthogonale 597
 Routing-Test 514
- S**
- Sampling-Algorithmus 476
 Satisficing 79
 Scaling Correction 609
 Schätzverfahren
 – iteratives 557
 – robustes 609
 Schiefe 161
 Schlüsselwort 95
 Schlüsselwortergänzungsaufgabe 95
 Schwellenparameter 265
 Schwellenwert 180, 234, 542
 Schwierigkeit, multidimensionale 439
 Schwierigkeitsanalyse 146
 Schwierigkeitsindex 146, 379
 Schwierigkeitsparameter 147, 379
 Score Category Response Function (SCRF) 432
 Score-Funktion 390, 430, 452, 462
 Screening, diagnostisches 541
 Screening-Test 331
 Scree-Plot 591
 Selbsteinschätzung 54
 Selbsttäuschung 82
 Self-Assessment 136
 Self-deceptive Enhancement 82
 Semiprojektives Verfahren 49
 Sensitivität 180
 Sequential Model 436
 Sequenzierung 132
 Shadow Testing 131
 Shadow-Testing-Methode 515
 Shrinkage-Effekt 495
 Situation 692
- Situationismus 689
 Situationsspezifität 299
 Skalenpunkt 107
 Skalenstufe 106
 Skalierung 20
 Sokrates-Effekt 735
 Sortieraufgabe 125
 Soziale Erwünschtheit 27
 Spannweite 161
 Spearman-Brown-Formel der Testverlängerung 290
 Speedtest 44, 148
 Spezifische Objektivität 255, 259, 261, 262, 380
 Spezifität 180, 583, 703
 Spezifitätskoeffizient 695
 Split-Half-Reliabilität 29, 322, 324
 – Probleme 328
 Sprachverarbeitung, natürliche 130
 Sprungregel 132
 Stabilität 730
 Stakeholder 224
 Standard
 – ethischer 244
 – forschungsethischer 244
 – für pädagogisches Testen 238
 – kriteriumsorientierter 187
 – normorientierter 187
 Standardfehler 291
 Standardfehlerschätzung 390
 Standardized Root Mean Square Residual (SRMR) 649
 Standardmessfehler 295
 Standardmethode 716
 Standardnorm 178
 Standards for Educational and Psychological Testing (SEPT) 198, 219, 538
 Standards für pädagogisches Testen 220
 Standardsetting 235
 Standardwert 177
 Stanine-Norm 178
 Startwert 557
 State 43, 688
 State-Angst 689, 690
 State-Residuum-Variable, latente 693
 State-Trait Anxiety Inventory (STAII) 690
 State-Trait-Ärgerausdrucks-Inventar (STAXI) 43, 571
 State-Variablen, latente 693, 700
 State-Wert, latenter 693
 Statistik, suffiziente 390
 Stichprobe
 – anfallende 191
 – geschichtete 191
 – mathematische 451
 Stichprobengröße 507
 Stichprobenumfang 192
 Stichprobenunabhängigkeit 264, 265, 396
 Störvariable 81
 – potenzielle 518
 Structural Equation Modeling 617
 Strukturgleichungsmodell 540, 617
 – exploratorisches 608
 – für dichotome und ordinale Variablen 483
 Strukturkoeffizient 637
 Strukturmatrix 599
 Strukturmodell 617
 Strukturmodellgleichung 637
 Stufenantwortaufgabe 106
 Subskalenwert 635

Stichwortverzeichnis

Subtraktionseffekt 85

Suffizienz 457

Suggestion 78

Summennormierung 379

Summenscore 257, 263, 390

Symbolskala 108

Symmetrische Grundfunktion 459

T

Tailored Testing 413, 510

Target-Rotation 600

Tau-Äquivalenz (τ -Äquivalenz) 698

– essentielle 256, 259, 284, 290, 338, 342, 632

Tau-Kongeneritt (τ -Kongeneritt) 256, 284, 290, 338, 342, 629, 698Tau-Parallelitt (τ -Parallelitt) 698, 705

– essentielle 256, 258, 285, 290, 324, 632

Taxonomie 230

Template 133

Tendenz

– zum extremen Urteil 84, 109

– zur Mitte 84, 109, 110

Test 16

– computeradministrierter 53

– computerbasierter 53, 208

– computerisierter 53

– computerisierter adaptiver 413, 502

– fest verzweigter 513

– Item-by-Item-adaptiver 131

– mehrstufiger verzweigter 514

– multidimensionaler 43

– psychometrischer 42

– unidimensionaler 43

Test Accommodation 128

Test Anxiety Inventory 704

Testadaptation 198

Testadministration 53, 132

– computerbasierte 516

Testanwendung 198, 206

Teststart 44

Testaufbau 55

Testaufgabenbearbeitungsprozess 233

Testauslieferung 133

Testauswertung 209

Testbeurteilung 211

Testbeurteilungssystem des Testkuratoriums (TBS-TK) 210

Testdokumentation 202

Testdurchfhrung 209

Testeichung 189

Testeinsatz 228

Testen

– adaptives 24, 53, 131–133, 264, 270, 331, 381, 396, 406, 413, 502, 516, 520

– mageschneidertes 510

– summatives 223

Testentwicklung 201

Testerleichterung 128

Testfairness 209

Testform 130, 131, 240

– adaptive 518

– kollaborierende 241

– nicht adaptive 518

– parallele 412

Testgutekriterium 17

Testhalbierungs-Reliabilitt 29, 324

Testimplementierung 228

Testinhalt 232

Testkonstruktion 198

Testkonstruktionsrahmen 228

Testlnge 51

Testlet 70, 513

Testlet-Effekt 86

Testmanual 18, 36, 202

Testmerkmal 41

– multidimensionales 43

– qualitatives 42

– unidimensionales 43

Testnorm 202

Testrevision 163

Testrezensionen 210

Testroutine 26

Testspezifikation 130

Teststandard 198, 220

Test-Test-Korrelation 322

– Probleme 327

– Voraussetzungen 324

– Voraussetzungstestung 326

Testvorbereitung 206

Testwert 153, 172

– manifester 294

Testwertbeschreibung 418

Testwertbestimmung 156

Testwertermittlung, vorlufige 153

Testwertinterpretation 234, 530, 534, 536

– kriteriumsorientierte 173, 179, 413, 414

– normorientierte 173, 413

– spezifische 535

Testwertnormierung 234

Testwertvariable 257, 280, 293, 308

Testwertvarianz 161, 287

Testwertverteilung 160, 163

– leptokurtische 163

– linksschiefe 162, 163

– mesokurtische 163

– platykurtische 163

– rechtsschiefe 162, 163

Testzeit 52

Testzusammenstellung 130

Think aloud 59

Traceplot 481

Trait 43, 688

Trait-Angst 689, 690, 698

Trait-Faktor 673

– indikatorspezifischer 696, 702

Trait-Methoden-Einheit 664, 679, 695, 722, 726

Trait-Variabile

– indikatorspezifische 702

– latente 693

Transparenz 240

Treﬀsicherheit 561

Trends in International Mathematics and Science Study (TIMSS) 203, 406, 490

Trennschrfe 561

Trennschrfeindex 154

True-Score 278, 691

True-Score-Variable 254, 279

True-Score-Varianz 257, 293, 313

Tutoring, computerbasiertes 137

U

Ubereinstimmungsvaliditt 33, 531

- Übertragungseffekt 328
 Umordnungsaufgabe 98
 Unabhängigkeit
 – korrelative 270
 – lokale stochastische 261, 269, 381, 384, 454, 553
 Unconditional Maximum Likelihood (UML) 393
 Uncorrelated-Trait-Correlated-Method-Modell (UTCM-Modell) 673
 Uncorrelated-Trait-Uncorrelated-Method-Modell
 (UTUM-Modell) 673
 Uneindeutigkeit von Testwerten 173
 Unique Maximum Condition 433
 Urteil 80
- Vergleichsmaßstab 173
 Verhaltenskodierung 59
 Verständlichkeit, sprachliche 75
 Verständlichkeitsprüfung 59
 Verständnis 80
 Verteilung, intraindividuelle 691
 Verteilungseigenschaft 192
 Verteilungsfunktion der Standardnormalverteilung 166
 Videomaterial 128
 Vorhersagevalidität 33
 Vortesten, kognitives 59

V

- Validierung 535
 Validierungsansatz, argumentationsbasierter 535
 Validierungsprozess 539
 Validität 30, 200, 229, 254, 309, 530
 – diskriminante 34, 531, 663, 669, 715
 – faktorielle 531
 – inkrementelle 531
 – konvergente 34, 531, 663, 669, 715
 – prognostische 531
 Validitätsargument 535, 538
 Validitätsdiagonale 668
 Validitätsevidenz 229, 232
 Validitätsprüfung 61
 Validitätsstandard 229
 Variabilität 730
 Variable
 – geordnete kategoriale 436
 – latente 621
 – latente diskrete 444
 Varianz, konstruktirrelevante 537
 Varianzerlegung 581
 Varianzfunktion, bedingte 492
 Varianzkomponente 728
 Varianzzerlegung 282, 285, 300, 582, 707
 Varimax-Rotation 598
 Verbundwahrscheinlichkeit 381
 Verfahren
 – personenzentriertes 236
 – semiprojektives 49
 – testzentriertes 236
 Vergleich, spezifisch objektiver 259

W

- Wahrer Wert 278, 295, 309, 692
 Wahrscheinlichkeitsbegriff
 – objektiver 467
 – subjektiver 466
 Wahrscheinlichkeitsfunktion 371
 Wald-Test 264, 397
 Weighted Maximum Likelihood (WML) 393
 Weiß-nicht-Kategorie 110
 Wettquotient 380, 413
 Wiedergabetreue 126
 Wissenschaftsethik 244

X

- Ξ -Äquivalenz (ξ -Äquivalenz) 705

Y

- Youden-Index 183

Z

- Zielgruppe 50, 69
 Zielpopulation 190
 z -Tabelle 166
 Zufallsstichprobe 190, 191
 Zumutbarkeit 25
 Zuordnungsaufgabe 97, 125
 Zustand, variabler 720
 Zustimmungstendenz 83
 z_v -Normwert 177