

Einführung in die Forschungsmethoden der Psychologie und Psychotherapie

Einheit 2: Messen in der Psychologie

16.04.2024 | Dr. Caroline Zygar-Hoffmann

Heutige Themen

Unterteilung, Auswahl und Einsatz von psychologischen Erhebungsmethoden

Gütekriterien

Reaktivität

Take-Aways

Unterteilung, Auswahl und Einsatz von psychologischen Erhebungsmethoden

Frage nach der sogenannten **Operationalisierung** von psychologischen Variablen: Wie kann und möchte ich die psychologische Variable abbilden, die mich interessiert?

Variablen in der Psychologie unterscheiden sich in ihrer empirischen Zugänglichkeit:

Manifeste Variablen:

- der Sinneserfahrung direkt zugänglich, direkt beobachtbar
- Beispiele: sichtbares Verhalten, Gesichtsausdruck, Blutdruck, Reaktionszeiten, Inhalt von Aussagen/Antworten, Anzahl gelöster Aufgaben

→ leicht feststellbar, theoretische Bedeutung meist direkt ersichtlich

| Was für weitere manifeste oder latente Variablen fallen Ihnen ein?

Latente Variablen = Konstrukte:

- nur indirekt mit Beobachtungssachverhalten in Verbindung zu bringen, nicht direkt beobachtbar
- Beispiele: Persönlichkeit, Intelligenz, Emotion, Depression

→ man schließt von manifesten Indikatoren auf das latente Merkmal auf Basis von theoretischen Überlegungen (Annahme: latente Variable beeinflusst den manifesten Indikator)

Unterteilung von Erhebungsmethoden

1. Unterscheidung nach Vorgehen bei der Methode:

- Verhaltensbeobachtungsverfahren → siehe Einheit 3
- Verfahren des Selbst- bzw. Fremdberichts (Fragebögen, Interviews) → siehe Einheit 4
- Psychologische Tests (z.B. Leistungstests) → siehe Einheit 4
- Biopsychologische bzw. neurowissenschaftliche Messungen → siehe Einheit 6
- Computerbasierte Verfahren → siehe Einheit 4
- Implizite Verfahren → siehe Einheit 4
- Dokumentenanalyse → siehe Einheit 5

→ Verfahren sind nicht völlig distinkt, sondern weisen Überschneidungen auf (z.B. gibt es psychologische Tests als computerbasierte Verfahren)

Unterteilung von Erhebungsmethoden

2. Unterscheidung nach Art der Daten, die generiert werden:

Quantitative Erhebungsmethoden:

- Erfahrungsrealität wird in Zahlen erfasst bzw. übersetzt
- Ergebnis: primär numerische Daten
- Beispiele: Antworten auf eine Frage mit einer Antwortskala von 1-5, Anzahl gelöster Aufgaben, Amplitudenhöhe bei einer EEG-Messung
- Häufig stark standardisiert
- Dominante Erhebungsmethode in der Psychologie

→ Viele der auf der vorherigen Folie genannten Methoden gibt es in quantitativen und qualitativen Varianten

Qualitative Erhebungsmethoden:

- Erfahrungsrealität wird in Wörtern oder anderen nicht-numerischen Repräsentationen (z.B. Abbildungen) erfasst bzw. übersetzt
- Ergebnis: primär verbale Daten
- Häufig weniger stark bis gar nicht standardisiert
- Beispiele: gesprochene Inhalte, geschriebene Texte, Beobachtungsprotokolle

Auswahl und Einsatz von Erhebungsmethoden

Übergeordnete Perspektive

- **Ziel:** wissenschaftliche Fragestellungen in der Psychologie in Breite und Tiefe beantworten
- **Lösung:** Prinzipiell anstreben, Fragestellungen mit mehreren Datenerhebungsverfahren bzw. basierend auf unterschiedlichen Datenquellen zu untersuchen ("**multimodale** Erfassung" oder "**multimethodale** Erfassung")

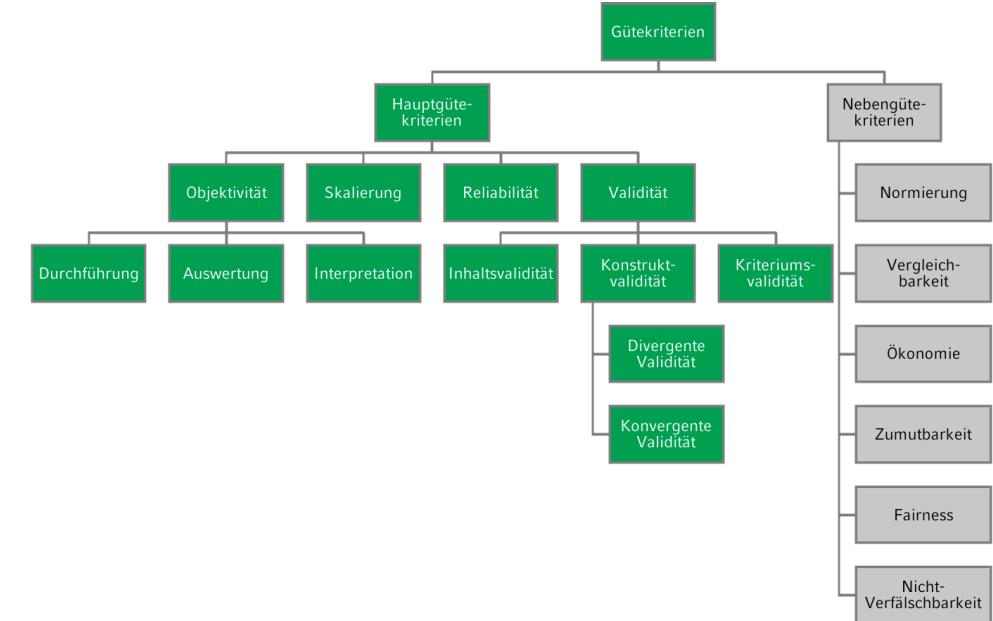
Perspektive der Einzelstudie

- häufig nicht möglich oder sinnvoll (ökonomischen oder versuchsplanerische Gründe), innerhalb einer einzigen Studie verschiedene Methoden einzusetzen
- konkrete einzelne Untersuchung → gezielte Auswahl weniger Methoden

Erhebungsmethoden der psychologischen Forschung

Gütekriterien von Erhebungsmethoden

- Zur Auswahl und Bewertung psychologischer Erhebungsmethoden müssen Qualitätskriterien berücksichtigt werden.
- Für quantitative Methoden kann man rechtstehende Gütekriterien betrachten.
- Skalierbarkeit manchmal auch als Teilespekt der Validität ("Strukturelle Validität") oder als Nebengütekriterium
- Für Gütekriterien von qualitativen Methoden → siehe Einheit 5



Gütekriterien von Erhebungsmethoden

Objektivität

Definition: Eine Erhebungsmethode ist objektiv, wenn sie das Merkmal unabhängig von Testleiter:in, Testauswerter:in und von Ergebnisinterpretation misst.

3 Bereiche lassen sich unterscheiden:

1. Durchführungsobjektivität (Testleiterunabhängigkeit)
2. Auswertungsobjektivität (Verrechnungssicherheit)
3. Interpretationsobjektivität (Interpretationseindeutigkeit)

Gütekriterien von Erhebungsmethoden

Objektivität

Durchführungsobjektivität

Definition: Testergebnis soll nicht davon abhängen, welche Testleiter:in Test durchführt → Erhebung sollte unter möglichst standardisierten Bedingungen stattfinden, Testperson als einzige Variationsquelle

Kann bei Verhaltensbeobachtung z.B. durch statistische Kennzahlen zur *Beobachtungsübereinstimmung* erfasst werden

Standardisierung wird optimiert durch:

- Instruktionen der Testleiter und Ablauf schriftlich festhalten (z.B. auch durch Interviewleitfäden bei Interviews, oder Beobachtungsplan bei Verhaltensbeobachtung)
- soziale (nicht-testbezogene) Interaktion zwischen Versuchsleiter und Testperson gering halten
- für möglichst ähnliche Untersuchungssituation sorgen (z.B. Einzel vs. Gruppentestung, Zeitbegrenzungen)
- eindeutige Anweisungen für Umgang mit Nachfragen, Störungen im Testablauf
- Bei Verhaltensbeobachtung: Training von Beobachtern

Gütekriterien von Erhebungsmethoden

Objektivität

Auswertungsobjektivität

Definition: Beim Vorliegen der Antworten/Beobachtungen einer Person soll jede Auswerter:in zum selben Ergebnis kommen

Kann z.B. durch statistische Kennzahlen zur *Beurteilungsübereinstimmung* erfasst werden

Auswertungsobjektivität wird optimiert durch:

- Vermeiden freier Antwortformate
- klare Auswertungsregeln
- Hilfsmittel, wie z.B. Auswertungsschablonen oder computergestützte Auswertung
- Festgelegte Ausschlusskriterien
- Informationen zum Umgang mit fehlenden Werten
- Festlegung von Antwortmöglichkeiten z.B. bei Interviews
- Verhaltensverankerte Ratingskalen, z.B. bei Verhaltensbeobachtungen
- Training von Beurteilern

Gütekriterien von Erhebungsmethoden

Objektivität

Interpretationsobjektivität

Definition: Unterschiedliche Erheber:innen sollen beim Vorliegen der Ergebnisse zum selben Schluss kommen.

Interpretationsobjektivität kann erhöht werden durch:

- klare Regeln für die Interpretation, z.B. durch Vorgaben zur Benennung und Beschreibung des erhobenen Merkmals, sowie der Bedeutung seiner Ausprägungen
- Vorhandensein von Normen/Normwerten inkl. Informationen zu den darin verwendeten Stichproben
- Hinweise auf Interpretation auf Basis von Konfidenzintervallen (Vertrauensbereiche, siehe VL Quantitative Methoden) und Klassifikation in Kategorien (z.B. „durchschnittlich“)
- Fallbeispiele

Gütekriterien von Erhebungsmethoden

Reliabilität

Definition: Eine Erhebungsmethode ist (vollständig) reliabel (zuverlässig), wenn sie das Merkmal ohne Messfehler misst → Reliabilität gibt den Grad der Genauigkeit an, mit der eine Erhebungsmethode misst

- Formal: Anteil der Varianz der wahren Werte an der Gesamtvarianz (siehe Vorlesung Testtheorie)
- Wichtige Einflussgröße auf Breite der Konfidenzintervalle in der Einzelfalldiagnostik (wie sicher kann ich mir bei einer einzelnen Messung sein)

Gütekriterien von Erhebungsmethoden

Reliabilität

Es lassen sich verschiedene Arten zur Schätzung der Reliabilität unterscheiden:

- **Retest-Reliabilität** → Erhebungsmethode kommt bei Wiederholung zum selben Ergebnis
- **Paralleltest-Reliabilität** → Erhebungsmethode kommt unter vergleichbaren Bedingungen bzw. vergleichbaren Erhebungsmethoden (z.B. bei Durchführung mit einer Parallelform) zum selben Ergebnis
- **Innere Konsistenz** → Einzelteile der Erhebungsmethode (z.B. Items eines Fragebogens) kommen alle zu ähnlichen Ergebnissen
- **Testhalbierungs- (Split Half-)Reliabilität** → analog zur inneren Konsistenz: Trennung der Erhebungsmethode in genau zwei Hälften, und Vergleich der Ergebnisse

Gütekriterien von Erhebungsmethoden

Validität

Definition: Eine Erhebungsmethode ist valide, wenn sie das Merkmal, das sie messen soll, auch wirklich misst (und nichts anderes).

Zwei wichtige Aspekte:

Kausale Validität

- Verursache Variation im Merkmal eine Variation im Testwert?
- Bei kausaler Validität geht es nicht um Korrelation (ungerichtete Zusammenhänge), sondern um Kausalität (gerichtete Zusammenhänge)
- Kausale Validität ist das eigentliche Herzstück, wenn man von "Validität" spricht

Inhaltsvalidität

- Erhebungsmethode erfasst repräsentativ alle relevanten Bestandteile des erhobenen Konstrukt
- Beispiel: Depressionsfragebogen sollte alle relevanten Depressionssymptome und keine nicht für Depression relevanten Symptome enthalten

Gütekriterien von Erhebungsmethoden

Validität

Häufig angewandte **Validierungsmöglichkeiten** auf Basis von Korrelationen:

1) Untersuchung der **Konstruktvalidität**

- Erhebungsmethode erzeugt Daten, die mit Daten anderer Erhebungsmethoden zusammenhängen, die dasselbe oder etwas sehr ähnliches messen sollen (**konvergente Validität**)
- Erhebungsmethode erzeugt Daten, die mit Daten anderer Erhebungsmethoden, die *nicht* dasselbe messen sollen, weniger stark zusammenhängen (**divergente/diskriminante Validität**)
- (Teilweise wird hier auch **Faktorielle Validität** verordnet: Erwartungsgemäße statistische Faktorenstruktur des Tests → siehe Vorlesung Testtheorie)

Gütekriterien von Erhebungsmethoden

Validität

- Basis für Erwartungen im Rahmen der Konstruktvalidität sind theoretische Überlegungen, die a priori (vor Kenntnis der empirischen Zusammenhänge) aufgestellt werden sollten
- Mit der sogenannten "**Multitrait-Multimethod-Methode**" können diese a priori Erwartungen systematisiert und entsprechende empirische Belege verordnet werden
- Problem sind Zirkelschlüsse:
 - Test A: „Test A korreliert (erwartungsgemäß) mit Test B, also ist A valide!“
 - Test B: „Test B korreliert (erwartungsgemäß) mit Test A, also ist B valide!“
 - Was wäre denn, wenn Test A und Test B beide etwas völlig anderes messen würden? (Intelligenz vs. Arbeitsgedächtnis)

		Methode 1			Methode 2			Methode 3		
		Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
Methode 1	Trait 1	(Rel.)								
	Trait 2		(Rel.)							
	Trait 3			(Rel.)						
Methode 2	Trait 1				(Rel.)					
	Trait 2					(Rel.)				
	Trait 3						(Rel.)			
Methode 3	Trait 1							(Rel.)		
	Trait 2								(Rel.)	
	Trait 3									(Rel.)

Anmerkungen.
»Reliabilitätsdiagonale«: In der Hauptdiagonalen stehen die Reliabilitäten (Rel.) der Verfahren.
Graue Felder = »Validitätsdiagonalen« (»monotrait-heteromethod«): Ein Merkmal wird mit verschiedenen Methoden gemessen.
Blaue Felder = »Heterotrait-Monomethod-Dreiecke«: Verschiedene Merkmale werden mit der gleichen Methode erfasst.
Alle weißen Felder unter der Reliabilitätsdiagonalen = »Heterotrait-Heteromethod-Dreiecke«: Korrelation zwischen verschiedenen Merkmalen, die zudem mit unterschiedlichen Methoden gemessen wurden.
Die Felder über der Hauptdiagonalen bleiben leer.

Gütekriterien von Erhebungsmethoden

Validität

Häufig angewandte **Validierungsmöglichkeiten** auf Basis von Korrelationen:

2) Untersuchung der **Kriteriumsvalidität**

- Erhebungsmethode erzeugt Daten, die mit relevanten, konkreten, externen Kriterien (außerhalb der unmittelbaren Testsituation) in Zusammenhang stehen (z.B. Intelligenztest mit Kriterien für Berufserfolg, z.B. Gehalt, Karrierestufe, Abschluss)
- Auch Untersuchung über (**Extrem**) -gruppenvergleiche, die erwartungsmäßige Muster zeigen, d.h. Studien zu Mittelwertsunterschieden zwischen Gruppen bei denen ein Unterschied erwartet wird (z.B. Test der Einstellung gegenüber der Kirche misst sollte bei Kirchengängern höher ausfallen als bei Nicht-Kirchengängern)
- **Inkrementelle (Kriteriums-)validität:** Beitrag einer Erhebungsmethode zur Verbesserung der Vorhersage eines Kriteriums über andere Erhebungsmethoden hinaus

Gütekriterien von Erhebungsmethoden

Validität

■ Tab. 2.19 Beispiele für Kriterien zur Validierung von Tests

Diagnostisches Verfahren (Verwendungszweck)	Mögliches Kriterium	Begründung
Depressionsfragebogen (soll Schwere der Depression erfassen)	Dauer des Aufenthaltes in einer psychiatrischen Klinik	Je schwerer die Störung, desto länger sollte die Behandlung im Krankenhaus dauern.
Intelligenztest (soll Schulerfolg vorhersagen)	Abiturnote drei Jahre nach Testdurchführung	Die Abiturnote ist ein anerkanntes Maß für Schulerfolg; da prognostische Validität angestrebt wird, muss das Kriterium deutlich später erhoben werden.
Aufmerksamkeitstest (soll Fahreignung erfassen)	Fehler in einer standardisierten Fahrprobe	Aufmerksamkeitsdefizite sollten sich in bestimmten Fehlern wie Übersehen von Verkehrszeichen, Gefahren oder der Geschwindigkeitsanzeige im Auto niederschlagen. Das Verhalten sollte im Straßenverkehr erfasst werden, weil der Test für diesen Bereich eingesetzt wird.

Erhebungsmethoden der psychologischen Forschung

Gütekriterien von Erhebungsmethoden

Validität

DOI: 10.1111/nyas.15081
ORIGINAL ARTICLE **ANNALS** OF THE NEW YORK
ACADEMY OF SCIENCES

How accurate are self-evaluations of singing ability?

Daniel Yeom^{1,†} | Kendall S. Stead^{1,2,‡} | Yi Ting Tan³ | Gary E. McPherson³ |
Sarah J. Wilson^{1,4}

¹Melbourne School of Psychological Sciences,
University of Melbourne, Melbourne, Victoria,
Australia

²School of Psychological Sciences, Macquarie
University, Sydney, New South Wales,
Australia

³Melbourne Conservatorium of Music,
University of Melbourne, Melbourne, Victoria,
Australia

⁴Department of Medicine, Epilepsy Research
Centre, University of Melbourne, Austin
Health, Heidelberg, Victoria, Australia

Correspondence
Sarah J. Wilson, Melbourne School of
Psychological Sciences, University of
Melbourne, Redmond Barry Building, Parkville,
VIC 3010, Australia. Email:
sarahw@unimelb.edu.au

[†]These authors contributed to the work
equally and share first-author status.

Funding information
Australian Government Research Training
Program Scholarship; Australian Research
Council, Grant/Award Numbers:
DP170102479, DP200100961

Abstract

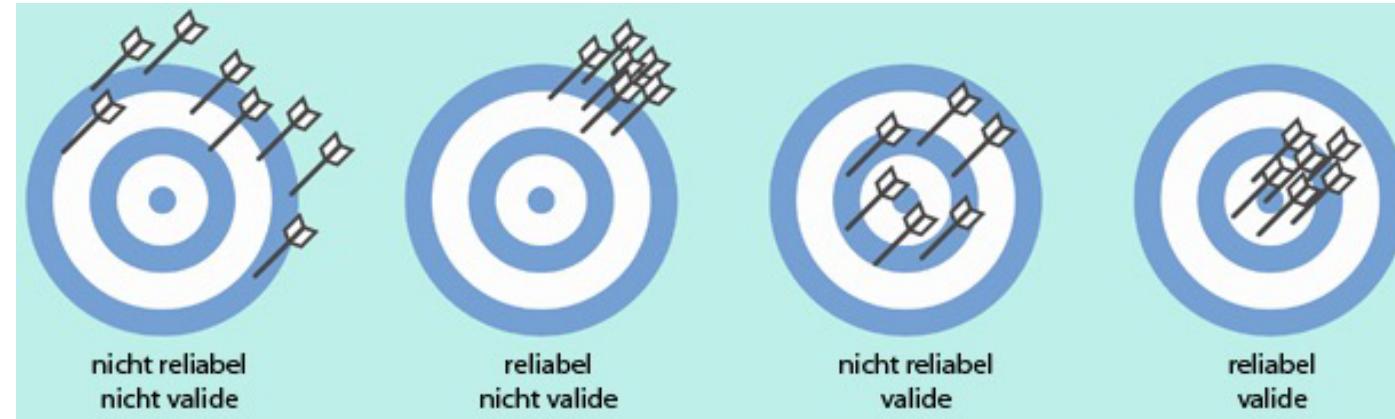
Research has shown that people inaccurately assess their own abilities on self-report measures, including academic, athletic, and music ability. Evidence suggests this is also true for singing, with individuals either overestimating or underestimating their level of singing competency. In this paper, we present the *Melbourne Singing Tool Questionnaire (MST-Q)*, a brief 16-item measure exploring people's self-perceptions of singing ability and engagement with singing. Using a large sample of Australian twins ($n = 996$), we identified three latent factors underlying MST-Q items and examined whether these factors were related to an objective phenotypic measure of singing ability. The three factors were identified as Personal Engagement, Social Engagement, and Self-Evaluation. All factors were positively associated with objective singing performance, with the Self-Evaluation factor yielding the strongest correlation ($r = 0.66$). Both the Self-Evaluation factor and a single self-report item of singing ability shared the same predictive strength. Contrary to expectations, our findings suggest that self-evaluation strongly predicts singing ability, and this self-evaluation is of higher predictive value than self-reported engagement with music and singing.

KEY WORDS

self-assessment, self-report, singing, singing engagement, singing questionnaire

Gütekriterien von Erhebungsmethoden

Veranschaulichung unterschiedlicher Reliabilität und Validität



Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Bei der Skalierbarkeit geht es um die **statistische Modellierung von Antwortprozessen**: Ein psychologischer Test gilt als skalierbar, wenn die Zuordnung der Messwerte zu den Personen auf der Basis eines empirisch nachgewiesenen testtheoretischen Modells geschieht.

Zwei Aspekte:

- Empirischer "Nachweis", dass ein bestimmtes testtheoretisches Modell gilt, also das "Richtige" ist
- Auf der Basis des "nachgewiesenen" testtheoretischen Modells Personen Werte auf den latenten Variablen zuweisen (und nicht irgendwie anders)

Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Klassische Testtheorie:

- Die meisten Fragebögenskalen und Tests basieren auf der klassischen Testtheorie (daher werden Sie dazu bereits im Bachelor eine eigene Vorlesung hören)
- Zentrale Annahme der Klassischen Testtheorie: Jeder Testwert einer Person auf einem konkreten Item (z.B. einer Frage) ist aus zwei Komponenten zusammengesetzt

1. Wahrer Wert:

- mittlerer Testwert, den eine Person in einer unendlichen Serie von Testwiederholungen erzielen würde (Erwartungswert)
- keine praktisch erzielbare, sondern eine theoretische Größe
- kann durch die konkrete empirische Antwort einer Person geschätzt werden
- Es existieren verschiedene testtheoretische Modelle je nach Annahme wie sich die Items in Hinblick auf die Messung der latenten Variable unterscheiden

2. Fehleranteil (Messfehler):

- Abweichung dieses empirischen Schätzwerts vom wahren Wert

Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Klassische Testtheorie:

- Anwendung auf Ratingskalen möglich (aber eigentlich für stetige Variablen entwickelt, z.B. Reaktionszeiten)
- Ziel: möglichst direkte und präzise Schätzung des wahren Werts auf einem Item, um darauf basierend einen Rückschluss auf eine latente Variable machen zu können
- Durch den Einsatz mehrerer Testitems soll der Fehleranteil insgesamt minimiert werden
- Anders ausgedrückt: Mehrere Items ermöglichen eine bessere Annäherung an die latente Variable einer Person
- klassische Testtheorie setzt somit voraus, dass wahre Werte und Fehlerwerte (und ihre jeweiligen Varianzen in Stichproben) getrennt bestimmt werden können → wertvoll zur Bestimmung der Reliabilität

Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Axiome (definitorische Festlegungen/Annahmen) der klassischen Testtheorie und die Folgerungen daraus

1. **Verknüpfungsaxiom:** Ein Testwert setzt sich zusammen aus der Summe von wahren Wert (Erwartungswert über unendlich viele Messungen) und Messfehler (z.B. Störeinflüsse der Umwelt).
 2. **Existenzaxiom:** Der mittlere Messfehler ist gleich null. Bei wiederholten Testanwendungen gleichen sich die verschiedenen Messfehler sozusagen aus.
 3. **Unabhängigkeitsaxiom:** Wahre Werte und Messfehler eines Items/Tests sind nicht miteinander korreliert (voneinander unabhängig) → Es werden nicht in bestimmten Ausprägungsbereichen des Items/Merkals mehr oder weniger Messfehler gemacht.
 4. **Lokale stochastische Unabhängigkeit:** Die Messfehler von verschiedenen Items/Tests sind nicht miteinander korreliert (voneinander unabhängig).
 5. **Zusatzannahme:** Messfehler eines Items/Tests sind nicht mit den wahren Werten eines anderen Items/Tests korreliert → Die Messfehler hängen nicht von bestimmten Eigenschaften ab.
- **Mit Messfehler werden in der KTT nur unsystematische Messfehler gemeint/berücksichtigt. Es kann jedoch auch systematische Messfehler geben...**

Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Unsystematische Messfehler

- zufällige Fehler
- Beispiel: vorübergehende, nicht systematisch auftretende Unaufmerksamkeit, die bei einem Mal zu einer höheren und beim anderen Mal zu einer niedrigeren Itemantwort führt

Systematische Messfehler

- spezifische Situations- oder Persönlichkeitseffekte, die die Antworten systematisch in eine bestimmte Richtung verzerren
- Beispiele: extremer Antwortstil, sozial erwünschte Antworten, mangelnde Motivation bei der Bearbeitung des Tests, Methodeneffekte...

Gütekriterien von Erhebungsmethoden

Skalierbarkeit

Zusammenfassung: Klassische Testtheorie:

- Schätzung des wahren Werts unter Berücksichtigung des (unsystematischen) Messfehlers
- Es existieren verschiedene testtheoretische Modelle je nach Annahme wie sich die Items in Hinblick auf die Messung der latenten Variable unterscheiden

Ausblick: Probabilistische Testtheorie aka. Item-Response-Theorie (IRT, Embretson & Reise, 2000; Rasch, 1980)

- Schätzung der Wahrscheinlichkeit, mit der eine Person mit einer bestimmten Merkmalsausprägung ein Item auf eine bestimmte Art beantwortet (bzw. ein Item löst)
- Es existieren verschiedene testtheoretische Modelle je nach Skalenniveau der Antworten (d.h. auch für dichotome Antworten einsetzbar, z.B. ja/nein, richtig/falsch)
- Detailliertere Abbildung des Antwortprozesses möglich als bei der KTT (z.B. Berücksichtigung von Ratewahrscheinlichkeiten)
- Ausführliche Behandlung im Masterstudium

Gütekriterien von Erhebungsmethoden

Nebengütekriterien

- **Normierung:** Bezugssystem vorhanden, auf Basis aktueller, repräsentativer und großer Stichprobe, um die individuellen Testwerte vergleichend einordnen zu können
- **Vergleichbarkeit:** Parallelform bzw. andere Verfahren mit gleichem Gültigkeitsbereich vorhanden
- **Ökonomie:** Verfahren ist kurz, einfach in der Handhabung, für Gruppenuntersuchung geeignet, wenig materialintensiv und schnell auswertbar (Verhältnis von Kosten und Nutzen im Vergleich zu anderen Verfahren relevant)
- **Zumutbarkeit:** Schonung der untersuchten Personen in zeitlicher, psychischer und körperlicher Hinsicht
- **Fairness:** Einzelne Gruppen werden durch die erhaltenen Ergebnisse eines Verfahrens nicht diskriminiert, d.h. nicht aufgrund einer testirrelevanten Eigenschaft systematisch benachteiligt
- **Nicht-Verfälschbarkeit:** keine willentliche Beeinflussung der Testleitung zum Erlangen eines ungerechtfertigten Vorteils
- **Nutzen** (wird nicht immer aufgeführt): Erfüllung eines praktischen Bedürfnisses durch den Test (der nicht schon besser durch andere Verfahren abgedeckt ist)

Problem der Reaktivität

Definition: Veränderung/Verzerrung der erhobenen Daten aufgrund der Kenntnis der untersuchten Person darüber, dass sie Gegenstand einer Untersuchung ist

Mögliche Ursachen: Veränderte Aufmerksamkeitslenkung, Relevanz sozialer Erwünschtheit (Selbstdarstellerisches, sozial konformes (Antwort-)Verhalten)

→ Datenerhebungen in Psychologie verändern oft bereits den Gegenstand

Beispiel: Hawthorne-Effekt

- klassische Studie in den Hawthorne-Werken der Western Electric Company (Roethlisberger und Dickson, 1939)
- Aussage: bloße wissenschaftliche Untersuchung der Arbeiter:innen führte zu Steigerung der Produktivität, Produktivitätssteigerung war **unabhängig** von den durch die Forscher implementierten Veränderungen der Arbeitsbedingungen
- Es folgte methodische Kritik an dieser Studie (z.B. Adair, 1984; Jones, 1992)
- Trotzdem: Mögliche **Bewertungserwartung** der untersuchten Personen aka **Aufforderungscharakteristika** der Untersuchung bekam dadurch Aufmerksamkeit

Problem der Reaktivität

Einschätzung des Ausmaßes der Reaktivität bzw. Nichtreakтивität abhängig von Informationslage der Versuchspersonen (Vpn)

Setting vom Forschenden für die Untersuchung geschaffen/ausgewählt?	ja	ja	ja	ja	ja	nein
Sind sich die Vpn des Forschungssettings bewusst?	ja	ja	ja	ja	nein	nein
Kennen die Vpn das Forschungsthema?	ja	ja	ja	nein	nein	nein
Sind sich die Vpn der Hypothese bewusst?	ja	ja	nein	nein	nein	nein
Sind sich die Vpn der Manipulierbarkeit der Erhebung bewusst?	ja	nein	nein	nein	nein	nein
Typ (Nichtreakтивität):	0	1	2	3	4	5

Beispiel für Typ 4: "Lost-Letter-Technique", bei der Briefe frankiert an öffentlichen Orten fallen gelassen werden, und geprüft (und dadurch zeitversetzt "beobachtet" wird), ob sie abgesendet werden (Milgram, Mann & Harter, 1965)

Beispiel für Typ 5: Spontane Nähe-vs. Distanzregulation im Hörsaal (Campbell, Kruskal & Wallace, 1966) oder in einer Cafeteria (Clack, Dixon & Tredoux, 2005)

Maßnahmen zur Reduktion von Reaktivität

Maßnahme	Bewertung
Versuchspersonen in Unkenntnis darüber lassen, dass sie untersucht werden	Nur in Feld-, Archiv- oder Internetstudien praktikabel, nicht in Laborstudien; kann ethisch problematisch sein
Versuchspersonen Anonymität zusichern	Besonders wichtig bei der Erhebung von persönlichen/sensiblen Daten; reduziert sozial-erwünschtes Antworten
Coverstory über den Untersuchungszweck entwickeln	Wichtig in hypothesenprüfenden Studien, in denen die Versuchspersonen erforschtes Verhalten kontrollieren können; Täuschung ethisch zu reflektieren
Maße einsetzen, die von Versuchspersonen nicht kontrolliert oder beeinflusst werden können (nicht-reaktive Messverfahren)	Angenommen für biopsychologische Maße, die kaum kontrollierbare physiologische Vorgänge erfassen (z.B. Messung von Hormonspiegeln oder Verfahren zur Messung der Gehirnaktivität)
Indirekte/implizite Messverfahren einsetzen	Versuchspersonen können aus gemessenen Verhaltensweisen (oft Reaktionszeiten) nur schwer auf das untersuchte Konstrukt schließen

Warum ist das für mich wichtig?

"Methoden, mit denen die Psychologie ihre Daten gewinnt und auswertet, gehören zu den »**Regeln der Kunst**«, auf die uns die **ethischen Richtlinien unserer Profession** verpflichten (Deutsche Gesellschaft für Psychologie und Berufsverband Deutscher Psychologinnen und Psychologen, 2005).

Es ist ein verbreitetes Missverständnis, dass gute Methodenkenntnisse nur in der psychologischen Forschung benötigt werden. Auch in der psychologischen Anwendungspraxis ist gutes Methodenwissen unverzichtbar. Nur wer über dieses Wissen verfügt, ist beispielsweise in der Lage, die wissenschaftliche Literatur kritisch zu beurteilen, zu entscheiden, welches diagnostische Verfahren welche Gütekriterien wie gut erfüllt und wie die psychologische Praxis zur Generierung von Wissen und somit für den wissenschaftlichen Fortschritt genutzt werden kann.

Da mit Psychologie auch Geld verdient wird, gibt es ein Spannungsverhältnis zwischen dem wirtschaftlichen Gewinn und der wissenschaftlichen Qualität einer psychologischen Dienstleistung. Nur wer über methodisches Wissen verfügt, kann psychologische Dienstleistungsangebote in diesem Spannungsverhältnis verorten und unseriöse Angebote von seriösen unterscheiden."

(Eid, Gollwitzer & Schmitt, S.37)

Take-Aways

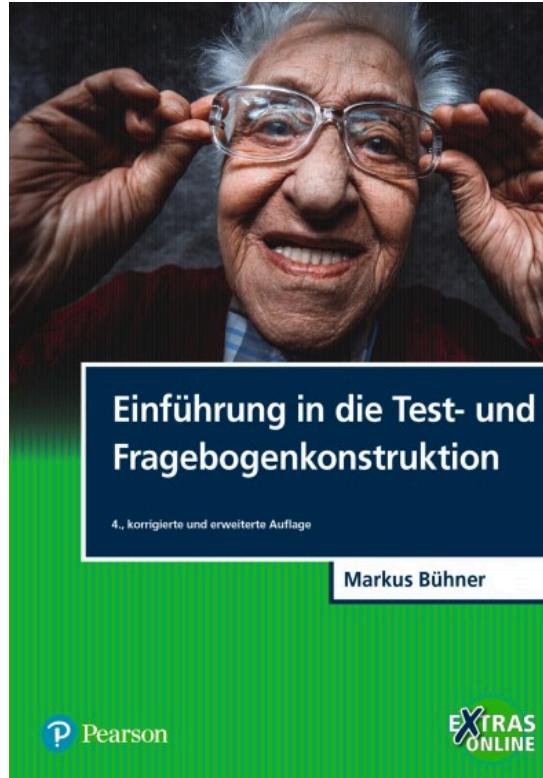
- **Hauptgütekriterien** für psychologische Erhebungen sind Objektivität, Reliabilität und Validität (je nach Autor auch Skalierbarkeit)
- **Objektivität:** Erhebung, Ergebnis und Interpretation unabhängig von Testleiter:in und Testauswerter:in
- **Reliabilität:** Messfehlerfreie und zuverlässige Erhebung des Merkmals
- **Validität:** Methode erhebt wirklich das interessierende Merkmal und nicht etwas anderes
- **Klassische Testtheorie:** Schätzung des wahren Werts unter Berücksichtigung des Messfehlers
- **Probabilistischen Testtheorie:** Schätzung der Wahrscheinlichkeit für eine bestimmte Itemantwort
- **Reaktivität:** Veränderung/Verzerrung der erhobenen Daten aufgrund der Kenntnis der untersuchten Person darüber, dass sie Gegenstand einer Untersuchung ist

Schlüssel-/Fachbegriffe der heutigen Vorlesung

Operationalisierung	Gütekriterien	Durchführungsobjektivität	Kausale Validität
manifest	Objektivität	Auswertungsobjektivität	Inhaltsvalidität
latent	Reliabilität	Interpretationsobjektivität	konvergente / divergente Konstruktvalidität
Konstrukt	Validität	Retest-Reliabilität	Kriteriumsvalidität
quantitative	Skalierung	Paralleltest-Reliabilität	Faktorielle Validität
qualitativ	unsystematische / systematische Messfehler	Innere Konsistenz	Multitrait-Multimethod-Methode
multimodal / multimethodal		Testhalbierungs-Reliabilität / Split-Half-Reliabilität	Reaktivität

[zurück zur heutigen Übersicht der Vorlesung →](#)
[zum Quiz zur Wissensprüfung →](#)

Literatur für die heutige Sitzung



Kapitel 8 in Bühner, M. (2021). Einführung in die Test- und Fragebogenkonstruktion. Pearson.

Materialien: Vielen Dank an Prof. Dr. Stephan Goerigk, Prof. Dr. Mario Gollwitzer und den Lehrstuhl für Psychologische Methodenlehre und Diagnostik an der LMU für Bereitstellung der Grundlage für die Materialien