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

H. Moosbrugger, A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion*,
https://doi.org/10.1007/978-3-662-61532-4_2

2.6 Zusammenfassung – 36

2.7 Kontrollfragen – 36

Literatur – 37

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/Aufgabestellungen/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.

Merkmalsindikatoren zur Operationalisierung latenter Konstrukte

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:

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 Definition[1] 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 Parteipräfenzen 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.

Testgütekriterien

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:

Allgemeine Gütekriterien

- 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).

Spezielle testtheoriebasierte Gütekriterien

- 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 (KTT; ▶ 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).

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

Eliminierung von Störeinflüssen

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.

Standardisierung

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:

Festlegung von Testmaterialien, Zeitdauer und Instruktion

- *Konstanz der Fragen/Aufgabestellungen/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

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.

Vergabe von Punktwerten/Itemwerten

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 Mehrfachwahlaufgaben, ►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

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

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 „Konkordanzkoeffizienten W" nach Kendall (1962) berechnet werden. (Für weitere Übereinstimmungsmaße sei im Überblick auf Wirtz und Caspar 2002, verwiesen.)

Statistische Kontrolle systematischer Abweichungen

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.

Gütekriterium der Skalierung

Definition

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

Adäquate Entsprechung von Merkmalsausprägungen und Testwerten

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

symptomatischen Antworten und einem entsprechend höheren Testwert einhergehen muss.

Vergleichsmöglichkeiten von Testwerten in Abhängigkeit vom Skalenniveau

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 Rationalskalen- 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 Intervallskala 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.

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.

Vergleich mit Ergebnissen der Eichstichprobe

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).

Prozentrangnormen

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.

Standardnormen

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.

Verteilungsform der Eichstichprobe beachten

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).

Anmerkung: Um eine angemessene Vergleichbarkeit der Personen zu gewährleisten, dürfen die Normtabellen nicht veraltet sind. 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

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

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.

Berücksichtigung weiterer allgemeiner Gütekriterien

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.

2.3.2 Ökonomie

Wirtschaftlichkeit eines Tests

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.

Definition

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

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.

Finanzieller Aufwand

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.

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

Ökonomievorteile durch adaptives Testen

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

Zeitliche, physische und psychische Beanspruchung der Testpersonen

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.

Zumutbarkeit hängt von der Fragestellung ab

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.

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 Augenscheinvalidität (► Abschn. 2.4.2.1) besitzen.

Ergebnisverfälschendes Verhalten durch Soziale Erwünschtheit

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

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:

Messgenauigkeit des Tests

Definition

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

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

Zwei Messzeitpunkte, derselbe Test

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-

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

■■ Paralleltest-Reliabilität

Zwei Messzeitpunkte, parallele Testformen

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. Parallele 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.

Ein Messzeitpunkt, zwei Testhälften

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 Gesamtreliabilität der ursprünglichen Testlänge („Split-Half-Reliabilität") hochzurechnen.

■■ Cronbachs Alpha (α)

Verallgemeinerung der Testhalbierungsmethode

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.

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

2.4.1.2 Modellbasierte Methoden der Reliabilitätsschätzung

Weniger strenge Annahmen erforderlich

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).

Kritik an Cronbachs Alpha

So wurde erst in jüngerer Zeit genauer erkannt, dass Cronbachs Alpha die Erfüllung (zu) strenger Annahmen voraussetzt und zudem eine Reihe von problematischen Eigenschaften bei der Schätzung der Reliabilität aufweist, sodass inzwischen von einer unkritischen Verwendung des Koeffizienten Alpha zur Schätzung der Reliabilität 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).

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. *Augenscheinvalidität* 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. Augenscheinvalidität 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. Augenscheinvalidität,
b. Inhaltsvalidität,
c. Kriteriumsvalidität und
d. Konstruktvalidität

berücksichtigen.

2.4.2.1 Augenschein- und Inhaltsvalidität

Akzeptanz des Tests vonseiten der Testperson

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

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

Augenscheinvalidität 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 Augenscheinvalidität, 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 Augenscheinvalidität nicht mit Inhaltsvalidität verwechselt werden darf, obwohl dies leicht passieren kann (vgl. Tent und Stelzl 1993), da augenscheinvaliden 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 extrapolierende Testwertinterpretationen

Extrapolierende 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".

„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.

Feststellung der Kriteriumsvalidität als Test-Kriterium-Korrelation

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.

Stützung der Interpretation durch theoretische Annahmen und empirische Belege

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.

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-)Merkmalen („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

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

Struktursuchendes vs. strukturprüfendes Vorgehen

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

Exploratorische Faktorenanalyse

- 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.

Nomologisches Netzwerk: Beziehungen zu konstruktverwandten bzw. konstruktfremden Merkmalen

Konvergente Validität

Diskriminante Validität

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 *konvergenten 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/Merkmals 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

Strukturprüfendes Vorgehen

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).

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.

2.5 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.