



Leseprobe aus: Schmitt, Gerstenberg, Psychologische Diagnostik kompakt, ISBN 978-3-621-28143-0
© 2014 Beltz Verlag, Weinheim Basel
<http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-621-28143-0>

3 Gütekriterien diagnostischer Instrumente

Was Sie in diesem Kapitel erwartet

Beim gemeinsamen Mittagessen in der Mensa erzählt ein Kommilitone Ihnen, dass er in seiner Freizeit einen Fragebogen zur Erfassung der Ärgerneigung von Personen entworfen hat.

Der Fragebogen besteht aus 20 Items mit einem dichotomen Antwortformat (Zustimmung/Ablehnung), da dies die Auswertung erleichtern soll. Außerdem hat ihr Kommilitone auf Instruktionen verzichtet, da der Fragebogen seiner Meinung nach selbsterklärend ist. Die Objektivität des Fragebogens sei damit gesichert. Zur Überprüfung der Reliabilität hat er den Fragebogen von seinen Eltern und seinen sechs Geschwistern zweimal im Abstand von drei Tagen ausfüllen lassen und daraus eine Retestreliabilität von .75 errechnet.

Des Weiteren nennt er Ihnen zwei Beispielimens: »Wenn jemand so viel Geld hat, dass er nicht arbeiten muss, finde ich das ungerecht« und »Ich finde es nicht gut, wenn andere Leute sich vordrängeln«.

Ihr Kommilitone erklärt Ihnen, dass er den Fragebogen bald veröffentlichen möchte und bittet Sie um eine Einschätzung der Wahrscheinlichkeit, dass der Hogrefe Verlag den Fragebogen kaufen wird. Was erwidern Sie ihm?

Keine Sorge, falls Sie jetzt noch keine Antwort wissen. Nach der Lektüre dieses Kapitels zu den Gütekriterien, denen ein diagnostischer Test genügen sollte, werden Ihnen viele Argumente einfallen, um Ihrem Kommilitonen von der Einreichung des Fragebogens beim Verlag abzuraten.

Wenn wir im Supermarkt unser Obst und Gemüse wiegen oder an der Kasse gewogen bekommen, verlassen wir uns auf die Genauigkeit der verwendeten Waage. Wir wären empört, wenn wir zu Hause feststellen würden, dass beim Nachwiegen der angeblich 1321 Gramm Tomaten, für die wir im Gemüseladen 2,95 Euro bezahlt haben, unsere elektronische Hauswaage lediglich 909 Gramm anzeigen würde. Dies gilt zumindest, wenn wir unterwegs keine Tomaten gegessen oder verloren haben. Ebenso wären wir empört, wenn wir an einer Tankstelle laut Anzeige der Zapfsäule 48 Liter Benzin getankt hätten, unser Kleinwagen mit einem Durchschnittsverbrauch von 6,2 L/100 km aber bereits nach 108 km mit leerem Tank stehen bleiben würde. Unsere Empörung zeigt uns, dass wir uns im Alltag auf die Genauigkeit von Messinstrumenten verlassen. Ungenauigkeiten halten wir nicht für akzeptabel oder nur dann, wenn sie sehr gering sind oder keine Nachteile mit sich bringen.

Nicht anders ist die Situation in der Psychologie. Nehmen wir an, eine Mutter wird beim Schulpsychologen mit der Frage vorstellig, ob ihre fünfjährige Tochter, die bereits lesen und einfache Rechenaufgaben lösen kann, hochbegabt ist. Der Psychologe testet das Mädchen mit einem Intelligenztest und stellt fest, dass das Kind laut Testergebnis unterdurchschnittlich intelligent ist. Die Mutter wäre verwundert und würde wahrscheinlich einen zweiten Psychologen aufsuchen. Wenn dieser mit einem anderen Intelligenztest das Ergebnis des ersten Diagnostikers bestätigt, wird die Mutter entweder betrübt nach Hause gehen oder sich in einem dritten Versuch um eine Bestätigung ihrer Einschätzung ihrer Tochter bemühen. Sollte der zweite Diagnostiker jedoch bei Verwendung des gleichen oder eines anderen Intelligenztests das Mädchen als weit überdurchschnittlich intelligent bezeichnen, wäre die Mutter vermutlich zutiefst verunsichert und ratlos, welchem Diagnostiker sie wohl glauben könne.

Das Gedankenbeispiel zeigt, dass sich nicht nur die Kunden von Supermärkten, sondern auch die Kunden psychologischer Expertise darauf verlassen, dass genau gemessen bzw. genau psychologisch diagnostiziert wird. Es überrascht deshalb nicht, dass hochwertige Diagnostik berufsethischen Geboten der Psychologie unterliegt (Deutsche Gesellschaft für Psychologie und Berufsverband Deutscher Psychologinnen und Psychologen, 2005). Die beiden in Klammern genannten Standesvertretungen haben zur Qualitätssicherung im Bereich der psychologischen Diagnostik eine ständige Kommission gebildet, das Diagnostik- und Testkuratorium. Nähere Informationen finden sich online unter: www.zpid.de/testkuratorium. In Kapitel 5 werden wir das Thema Qualitätssicherung in der psychologischen Diagnostik noch einmal aufgreifen und vertiefen.

In diesem Kapitel wollen wir uns mit der Frage befassen, was in der psychologischen Diagnostik unter Messgenauigkeit verstanden wird und wie sich die Messgenauigkeit diagnostischer Instrumente prüfen und verbessern lässt. Es hat sich im Laufe der Geschichte der psychologischen Diagnostik eingebürgert, die Messgenauigkeit an sogenannten Gütekriterien festzumachen. Diese Gütekriterien werden in drei Hauptgütekriterien und mindestens fünf Nebengütekriterien unterteilt.

Hauptgütekriterien

- (1) Objektivität
- (2) Reliabilität
- (3) Validität

Nebengütekriterien

- (4) Normierung
- (5) Fairness
- (6) Ökonomie
- (7) Nützlichkeit
- (8) Akzeptanz

3.1 Objektivität

Definition

Objektivität

Ein diagnostisches Instrument ist in dem Maße objektiv, in dem das diagnostische Ergebnis unabhängig von der diagnostizierenden Person (dem Diagnostiker) ist.

Objektivität wird unterteilt in Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität.

Durchführungsobjektivität

Durchführungsobjektivität ist gegeben, wenn der Ablauf der diagnostischen Erhebung unabhängig von der diagnostizierenden Person erfolgt. Durchführungsobjektivität wäre z. B. dann nicht gegeben, wenn ein Diagnostiker einem Diagnostikanden bei der Durchführung eines Leistungstests Hilfestellung geben würde, ein anderer Diagnostiker hingegen nicht. Bei standardisierten Verfahren ist die Voraussetzung zur Durchführungsobjektivität durch eine detaillierte Instruktion gegeben. Allerdings müssen sich die diagnostizierenden Personen auch genau an die Instruktion halten. Leider wird diese Vorschrift in der Praxis häufig verletzt. Zu Einschränkungen der Durchführungsobjektivität kommt es z. B., wenn diagnostische Erhebungen an angelerntes Hilfspersonal delegiert werden, diesem die wichtige Bedeutung der exakten Befolgung von Instruktionen aber nicht bewusst ist.

Bei nicht standardisierten Verfahren wie Anamnesegesprächen ist die Durchführungsobjektivität grundsätzlich gering. Sie kann aber durch einen strukturierenden Gesprächsleitfaden erhöht werden (vgl. Kap. 2, Gesprächsmethoden). Untersuchungen zur Messgenauigkeit von Einstellungsinterviews zeigen bspw., dass deren Aussagekraft durch Strukturierung deutlich gesteigert werden kann (Schuler, 2002; s. Kap. 2).

Besonders hoch ist die Durchführungsobjektivität bei allen computergestützten Verfahren, bei denen der Ablauf durch das Programm definiert wird und deshalb im Regelfall nicht zwischen diagnostischen Erhebungen variieren kann. In Kapitel 5 werden wir erklären, wie man die Durchführungsobjektivität eines diagnostischen Verfahrens empirisch ermitteln kann.

Auswertungsobjektivität

Auswertungsobjektivität ist gegeben, wenn die Übersetzung diagnostischen Verhaltens in eine Symbolsprache (z. B. Zahlen) unabhängig vom Diagnostiker erfolgt, verschiedene Diagnostiker also bspw. identische Summenwerte anhand eines ausgefüllten Fragebogens bilden. Bei standardisierten Verfahren ist die Auswertungsobjektivität in der Regel hoch: Die Vorschriften der Übersetzung von Verhalten (z. B. Antworten auf Fragebogenitems) in Zahlen ist genau festgelegt und wird bei schriftlichen Verfahren häufig durch Auswertungsschablonen erleichtert. Allerdings kommt es trotz solcher Schablonen manchmal zu Übertragungsfehlern.

Gefeit vor solchen Fehlern sind computergestützte Verfahren (vorausgesetzt, sie wurden korrekt programmiert), weil dort das Verhalten der Person gespeichert und nach festgelegten Regeln ausgewertet wird.

Gering ist die Auswertungsobjektivität hingegen bei Verfahren mit einem offenen Antwortformat. Mit diesem Problem sind z. B. nichtstandardisierte Interviews und projektive Verfahren konfrontiert. Zwar wurden für diese Verfahren Auswertungsschlüssel entwickelt, die teilweise sehr differenziert sind, dennoch gibt es auf jede offene Frage nahezu unendlich viele Antworten. Selbst der beste Auswertungsschlüssel wird deshalb an Grenzen stoßen. In Kapitel 5 werden wir erklären, wie die Auswertungsobjektivität eines diagnostischen Verfahrens empirisch festgestellt werden kann.

Interpretationsobjektivität

Interpretationsobjektivität ist gegeben, wenn unterschiedliche Diagnostiker aus dem diagnostischen Verhalten dieselben Schlüsse ziehen. Betrachten wir drei Beispiele, die sich in diesem Kriterium unterscheiden. Die Frage, ob eine Person hochbegabt ist oder nicht, wird häufig durch Vereinbarung gelöst. Man definiert einen bestimmten Prozentsatz von Personen, den man als hochbegabt bezeichnen möchte. Beispielsweise könnte man sich darauf verständigen, das oberste Prozent der Begabungsverteilung als hochbegabt einzustufen. Wenn alle Diagnostiker dieses Kriterium anwenden und das gleiche Diagnoseinstrument verwenden, ist Interpretationsobjektivität leicht zu erreichen.

Viel schwieriger ist das Kriterium in der klinisch-psychologischen Diagnostik zu erfüllen. Zwar genügen Diagnose-Checklisten (vgl. Kap. 8) den Gütekriterien der Durchführungs- und Auswertungsobjektivität und lassen es daher zu, eine bestimmte Diagnose wie »Generalisierte Angststörung« zu fällen. Dennoch variieren die Auffassungen klinischer Psychologen, wie eine diagnostizierte Störung bei einem bestimmten Klienten am besten therapiert werden kann.

Noch schwieriger ist das Kriterium der Interpretationsobjektivität in der forensischen Diagnostik zu erfüllen. Wenn bspw. beurteilt werden soll, mit welcher Wahrscheinlichkeit ein Sexualstraftäter nach einer vorzeitigen Entlassung aus dem Gefängnis rückfällig wird, dürfte eine zufriedenstellend hohe Interpretationsobjektivität trotz identischer Informationen über den Täter kaum zu erreichen sein. Umso wichtiger ist es in solchen Fällen, den diagnostischen Entscheidungsprozess und die

Entscheidungsbegründung minutiös zu dokumentieren. Nur so ist es möglich, sich gegen Kritik und Anfeindungen im Falle eines Fehlurteils zur Wehr setzen zu können. In Kapitel 5 werden wir erklären, wie man die Interpretationsobjektivität eines diagnostischen Verfahrens empirisch überprüfen kann.

3.2 Reliabilität

Definition

Reliabilität

Reliabel oder zuverlässig ist ein diagnostisches Instrument in dem Maße, in dem es das genau misst, was es misst. Reliabilität ist gegeben, wenn ein Instrument bei wiederholter Messung der gleichen Objekte oder Merkmalsträger unter gleichen Bedingungen die gleichen Messergebnisse anzeigt. Reliabilität ist eine kontinuierliche Variable. Ein Messinstrument kann mehr oder weniger reliabel sein. Als Maßeinheit der Reliabilität werden Anteile (von 1) verwendet. Dies rührt daher, dass Reliabilität in der Klassischen Testtheorie als Anteil der systematischen (replizierbaren) Varianz an der Gesamtvarianz der Messwerte definiert wird (s. Kap. 4). Die Reliabilität

kann somit Werte zwischen 0 und 1 annehmen. Ein perfekt zuverlässiges Messinstrument hat eine Reliabilität von 1. Es misst ein Merkmal fehlerfrei. Ein vollkommen unzuverlässiges Messinstrument hat eine Reliabilität von 0. Es misst nur unsystematische Unterschiede zwischen den Merkmalsträgern (Messfehler).

Merke: Ein diagnostisches Instrument hat eine einzige Reliabilität. Es gibt jedoch mehrere Methoden, seine Reliabilität zu schätzen. Diese Schätzmethoden basieren auf bestimmten Annahmen (s. Kap. 4). Wenn diese verletzt sind, ist die Reliabilitätsschätzung fehlerhaft.

Bei der Definition der Reliabilität gibt es mehrere Dinge zu beachten. Erstens ist die Reliabilität eines diagnostischen Instruments unabhängig davon, ob das Instrument das misst, was es messen soll. Für die Beurteilung der Reliabilität ist es also bspw. unerheblich, ob ein Kreativitätstest Kreativität oder Intelligenz misst. Erheblich ist nur, ob er zuverlässig das misst, was er misst – was immer dieses ist. Zweitens ist es wichtig zu bedenken, dass von einem diagnostischen Instrument nur dann ein gleichbleibendes Ergebnis erwartet werden kann, wenn auch das gemessene Objekt und die Umstände der Messung gleich geblieben sind. Zur Illustration dieses wichtigen Aspekts wählen wir zunächst ein physikalisches Beispiel: Die Personenwaage.

Wir wiegen mehrere Personen mit einer Waage, deren Zuverlässigkeit wir beurteilen wollen. Jede Person wiegen wir zweimal. Zuverlässig wäre die Waage, wenn sie bei derselben Person zweimal jeweils exakt das gleiche Gewicht anzeigen würde. Unzuverlässig wäre die Waage, wenn das Gewicht derselben Person über die beiden Wiegevorgänge variieren würde, die Waage bei einer bestimmten Person bspw. beim ersten Wiegen 70 kg und beim zweiten Wiegen 71 kg anzeigen würde. Der Schluss, die Waage sei unzuverlässig, wäre jedoch nicht mehr berechtigt, wenn die gewogene Person zwischen den beiden Wiegevorgängen einen Liter Wasser getrunken hätte. In diesem Fall hätte sich das Objekt verändert und deshalb könnte aus dem veränderten Messwert nicht auf die Unzuverlässigkeit der Waage geschlossen werden. Dieses Beispiel zeigt, was mit der Bedingung »gleichbleibendes Objekt« gemeint ist.

Stellen wir uns nun vor, dass der erste Wiegevorgang auf der Erde, der zweite auf dem Mond stattgefunden hat, die Person zwischen den beiden Wiegevorgängen aber kein Wasser getrunken hat. Die Person ist also gleich schwer geblieben. Dennoch wird die Waage auf dem Mond einen anderen Wert anzeigen als auf der Erde, weil die Gravitationskraft der Erde größer ist als die des Mondes und sich damit die Messbedingungen geändert haben.

Reliabilität lässt sich also nur korrekt bestimmen, wenn das diagnostizierte Objekt unverändert geblieben ist und wenn alle Bedingungen, die sich auf das Ergebnis auswirken könnten, ebenfalls unverändert geblieben sind. Haben sich das Objekt oder die Bedingungen geändert und sind die Auswirkungen dieser Veränderungen bekannt, kann Reliabilität ebenfalls bestimmt werden. Wenn wir bspw. wissen, dass eine Person zwischen zwei Wiegevorgängen einen Liter Wasser getrunken hat, dann erwarten wir von einer zuverlässigen Waage, dass sie beim zweiten Wiegen genau 1 kg mehr anzeigt als beim ersten Wiegen.

Das Beispiel der Personenwaage lässt sich sinngemäß auf alle psychologischen Diagnoseinstrumente übertragen. Auch hier gilt, dass die Bestimmung der Reliabilität eines Instruments voraussetzt, dass sich weder das gemessene Merkmal noch die diagnostischen Bedingungen geändert haben. Zur Illustration wählen wir einen Stimmungsfragebogen und einen Konzentrationstest.

Unveränderte diagnostische Objekte. Nehmen wir an, mit einem Stimmungsfragebogen solle ermittelt werden, wie gut die aktuelle Stimmung einer Person ist. Nehmen wir weiterhin an, dass die Stimmung mehrerer Personen an zwei aufeinanderfolgenden Tagen gemessen wurde. Wenn sich das Messergebnis einer Person von einem auf den anderen Tag ändert, bedeutet das dann eine mangelnde Reliabilität des Fragebogens? Nicht unbedingt, da Stimmungen kurzlebige Phänomene sind und sich durchaus von einem auf den anderen Tag ändern können. Wichtig zur Beurteilung der Reliabilität wäre es also zu wissen, ob und wie sehr sich die Stimmung der Person tatsächlich geändert hat. Erst dann können wir beurteilen, ob der Stimmungsfragebogen zuverlässig war. Wir werden dieses Thema in Kapitel 7 wieder aufgreifen und vertiefen.

Gleiche Bedingungen. Nehmen wir nun an, die Zuverlässigkeit eines Konzentrationstests soll geprüft werden, indem der Test einer großen Stichprobe von Studierenden im Abstand von einer Woche zweimal vorgegeben wird. Die Testung erfolgt aus Platzmangel in kleinen Gruppen in einem Seminarraum einer Universität. Nehmen wir weiterhin an, dass in unmittelbarer Nähe des Seminarraums ein neues Gebäude errichtet wird und der Baulärm sehr stark schwankt. Da Lärm die Konzentrationsfähigkeit von Menschen beeinträchtigt, dieser Effekt aber von Person zu Person variiert und überdies der Lärm selbst nicht bei allen diagnostizierten Personen konstant war, müssen wir mit divergierenden Testergebnissen zwischen beiden Messungen rechnen. Aus diesen zu schließen, der Konzentrationstest sei unzuverlässig, wäre jedoch falsch, da die Bedingungen der Messung nicht konstant waren.

In Kapitel 7 werden wir zeigen, dass trotz veränderlicher Objekte und variabler Bedingungen die Reliabilität von diagnostischen Instrumenten geschätzt werden kann, wenn die relevanten Veränderungen und der Einfluss der veränderlichen Bedingungen berücksichtigt werden. Betrachten wir nun die wichtigsten Strategien der Reliabilitätsschätzung, die unveränderte Objekte und konstante Bedingungen voraussetzen: (1) die Paralleltestmethode, (2) die Retestmethode und die (3) die Testzerlegungsmethode. Der Testbegriff in diesen drei Bezeichnungen bedeutet nicht, dass sich diese Methoden zur Schätzung der Zuverlässigkeit auf Tests im engeren Sinn (vgl. Kap. 2) beschränken. Vielmehr steht dieser Begriff stellvertretend für prinzipiell alle diagnostischen Verfahren.

Paralleltestmethode

Die Paralleltestmethode geht davon aus, dass es außer demjenigen diagnostischen Verfahren, dessen Zuverlässigkeit man schätzen möchte, noch mindestens ein zweites (paralleles bzw. äquivalentes) Verfahren zur Erfassung des gleichen Merkmals gibt. Unter einem zweiten (parallelen bzw. äquivalenten) Verfahren versteht man dabei einen Test, der auf ähnlichen, aber nicht identischen Items basiert, die jedoch die gleichen Messeigenschaften aufweisen wie die Items des

ersten Tests. Was »gleiche Messeigenschaften« genau heißt, werden wir in Kapitel 4 bei der Vorstellung der Klassischen Testtheorie erklären. Durch den Vergleich der Messwerte, die mit beiden Verfahren gewonnen wurden, kann nun ihre Zuverlässigkeit geschätzt werden. Wenn die beiden Tests mindestens intervallskalierte Messwerte liefern und weitere Voraussetzungen erfüllt sind, die wir in Kapitel 4 genauer erläutern, kann die Schätzung anhand der Produktmomentkorrelation erfolgen. Das Merkmal, z. B. die Intelligenz, wird also nicht nur bei einem Merkmals-träger (einer Person) gemessen, sondern bei vielen. Je höher die Korrelation zwischen den Messwertreihen, die mit den beiden Verfahren gewonnen wurden, desto höher ist auch die Reliabilität. Zu beachten ist bei der Bewertung einer solchen Korrelation zwischen zwei diagnostischen Verfahren, dass sie nur dann die Zuverlässigkeit zutreffend angibt, wenn die Verfahren tatsächlich das gleiche Merkmal messen. Wenn die Verfahren nicht das gleiche Merkmal erfassen, wird die Reliabilität unterschätzt. Was »gleich« in diesem Zusammenhang genau bedeutet, werden wir in Kapitel 4 erklären.

Retestmethode

Die Retestmethode, die auch als Testwiederholungsmethode und Messwiederholungsmethode bezeichnet wird, hatten wir bei der Erläuterung der Reliabilität am Beispiel einer Personenwaage bereits kennengelernt. Von einem zuverlässigen Instrument erwarten wir, dass es bei mehrmaligem Messen des gleichen Objekts unter gleichen Bedingungen das gleiche Ergebnis anzeigt. Auch zur Schätzung der Reliabilität mit der Retestmethode werden Korrelationskoeffizienten berechnet. Je höher die Retestkorrelation zwischen zwei Messungen, desto zuverlässiger ist das Messinstrument. Allerdings wird, wie wir bereits gesehen haben, die Zuverlässigkeit nur dann zutreffend geschätzt, wenn die gemessenen Objekte sich zwischen den beiden Messungen nicht verändert haben. Diese Voraussetzung ist in der Psychologie häufig nicht erfüllt. Wenn wir bspw. die Stimmung einer Gruppe von Personen zweimal im Abstand von einem Tag messen, sind individuelle Stimmungsveränderungen höchst wahrscheinlich. Wir messen also bei vielen Personen zu den beiden Zeitpunkten verschiedene Stimmungen. In einem solchen Fall unterschätzt die Retestkorrelation die Reliabilität, sodass komplexere Messpläne und Modelle verwendet werden müssen, z. B. das Latent-State-Trait-Modell, das wir in Kapitel 7 kennenlernen werden.

Testzerlegungsmethode

Die Testzerlegungsmethode kann man als Variante der Paralleltestmethode auffassen. Gibt es kein zweites Instrument für das zu messende Merkmal, kann das vorhandene Instrument in Teile zerlegt und diese Teile können miteinander verglichen werden. Die meisten diagnostischen Verfahren, insbesondere nahezu alle Fragebogen und Tests, bestehen aus mehreren Items (Fragen, Aufgaben), die sich gut aufteilen lassen.

Testhalbierungsmethode

Die Items können entweder nach Zufall oder abwechselnd nach der Itemnummer zwei Gruppen zugewiesen werden. Das Ergebnis sind dann Testhälften. Unter der Voraussetzung, dass die beiden Testhälften das gleiche Merkmal messen, wird Reliabilität anhand der Korrelation der beiden Testhälften geschätzt.

Parcelmethode. Statt aus den Items eines Verfahrens zwei Testhälften zu bilden, kann man sie auch in mehrere Gruppen einteilen. Die entstehenden Testteile werden als Parcels (engl. für Päckchen) bezeichnet. Diese Methode wird v. a. in konfirmatorischen Faktorenanalysen und Strukturgleichungsanalysen verwendet, weil sich mit Parcels die Voraussetzungen dieser Verfahren leichter erfüllen lassen als mit einzelnen Items (Bandalos & Finney, 2001; Little et al., 2002). Ebenso wie bei

Testhälften kann man bei dieser Methode die Parcels miteinander korrelieren, um die Zuverlässigkeit des Instruments zu schätzen.

Interne Konsistenzanalyse. Schließlich ist es auch möglich, jedes einzelne Item als Maß für das zu messende Merkmal aufzufassen und die Zuverlässigkeit des gesamten diagnostischen Instruments aus der Korrelation der Items miteinander zu schätzen. Ein häufig verwendeter statistischer Koeffizient, der sich dieser Vorgehensweise bedient, ist der von Cronbach (1951) vorgeschlagene Koeffizient Alpha.

Cronbachs Alpha und interne Konsistenz

Dieser Koeffizient ist ein Maß für die sogenannte interne Konsistenz eines diagnostischen Verfahrens, bspw. eines Tests. Interne Konsistenz bedeutet in diesem Fall »Zusammenhang zwischen den Items«. Unter der Voraussetzung, dass die Items das gleiche Merkmal messen und nur dieses, bedeuten hohe Korrelationen zwischen den Items, dass diese zuverlässig sind. Wenn die einzelnen Items, die das Merkmal messen, zuverlässig sind, dann ist auch das aus ihnen zusammengesetzte diagnostische Instrument zuverlässig. Die Formel zur Berechnung von Alpha lautet:

$$\text{Alpha} = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^p \text{Var}(Y_i)}{\text{Var}(S)} \right) \quad (\text{Formel 3.1})$$

Die Symbole in Formel 3.1 haben folgende Bedeutung:

p Anzahl der Items, aus denen der Test besteht

$\text{Var}(Y_i)$ Varianz der Items i des Tests

$\text{Var}(S)$ Varianz der Summe der Testitems

Cronbachs Alpha ist einer der in der Psychologie am häufigsten berichteten Koeffizienten. Allerdings wird oftmals versäumt zu überprüfen, ob überhaupt die Voraussetzungen erfüllt sind, damit Alpha die Reliabilität eines diagnostischen Verfahrens unverfälscht schätzt (Eid et al., 2013). Die wichtigste Voraussetzung haben wir bereits genannt: Alle Items müssen das gleiche Merkmal messen. Was dies genau bedeutet, werden wir in Kapitel 4 erfahren.

Reliabilitätssteigerung durch Testverlängerung

Bei der Vorstellung der drei Testzerlegungsmethoden (Testhalbierungsmethode, Parcelmethode, interne Konsistenz) hatten wir davon gesprochen, dass sich die Korrelation zwischen den Testteilen zur Schätzung der Reliabilität heranziehen lässt, sofern alle Testteile das gleiche Merkmal messen. Nun ist es aber so, dass die Korrelation zwischen den Testteilen nicht die Reliabilität des Tests selbst schätzt, sondern die Reliabilität der Testteile. Nur bei der Paralleltestmethode und bei der Retestmethode schätzt die Korrelation die Reliabilität des gesamten Tests, sofern die Voraussetzungen (die Tests messen das gleiche Merkmal, die gemessenen Objekte haben sich nicht verändert) erfüllt sind. Bei allen Testzerlegungsmethoden wird zunächst die Reliabilität der Testteile aus den Korrelationen dieser Teile geschätzt. Dann muss aus diesen Schätzungen die Reliabilität des Gesamttests bestimmt werden.

Die Reliabilität des Gesamttests ist höher als die Reliabilität seiner einzelnen Teile. Dieses Prinzip ist uns aus vielen alltäglichen Zusammenhängen bekannt. Wir nutzen es bspw. bei der Leistungsbewertung in der Schule oder im Sport. In der Schule ist es üblich, die Leistung von Schülern nicht mittels einer einzigen Prüfung zu bestimmen, sondern auf der Grundlage mehrerer Prüfungen, deren Ergebnis gemittelt wird. Hinter diesem Vorgehen steht die begründete Überlegung, dass jede einzelne Leistungsmessung fehleranfälliger ist als der Durchschnitt aus mehreren Leistungsmessungen. Auch im Sport verlässt man sich nicht auf einzelne Leistungsmessungen, wenn diese nicht fehlerfrei möglich sind. Die Zeit, die jemand braucht, um 100 m zu laufen, kann mit modernen Messapparaturen fast fehlerfrei gemessen werden. Ebenso kann man annähernd

fehlerfrei messen, wie weit jemand eine Kugel, einen Speer, einen Diskus oder einen Hammer wirft. Anders ist die Situation bei der Vergabe der B-Note im Eiskunstlauf und beim Geräteturnen. Mit dieser Note wird die künstlerische Qualität der Leistung bewertet. Diese lässt sich nicht objektiv messen. Vielmehr wird der künstlerische Ausdruck durch die subjektive Beurteilung von Preisrichtern ermittelt. Jedes Einzelurteil ist dabei aus den unterschiedlichsten Gründen fehlerbehaftet. Beispielsweise ist es denkbar, dass ein Preisrichter bei der Beobachtung einer Übung einen Moment unaufmerksam war und ihm ein wichtiges Detail entgangen ist. Um solche Fehler zu minimieren, werden die Einschätzungen mehrerer Experten zusammengefasst. Diese durchschnittliche Einschätzung ist weniger fehlerbehaftet, allerdings nur dann, wenn die einzelnen Einschätzungen und die dabei gemachten Fehler unabhängig voneinander sind. Es darf beispielsweise nicht so sein, dass Preisrichter 1 absichtlich eine zu gute Note gibt und Preisrichter 2 dazu überredet, ebenfalls eine zu gute Note zu geben. In diesem Fall würde die Zuverlässigkeitssteigerung durch Testverlängerung versagen.

Spearman-Brown-Formel

Unter der Voraussetzung, dass die einzelnen Testteile (Hälften, Parcels, Items) das gleiche Merkmal messen, kann man mittels der sogenannten Spearman-Brown-Formel der Testverlängerung bestimmen, wie die Reliabilität eines Tests mit seiner Verlängerung zunimmt. Was unter einem gleichen Merkmal zu verstehen ist, werden wir in Kapitel 4 näher erläutern.

$$Rel(Y2) = \frac{p \cdot Rel(Y1)}{1 + (p - 1) \cdot Rel(Y1)} \quad (\text{Formel 3.2})$$

Die Symbole in Formel 3.2 haben folgende Bedeutung:

$Rel(Y2)$ Reliabilität des um Faktor p verlängerten Tests

$Rel(Y1)$ Reliabilität des Tests, der um Faktor p verlängert wurde

p Verlängerungsfaktor

Betrachten wir zunächst den Fall der Reliabilitätsschätzung aus Testhälften. Sofern diese das gleiche Merkmal messen und zu $r = .80$ miteinander korrelieren, beträgt ihre Reliabilität $Rel(Y1) = .80$. Wenn wir diesen Wert in Formel 3.2 einsetzen, ergibt sich als Schätzwert für die Reliabilität des Gesamttests $Rel(Y2) = .89$. Diesen Wert erhält man, da der Verlängerungsfaktor in diesem Fall $p = 2$ beträgt. Betrachten wir als nächstes den Fall der Reliabilitätsschätzung aus Parcels. Die Formel lässt sich in diesem Fall nur anwenden, wenn die Parcels identische Reliabilitäten haben. Nehmen wir an, wir hätten aus den

21 Items eines Fragebogens drei Parcels gebildet und alle Korrelationen zwischen zwei verschiedenen Parcels würden $r = .50$ betragen. Somit schätzen wir für die Reliabilitäten aller drei Parcels einen Wert von $Rel(Y1) = .50$. Da bei drei Parcels der Verlängerungsfaktor $p = 3$ beträgt, errechnet sich als geschätzte Reliabilität für den Gesamttest, also die Summe der drei Parcels, ein Wert von $Rel(Y2) = .75$. Den gleichen Wert würden wir erhalten, wenn die 21 Items untereinander alle zu $r = .125$ korrelieren würden.

Offensichtlich kann die Spearman-Brown-Formel auch verwendet werden, um zu ermitteln, wie sehr ein Test an Reliabilität verliert, wenn wir ihn kürzen. In diesem Fall ist $p < 1$. Aus den Beispielen, die wir gegeben haben, kann man sich die Effekte von Kürzungen leicht klar machen. Beispielsweise sinkt die Reliabilität eines Test von $Rel(Y2) = .89$ auf $Rel(Y1) = .80$, wenn wir ihn halbieren usw. Schließlich kann die Spearman-Brown-Formel auch verwendet werden, um zu ermitteln, um welchen Faktor ein Test mit einer unzureichenden Reliabilität verlängert werden muss, damit eine geforderte Mindestreliabilität erreicht wird. Wir müssen für diese Berechnung lediglich den (ungenügenden) Reliabilitätswert $Rel(Y1)$ für den kurzen Test und die gewünschte Reliabilität $Rel(Y2)$ für die gewünschte Reliabilität eintragen und Formel 3.2 nach p auflösen.

Weitere Möglichkeiten der Reliabilitätssteigerung

Außer der Testverlängerung gibt es weitere Möglichkeiten, die Zuverlässigkeit eines diagnostischen Verfahrens zu maximieren. Dazu gehört insbesondere eine möglichst umsichtige Konstruktion des Verfahrens. Beispielsweise trägt es zur Zuverlässigkeit eines Fragebogens bei, wenn

Fragen und Aussagen so einfach wie möglich formuliert werden. Schon aufgrund theoretischer Überlegungen zum Textverständnis ist zu befürchten, dass sich kompliziert formulierte Fragen oder lange Aussagen negativ auf die Zuverlässigkeit eines Fragebogens auswirken, weil sie höchste Konzentration erfordern und sich schon die kleinste Missachtung oder Fehlinterpretation von Teilen der Aussage auf die Antwort auswirken können. Da die Konzentration von Menschen Schwankungen unterliegt, muss bei kompliziert formulierten Items damit gerechnet werden, dass die Antworten der Diagnostikanden nicht über wiederholte Messungen stabil bleiben. Vergleichen wir zur Veranschaulichung die beiden folgenden Aussagen, die man zur Messung von Neurotizismus verwenden könnte.

- (1) Ich bin leicht beunruhigt.
- (2) Ich denke, dass ich wohl eher nicht zu den Menschen gehöre, an denen unerfreuliche, störende, ärgerliche und empörende Vorkommnisse spurlos abprallen, sondern eher zu den Menschen, die, wenn etwas viel zusammenkommt, zwar nicht immer, aber doch oft und öfter als die meisten anderen Menschen Angst bekommen, dass sie den Überblick verlieren oder überfordert sein oder in anderer Weise unter den unerfreulichen Dingen, mit denen sie konfrontiert sind, leiden könnten.

Das zweite Item ist frei erfunden und bewusst übertrieben ungünstig formuliert, um die Problematik zu verdeutlichen. Wenngleich Fragebögen derart ungünstige Items nur selten enthalten, stößt man selbst in etablierten Verfahren immer wieder auf viel zu komplexe und komplizierte Items, die sich ungünstig auf die Reliabilität auswirken und auch auf das Gütekriterium, das wir anschließend vorstellen, die Validität.

3.3 Validität

Wenn ein diagnostisches Verfahren zuverlässig ist, kann daraus nicht geschlossen werden, dass es auch das misst, was es messen soll. Ein Test, der entwickelt wurde, um Kreativität zu erfassen, mag sehr zuverlässig sein, aber vielleicht nicht nur Kreativität messen, sondern auch Intelligenz, Motivation und Ausdauer. Ein Mathematiktest, der aus Textaufgaben besteht, mag sehr zuverlässig sein, aber nicht nur die mathematische Leistung messen, sondern auch das Sprachverständnis. Textaufgaben kann nur derjenige lösen, der den Text versteht. Wer den Text einer Aufgabe falsch versteht, kommt zwar möglicherweise immer wieder zum gleichen Ergebnis (Reliabilität), aber das Ergebnis ist falsch, weil der Text falsch verstanden wurde.

Definition

Validität

Valide oder gültig ist ein diagnostisches Instrument in dem Maße, in dem es misst, was es messen soll. Validität ist also ebenso wie Reliabilität eine Variable. Ein Messinstrument kann mehr oder weniger valide sein. Es ist maximal valide, wenn die mit ihm gewonnenen Messwerte mit den wahren Merkmalswerten exakt übereinstimmen. Da zur Schätzung der Validität häufig die Produktmomentkorrelation verwendet wird, die Werte zwischen -1 und $+1$ annehmen kann, wird auch die Validität meistens auf dieser Skala angegeben. Eine negative Korrelation bedeutet, dass das Messinstrument (mehr oder weniger) das

Gegenteil dessen misst, was es messen soll. Beispielsweise wäre es prinzipiell denkbar, dass ein Fragebogen, der zur Messung von Hilfsbereitschaft entwickelt wurde, in Wahrheit Egoismus misst. Solche Fälle negativer Validität sind selten, aber nicht unmöglich.

Merke: Ein diagnostisches Instrument hat eine einzige Validität. Es gibt jedoch mehrere Methoden, die Validität eines diagnostischen Instruments zu schätzen. Diese Schätzmethoden basieren auf bestimmten Voraussetzungen. Sind diese nicht erfüllt, ist die Validitätsschätzung fehlerhaft.

In der Psychologie gibt es bis auf wenige Ausnahmen keine diagnostischen Instrumente, die vollkommen valide sind. Menschliches Erleben und Verhalten ist multideterminiert und hat fast nie nur eine Ursache, sondern meistens mehrere. Wie gut beispielsweise eine Person bei einem Intelligenztest abschneidet, hängt nicht nur von ihrer Intelligenz ab, sondern auch von ihrer Motivation, ihrer momentanen Konzentration, vom Selbstvertrauen und anderen Faktoren. Die Psychologie müsste scheitern, wenn sie nach vollständig validen diagnostischen Verfahren streben würde. Sie kann nur so gut es geht versuchen, den Einfluss diagnostisch irrelevanter Einflüsse, die die Validität eines Messergebnisses reduzieren, zu minimieren. Bei Intelligenztests ist dies relativ gelungen. Die oben genannten diagnostisch irrelevanten Einflüsse sind im Vergleich zum Einfluss der Intelligenz auf das Messergebnis gering. Bei vielen anderen Verfahren ist die Situation weniger günstig. Es ist oftmals auch nicht einfach, die Validität eines Instrumentes gut zu schätzen. Diese Einschränkung gilt selbst bei Verfahren, deren Validität auf den ersten Blick unstrittig zu sein scheint. Ein gutes Beispiel hierfür sind Reaktionszeitmaße. Wir können die Zeit, die eine Person benötigt, um auf einen Reiz (z. B. ein auf dem Bildschirm erscheinendes Bild) zu reagieren (z. B. mit dem Druck einer Taste der Tastatur), zwar sehr exakt bestimmen. Aber welche Faktoren die Reaktionszeit genau bedingen, wissen wir in den meisten Fällen nicht so genau. Vielmehr müssen wir auch bei so präzisen Maßen wie der Reaktionszeit theoretische Überlegungen zu den Ursachen des Verhaltens anstellen (s. Kap. 2). Diese theoretischen Überlegungen begründen Annahmen, unter denen wir die Validität schätzen. Betrachten wir nun die wichtigsten Methoden der Validitätsschätzung. Die Bezeichnungen dieser Methoden enthalten den Begriff »Validität«. Streng genommen müsste man statt von »Validität« von »Validierung« oder von »Validitätsschätzung« sprechen, da jedes Verfahren nur eine Validität hat. Da sich diese treffenderen Begriffe in der Literatur jedoch nicht durchgesetzt haben, verwenden wir hier die gebräuchlichen Begriffe, obwohl sie das Prinzip der Validitätsschätzung nicht angemessen wiedergeben.

Inhaltliche, logische oder Augenscheinvalidität

Die einfachste Art, die Validität eines diagnostischen Instruments abzuschätzen, wird inhaltliche, logische oder Augenscheinvalidität genannt. Diese Bezeichnungen sollen zum Ausdruck bringen, dass dem Instrument sozusagen anzusehen ist, was es misst. Im Sport ist dies relativ einfach möglich. Wenn wir eine Gruppe von Personen 1000 m laufen lassen, ihnen sagen, dass sie so schnell wie möglich laufen sollen und für den Sieger einen attraktiven Preis ausloben, dann können wir uns ziemlich sicher sein, was unser Test misst. In der Psychologie haben wir selten mit derart einfachen Aufgaben zu tun. Aber auch in der Psychologie gibt es diagnostische Verfahren, deren Validität kaum strittig sein dürfte. Dies trifft beispielsweise auf Gedächtnistests zu. Wenn wir eine Person 50 sinnlose Silben lernen lassen (z. B. fim, pez, ofa) und sie eine Woche später bitten, alle gelernten Silben aufzuschreiben, können wir uns relativ sicher sein, die Gedächtnisleistung der Person gemessen zu haben. Gleiches gilt für Verfahren zum Messen von Wahrnehmungsschwellen. Wenn wir einer Person wiederholt zwei Töne vorgeben, die sich nur in der Tonhöhe (Frequenz) unterscheiden und dann die Tonhöhendifferenz von 0 Hertz kleinschrittig so lange erhöhen, bis die Person einen Unterschied hört, haben wir vermutlich in den meisten Fällen das gemessen, was wir messen wollen: Die Unterschiedsschwelle für Tonfrequenzen. Diese relativ klaren Beispiele sind in der Psychologie jedoch Ausnahmen. Den meisten diagnostischen Instrumenten ist nicht auf Anhieb anzusehen, was sie messen.

In diesen Fällen greift man auf das Urteil von Experten zurück. Dabei kann es sich um erfahrene Wissenschaftler handeln, aber auch um erfahrene Psychologen aus der Anwendungspraxis oder sogar fortgeschrittene Studierende, die sich mit dem zu messenden Merkmal im Rahmen ihres Studiums beschäftigt haben. Stellen wir uns beispielsweise vor, eine Psychologiestudentin habe

sich dazu entschieden, im Rahmen ihrer Masterarbeit einen Fragebogen zur Messung des Leistungsmotivs, des Anschlussmotivs und des Machtmotivs zu entwickeln. Sie denkt sich u. a. die folgenden Items aus:

- (1) Ich will immer die/der Beste sein.
- (2) Ich gebe gerne den Ton an.
- (3) Ich bin nicht gerne alleine.
- (4) Ich halte gerne Vorträge.

Sie könnte nun 20 Kommilitonen bitten, für jedes Item einzuschätzen, wie gut man von einer Bejahung des Items auf die Motive zurückschließen kann. Vermutlich werden die meisten Kommilitonen sagen, dass man von Item 1 sehr gut auf das Leistungsmotiv schließen kann, von Item 2 sehr gut auf das Machtmotiv und von Item 3 sehr gut auf das Anschlussmotiv. Bei Item 4 dürften die Urteile weniger eindeutig ausfallen. Ein Vortrag ist eine Leistung, er kann eine Machtdemonstration sein (ich bestimme!) und schließlich das Bemühen einer Person wider spiegeln, Kontakt zu anderen Menschen (aus dem Auditorium) zu bekommen.

Wir möchten darauf hinweisen, dass in anderen Lehrbüchern (vgl. Asendorpf, 2011; Moosbrugger & Kelava, 2008) zwischen Inhalts- und Augenscheinvalidität unterschieden wird. Wir verzichten hier auf diese Unterscheidungen der Validierungsansätze, da sie für das Verständnis der Logik des Vorgehens nicht erheblich sind.

Systematische und unsystematische Messfehler

Item 4 dürfte somit von den Experten die geringste Validität attestiert werden. Es misst nicht nur eines der drei Motive, sondern wahrscheinlich alle drei Motive gleichzeitig. In der Sprache der Faktorenanalyse werden solche multideterminierten Items als multifaktoriell oder faktoriell komplex bezeichnet. Valide Items, die nur ein Merkmal (einen Faktor) messen, werden hingegen als faktoriell einfach oder faktoriell rein bezeichnet. Wollte man mit Item 4 das Anschlussmotiv messen, würde man sich höchstwahrscheinlich Messfehler einhandeln. Diese Messfehler reduzieren die Validität des

Items. Es ist zu beachten, dass der Messfehler systematischer Natur wäre, weil das Item nicht nur das Anschlussmotiv misst, sondern auch die beiden anderen Motive. Die Zuverlässigkeit des Items ist dadurch nicht beeinträchtigt.

Merke: Systematische Messfehler beeinträchtigen die Validität eines diagnostischen Instrumentes, nicht aber dessen Reliabilität. Unsystematische Messfehler beeinträchtigen sowohl die Reliabilität als auch die Validität eines diagnostischen Instrumentes.

Kriteriumsbezogene Validität oder Kriteriumsvalidität

Diagnostische Instrumente werden in der Praxis häufig dazu verwendet, ein relevantes Ergebnis oder Ereignis vorherzusagen. Relevante Ereignisse sind z. B. die Rückfälligkeit eines vorzeitig aus dem Gefängnis entlassenen Straftäters, eine Eheschließung oder der Eintritt einer Krankheit. Relevante Ergebnisse sind z. B. der Erfolg in Schule, Studium und Beruf. Solche Ereignisse und Ergebnisse bezeichnet man als Kriterien. Wenn ein diagnostisches Instrument entwickelt wird, um ein solches Kriterium vorherzusagen, wird seine Validität häufig über die Güte, mit der es das Kriterium vorhersagt, geschätzt. Das entwickelte Instrument fungiert dabei als der Prädiktor. Seine Validität schätzt man über die Korrelation zwischen Prädiktor und Kriterium.

Inkrementelle Validität

Ist bekannt oder anzunehmen, dass das Kriterium auch durch andere Prädiktoren vorhergesagt werden kann, schätzt man die sogenannte inkrementelle (zusätzliche) Validität des zu validierenden Instruments, indem man auch die anderen Prädiktoren mit geeigneten Instrumenten misst und mittels einer multiplen Regressionsanalyse (s. Eid et al., 2013) prüft, welcher Vorhersagegewinn sich mit

dem zu validierenden Instrument über die anderen Prädiktoren hinaus erzielen lässt. Üblicherweise wird die inkrementelle Validität als derjenige Anteil der Varianz des Kriteriums beziffert, der nur durch das zu validierende Instrument vorhersagbar ist. Ein Beispiel hierfür hatten wir in Kapitel 2 kennengelernt (s. Kasten »Studierfähigkeit im Fach Psychologie«).

Prädiktive, konkurrenente und postdiktive Validität

Die Unterscheidung zwischen prädiktiver, konkurrenenter und postdiktiver Validität bezieht sich nicht auf die Schätzmethode oder die Art des Validitätskriteriums, das zur Validierung herangezogen wird, sondern auf die zeitliche Abfolge der Erhebung der Daten. Bei der prädiktiven (vorhersagenden) Validierung wird das Kriterium in einem zeitlichen Abstand nach der Anwendung des zu validierenden Instruments erhoben. Die Validierung von Studierfähigkeitstests ist hierfür ein gutes Beispiel. Zunächst wird die Studierfähigkeit gemessen, einige Jahre später der Studienerfolg, z. B. in Form der Bachelor-Abschlussnote.

Anders als bei der prädiktiven Validierung wird bei der konkurrenten Validierung das Validierungskriterium nicht mit zeitlicher Verzögerung, sondern gleichzeitig gemessen. Wenn man z. B. einen neu entwickelten Fragebogen zur Messung der drei oben genannten Motive validieren möchte, könnte man zur Messung relevanter Kriterien unmittelbar anschließend einen Wettbewerb inszenieren (Leistungsmotiv), eine Gruppendiskussion anzetteln (Machtmotiv) und die Testpersonen in eine Situation bringen, in der sie Kontakt zu anderen Testpersonen aufnehmen oder vermeiden können (Anschlussmotiv).

Bei der postdiktiven (rückblickenden) Validierung ist das Validierungskriterium bekannt, bevor das zu validierende Instrument angewandt wird. Bei der Validierung von klinisch-psychologischen Diagnoseinstrumenten greift man häufig auf diese Strategie zurück. Bei der Validierung eines der bekanntesten klinisch-psychologischen Persönlichkeitsinventare, des Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943; deutsch: Spreen, 1963), wurde beispielsweise so vorgegangen. Dieses Inventar wurde entwickelt, um klinisch-psychologische Störungen zu diagnostizieren. Bei der Validierung an einer großen Zahl von Patienten war deren Diagnose bereits bekannt. Man hat somit die diagnostizierte Störung mit dem zu validierenden Inventar nicht vorhergesagt, sondern rekonstruiert.

Konvergente und diskriminante/divergente Validität

Diesen beiden Begriffen begegnet man in der diagnostischen Literatur besonders häufig. Von konvergenter Validität spricht man, wenn ein zu validierendes Instrument mit anderen Instrumenten, deren Validität bereits bekannt ist und die das gleiche oder ein ähnliches Merkmal messen, zusammenhängt (korreliert, konvergiert). Von diskriminanter oder divergenter Validität spricht man, wenn ein zu validierendes Instrument mit anderen Instrumenten, deren Validität bereits bekannt ist und die ein anderes Merkmal messen, nicht zusammenhängt (divergiert).

Probleme bei der konvergenten und diskriminanten Validierung

Obwohl mit der konvergenten und diskriminanten Validierung sehr häufig gearbeitet wird, gibt es wohl bei keiner anderen Validierungsmethode so große Interpretationsspielräume wie bei diesen beiden Formen der Validierung. Man könnte statt von Interpretationsspielraum kritischer auch von Beliebigkeit sprechen. Das Problem rührt daher, dass »ähnliches Merkmal« und »anderes Merkmal« unscharfe Begriffe sind. Wie unähnlich dürfen Merkmale sein, dass sie (gerade) noch als ähnlich gelten können? Wie unähnlich müssen Merkmale sein, dass sie (gerade schon) als andere Merkmale gelten dürfen? Auf diese Fragen gibt es keine verbindlichen Antworten. Nehmen wir das bereits vertraute Beispiel von Studierfähigkeitstests. Zur Validierung dieser Tests kommen der Studienerfolg und der Schulerfolg in Betracht. Kaum jemand würde wohl anzweifeln, dass sich der Studienerfolg zur konvergenten Validierung von Studierfähigkeitstests eignet. Schließlich ist es genau

dieses Kriterium, das der Test vorhersagen soll. Aber wie ist es um den Schulerfolg bestellt? Kann man den Schulerfolg ebenfalls als Kriterium der konvergenten Validität heranziehen? Oder könnte man argumentieren, dass von einem Studierfähigkeitstest v. a. die Prognose des Studienerfolgs erwartet werden kann, nicht aber oder zumindest weniger gut die Prognose des Schulerfolgs? Dieser Argumentation zufolge könnte der Schulerfolg als Kriterium für diskriminante Validität herangezogen werden. Was aber, wenn sich herausstellen sollte, dass ein Studierfähigkeitstest mit dem Schulerfolg enger zusammenhängt als mit dem Studienerfolg (Formazin et al., 2011)? Wäre daraus dann zu schließen, dass der Studierfähigkeitstest nicht valide ist? Das Beispiel zeigt, dass man konvergente und diskriminante Validierung nur praktizieren kann, wenn sehr präzise Vorstellungen der Ähnlichkeit und der Unähnlichkeit von Merkmalen bestehen.

Konstruktvalidität

Diese Präzisierung von Ähnlichkeit und Unähnlichkeit leisten psychologische Theorien. Ohne eine Theorie des Schulerfolgs und des Studienerfolgs lassen sich keine Annahmen über den Grad der Ähnlichkeit beider Merkmale und ihrer Eignung zur konvergenten und diskriminanten Validierung von Studierfähigkeitstests treffen. Konstruktvalidierung nimmt diese Forderung nach Theorie ernst.

Definition

Konstruktvalidität

Ein Instrument zur Messung eines Merkmals (Konstrukts) ist konstruktvalide in dem Maße, in dem es Daten liefert, die von einer bewährten Theorie vorher-

gesagt werden, in die das Merkmal (Konstrukt) eingebunden ist.

Konstruktvalidität ist die überzeugendste, aber auch die anspruchsvollste Form der Validierung. Überzeugend ist sie, weil Erwartungen über (fehlende) Zusammenhänge zwischen dem zu validierenden Instrument und Validitätskriterien nicht auf der Basis oberflächlicher und augenscheinlicher Ähnlichkeiten und Unähnlichkeiten vorgenommen werden, sondern auf der Basis einer Theorie, die Zusammenhänge zwischen Merkmalen mit Prozessen erklärt, die sich dem Augenschein oft verschließen. Nehmen wir die Persönlichkeitstheorie von Eysenck (1953) als Beispiel. Diese Theorie besagt, dass Extraversion mit dem Erregungszustand (Arousal) des zentralen Nervensystems zusammenhängt. Bei extravertierten Personen liegt der Erregungszustand häufig unter dem Optimum, dem sogenannten hedonischen Tonus (vgl. Schmitt & Altstötter-Gleich, 2010). Stark extravertierte Personen sind also chronisch kortikal untererregt. Bei stark introvertierten Personen ist es genau umgekehrt. Ihr kortikaler Erregungszustand liegt häufiger über dem Optimum. Das Niveau der kortikalen Erregung kann der Theorie Eysencks

zufolge durch sensorische und kognitive Reize erhöht werden. Die Theorie sagt deshalb vorher, dass Extravertierte gerne in Gesellschaft sind, gerne Musik hören und Einsamkeit und Stille eher nicht mögen. Aus diesen theoretischen Annahmen folgt, dass die Vorliebe für das Hören von Musik herangezogen werden kann, um einen Extraversionsfragebogen zu validieren. Es wird hier also ein Zusammenhang zwischen einem zu validierenden Instrument und einem Validierungskriterium vorhergesagt, auf den man ohne Kenntnis der Theorie Eysencks niemals kommen würde. Gerade dadurch bezieht Konstruktvalidierung ihre Überzeugungskraft. Die Validierungshypothesen sind theoretisch gehaltvoll und können, anders als bei der Augenscheinvalidierung, nicht als trivial oder zirkulär abgetan werden.

Genau deswegen ist Konstruktvalidierung nicht nur die überzeugendste Form der Validierung von diagnostischen Instrumenten, sondern auch die schwierigste. Nur bewährte Theorien eignen sich für die Konstruktvalidierung. Erst wenn weitgehend Sicherheit darüber besteht, dass eine Theorie gilt, kann man sie heranziehen, um die Validität von Instrumenten zu beurteilen. Wenn die Gültigkeit der Theorie noch unklar ist, stellt sich folgendes Problem: Tritt das aus der Theorie vorhergesagte Ergebnis nicht ein, weiß man nicht, ob die Theorie falsch, das diagnostische Instrument nicht valide oder ob beides der Fall ist. Trotz dieser Schwierigkeit plädieren wir nachdrücklich dafür, Validierung grundsätzlich auf der Grundlage einer sorgfältigen theoretischen Begründung von Zusammenhangserwartungen vorzunehmen.

Strategien der Validitätssteigerung

Nichts ist bei der Konstruktion eines diagnostischen Instruments so wichtig wie eine möglichst bewährte oder doch zumindest plausible Theorie über die psychologischen Ursachen des Verhaltens, das Diagnostikanden gegenüber dem Verfahren zeigen. Deshalb wirkt es sich auf die Validität eines Instruments positiv aus, wenn bereits bei der Entwicklung von Items sorgfältige Überlegungen angestellt werden, welche systematischen Störeinflüsse theoretisch zu befürchten sind. Bei der Formulierung von Fragebogenitems ist es wichtig darauf zu achten, dass das beschriebene Verhalten nicht gegen allgemein akzeptierte Normen und Werte verstößt, auch wenn solche Verstöße relativ häufig vorkommen. Es gilt heutzutage als diskriminierend zu sagen, dass Frauen in die Küche gehören und Männern die Rolle des Familienoberhaupts zusteht. Vor 50 Jahren konnte man solche Einstellungen noch unverblümt äußern, ohne Missbilligung befürchten zu müssen. Das ist heute anders. Trotzdem gibt es auch heute noch viele Menschen, die gerne an der traditionellen Rollenverteilung zwischen Mann und Frau festhalten möchten. Es gibt nach wie vor auch Menschen, die Frauen für minderwertig halten. Will man heute diskriminierende Einstellung messen, müssen subtilere Aussagen verwendet werden (Swim & Cohen, 1997), wie: »Frauen werden heutzutage nicht mehr wirklich benachteiligt.« Diese Aussage ist objektiv falsch. Benachteiligungen von Frauen wie z. B. ihre schlechtere Bezahlung im Vergleich zu Männern sind allgemein bekannt. Wer der Aussage dennoch zustimmt, bestreitet, dass es ein Problem gibt und dies lässt auf eine zugrunde liegende Einstellung zu Frauen und Geschlechterrollen schließen. Viele Personen, die dieser subtilen Aussage zustimmen, würden sich jedoch nicht trauen, die Aussage zu bejahen, dass Frauen an den Herd gehören und gegenüber ihren Ehemännern gehorsampflichtig sind. Das Beispiel zeigt, wie wichtig es ist, theoretische Überlegungen darüber anzustellen, welche Faktoren das Verhalten einer Person in einer diagnostischen Situation beeinflussen und welche dieser Einflüsse die Validität des Verfahrens reduzieren. Wir werden dieses wichtige Thema in Kapitel 5 vertiefen.

Auch auf die Maximierung der Nebengütekriterien, denen wir uns nun zuwenden werden, kann bereits bei der Entwicklung eines diagnostischen Instruments geachtet werden. Wir werden dies am Beispiel des Nebengütekriteriums der Akzeptanz später verdeutlichen.