Der Berufsverband für Trainer, Berater und Coaches

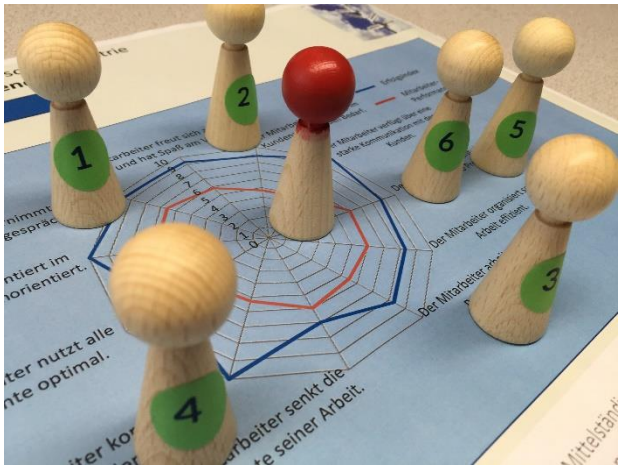
News & Facts

Qualitätskriterium der Messbarkeit

Bruno Schmalen, Vizepräsident BDVT

Erfolgskontrolle beobachtet misst und dokumentiert Personal- und Organisationsentwicklungen. In Training, Berstung und Coaching weist Erfolgskontrolle Wirkung.
Um dies zu erreichen, muss sich die Meythode Er:Kon mit der Qualität der Messkriterien beschäftigen. Dabei spielen in der Literatur drei Qualitätskriterien nach, an denen die Redlichkeit des Messens deutlich wird: die Objektivität, die Reabilität und die Validität. Alle drei Kriterien werden hier vorgestellt.

Objektivität



Die Objektivität als eins der drei Qualitätskriterien von Tests. Es gilt dabei die Grundregel:
Eine Aussage (v.a. Messaussagen) ist dann objektiv, wenn sie vom Untersuchungsleiter unabhängig ist. Man unterscheidet die Durchführungsobjektivität (keine Beeinflussung der Untersuchungsergebnisse durch das äußere Erscheinungsbild, das Ziel- und Wertsystem des Durchführenden bzw. Interviewers), die Auswertungsobjektivität (v.a. gegeben bei standardisierten Frage-Items) und die Interpretationsobjektivität (wenig Spielraum für die subjektive Interpretation durch den Untersuchungsleiter).

Objektivität kann in drei Bereiche unterteilt werden:

Durchführungsobjektivität
Hiermit ist das Ausmaß gemeint, in dem die Testergebnisse unabhängig von der Person des Versuchsleiters und den räumlichen Bedingungen sind. Daraus folgt, dass eine maximale Standardisierung der Testsituation und eine minimale soziale Interaktion zwischen Versuchsleiter und Testteilnehmer angestrebt werden sollten.

Auswertungsobjektivität
Gemeint ist das Ausmaß, in dem gleiches Verhalten einer Testperson stets auf die gleiche Weise ausgewertet wird. Hier ist die Objektivität bei projektiven Tests eher als gering, bei standardisierten Intelligenztests, wie dem WAIS-IV (Wechsler Adult Intelligence Scale IV), hingegen durch die standardisierte Auswertung als hoch bzw. ausreichend zu betrachten. Untersuchungen legen nahe, dass die auch Auswertung von Papier-und-Bleistift-Tests mit den üblichen Schablonen sehr fehleranfällig sein kann und fast selbst schon ein Konzentrationstest ist. Je weniger Freiheitsgrade der Auswerter bei der Auszählung der Testergebnisse hat, desto höher ist die Auswertungsobjektivität.

Interpretationsobjektivität
Hier ist das Ausmaß gemeint, in dem gleiche Testwerte auf die gleiche Weise interpretiert werden: Existieren klare Regeln, inwiefern aus einem Ergebnis eine konkrete diagnostische Entscheidung abgeleitet werden kann? Je weniger definiert die Schlussfolgerungen aus dem Ergebnis sind, umso mehr hängt die Interpretation von den subjektiven Fähigkeiten und Erfahrungen der Fachperson ab. Bei der

Interpretationsobjektivität geht es praktisch um die Frage, ob ein Test normiert ist oder nicht. Werden unterschiedliche Normen verwendet, entstehen verschiedene Interpretationen. Normentabellen für unterschiedliche Populationen (z.B. Schülernormen, Studentennormen usw.) erhöhen die Interpretationsobjektivität. In der DIN 33430 (!) für Eignungsdiagnostik werden klare und wissenschaftlich überprüfte Regeln für die Interpretation gefordert und es werden auch Aus- und Weiterbildungsstandards für die Fachpersonen definiert.
Als Beispiel soll die Schulnote 1 in allen Bundesländern eine sehr gute Leistung anzeigen und die Note 5 soll als mangelhaft und nicht ausreichend zum Bestehen betrachtet werden. Interpretationsobjektivität sagt aber nichts über die inhaltliche Güte aus. Wenn Durchführungs- oder Auswertungsobjektivität verletzt sind, verfälscht sich dadurch die Aussage der Interpretation. Somit wird gesagt, wer eine 2 in Bayern hat, hat vielleicht eine 1 in NRW (Problem der mangelnden Standardisierung im Sinne der Durchführungsobjektivität; jeder macht seinen eigenen Test). Interpretationsobjektivität ist bei Normskalen immer gegeben, da eine Standardabweichung "+/-1" immer 68,3 % der Testwerte umfasst.

Reliabilität

Die **Reliabilität** (dt.: Zuverlässigkeit) ist ein Maß für die formale Genauigkeit bzw. Verlässlichkeit wissenschaftlicher Messungen. Sie ist derjenige Anteil an der Varianz, der durch tatsächliche Unterschiede im zu messenden Merkmal und nicht durch Messfehler erklärt werden kann. Hochreliable Ergebnisse müssen weitgehend frei von Zufallsfehlern sein, d.h. bei Wiederholung der Messung unter gleichen Rahmenbedingungen würde das gleiche Messergebnis erzielt werden (Replizierbarkeit von Ergebnissen unter gleichen Bedingungen).

Die Reliabilität stellt neben der Validität und der Objektivität eines der drei wichtigsten Qualitätskriterien für empirische Untersuchungen dar.

Reliabilität umfasst drei Aspekte:
- Stabilität (Gleichheit bzw. Ähnlichkeit der Messergebnisse bei Anwendung zu unterschiedlichen Zeitpunkten)
- Konsistenz (Ausmaß, nach dem alle Items, die in einem Test zu einem Merkmal zusammengefasst werden, dasselbe Merkmal messen)
- Äquivalenz (Gleichwertigkeit von Messungen)

Die Reliabilität kann mit verschiedenen Methoden geschätzt werden. Je nach Methode wird von anderen Reliabilitäts-Typen gesprochen.

Paralleltest-Reliabilität

Denselben Versuchspersonen werden zwei einander stark ähnelnde Tests (entweder unmittelbar hintereinander oder zeitlich versetzt) dargeboten. Die Paralleltest-Reliabilität wird im Paralleltest-Verfahren bestimmt. Sie gibt an, ob ein vergleichbares Messverfahren identische Ergebnisse liefert. Anstelle gleichwertiger Testverfahren können auch *Parallelformen* des Tests verwendet werden (zum Beispiel dürften die Aufgaben 3 + 4 = ? und 5 + 2 = ? gleichermaßen dazu geeignet sein, die Fähigkeit zur einfachen Addition zu messen).

Retest-Reliabilität

Der gleiche Test wird den Versuchspersonen zu verschiedenen Zeitpunkten dargeboten. Die Ergebnisse der ersten und zweiten Messung werden korreliert. Beim Test-Retest-Verfahren wird geprüft, ob eine Wiederholung der Messung bei Konstanz der zu messenden Eigenschaft die gleichen Messwerte liefert. Die Retest-Reliabilität gibt den Grad der Übereinstimmung an. Für viele Tests ist eine Wiederholung entsprechend dem Test-Retest-Verfahren nur theoretisch möglich, da die mit dem Test einhergehenden Erinnerungs-, Lern- oder Übungseffekte das Ergebnis beeinflussen und eine „Scheinreliabilität“ vortäuschen können. So ist eine mathematische Aufgabe in einem Intelligenztest nicht zweimal zu lösen, da der Proband sich an die Lösung der ersten Aufgabe erinnert. Das Zeitintervall zwischen den Messungen muss also groß genug sein, um Gedächtniseffekte auszuschließen, gleichzeitig aber kurz genug, um Merkmalskonstanz zu

gewährleisten. Mit der Retest-Reliabilität können keine systematischen, versuchsbedingten Fehler entdeckt werden.

Validität

Die Validität bestimmt das Ausmaß, in dem eine Messmethode tatsächlich das Konstrukt misst, das gemessen werden soll (misst z.B. die durch Befragung gemessene Kaufabsicht das tatsächliche Kaufverhalten?). Dies betrifft besonders die Relevanz bei der Messung von nicht direkt beobachtbaren theoretischen Konstruktionen (Motivation, Einstellung, Preisbereitschaft etc.).

Psychologische Tests können als Messinstrumente betrachtet werden; daher sind Objektivität, Reliabilität und Validität auch die wichtigsten Gütekriterien psychodiagnostischer Verfahren. In ihren Technical recommendations for psychological tests and diagnostic techniques (1954) schlug die American Psychological Association vier Arten der Validität vor, diese sind Inhaltsvalidität, Konstruktvalidität und prognostische und diagnostische Kriteriumsvalidität, von denen „historisch und praktisch gesehen […] die kriteriumsbezogene Validität der bedeutsamste Aspekt" ist. „Die Übereinkunft durch ein Rating ist wie alle Übereinkünfte nicht etwas Abgeschlossenes, sondern kann einem ständigen Wandel unterworfen sein. […] Es bleibt dabei jedem Testinterpreten überlassen, dieses Kriterium anzuerkennen oder zu verwerfen bzw. nach einem besseren zu suchen. (wikipedia)

Beispiel
Als Standardbeispiel wird oft der Intelligenztest herangezogen. Betrachtet man die drei Gütekriterien Objektivität, Reliabilität und Validität, so sei für dieses Beispiel angenommen, dass die ersten beiden Gütekriterien gut erfüllt seien: Der Intelligenztest ist so konstruiert, dass sein Ergebnis (fast) unabhängig vom Testleiter ist (Objektivität) und das Testergebnis sich auch wiederholen lässt (Reliabilität). Die Validität, also die Gültigkeit, des Testverfahrens wird aber oft bezweifelt, wenn kritisiert wird, dass der Intelligenztest keine (genaue) Aussage über die wahre Intelligenz (das heißt das Konstrukt „Intelligenz") mache und sich Intelligenz also gar nicht auf diese Weise messen lasse (siehe auch Kritik am Intelligenzbegriff).

Arten

a. ***Inhaltsvalidität*** *(Content Validity):* Sie gibt an, inwieweit die beobachtete Wirkung auch für die relevante Grundgesamtheit gilt. Sie wird angenommen, wenn ein Verfahren zur Messung eines bestimmten Konstrukts oder Merkmals die bestmögliche Operationalisierung dieses Konstrukts ist. Das ist zum Beispiel bei Interessen- und Kenntnistests der Fall: Eine Klassenarbeit oder Führerscheinprüfung repräsentieren direkt die zu messenden Fähigkeiten. Daher spricht man auch von logischer oder trivialer Validität. Ob Inhaltsvalidität gegeben ist oder nicht, entscheiden Experten per Rating.

b. ***Kriterien-Validität*** *(Criterion Validity):* Die Validität wird durch einen Vergleich mit einem beobachtbaren Kriterium (z.B. Kaufverhalten) überprüft. Korreliert man beobachtetes Verhalten mit dem Verhalten, das aus der Messung von Einstellung prognostiziert wurde, spricht man von *Vorhersage-Validität (Predictive Validity).* Werden Einstellung und Verhalten gleichzeitig gemessen, handelt es sich um *Übereinstimmungs-Validität (Concurrent Validity).* Kriteriumsvalidität bezieht sich auf den Zusammenhang zwischen den Ergebnissen des Messinstruments und einem empirischen Kriterium. Zum Beispiel: Ein Forscher untersucht den Zusammenhang seines neuen Intelligenztests mit den Schulnoten der Probanden, um die Gültigkeit seines Tests zu prüfen. Von „innerer (Kriteriums)validität" wird dabei dann gesprochen, wenn als Kriterium ein anderer, als valide anerkannter Test herangezogen wird. Sofern als Kriterium ein objektives Maß (zum Beispiel psychophysiologische Maße oder ökonomische Größen) oder ein Expertenrating herangezogen wird, wird von äußerer (Kriteriums)validität gesprochen. Auch lässt sich unterscheiden nach dem Zeitpunkt, zu dem Übereinstimmung mit dem Kriterium vorliegen soll:

   Diagnostische Validität/Übereinstimmungsvalidität (concurrent validity)
   Das Außenkriterium, das bereits valide sein muss (z. B. ein anderer Test) wird zeitgleich mit dem zu validierenden Messinstrument den gleichen Versuchspersonen dargeboten. Die Ergebnisse der

beiden Messinstrumente werden korreliert. Die Höhe der Korrelation ist das Maß für die Übereinstimmungsvalidität. Das Vorgehen zur Ermittlung der konvergenten und der diskriminanten Testvalidität sind Spezialfälle dieser Kategorie.

Prognostische Validität/Vorhersagevalidität (predictive validity)
Die Messdaten werden zu einem Zeitpunkt erhoben, der vor der Erhebung des Außenkriteriums liegt. Im Unterschied zur Übereinstimmungsvalidität liegt bei der Bestimmung der Vorhersagevalidität das Prognoseintervall zwischen den beiden Messungen. So kann der Grad bestimmt werden, in dem die Messdaten das Kriterium vorhersagen. Zum Beispiel kann im Rahmen eines Assessment-Centers eine Prognose für beruflichen Erfolg gestellt werden, oder aus der Leistung in einem Intelligenztest wird der spätere Schulerfolg vorhergesagt. Ein Test erfüllt die Vorhersagevalidität, wenn seine Vorhersagen hoch mit dem tatsächlich später eingetretenen Ergebnis korrelieren.

c. ***Konstruktvalidität*** *(Construct Validity):* liegt vor, wenn man die Ergebnisse aus mehreren Messungen eines theoretischen Konstrukts bei Verwendung verschiedener Methoden korreliert *(Convergent Validity)* oder die Ergebnisse aus mehreren Messungen verschiedener Konstrukte korreliert *(Discriminant Validity).*

d. Im Unterschied zu der bisher behandelten wissenschaftlichen Validität versteht man unter *Anschauungs-Validität (Face Validity)* die Übereinstimmung der Ergebnisse mit den subjektiven Einschätzungen von Experten (Expertenbefragung).

Die Validität ein Qualitätskriterium für die Belastbarkeit einer bestimmten Aussage. Validität ist also einerseits die Belastbarkeit der Operationalisierung („Inwieweit misst das Testinstrument das, was es messen soll?“), andererseits die Belastbarkeit der auf den Messungen beruhenden Aussagen oder Schlussfolgerungen („Inwieweit trifft es zu, dass X Y beeinflusst?“).
Bei guten Messinstrumenten sind die Messwerte unabhängig vom Messenden; dieses Gütekriterium heißt Objektivität oder Beobachterübereinstimmung. Auch liefern gute Messinstrumente zuverlässig von denselben Objekten dieselben Messwerte; dieses Kriterium heißt Reliabilität oder Reproduzierbarkeit. Das dritte Gütekriterium, die Validität, ist ein Maß dafür, ob die bei der Messung erzeugten Daten wie beabsichtigt die zu messende Größe repräsentieren. Nur dann können die Daten sinnvoll interpretiert werden. Die Validität wird durch Experten-Schätzung festgelegt. Die Gütekriterien bauen aufeinander auf; ohne Objektivität keine Reliabilität, ohne Reliabilität keine Validität.

Quelle: Wikipadia, Gabler

Bruno Schmalen
SCHMALEN-Kommunikation und Training
E-Mail: schmalen@schmalen-online.de
www.schmalen-online.de