

Aus der Forschung

Was ist Messvalidität?

Messvalidität heißt: Ein Score misst, wofür eine Entscheidung ihn hält. Die Klassiker streiten darüber, was valide ist: der Test oder seine Deutung.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic)	18. Oktober 2026	20 Min.	4.212
Führungskraft für Umsatzwachstum, Dozent und Doktorand			

Management Summary Messvalidität heißt, dass ein Score misst, wofür eine Entscheidung ihn hält. Dieser Konzeptleitfaden stellt drei klassische Antworten auf die Frage vor, was valide ist: die Deutung eines Tests im Netz einer Theorie (Cronbach und Meehl), Bedeutung und Verwendung von Scores als Gesamturteil (Messick) oder der Test selbst, wenn das Merkmal existiert und die Schwankung des Scores verursacht (Borsboom, Mellenbergh und van Heerden). Er stellt die Arten von Validitätsbelegen zusammen und sagt, was jede nicht zeigen kann, erklärt, warum eine hohe Korrelation kein Beweis ist, grenzt Validität von Reliabilität und von zusammengesetzten Indizes ab und gibt ein Blatt auf einer Seite, auf dem eine Validitätsaussage mit der Bezeichnung festgehalten wird, die die Belege tragen.

Inhalt	§ 1	Was ist Messvalidität?	2
	§ 2	Warum widersprechen sich die klassischen Definitionen?	3
	Abbildung 1	Drei Antworten auf die Frage, was valide ist	5
	§ 3	Was zählt als Beleg für Validität?	6
	Abbildung 2	Arten von Validitätsbelegen und was jede allein nicht zeigen kann	8
	§ 4	Warum ist eine hohe Korrelation kein Beweis für Validität?	10
	§ 5	Wie unterscheidet sich Validität von Reliabilität?	11
	§ 6	Gilt Validität auch für einen zusammengesetzten Score?	12
	§ 7	Was ist Messvalidität nicht?	13
	§ 8	Wie hält man eine Validitätsaussage fest?	13
	Abbildung 3	Blatt für eine Validitätsaussage	14
	§ 9	Wo endet die Evidenz zur Messvalidität?	15
		Referenzen	16

Was ist Messvalidität?

§ 1 von 9

Messvalidität ist die Frage, ob ein Score misst, wofür er verwendet wird. In ihrer schlichten Form ist diese Frage alt. Borsboom et al. (2004)¹ sehen sie unter anderem bei Kelley im Jahr 1927 zutreffend formuliert: Ein Test sei valide, wenn er misst, was er zu messen vorgibt („when he stated that a test is valid if it measures what it purports to measure“).

¹ Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071.

Die Definitionen der Messtheorie sind weniger schlicht, und sie widersprechen einander. Messick (1995)⁴ definiert Validität als Gesamturteil darüber, wie weit empirische Belege und theoretische Begründungen die Deutungen und Handlungen stützen, die auf Scores beruhen („an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment“). Validität sei keine Eigenschaft des Tests, sondern der Bedeutung seiner Scores („Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores.“). Borsboom et al. (2004)¹ halten der Fassung derselben Definition von 1989 entgegen, Validität sei gar kein Urteil, sondern die beurteilte Eigenschaft („Validity is not a judgment at all. It is the property being judged.“). Nach ihrer Auffassung ist ein Test genau dann valide für ein Merkmal, wenn das Merkmal existiert und seine Schwankungen die Messergebnisse ursächlich schwanken lassen („a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure“).

⁴ Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

In diesem Leitfaden heißt Messvalidität: Ein Score kann die Deutung tragen, die eine Entscheidung ihm gibt. Er schwankt, weil das Merkmal schwankt, nach dem er benannt ist, und nicht hauptsächlich aus anderen Gründen. Diese Arbeitsdefinition übernimmt den kausalen Test von Borsboom et al.¹ und von Messick den Blick auf die Deutung im Gebrauch. Das Wort hauptsächlich stammt vom Leitfaden selbst: Es macht Validität zu einer Frage des Grades, und genau das bezweifeln Borsboom et al. Den Streit zwischen beiden entscheidet die Definition nicht.

Die Frage reicht über die Psychologie hinaus. Messick (1995)⁴ wendet die Grundsätze der Validität auf Schlüsse aus jeder Art an, beständiges Verhalten oder Merkmale zu beobachten oder festzuhalten („inferences based on any means of observing or documenting consistent behaviors or attributes“), auch auf Merkmale von Gruppen, Situationen, Objekten und Institutionen („Scores may also refer to functional consistencies and attributes of groups, situations or environments, and objects or institutions“), mit der Schulabbrecherquote als einem seiner Beispiele. So gelesen, und das ist die Übertragung dieses Leitfadens auf kommerzielle Scores, steht ein Lead-Score, ein Kunden-Gesundheits-Score oder ein Engagement-Index für Accounts vor denselben Fragen wie ein Test.

Der Leitfaden zum Messfehler behandelt die Lücke zwischen einem erklärten Konstrukt und seinem Indikator sowie die Mechanismen, die sie öffnen. Diese Seite stellt die Frage, in die diese Lücke mündet: Kann der Score die Deutung noch tragen, wenn die Lücke beschrieben ist?

Warum widersprechen sich die klassischen Definitionen?

§ 2 von 9

Weil sie uneins darüber sind, was valide ist: der Test oder die Deutung seiner Scores. Borsboom et al. (2004)¹ beschreiben die Geschichte als Weg von der Frage, ob man misst, was man messen will, zur Frage, ob die empirischen Beziehungen zwischen Testwerten den theoretischen Beziehungen in einem nomologischen Netz entsprechen („to the question of whether the empirical relations between test scores match theoretical relations in a nomological network“), und schließlich zur Frage, ob Deutungen und Handlungen auf Basis der Scores gerechtfertigt sind, soziale Folgen eingeschlossen.

Cronbach und Meehl (1955)² lieferten den mittleren Schritt. Ein Konstrukt ist ein postuliertes Merkmal von Personen, von dem angenommen wird, dass es sich in der Testleistung niederschlägt („A construct is some postulated attribute of people, assumed to be reflected in test performance.“). Seine Bedeutung kommt aus dem nomologischen Netz, dem verzahnten System von Gesetzen, das eine Theorie bildet („We shall refer to the interlocking system of laws which constitute a theory as a nomological network.“). Validierung prüft Vorhersagen aus diesem Netz. Ein Konstrukt, das sich bewährt, wird übernommen, nicht bewiesen, und eine gescheiterte Vorhersage sagt nicht, wo der Fehler liegt („If the prediction is not confirmed, any link in the chain may be wrong.“).

² Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

Messick (1995)⁴ weitete den Begriff aus. Die herkömmliche Aufteilung in Inhalts-, Kriteriums- und Konstruktvalidität nennt er zersplittert und unvollständig, vor allem weil sie die Wertimplikationen der Score-Bedeutung und die sozialen Folgen der Verwendung auslässt („The traditional conception of validity divides it into three separate and substitutable types“; „This view is fragmented and incomplete“). An ihre Stelle setzt er eine einzige Konstruktvalidität mit sechs Aspekten: inhaltlich, substanziell, strukturell, Generalisierbarkeit, extern und folgenbezogen („These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity.“). Validität sei eine sich entwickelnde Eigenschaft und Validierung ein fortlaufender Prozess („validity is an evolving property and validation a continuing process“).

Borsboom et al. (2004)¹ verengten ihn wieder. Sie trennen die Eigenschaft von der Arbeit, sie zu finden („validity is a property, whereas validation is an activity“). Validierung solle die Prozesse untersuchen, die die Wirkung des gemessenen Merkmals auf die Testwerte übertragen („the processes that convey the effect of the measured attribute on the test scores“), nicht die Beziehungen des Merkmals zu anderen Merkmalen. Und sie bezweifeln, dass Validität überhaupt Abstufungen kennt: Ihre Theorie sei mit der Möglichkeit vereinbar, Validität als qualitativen statt quantitativen Begriff zu fassen („naturally compatible with the possibility of conceptualizing validity as a qualitative, rather than quantitative, concept“), eine Möglichkeit, die sie weiterer Untersuchung überlassen. Cronbach und Meehl (1955)² hatten verlangt, den Grad der Validität so bestimmt wie möglich anzugeben („to state as definitely as possible the degree of validity the test is presumed to have“).

Abbildung 1 Drei Antworten auf die Frage, was valide ist

Die drei klassischen Quellen sind uneins, was valide ist, wo die Belege liegen und ob Validität Abstufungen kennt. Jede liefert dennoch eine brauchbare Frage.

Frage	Cronbach und Meehl (1955)	Messick (1995)	Borsboom et al. (2004)
Was valide ist	Die Deutung eines Tests im Sinne eines Konstrukts	Die Bedeutung der Scores und die darauf gestützten Handlungen	Der Test, für die Messung eines benannten Merkmals
Wo die Belege liegen	Bestätigte Vorhersagen aus dem nomologischen Netz, geprüft mit vielen Methoden, darunter Prozessstudien	Sechs Aspekte, Folgen eingeschlossen, zu einem Urteil verbunden	Der Prozess, der die Wirkung des Merkmals auf den Score überträgt
Grad oder Ja und Nein	Ein Grad, so bestimmt wie möglich angegeben	Ein Urteil über einen Grad	Ja oder nein, als zu prüfende Möglichkeit
Die Frage an einen Score	Entsprechen seine Beziehungen zu anderen Maßen dem, was die Theorie vorhersagt?	Stützen Belege und Begründung diese Deutung und diese Verwendung?	Existiert das Merkmal, und lässt seine Schwankung diesen Score schwanken?

Man muss sich nicht für eine Seite entscheiden, um die Tabelle zu nutzen. Eine Aussage, die alle drei Fragen besteht, ist stärker als eine, die nur eine besteht.

Quelle Cronbach und Meehl (1955), S. 283, 287, 289, 290, 294; Messick (1995), S. 741, 744; Borsboom, Mellenbergh und van Heerden (2004), S. 1061, 1063, 1070. Die letzte Zeile ist die Umschreibung jeder Position als Frage durch den Autor.

Der Streit ist echt, und er hat praktische Folgen. Nach Messick ist eine Validitätsaussage eine Aussage über eine Deutung und, wo der Score verwendet wird, über eine Verwendung; derselbe Score kann also für eine Entscheidung valide sein und für eine andere nicht. Nach

Borsboom et al.¹ ist es eine Ja-oder-Nein-Frage, ob das Merkmal existiert und den Score bewegt, und wie gut er einer Entscheidung dient, ist eine eigene Frage. Für dieses Urteil schlagen sie einen schlichten Ausdruck vor, die Gesamtqualität („We suggest that to express this judgment, overall quality is not such a bad choice.“). Nebeneinander gelesen geben die drei Positionen drei Fragen, die man einem kommerziellen Score stellen kann, ohne vorher zu klären, welche Theorie recht hat.

Was zählt als Beleg für Validität?

§ 3 von 9

Mehrere Arten zählen, und keine allein entscheidet. Cronbach und Meehl (1955)² berichten die vier Arten, die ein Ausschuss der American Psychological Association benannt hatte: prädiktive Validität, Übereinstimmungsvalidität, Inhaltsvalidität und Konstruktvalidität („The categories into which the Recommendations divide validity studies are: predictive validity, concurrent validity, content validity, and construct validity.“). Die ersten beiden korrelieren den Score mit einem Kriterium und unterscheiden sich nur im Zeitpunkt: Wird das Kriterium später erhoben, geht es um prädiktive Validität („If the criterion is obtained some time after the test is given, he is studying predictive validity.“). Inhaltsvalidität zieht eine Stichprobe aus einem definierten Bereich. Konstruktvalidität ist nötig, wo beides nicht genügt, also wenn kein Kriterium und kein Inhaltsbereich als ausreichend gilt, um die zu messende Eigenschaft zu bestimmen („Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured.“). Sie fügen hinzu, das sei für fast jeden Test wünschenswert („Determining what psychological constructs account for test performance is desirable for almost any test.“).

Messick (1995)⁴ behandelt Inhalts-, Kriteriums- und Folgenbelege als ergänzende Beiträge zu einem einzigen Urteil. Keiner davon genügt allein, und keiner ist zwingend, sofern belastbare konvergente und diskriminante Belege die Bedeutung des Scores stützen („test validity cannot rely on any one of the supplementary forms of evidence just discussed. However, neither does validity require any one form, granted that there is defensible convergent and discriminant evidence supporting score meaning.“). Dieses Paar nennt er die Grundlage der Konstruktvalidierung („Both convergent and discriminant evidence are basic to construct validation.“).

MacKenzie et al. (2011)³, die für Forschende in Wirtschaftsinformatik und Verhaltensforschung schreiben, setzen die Validitätsprüfung als Schritt 8 in einen zehnstufigen Prozess der Skalenentwicklung, der von

der begrifflichen Definition bis zu Normwerten reicht. Schritt 8 fragt, ob die Indikatoren das zugrunde liegende Konstrukt zutreffend abbilden, geprüft durch experimentelle Manipulation oder den Vergleich von Gruppen, die sich im Konstrukt bekanntermaßen unterscheiden, und ob sie seine Mehrdimensionalität erfassen („of the underlying construct (through experimental manipulation or comparing groups known to differ on the construct), (2) adequately capture the multidimensional nature of the construct“). Er fragt außerdem nach der diskriminanten Validität, also ob sie sich von Indikatoren anderer Konstrukte unterscheiden, und nach der nomologischen Validität, also ob sie mit Maßen der Konstrukte zusammenhängen, die das theoretische Netz des Konstrukts vorsieht („(3) are distinguishable from the indicators of other constructs (discriminant validity), and (4) are related to the measures of other constructs“).

3 MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.

Abbildung 2 Arten von Validitätsbelegen und was jede allein nicht zeigen kann

Jede Art von Beleg beantwortet eine Frage. Keine schließt allein jede andere Deutung des Scores aus.

Beleg	Die Frage, die er beantwortet	Was er allein nicht zeigen kann
Inhalt	Decken die Items oder Eingangsgrößen den definierten Bereich ab?	Dass sich der Score mit dem Merkmal bewegt und nicht mit der Auswahl der Items
Antwortprozess	Wirkt das Merkmal über einen bekannten Prozess auf den Score?	Dass der behauptete Prozess der tatsächliche ist, solange er nicht an den Antworten geprüft wurde
Kriterium: prädiktiv oder übereinstimmend	Korreliert der Score mit einem Kriterium, das später oder gleichzeitig erhoben wird?	Dass er das Merkmal misst und nicht etwas anderes, das das Kriterium ebenfalls vorhersagt
Konvergent und diskriminant	Stimmt er mit Maßen desselben Konstrukts überein und unterscheidet er sich von Maßen anderer?	Dass nicht eine gemeinsame Methode die Übereinstimmung erzeugt, sofern Merkmale und Methoden nicht gekreuzt werden
Bekannte Gruppen oder Manipulation	Schneiden Gruppen, die sich im Merkmal bekanntermaßen unterscheiden, unterschiedlich ab, und bewegt eine gezielte Änderung des Merkmals den Score?	Bei bekannten Gruppen, dass das Merkmal und kein anderer Gruppenunterschied den Score bewegt; bei einer Manipulation, dass sie nur das Merkmal verändert hat
Nomologisch	Entsprechen seine Beziehungen zu anderen Konstrukten der Theorie?	Dass keine Scheinursache wirkt: Methodenverzerrung bleibt ohne verfahrensbezogene oder statistische Kontrollen offen, andere konkurrierende Konstrukte bleiben offen, weil die Daten korrelativ sind

Beleg	Die Frage, die er beantwortet	Was er allein nicht zeigen kann
Folgen	Stützen die beabsichtigten und unbeabsichtigten Folgen einer Verwendung diese, und gehen nachteilige Wirkungen auf eine Quelle von Invalidität zurück?	Validität selbst, nach Borsboom et al., die Folgen aus ihr ausschließen

Die dritte Spalte ist die Liste der Gegenerklärungen, auf die eine Validitätsaussage noch antworten muss.

Quelle Arten und Fragen nach Cronbach und Meehl (1955), S. 281-282; Messick (1995), S. 742, 744, 745, 746, 748; MacKenzie, Podsakoff und Podsakoff (2011), S. 296, 317-320; Borsboom, Mellenbergh und van Heerden (2004), S. 1061, 1067, 1069. Die dritte Spalte ist eine Zusammenfassung des Autors, gestützt auf die Grenzen, die diese Quellen nennen.

MacKenzie et al. (2011)³ benennen die Schwäche der gängigsten Art. Prüfungen im nomologischen Netz könnten Scheinursachen als Erklärung nicht ausschließen („the disadvantage of this technique is that it cannot rule out spurious causes as an explanation for the findings“), Methodenverzerrungen eingeschlossen, solange keine verfahrensbezogenen oder statistischen Kontrollen greifen („unless the researcher implements procedural or statistical controls“). Sie verweisen auf zwei Alternativen, die in der Verhaltensforschung lange etabliert, in Wirtschaftsinformatik und Management aber wenig genutzt sind („are not used very frequently in the MIS and management literatures“): das Konstrukt direkt zu manipulieren und Gruppen zu vergleichen, die im Konstrukt bekanntermaßen hoch oder niedrig liegen („identify groups known to be high or low on the focal construct and then compare the scores of these groups on the measures of the focal construct“). Sie räumen auch ein, dass nicht jeder Schritt in eine einzige Studie passt („there may be practical limitations that prevent researchers from being able to follow all of the recommendations discussed in this paper in a single study“). Den Gruppenvergleich halten sie für die schwächere der beiden, weil er nur korrelative und keine kausalen Belege liefert („because it provides only correlational (rather than causal) evidence“).

Warum ist eine hohe Korrelation kein Beweis für Validität? § 4 von 9

Weil ein Score mit vielem korrelieren kann, was er nicht misst. Borsboom et al. (2004)¹ halten Korrelationstabellen höchstens für Indizien („tables of correlations between test scores and other measures cannot provide more than circumstantial evidence for validity“) und geben ein Beispiel: Größe und Gewicht korrelieren in der Bevölkerung mit etwa 0,80, und doch misst eine Waage nicht die Körpergröße („Height and weight correlate about .80 in the general population, but this does not mean that the process of letting people stand on a scale and reading off their weight gives one valid measurements of their height.“).

Cronbach und Meehl (1955)² machen einen verwandten Punkt von der anderen Seite: Hohe Korrelationen und hohe Stabilität können je nach Theorie für oder gegen eine Deutung sprechen („High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.“). In ihrem ausgeführten Beispiel, einem hypothetischen Test der Ängstlichkeit, machen vernachlässigbare Korrelationen mit sozialer Schicht, Berufsziel und Wertorientierung es recht sicher, die Lesart zu verwerfen, der Test messe akademischen Ehrgeiz; negative Ergebnisse schließen konkurrierende Erklärungen aus („Negative results eliminate competing explanations of the X score“). Messick (1995)⁴ formuliert die allgemeine Regel: Diskriminante Belege seien besonders wichtig, um plausible Gegendeutungen zu entkräften („Discriminant evidence is particularly critical for discounting plausible rival alternatives to the focal construct interpretation.“).

Am schärfsten wirkt der Punkt, wo ein Score zur Vorhersage gebaut wird. Borsboom et al. (2004)¹ argumentieren, wer auf Vorhersage optimiere, wähle Items, die untereinander kaum, mit dem Kriterium aber stark korrelieren („What one selects when optimizing predictive utility are items that are mutually uncorrelated but highly correlated with the criterion“), während Items, die ein Merkmal messen, miteinander korrelieren sollten. Messen und Vorhersagen seien daher keine gleichgerichteten Wege der Testkonstruktion („optimizing measurement properties and optimizing predictive properties are not convergent lines of test construction“).

Die praktische Lesart stammt von diesem Leitfaden, nicht von den Autoren. Ein Lead-Score, der auf Konversion getrimmt ist, oder ein Kunden-Gesundheits-Score, der wie dort beschrieben an Verlängerungen kalibriert wird, kann gut vorhersagen und trotzdem

weder Kaufabsicht noch die Gesundheit eines Accounts messen. Sind die Belege nur prädiktiv, ist die engere Bezeichnung die vertretbare (sagt Konversion in diesem Segment für diesen Zeitraum vorher), nicht die weitere (misst Kaufabsicht).

Wie unterscheidet sich Validität von Reliabilität? § 5 von 9

Reliabilität betrifft, wie beständig oder genau ein Score ist; Validität betrifft, was er misst. Messick (1995) ordnet Reliabilität der Validität selbst zu: Sie gehört zum Generalisierbarkeitsaspekt der Konstruktvalidität; Messfehler aus der Auswahl von Aufgaben, Gelegenheiten und Beurteilenden seien das, worum es bei Reliabilität traditionell geht („Such sources of measurement error associated with the sampling of tasks, occasions, and scorers underlie traditional reliability concerns“). Borsboom et al. (2004)¹ halten beides auseinander: Reliabilität sei ein Maß der Messgenauigkeit („reliability is an index of measurement precision“), zwei Tests könnten beide valide sein und einer davon reliabler („Two tests may both be valid, but one may be more reliable.“), und wer Validität behaupte, behaupte damit weder Reliabilität noch Vorhersagekraft noch Freiheit von Verzerrung („when one claims validity, one is not thereby claiming reliability, predictive adequacy, or absence of bias“). Fasse man Validität als qualitativen Ja-oder-Nein-Begriff, so schlagen sie vor, sei Reliabilität nicht mehr sinnvoll als Obergrenze der Validität zu verstehen; eher sei Validität eine notwendige Bedingung dafür, dass Reliabilitätsschätzungen überhaupt Sinn ergeben („validity is a necessary condition for reliability estimates to make sense“).

Messick (1995)⁴ benennt auch die zwei großen Gefahren für die Konstruktvalidität („There are two major threats to construct validity“). Die erste ist Konstruktunterrepräsentation: Die Messung ist zu eng und lässt wichtige Dimensionen des Konstrukts aus („In the one known as construct underrepresentation, the assessment is too narrow and fails to include important dimensions or facets of the construct.“). Die zweite ist konstruktirrelevante Varianz: Die Messung ist zu breit und enthält überschüssige reliable Varianz aus anderen Konstrukten sowie Methodenvarianz wie Antworttendenzen oder Rateneigung („the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities“). Das Wort reliable in diesem Satz ist die Warnung: Ein Score kann sehr beständig sein und beständig etwas anderes messen, als sein Name sagt.

Gilt Validität auch für einen zusammengesetzten Score? § 6 von 9

Das hängt davon ab, wie die Eingangsgrößen mit dem Merkmal zusammenhängen, nach Borsboom et al. (2004)¹. Eine Summe von Items, die alle auf ein Merkmal ansprechen, bleibt ein Maß dieses Merkmals, und ihre Definition gilt direkt für reflektive Modelle („can be applied directly to reflective latent variable models“). Ein Score aus ungleichen Eingangsgrößen wie Nutzung, Support-Tickets und Umfrageantworten wird dagegen meist so gebaut, dass die Eingangsgrößen ihn bilden, statt ein gemeinsames Merkmal widerzuspiegeln. Auf solche formativen Modelle, in denen die Indikatoren die latente Variable verursachen und nicht umgekehrt, lässt sich ihre Definition nach Borsboom et al. offenbar nicht anwenden; sie ließen sich womöglich besser als Indizes oder Zusammenfassungen der Indikatoren verstehen („might be better conceptualized as models for indexing or summarizing the indicators“). An jede einzelne Eingangsgröße lässt sich die Validitätsfrage aber weiter stellen („And at this level, one can consistently ask the question of validity“).

MacKenzie et al. (2011)³ ergänzen, dass die Richtung nicht durch das Konstrukt festgelegt ist: Konstrukte seien nicht von Natur aus formativ oder reflektiv („Constructs are not inherently formative or reflective in nature, and most can be modeled as having either formative or reflective indicators“). Für die experimentelle Prüfung eines Konstrukts mit formativen Indikatoren empfehlen sie, die Validität an jedem einzelnen Indikator zu prüfen, indem das Merkmal manipuliert wird, das der jeweilige Indikator misst („we recommend that validity be assessed at the level of the formative indicators“). Das zusammengesetzte Konstrukt prüfen sie trotzdem an seinen angenommenen Ursachen, Korrelaten und Folgen, für formative wie reflektive Indikatoren nach demselben Verfahren („Regardless of whether one is assessing the nomological validity of constructs with formative or reflective indicators, the procedure is the same.“).

In der Lesart dieses Leitfadens wandert die erste Validitätsfrage bei einem zusammengesetzten Geschäftsscore damit zu jeder Eingangsgröße: Ändert sich die Nutzungszahl, wenn sich die Nutzung ändert, und die Ticketzahl, wenn sich die Probleme ändern? Der zusammengesetzte Score selbst braucht dann eine festgelegte Regel und eine festgelegte Verwendung. Die Behauptung, er messe Gesundheit, braucht mehr: nach der kausalen Definition dieses Leitfadens Belege dafür, dass die Gesundheit des Accounts den Score

bewegt; die Eingangsgrößen wären dann reflektiv und nicht formativ. Eine Prüfung auf der Ebene des Konstrukts, etwa die nomologische Prüfung, die MacKenzie et al.³ beschreiben, gehört zu diesen Belegen, kann Scheinursachen allein aber nicht ausschließen (Abbildung 2).

Was ist Messvalidität nicht?

§ 7 von 9

Sie ist keine Messinvarianz, keine externe Validität, keine statistische Schlussfolgerungsvalidität und kein Common Method Bias, auch wenn jedes davon eine Validitätsaussage schwächen kann. Messinvarianz stellt eine vergleichende Frage: ob ein Instrument über Gruppen, Sprachen oder Zeitpunkte hinweg gleich funktioniert, sodass sich deren Werte vergleichen lassen. Borsboom et al. (2004)¹ lassen offen, wie sich Validität zu Reliabilität, Eindimensionalität, Messinvarianz und Verzerrung verhält („how does validity line up with other test theoretic concepts like reliability, unidimensionality, measurement invariance, and bias“).

Externe Validität fragt, ob ein Ergebnis auf eine andere Population oder Umgebung übertragbar ist. Auf die Deutung eines Scores angewandt, ist dieselbe Frage das, was Messick (1995)⁴ den Generalisierbarkeitsaspekt der Konstruktvalidität nennt („score properties and interpretations generalize to and across population groups, settings, and tasks“). Statistische Schlussfolgerungsvalidität fragt, ob ein Schluss aus Daten tragfähig ist, etwa ein Unterschied zwischen zwei Gruppen. Common Method Bias ist eine Quelle konstruktirrelevanter Varianz: Übereinstimmung zwischen Maßen, die aus der gemeinsamen Art ihrer Erhebung stammt.

Wie hält man eine Validitätsaussage fest?

§ 8 von 9

Als ein Blatt, das den Score nennt, die Deutung, die er tragen soll, die Belege nach Art, die erwogenen Gegenerklärungen und die Bezeichnung, die die Belege tragen. Cronbach und Meehl (1955)² haben die Pflicht beschrieben: Ein Bericht über Konstruktvalidierung soll klarmachen, welche Deutung vorgeschlagen wird, wie gut sie nach Ansicht des Verfassers belegt ist und welche Belege und Überlegungen dazu führen („The proper goals in reporting construct validation are to make clear (a) what interpretation is proposed, (b) how adequately the writer believes this interpretation is substantiated, and (c) what evidence and reasoning lead him to this belief.“). Sie fügen hinzu, der Autor trage die ethische Verantwortung, zu verhindern, dass unbelegte Deutungen als Wahrheiten erscheinen („The author has an ethical responsibility to prevent unsubstantiated interpretations from appearing as truths.“).

Abbildung 3 Blatt für eine Validitätsaussage

Score und Regel, Merkmal und Entscheidung, Belege nach Art, erwogene Gegenerklärungen und die gestützte Bezeichnung in getrennten Feldern festhalten.

Score und Regel	Merkmal und Verwendung	Belege nach Art	Erwogene Gegenerklärungen	Gestützte Bezeichnung
Welcher Score, wie berechnet, aus welchen Daten, in welcher Version?	Als welches Merkmal gilt er, und für welche Entscheidung?	Inhalt, Antwortprozess, Kriterium, konvergent, diskriminant, bekannte Gruppen, Manipulation, nomologisch: Was zeigte jede?	Was könnte ihn sonst bewegen (Größe, Methode, Erfassung), was schließt es aus?	Welche Bezeichnung trägt heute: das Merkmal, ein engeres oder nur Vorhersage?

Ein Score, dessen Bezeichnung weiter reicht als seine Belege, ist eine Behauptung, kein Maß. Die Bezeichnung vor der Entscheidung einengen, nicht danach.

Eine Zeile pro Score und Verwendung. Ist das Feld für Gegenerklärungen leer, wurde die Aussage noch an nichts geprüft.

Quelle Entscheidungsblatt des Autors. Die Felder für Deutung, Belegstärke und Belege folgen den Berichtszielen bei Cronbach und Meehl (1955), S. 297; das Feld für Gegenerklärungen folgt Messick (1995), S. 746, zu diskriminanten Belegen, und Cronbach und Meehl (1955), S. 297, zu erwogenen und verworfenen Gegendeutungen; die Arten von Belegen folgen MacKenzie, Podsakoff und Podsakoff (2011), S. 317. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Zwei Regeln machen das Blatt nützlich. Erstens wird kein Feld mit einem einzelnen Koeffizienten gefüllt; Cronbach und Meehl (1955)² schreiben, Konstruktvalidität lasse sich in der Regel nicht als ein einfacher Koeffizient ausdrücken („Construct validity cannot generally be expressed in the form of a single simple coefficient.“). Zweitens darf die Bezeichnung enger sein als der Name des Scores. Sind die einzigen Belege prädiktiv, sagt die Bezeichnung genau das.

Am Anfang steht das Merkmal, nicht die Daten. Borsboom et al. (2004)¹ argumentieren, ein großer Teil der Validität müsse schon bei der Konstruktion in den Test gelegt werden („A large part of test validity must be put into the test at the stage of test construction“), und vermuten, dass die eigentliche Schwierigkeit dann davor liege: herauszufinden, was man messen will („the problem is that it is difficult to find out what one intends to measure“).

Ein ausgeführtes Beispiel ohne Unternehmensdaten: Ein Team nennt eine gewichtete Summe aus Logins, aktiven Lizenzen und Feature-Ereignissen einen Engagement-Score und wählt damit Accounts für einen Upsell-Anruf aus. Das Blatt fragt, was Engagement hier bedeutet und ob diese drei Größen es abdecken; was einen Login auslöst, Nutzen oder eine automatische Sitzungserneuerung; ob die Größe des Accounts allein den Score bewegt; und ob Accounts, die das Kundenteam unabhängig als engagiert einstuft, höher abschneiden. Solange die Gegenerklärungen offen sind, lautet die gestützte Bezeichnung Aktivitätsvolumen je Account, was als Eingangsgröße für die Anrufliste trotzdem nützlich sein kann.

Wo endet die Evidenz zur Messvalidität?

§ 9 von 9

Bei Theorie, die für Tests an Personen geschrieben wurde. Cronbach und Meehl (1955)², Messick (1995)⁴ und Borsboom et al. (2004)¹ sind konzeptionelle Arbeiten mit Beispielen aus der psychologischen und pädagogischen Diagnostik; keine prüft einen Geschäftsscore, und sie sind sich in Grundfragen uneinig. Borsboom et al. (2004) halten Fragen der Folgen für zentral bei der Verwendung eines Tests, aber für unerheblich für Begriff und Definition der Validität („central to test use“; „the concept and definition of validity itself“). Messick (1995) behält sie, mit einer Einschränkung: Nachteilige soziale Folgen machten eine Verwendung nicht schon ungültig, sie dürften nur nicht auf eine Quelle von Invalidität zurückgehen („it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity“). MacKenzie et al. (2011)³ ist eine Methodenübersicht für Wirtschaftsinformatik und Verhaltensforschung und räumt ein, dass praktische Grenzen eine einzelne Studie daran hindern können, allen Empfehlungen zu folgen. Keine Quelle liefert einen Schwellenwert für einen Validitätskoeffizienten im Sinne einer Korrelation mit einem Kriterium; MacKenzie et al. geben aber Faustregeln für konvergente und diskriminante Validität in Modellen mit latenten Variablen und reflektiven Indikatoren, etwa eine durchschnittlich erfasste Varianz über 0,50. Messick (1995) verweist bereits auf einen argumentbasierten Ansatz der Validierung, mit dem sich die nötigen Arten von Belegen priorisieren lassen („one can prioritize the forms of validity evidence needed“); dieser Leitfaden behandelt ihn nicht vollständig, das Blatt oben ist eine einfache Form davon. Die Übertragung auf Lead-, Gesundheits- und Engagement-Scores in diesem Leitfaden ist die Überlegung des Autors und wurde nicht geprüft.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1** Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071.
<https://doi.org/10.1037/0033-295X.111.4.1061>
- 2** Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
<https://doi.org/10.1037/h0040957>
- 3** MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.
<https://doi.org/10.2307/23044045>
- 4** Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
<https://doi.org/10.1037/0003-066X.50.9.741>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist Messvalidität? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-messvaliditaet/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-messvaliditaet/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten