

From the research bench

What is measurement validity?

Measurement validity is whether a score measures what a decision treats it as measuring. The classics disagree on what is valid: the test or its reading.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	17 min	3,651

Management summary Measurement validity is whether a score measures what a decision treats it as measuring. This concept guide sets out three classic answers to what is valid: the interpretation of a test in a theory's network (Cronbach and Meehl), the meaning and use of scores judged as a whole (Messick), or the test itself, if the attribute exists and causes the score to vary (Borsboom, Mellenbergh and van Heerden). It lists the kinds of validity evidence and what each cannot show, explains why a strong correlation is not proof, separates validity from reliability and composite indexes, and gives a one-page record for writing a validity claim down with the label the evidence supports.

Contents	§ 1	What is measurement validity?	2
	§ 2	Why do the classic definitions disagree?	3
	Figure 1	Three answers to what validity is	4
	§ 3	What counts as evidence of validity?	5
	Figure 2	Kinds of validity evidence, and what each cannot show alone	7
	§ 4	Why is a strong correlation not proof of validity?	8
	§ 5	How is validity different from reliability?	9
	§ 6	Does validity apply to a composite score?	10
	§ 7	What is measurement validity not?	11
	§ 8	How do you write a validity claim down?	11
	Figure 3	A validity claim record	12
	§ 9	Where does the evidence on measurement validity stop?	13
		References	14

What is measurement validity?

§ 1 of 9

Measurement validity is whether a score measures what it is used to measure. The plain form of that question is old. Borsboom et al. (2004)¹ hold that it was formulated correctly, among others by Kelley in 1927, “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.

The definitions in measurement theory are less plain, and they conflict. Messick (1995)⁴ defines validity as “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”, and adds: “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, answering the 1989 version of the same definition, reply that “Validity is not a judgment at all. It is the property being judged.” On their account, “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 this guide, measurement validity means that a score can carry the interpretation a decision puts on it: the score varies because the attribute it is named after varies, and not mainly because of something else. That working definition borrows the causal test from Borsboom et al.¹ and the attention to the interpretation in use from Messick. The word mainly is the guide's own: it makes validity a matter of degree, which Borsboom et al. question. The definition does not settle the dispute between them.

The question reaches beyond psychology. Messick (1995)⁴ applies the principles of validity to “inferences based on any means of observing or documenting consistent behaviors or attributes”, and notes that “Scores may also refer to functional consistencies and attributes of groups, situations or environments, and objects or institutions”, with school dropout rate among his examples. On that reading, which extends his examples to commercial scores, a lead score, a customer

health score or an account engagement index faces the same questions as a test.

The measurement-error guide treats the gap between a declared construct and its indicator, and the mechanisms that open it. This page asks the question that gap feeds: once the gap is described, can the score still carry the reading put on it?

Why do the classic definitions disagree?

§ 2 of 9

Because they disagree about what is valid: the test, or the interpretation of its scores. Borsboom et al. (2004) describe the history as a move from asking whether one measures what one intends, “to the question of whether the empirical relations between test scores match theoretical relations in a nomological network”, and finally to whether interpretations and actions based on scores are justified, social consequences included.

Cronbach and Meehl (1955)² supplied the middle step. “A construct is some postulated attribute of people, assumed to be reflected in test performance.” Its meaning comes from the nomological network: “We shall refer to the interlocking system of laws which constitute a theory as a nomological network.” Validation tests predictions derived from that network. A construct that survives is adopted rather than proved, and a failed prediction does not say where the fault lies: “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 widened the concept. He calls the traditional split into content, criterion and construct validity fragmented and incomplete, chiefly because it leaves out the value implications of score meaning and the social consequences of score use: “The traditional conception of validity divides it into three separate and substitutable types”, and “This view is fragmented and incomplete”. In its place he puts one construct validity with six aspects: “These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity.” On his account, “validity is an evolving property and validation a continuing process”.

Borsboom et al. (2004)¹ narrowed it again. They separate the property from the work of finding it: “validity is a property, whereas validation is an activity”. Validation, they argue, should study “the processes that convey the effect of the measured attribute on the test scores” rather than the attribute’s relations with other attributes. And they question

whether validity comes in degrees at all: their theory, they write, is “naturally compatible with the possibility of conceptualizing validity as a qualitative, rather than quantitative, concept”, a possibility they leave for further study. Cronbach and Meehl (1955)² had asked researchers “to state as definitely as possible the degree of validity the test is presumed to have”.

Figure 1 Three answers to what validity is

The three classic sources disagree on what is valid, where the evidence lies and whether validity comes in degrees. Each still gives a usable question.

Question	Cronbach and Meehl (1955)	Messick (1995)	Borsboom et al. (2004)
What is valid	The interpretation of a test in terms of a construct	The meaning of the scores and the actions based on them	The test, for measuring a named attribute
Where the evidence lies	Confirmed predictions from the nomological network, tested by many methods, studies of process among them	Six aspects, consequences included, integrated into one judgment	The process that carries the attribute's effect to the score
Degree or yes and no	A degree, stated as definitely as possible	A judgment of degree	Yes or no, offered as a possibility to investigate
The question it puts to a score	Do its relations with other measures match what the theory predicts?	Do evidence and rationale support this reading and this use?	Does the attribute exist, and does its variation cause this score to vary?

A reader does not have to choose a side to use the table. A claim that survives all three questions is stronger than one that survives one.

Source Cronbach and Meehl (1955), pp. 283, 287, 289, 290, 294; Messick (1995), pp. 741, 744; Borsboom, Mellenbergh and van Heerden (2004), pp. 1061, 1063, 1070. The last row is the author's paraphrase of each position as a question.

The dispute is real, and it matters for practice. On Messick's⁴ view, a validity claim is a claim about a reading, and about a use wherever the score is used, so the same score can be valid for one decision and not for another. On the view of Borsboom et al.¹, whether the attribute exists and drives the score is a yes-or-no question, and how well the score serves a decision is a separate matter; for that judgment they write, "We suggest that to express this judgment, overall quality is not such a bad choice." Read side by side, the three positions give three questions a commercial score can be put to without first settling which theory is right.

What counts as evidence of validity?

§ 3 of 9

Several kinds count, and no single kind settles the matter. Cronbach and Meehl (1955)² report the four types the American Psychological Association committee had named: "The categories into which the Recommendations divide validity studies are: predictive validity, concurrent validity, content validity, and construct validity." The first two correlate the score with a criterion and differ only by timing: "If the criterion is obtained some time after the test is given, he is studying predictive validity." Content validity samples a defined domain. Construct validity is needed where neither will do: "Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured." They add that it "is desirable for almost any test".

Messick (1995)⁴ treats content, criterion-related and consequential evidence as supplementary contributions to one judgment. No single one of them suffices and none is required, provided that convergent and discriminant evidence supports the reading of the score: 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." He calls that pair the base: "Both convergent and discriminant evidence are basic to construct validation."

MacKenzie et al. (2011)³, writing for MIS and behavioural researchers, place validity assessment at step 8 of a ten-step scale development process that runs from a conceptual definition to norms. Step 8 asks whether the indicators (1) are accurate representations "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", and whether they "(3) are distinguishable from the indicators of other constructs (discriminant

validity), and (4) are related to the measures of other constructs” in the construct’s theoretical network (nomological validity).

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.

Figure 2 Kinds of validity evidence, and what each cannot show alone

Each kind of evidence answers one question. None of them, alone, rules out every other reading of the score.

Evidence	The question it answers	What it cannot show alone
Content	Do the items or inputs cover the defined domain?	That the score moves with the attribute, rather than with how the items were chosen
Response process	Does the attribute act on the score through a known process?	That the stated process is the one at work, until it is tested against the responses
Criterion: predictive or concurrent	Does the score correlate with a criterion measured later or at the same time?	That it measures the attribute, rather than something else that also predicts the criterion
Convergent and discriminant	Does it agree with measures of the same construct, and differ from measures of others?	That a shared method is not producing the agreement, unless traits are crossed with methods
Known groups or manipulation	Do groups known to differ on the attribute score differently, and does a deliberate change in the attribute move the score?	For known groups, that the attribute and not another difference between the groups moves the score; for a manipulation, that it changed the attribute and nothing else
Nomological	Do its relations with other constructs match the theory?	That no spurious cause is at work: method bias stays open without procedural or statistical controls, and other rival constructs stay open because the data are correlational
Consequences	Do the intended and unintended consequences of a use support it, and do adverse effects trace back to a source of invalidity?	Validity itself, on the view of Borsboom et al., who exclude consequences from it

Use the third column as the list of rivals a validity claim still has to answer.

Source Kinds and questions from Cronbach and Meehl (1955), pp. 281–282; Messick (1995), pp. 742, 744, 745, 746, 748; MacKenzie, Podsakoff and Podsakoff (2011), pp. 296, 317–320; Borsboom, Mellenbergh and van Heerden (2004), pp. 1061, 1067, 1069. The third column is the author's summary, drawing on the limits these sources name.

MacKenzie et al. (2011)³ name the weakness of the most common kind. Of testing relations within the nomological network, they write that “the disadvantage of this technique is that it cannot rule out spurious causes as an explanation for the findings”, and that method bias stays a live explanation “unless the researcher implements procedural or statistical controls”. They point to two alternatives with a long history in behavioural research that, they note, “are not used very frequently in the MIS and management literatures”: manipulating the construct directly, and comparing groups, which means to “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”. They also concede that not every step fits in one study: “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”. They rate the group comparison as the weaker of the two “because it provides only correlational (rather than causal) evidence”.

Why is a strong correlation not proof of validity? § 4 of 9

Because a score can correlate with many things it does not measure. Borsboom et al. (2004)¹ hold that “tables of correlations between test scores and other measures cannot provide more than circumstantial evidence for validity”, and give an example: “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 and Meehl (1955)² make a related point from the other side: “High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.” In their worked example, a hypothetical test of anxiety proneness, negligible correlations with social class, vocational aim and value orientation make it fairly safe to reject the reading that the test measures academic aspiration: “Negative results eliminate competing explanations of the X score”. Messick (1995)⁴ names the general rule: “Discriminant evidence is particularly critical for discounting plausible rival alternatives to the focal construct interpretation.”

The point bites hardest where a score is built to predict. Borsboom et al. (2004)¹ argue that “What one selects when optimizing predictive utility are items that are mutually uncorrelated but highly correlated with the criterion”, while items that measure one attribute should correlate with each other, so “optimizing measurement properties and optimizing predictive properties are not convergent lines of test construction”.

The practical reading is this guide’s, not the authors’. A lead score tuned to predict conversion, or a customer health score calibrated against renewal as that guide describes, can predict well and still not measure buying intent or account health. Where the evidence is predictive only, the defensible label is the narrower one (predicts conversion in this segment over this period), not the broader one (measures intent).

How is validity different from reliability?

§ 5 of 9

Reliability is about how consistent or precise a score is; validity is about what the score measures. Messick (1995)⁴ folds reliability into validity: it belongs to the generalizability aspect of construct validity, where “Such sources of measurement error associated with the sampling of tasks, occasions, and scorers underlie traditional reliability concerns”. Borsboom et al. (2004)¹ keep the two apart: “reliability is an index of measurement precision”, “Two tests may both be valid, but one may be more reliable.”, and “when one claims validity, one is not thereby claiming reliability, predictive adequacy, or absence of bias”. If validity is a yes-or-no concept, they suggest, it would no longer make sense to treat reliability as an upper bound for validity; one would rather say that “validity is a necessary condition for reliability estimates to make sense”.

Messick (1995)⁴ also names “two major threats to construct validity”. The first is construct underrepresentation: “In the one known as construct underrepresentation, the assessment is too narrow and fails to include important dimensions or facets of the construct.” The second is construct-irrelevant variance: “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”. The word reliable in that sentence is the warning: a score can be highly consistent and consistently measure something other than its name.

Does validity apply to a composite score?

§ 6 of 9

It depends on how the inputs relate to the attribute, on the account of Borsboom et al. (2004)¹. A sum of items that all respond to one attribute is still a measure of it, and their definition “can be applied directly to reflective latent variable models”. A composite of unlike inputs, such as usage, support tickets and survey answers, is usually built so that the inputs form the score rather than reflect one attribute. Borsboom et al. say their definition seems not to apply to such formative models, in which the indicators cause the latent variable rather than the reverse, and suggest that these models “might be better conceptualized as models for indexing or summarizing the indicators”. The validity question can still be asked of each input: “And at this level, one can consistently ask the question of validity”.

MacKenzie et al. (2011)³ add that the direction is not fixed by the construct: “Constructs are not inherently formative or reflective in nature, and most can be modeled as having either formative or reflective indicators”. For the experimental test of a construct with formative indicators they write, “we recommend that validity be assessed at the level of the formative indicators”, by manipulating the attribute each indicator measures. The composite is still tested against its antecedents and consequences: “Regardless of whether one is assessing the nomological validity of constructs with formative or reflective indicators, the procedure is the same.”

In this guide’s reading, that moves the first validity question for a composite business score to each input: does the usage figure change when usage changes, does the ticket count change when problems do? The composite itself then needs a stated rule and a stated use. A claim that it measures health needs more: on the causal definition this guide uses, evidence that account health moves the score, which would make the inputs reflective of it rather than formative. A construct-level test such as the nomological one MacKenzie et al.³ describe is part of that evidence, but on its own it cannot rule out spurious causes (Figure 2).

What is measurement validity not?

§ 7 of 9

It is not measurement invariance, external validity, statistical conclusion validity or common-method bias, though each can weaken a validity claim. Measurement invariance asks a comparative question: whether an instrument works the same way across groups, languages or time, so that their scores can be compared. Borsboom et al. (2004)¹ leave open “how does validity line up with other test theoretic concepts like reliability, unidimensionality, measurement invariance, and bias”.

External validity asks whether a result transfers to another population or setting. Asked of a score’s interpretation, the same question is what Messick (1995) calls the generalizability aspect of construct validity: whether “score properties and interpretations generalize to and across population groups, settings, and tasks”. Statistical conclusion validity asks whether an inference from data, such as a difference between two groups, is sound. Common-method bias is one source of construct-irrelevant variance: agreement between measures that comes from the shared way they were collected.

How do you write a validity claim down?

§ 8 of 9

As one record that states the score, the reading it is meant to carry, the evidence by kind, the rivals considered and the label the evidence supports. Cronbach and Meehl (1955)² set out the duty: “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.” They add: “The author has an ethical responsibility to prevent unsubstantiated interpretations from appearing as truths.”

Figure 3 A validity claim record

Write the score and its rule, the attribute and the decision, the evidence by kind, the rivals considered and the label the evidence supports in separate fields.

Score and rule Which score, computed how, from which inputs, in which version?	Attribute and use Which attribute is it read as, and which decision does that feed?	Evidence by kind Content, response process, criterion, convergent, discriminant, known groups, manipulation, nomological: what did each show?	Rivals considered What else could move it (size, method, recording), and what rules that out?	Supported label Which label holds today: the attribute, a narrower one, or prediction only?

A score whose label is broader than its evidence is a claim, not a measure. Narrow the label before the decision, not after.

One row per score and use. If the rivals field is empty, the claim has not yet been tested against anything.

Source Author's decision worksheet. The interpretation, substantiation and evidence fields follow the reporting goals in Cronbach and Meehl (1955), p. 297; the rivals field follows Messick (1995), p. 746, on discriminant evidence, and Cronbach and Meehl (1955), p. 297, on alternative interpretations considered and rejected; the evidence kinds follow MacKenzie, Podsakoff and Podsakoff (2011), p. 317. The fields are reader inputs; no company data is supplied.

Two rules make the record useful. First, no field is filled with a single coefficient; Cronbach and Meehl (1955) say that “Construct validity cannot generally be expressed in the form of a single simple coefficient.” Second, the label field may be narrower than the name of the score. If the only evidence is predictive, the label says so.

Start with the attribute, not the data. Borsboom et al. (2004)¹ argue that “A large part of test validity must be put into the test at the stage of test construction”, and suggest that, if so, the hard part is upstream: “the problem is that it is difficult to find out what one intends to measure”.

A worked illustration, with no company data: a team calls a weighted sum of logins, active seats and feature events an engagement score and uses it to choose accounts for an upsell call. The record asks what engagement means here and whether those three inputs cover it; what makes a login happen, value or an automated session refresh; whether

account size alone moves the score; and whether accounts the customer team independently rates as engaged score higher. Until the rivals are answered, the supported label is activity volume per account, which may still be a useful input to the call list.

Where does the evidence on measurement validity stop?

§ 9 of 9

At theory written for tests of persons. Cronbach and Meehl (1955)², Messick (1995)⁴ and Borsboom et al.¹ (2004) are conceptual papers with examples from psychological and educational testing; none tests a business score, and they disagree on basic points. Borsboom et al. (2004) consider consequential issues “central to test use” but irrelevant to “the concept and definition of validity itself”. Messick (1995) keeps them in, with a caveat: “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 is a methods review for MIS and behavioural research and concedes that practical limits may keep a single study from following all of its recommendations. None supplies a threshold for a validity coefficient in the sense of a correlation with a criterion; MacKenzie et al. do give rules of thumb for convergent and discriminant validity in latent-variable models with reflective indicators, such as an average variance extracted above .50. Messick (1995) already points to an argument-based approach to validation, in which “one can prioritize the forms of validity evidence needed”; this guide does not cover it in full, and the record above is a light form of it. The application to lead, health and engagement scores in this guide is the author’s reasoning and has not been tested.

 End of essay

References

- 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>

How to cite this essay

Isoglu, S. (2026, October 18). What is measurement validity? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-measurement-validity/>
Online version: <https://isoglu.com/insights/journal/what-is-measurement-validity/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved