RESEARCH · QUANTITATIVE AND MIXED METHODS

Construct Validity and How to Establish It

What construct validity is, why it matters whenever research measures something that cannot be observed directly, and the practical steps used to establish it, from defining the construct to testing convergent, discriminant and nomological evidence.

ModQu research article · 1,800 words · 18 references (APA 7th edition)

Much of what researchers want to measure cannot be seen directly. Customer loyalty, employee engagement, trust, anxiety and brand image are **constructs**: concepts that researchers infer from observable behaviour or answers. **Construct validity** is the degree to which a measure actually captures the construct it is meant to capture. If a "loyalty" scale really measures habit, or an "engagement" survey really measures general mood, every conclusion built on those scores is at risk. This article explains what construct validity is and how researchers establish it in practice.

Origins of the idea

Cronbach and Meehl (1955) introduced construct validity in a paper written for the American Psychological Association's first technical recommendations for tests. They argued that when no single criterion exists against which a test can be checked, its validity must be established by showing that scores behave as the theory of the construct predicts. They called the set of theoretical relationships among a construct, other constructs and observable measures a **nomological network**. A test of anxiety, for example, should correlate with physiological signs of stress, rise in threatening situations and fall after effective treatment. If it does, both the test and the theory gain support; if it does not, either the test or the theory is wrong.

Loevinger (1957) argued that construct validity was not one type among several but the whole of validity, because every interpretation of test scores rests on assumptions about what the scores mean. This view gradually prevailed. Messick (1995) set out a unified concept of validity in which construct validity is central and other forms of evidence, such as content and criterion relationships, contribute to it.

Why it matters

Construct validity matters whenever research draws conclusions about a concept rather than about a specific set of answers. Shadish et al. (2002) listed construct validity as one of four kinds of validity for any study, concerned with whether the operations used in a study, including its measures, treatments and samples, correctly represent the concepts named in its conclusions. A study that finds "training increases engagement" has construct validity only if the training delivered was really the training described and the measure of engagement really measured engagement.

Flake and Fried (2020) showed how often this is neglected. Many published studies use measures created for the study, adapted without reporting changes, or applied without evidence that they work in the population studied. They argued that such **questionable measurement practices** threaten conclusions as seriously as better-known statistical problems.

Establishing construct validity

There is no single test of construct validity. Instead, researchers build a case from several kinds of evidence. Several frameworks describe the process, including those of Churchill (1979), Hinkin (1998) and MacKenzie et al. (2011). The diagram below summarises the common steps.

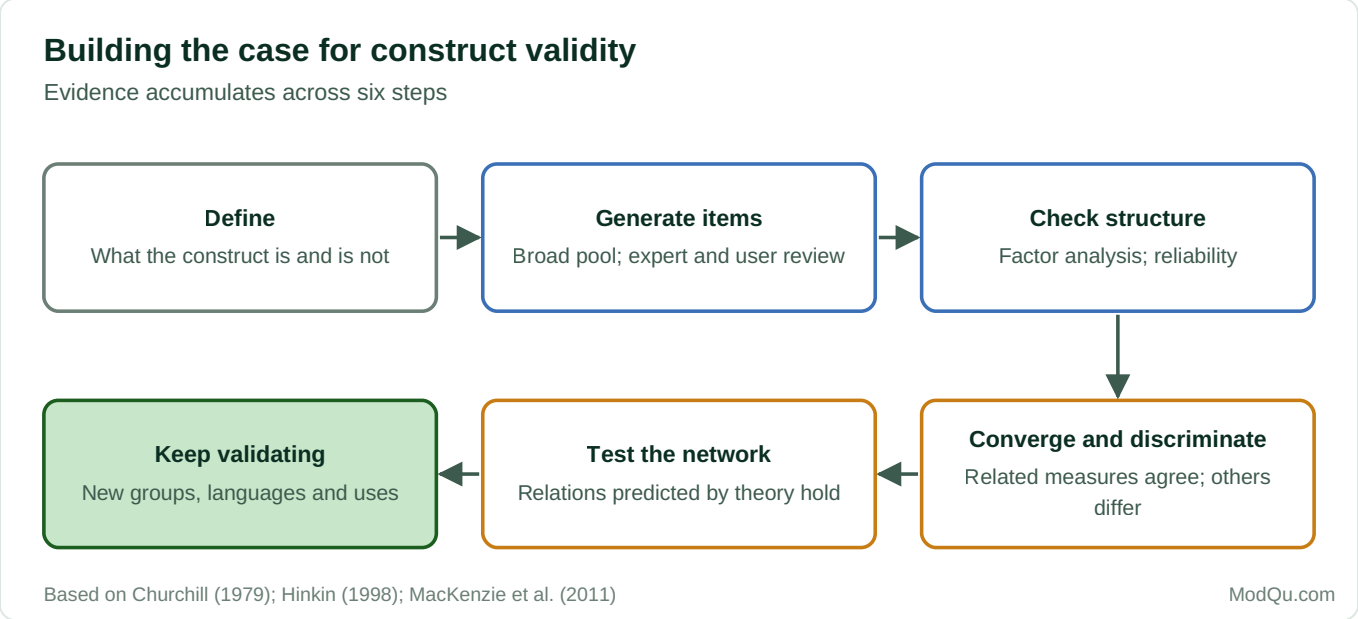


Figure 1. Six steps in establishing construct validity. Based on Churchill (1979), Hinkin (1998) and MacKenzie et al. (2011); diagram by ModQu.

1. Define the construct

Construct validation begins before any data are collected, with a clear definition. MacKenzie et al. (2011) argued that many measurement problems arise because constructs are poorly defined, so that it is unclear what items should cover. A good definition states what the construct is, what it is not, and whether it has distinct dimensions. Clark and Watson (1995) recommended reviewing the literature to understand how the construct has been defined and measured before, and writing a precise definition that distinguishes it from related constructs.

2. Generate and review items

Items should cover the full content of the construct and nothing else. Clark and Watson (1995) advised writing an initial pool of items broader than the final scale, including items that may turn out to be peripheral, because it is easier to remove items later than to add them. Experts and members of the target population then review the items for relevance, clarity and coverage. This step provides evidence of **content validity**. Hinkin (1998) described methods for checking whether items match the intended definition, such as asking judges to sort items into the constructs they seem to measure.

3. Examine internal structure

Once data are collected, researchers examine whether the items fit together as expected. Exploratory and confirmatory factor analysis test whether items load on the intended dimensions. If a scale meant to measure a single construct splits into two clusters, the definition or the items need revision. Internal consistency is also checked at this stage. Items that load weakly, load on the wrong dimension or are answered in unusual ways are candidates for removal, but each removal should be checked against the definition so that the scale does not lose coverage of an important part of the construct. Clark and Watson (1995) warned that very high correlations among items can signal that the scale has become too narrow.

A related question is whether items are **reflective** or **formative**. In a reflective measure, the construct causes the item responses, so items should be correlated and interchangeable. In a formative measure, the items together define the construct, as when socioeconomic status is formed from income, education and occupation; items need not correlate. Jarvis et al. (2003) showed that many marketing studies treated formative constructs as reflective, which can bias the results of structural models.

4. Test convergent and discriminant validity

Campbell and Fiske (1959) proposed that a valid measure should correlate highly with other measures of the same construct (**convergent validity**) and less highly with measures of different constructs (**discriminant validity**). Their **multitrait–multimethod matrix** measures several constructs by several methods, such as self-report and observer rating, so that the effect of the construct can be separated from the effect of the method.

In studies using structural equation modelling, Fornell and Larcker (1981) proposed a widely used criterion: a construct should share more variance with its own items than with any other construct. Henseler et al. (2015) showed that this criterion often fails to detect problems, and proposed the **heterotrait–monotrait ratio** as a more sensitive test of discriminant validity.

Method effects can also threaten validity. When all variables in a study are measured by the same survey from the same respondents, correlations may be inflated by shared method rather than by real relationships. Podsakoff et al. (2003) reviewed the sources of this **common method bias** and recommended remedies, such as collecting predictor and outcome measures from different sources or at different times.

5. Test the nomological network

Finally, the measure should relate to other variables as theory predicts. A measure of customer loyalty should predict repeat purchase, be higher among customers who have recommended the brand, and be lower among customers who have recently complained. Strauss and Smith (2009) described this kind of theory testing as the core of construct validation: each confirmed prediction supports both the measure and the theory, while each failure prompts a review of one or the other.

6. Continue validation

Construct validation is never finished. Kane (2013) argued that validation should be organised as an argument linking scores to their intended interpretations and uses, with evidence for each link. When a measure is used in a new population, translated into a new language or applied for a new purpose, new evidence is needed. Clark and Watson (2019) reviewed developments since their earlier paper and stressed that scale development is an iterative process in which theory and measurement improve together.

Common threats

Shadish et al. (2002) listed a number of threats to construct validity that apply to measures and to the other parts of a study. Several are especially common:

- **Inadequate explication.** The construct is poorly defined, so the measure captures too much, too little or the wrong thing.
- **Mono-operation bias.** Only one measure of the construct is used, so its particular quirks cannot be separated from the construct itself.
- **Mono-method bias.** All measures use the same method, such as a self-report survey, so the results may reflect the method as much as the construct.
- **Construct confounding.** The measure captures another construct as well as the intended one, as when a test of reasoning also depends heavily on reading ability.
- **Reactivity.** Respondents change their answers because they know they are being studied, or because they guess what the researcher hopes to find.

Each threat suggests a remedy: define constructs carefully, use several measures and methods, check for overlap with related constructs, and design data collection to reduce reactivity.

Debates about construct validity

Not everyone accepts the nomological network approach. Borsboom et al. (2004) argued that validity is simpler: a test is valid if the attribute it measures exists and causes variation in scores. On this view, showing that scores correlate with other variables as expected is useful, but the key question is whether the attribute is real and whether the test responds to it. Researchers should therefore be open to discovering that a supposed construct is really several things, or nothing coherent at all.

Reporting construct validity

Readers can only judge construct validity if the evidence is reported. A good report gives the definition of each construct, the source of each measure and any changes made to it, evidence about structure and reliability in the current sample, and the results of convergent and discriminant checks. When a measure is new, the report should describe how items

were generated and reviewed. Flake and Fried (2020) recommended that researchers state their measurement decisions openly, ideally before collecting data, so that readers can see that measures were not chosen or modified after the results were known.

An applied example

A hypothetical example: a market research team develops a scale of "brand warmth". They define warmth as consumers' perception that a brand is friendly, caring and well-intentioned, distinct from perceived quality. They write twenty items, which experts reduce to ten. Factor analysis shows a single factor. The scale correlates strongly with an existing measure of brand likeability (convergent) but only moderately with perceived quality (discriminant). As predicted, warmth scores rise after customers receive helpful service and predict willingness to forgive a product failure. The numbers in such a study would be specific to the data, but the pattern of evidence supports interpreting scores as brand warmth. If warmth had correlated as strongly with quality as with likeability, the team would need to reconsider whether the two constructs were really distinct.

Conclusion

Construct validity asks whether a measure captures the concept it claims to capture. It cannot be established by one statistic. It requires a clear definition, careful item development, evidence about internal structure, convergent and discriminant patterns, and relationships with other variables that fit the theory. Because each new use of a measure raises new questions, validation continues for as long as the measure is used. The effort is worthwhile: conclusions about concepts are only as sound as the measures that stand for them.

References

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(1), 64–73. https://doi.org/10.1177/002224377901600110

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104–121. https://doi.org/10.1177/109442819800100106

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218. https://doi.org/10.1086/376806

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

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

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

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639