Encyclopedia of Research Design

Construct Validity

Contributors: Keith A. Markus & Chia-ying Lin
Editors: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Construct Validity"
Pub. Date: 2010
Access Date: October 14, 2013
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412961271
Online ISBN: 9781412961288
DOI: http://dx.doi.org/10.4135/9781412961288.n72
Print pages: 230-234

This PDF has been generated from SAGE Research Methods. Please note that the pagination of the online version will vary from the pagination of the print book.

http://dx.doi.org/10.4135/9781412961288.n72
Construct validity refers to whether the scores of a test or instrument measure the distinct dimension (construct) they are intended to measure. The present entry discusses origins and definitions of construct validation, methods of construct validation, the role of construct validity evidence in the validity argument, and unresolved issues in construct validity.

Origins and Definitions

Construct validation generally refers to the collection and application of validity evidence intended to support the interpretation and use of test scores as measures of a particular construct. The term *construct* denotes a distinct dimension of individual variation, but use of this term typically carries the connotation that the construct does not allow for direct observation but rather depends on indirect means of measurement. As such, the term *construct* differs from the term *variable* with respect to this connotation. Moreover, the term *construct* is sometimes distinguished from the term *latent variable* because construct connotes a substantive interpretation typically embedded in a body of substantive theory. In contrast, the term *latent variable* refers to a dimension of variability included in a statistical model with or without a clear substantive or theoretical understanding of that dimension and thus can be used in a purely statistical sense. For example, the *latent traits* in item response theory analysis are often introduced as latent variables but not associated with a particular construct until validity evidence supports such an association.

The object of validation has evolved with validity theory. Initially, validation was construed in terms of the validity of a test. Lee Cronbach and others pointed out that validity depends on how a test is scored. For example, detailed content coding of essays might yield highly valid scores whereas general subjective judgments might not. As a result, validity theory shifted its focus from validating tests to validating test scores. In addition, it became clear that the same test scores could be used in more than one way and that the level of validity could vary across uses. For example, the same test scores might offer a highly valid measure of intelligence but only a moderately valid indicator of attention deficit/hyperactivity disorder. As a result, the emphasis of validity theory again shifted from test scores to test score interpretations. Yet a valid

interpretation often falls short of justifying a particular use. For example, an employment test might validly measure propensity for job success, but another available test might do as good a job at the same cost but with less adverse impact. In such an instance, the validity of the test score interpretation for the first test would not justify its use for employment testing. Thus, Samuel Messick has urged that test scores are rarely interpreted in a vacuum as a purely academic exercise but are rather collected for some purpose and put to some use. However, in common parlance, one frequently expands the notion of test to refer to the entire procedure of collecting test data (testing), assigning numeric values based on the test data (scoring), making inferences about the level of a construct on the basis of those scores (interpreting), and applying those inferences to practical decisions (use). Thus the term *test validity* lives on as shorthand for the validity of test score interpretations and uses.

Early on, tests were thought to divide into two types: signs and samples. If a test was interpreted as a sign of something else, the something else was understood as a construct, and construct validation was deemed appropriate. For example, responses to items on a personality inventory might be viewed as **[p. 230 ↓]** signs of personality characteristics, in which case the personality characteristic constitutes the construct of interest. In contrast, some tests were viewed as only samples and construct validation was not deemed necessary. For example, a typing test might sample someone's typing and assess its speed and accuracy. The scores on this one test (produced from a sampling of items that could appear on a test) were assumed to generalize merely on the basis of statistical generalization from a sample to a population. Jane Loevinger and others questioned this distinction by pointing out that the test sample could never be a random sample of all possible exemplars of the behavior in question. For example, a person with high test anxiety might type differently on a typing test from the way the person types at work, and someone else might type more consistently on a brief test than over a full workday. As a result, interpreting the sampled behavior in terms of the full range of generalization always extends beyond mere statistical sampling to broader validity issues. For this reason, all tests are signs as well as samples, and construct validation applies to all tests.

At one time, test validity was neatly divided into three types: content, criterion, and construct, with the idea that one of these three types of validity applied to any one type of test. However, criterion-related validity depends on the construct interpretation

of the criterion, and test fairness often turns on construct-irrelevant variance in the predictor scores. Likewise, content validation may offer valuable evidence in support of the interpretation of correct answers but typically will not provide as strong a line of evidence for the interpretation of incorrect answers. For example, someone might know the mathematical concepts but answer a math word problem incorrectly because of insufficient vocabulary or culturally inappropriate examples. Because all tests involve interpretation of the test scores in terms of what they are intended to measure, construct validation applies to all tests. In contemporary thinking, there is a suggestion that all validity should be of one type, construct validity.

This line of development has led to unified (but not unitary) conceptions of validity that elevate construct validity from one kind of validity among others to the whole of validity. Criterion-related evidence provides evidence of construct validity by showing that test scores relate to other variables (i.e., criterion variables) in the predicted ways. Content validity evidence provides evidence of construct validity because it shows that the test properly covers the intended domain of content related to the construct definition. As such, construct validity has grown from humble origins as one relatively esoteric form of validity to the whole of validity, and it has come to encompass other forms of validity evidence.

Messick distinguished two threats to construct validity. *Construct deficiency* applies when a test fails to measure some aspects of the construct that it should measure. For example, a mathematics test that failed to cover some portion of the curriculum for which it was intended would demonstrate this aspect of poor construct validity. In contrast, *construct-irrelevant variance* involves things that the test measures that are not related to the construct of interest and thus should not affect the test scores. The example of a math test that is sensitive to vocabulary level illustrates this aspect. A test with optimal construct validity therefore measures everything that it should measure but nothing that it should not.

Traditionally, validation has been directed toward a specific test, its scores, and their intended interpretation and use. However, construct validation increasingly conceptualizes validation as continuous with extended research programs into the construct measured by the test or tests in question. This shift reflects a broader shift in the behavioral sciences away from *operationalism*, in which a variable is theoretically

defined in terms of a single *operational definition*, in favor of *multioperationalism*, in which a variety of different measures triangulate on the same construct. As a field learns to measure a construct in various ways and learns more about how the construct relates to other variables through evidence collected using these measures, the overall understanding of the construct increases. The stronger this overall knowledge base about the construct, the more confidence one can have in interpreting the scores derived from a particular test as measuring this construct. Moreover, the more one knows about the construct, the more specific and varied are the consequences entailed by interpreting test scores as measures of that construct. As a result, one can conceptualize construct validity as broader than test validity because it involves the collection of evidence to validate theories about the underlying construct as measured by **[p. 231 ↓]** a variety of tests, rather than merely the interpretation of scores from one particular test.

Construct Validation Methodology

At its inception, when construct validity was considered one kind of validity appropriate to certain kinds of tests, inspection of patterns of correlations offered the primary evidence of construct validity. Lee Cronbach and Paul Meehl described a *nomological net* as a pattern of relationships between variables that partly fixed the meaning of a construct. Later, factor analysis established itself as a primary methodology for providing evidence of construct validity. Loevinger described a *structural aspect* of construct validity as the pattern of relationships between items that compose a test. Factor analysis allows the researcher to investigate the internal structure of item responses, and some combination of replication and confirmatory factor analysis allows the researcher to test theoretical hypotheses about that structure. Such hypotheses typically involve multiple dimensions of variation tapped by items on different subscales and therefore measuring different constructs. A higher order factor may reflect a more general construct that comprises these subscale constructs.

Item response theory typically models dichotomous or polytomous item responses in relation to an underlying latent trait. Although item response theory favors the term *trait*, the models apply to all kinds of constructs. Historically, the emphasis with item response theory has been much more heavily on unidimensional measures and

providing evidence that items in a set all measure the same dimension of variation. However, recent developments in factor analysis for dichotomous and polytomous items, coupled with expanded interest in multidimensional item response theory, have brought factor analysis and item response theory together under one umbrella. Item response theory models are generally equivalent to a factor analysis model with a threshold at which item responses change from one discrete response to another based on an underlying continuous dimension. Both factor analysis and item response theory depend on a shared assumption of *local independence*, which means that if one held constant the underlying latent variable, the items would no longer have any statistical association between them. Latent class analysis offers a similar measurement model based on the same basic assumption but applicable to situations in which the latent variable is itself categorical. All three methods typically offer tests of goodness of fit based on the assumption of local independence and the ability of the modeled latent variables to account for the relationships among the item responses.

An important aspect of the above types of evidence involves the separate analysis of various scales or subscales. Analyzing each scale separately does not provide evidence as strong as does analyzing them together. This is because separate analyses work only with local independence of items on the same scale. Analyzing multiple scales combines this evidence with evidence based on relationships between items on different scales. So, for example, three subscales might each fit a one-factor model very well, but a three-factor model might fail miserably when applied to all three sets of items together. Under a hypothetico-deductive framework, testing the stronger hypothesis of multiconstruct local independence offers more support to interpretations of sets of items that pass it than does testing a weaker piecemeal set of hypotheses.

The issue just noted provides some interest in returning to the earlier notion of a nomological net as a pattern of relationships among variables in which the construct of interest is embedded. The idea of a nomological net arose during a period when causation was suspect and laws (i.e., nomic relationships) were conceptualized in terms of patterns of association. In recent years, causation has made a comeback in the behavioral sciences, and methods of modeling networks of causal relations have become more popular. Path analysis can be used to test hypotheses about how a variable fits into such a network of observed variables, and thus path analysis provides construct validity evidence for test scores that fit into such a network as predicted by the

construct theory. Structural equation models allow the research to combine both ideas by including both *measurement models* relating items to latent variables (as in factor analysis) and *structural models* that embed the latent variables in a causal network (as in path analysis). These models allow researchers to test complex hypotheses and thus provide even stronger forms of construct validity evidence. When applied to passively observed data, however, such causal models **[p. 232 ↓]** contain no magic formula for spinning causation out of correlation. Different models will fit the same data, and the same model will fit data generated by different causal mechanisms. Nonetheless, such models allow researchers to construct highly falsifiable hypotheses from theories about the construct that they seek to measure.

Complementary to the above, experimental and quasi-experimental evidence also plays an important role in assessing construct validity. If a test measures a given construct, then efforts to manipulate the value of the construct should result in changes in test scores. For example, consider a standard program evaluation study that demonstrates a causal effect of a particular training program on performance of the targeted skill set. If the measure of performance is well validated and the quality of the training is under question, then this study primarily provides evidence in support of the training program. In contrast, however, if the training program is well validated but the performance measure is under question, then the same study primarily provides evidence in support of the construct validity of the measure. Such evidence can generally be strengthened by showing that the intervention affects the variables that it should but also does not affect the variables that it should not. Showing that a test is responsive to manipulation of a variable that should not affect it offers one way of demonstrating construct-irrelevant variance. For example, admissions tests sometimes provide information about test-taking skills in an effort to minimize the responsiveness of scores to further training in test taking.

Susan Embretson distinguished *construct representation* from *nomothetic span.* The latter refers to the external patterns of relationships with other variables and essentially means the same thing as nomological net. The former refers to the cognitive processes involved in answering test items. To the extent that answering test items involves the intended cognitive processes, the construct is properly represented, and the measurements have higher construct validity. As a result, explicitly modeling the cognitive operation involved in answering specific item types has blossomed as a

means of evaluating construct validity, at least in areas in which the underlying cognitive mechanisms are well understood. As an example, if one has a strong construct theory regarding the cognitive processing involved, one can manipulate various cognitive subtasks required to answer items and predict the difficulty of the resulting items from these manipulations.

Role in Validity Arguments

Modern validity theory generally structures the evaluation of validity on the basis of various strands of evidence in terms of the construction of a *validity argument.* The basic idea is to combine all available evidence into a single argument supporting the intended interpretation and use of the test scores. Recently, Michael Kane has distinguished an *interpretive argument* from the validity argument. The interpretive argument spells out the assumptions and rationale for the intended interpretation of the scores, and the validity argument supports the validity of the interpretive argument, particularly by providing evidence in support of key assumptions. For example, an interpretive argument might indicate that an educational performance mastery test assumes prior exposure and practice with the material. The validity argument might then provide evidence that given these assumptions, test scores correspond to the degree of mastery.

The key to developing an appropriate validity argument rests with identifying the most important and controversial premises that require evidential support. Rival hypotheses often guide this process. The two main threats to construct validity described above yield two main types of rival hypotheses addressed by construct validity evidence. For example, sensitivity to transient emotional states might offer a rival hypothesis to the validity of a personality scale related to construct-irrelevant variance. *Differential item functioning*, in which test items relate to the construct differently for different groups of test takers, also relates to construct-irrelevant variance, yielding rival hypotheses about test scores related to group characteristics. A rival hypothesis that a clinical depression inventory captures only one aspect of depressive symptoms involves a rival hypothesis about construct deficiency.

Unresolved Issues

A central controversy in contemporary validity theory involves the disagreements over the breadth **[p. 233 ↓]** of validity evidence. Construct validation provides an integrative framework that ties together all forms of validity evidence in a way continuous with empirical research into the construct, but some have suggested a less expansive view of validity as more practical. Construct validity evidence based on test consequences remains a continuing point of controversy, particularly with respect to the notion of *consequential validity* as a distinct form of validity. Finally, there remains a fundamental tension in modern validity theory between the traditional fact–value dichotomy and the fundamental role of values and evaluation in assessing the evidence in favor of specific tests, scores, interpretations, and uses.

Keith A. Markus and Chia-ying Lin

http://dx.doi.org/10.4135/9781412961288.n72
See also

Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. pp. 17–64). Westport, CT: American Council on Education and Praeger.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. pp. 13–103). New York: American Council on Education and Macmillan.

Wainer, H., & Braun, H. I. (1988). Test validity. Mahwah, NJ: Lawrence Erlbaum.