Encyclopedia of Research Design

Content Validity

Contributors: Keith A. Markus & Kellie M. Smith
Editors: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Content 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.n74
Print pages: 239-244

http://dx.doi.org/10.4135/9781412961288.n74

Content validity refers to the extent to which the items on a test are fairly representative of the entire domain the test seeks to measure. This entry discusses origins and definitions of content validation, methods of content validation, the role of **[p. 239 ↓]** content validity evidence in validity arguments, and unresolved issues in content validation.

Origins and Definitions

One of the strengths of content validation is the simple and intuitive nature of its basic idea, which holds that what a test seeks to measure constitutes a *content domain* and the items on the test should sample from that domain in a way that makes the test items representative of the entire domain. Content validation methods seek to assess this quality of the items on a test. Nonetheless, the underlying theory of content validation is fraught with controversies and conceptual challenges.

At one time, different forms of validation, and indeed validity, were thought to apply to different types of tests. Florence Goodenough made an influential distinction between tests that serve as *samples* and tests that serve as *signs.* From this view, personality tests offer the canonical example of tests as signs because personality tests do not sample from a domain of behavior that constitutes the personality variable but rather serve to indicate an underlying personality trait. In contrast, educational achievement tests offer the canonical example of tests as samples because the items sample from a knowledge or skill domain, operationally defined in terms of behaviors that demonstrate that corresponding knowledge or skill that the test measures achievement in. For example, if an addition test contains items representative of all combinations of single digits, then it may adequately represent addition of single-digit numbers, but it would not adequately represent addition of numbers with more than one digit.

Jane Loevinger and others have argued that the above distinction does not hold up because all tests actually function as signs. The inferences drawn from test scores always extend beyond the test-taking behaviors themselves, but it is impossible for the test to include anything beyond test-taking behaviors. Even work samples can extend only to samples of work gathered within the testing procedure (as opposed to portfolios, which lack the standardization of testing procedures). To return to the above example, one does not use an addition test to draw conclusions only about answering addition items on a test but seeks to generalize to the ability to add in contexts outside addition tests.

At the heart of the above issue lies the paradigmatic shift from discrete forms of validity, each appropriate to one kind of test, to a more unified approach to test validation. The term *content validity* initially differentiated one form of validity from *criterion validity* (divisible into concurrent validity and predictive validity, depending on the timing of the collection of the criterion data) and *construct validity* (which initially referred primarily to the pattern of correlations with other variables, the *nomological net*, and to the pattern of association between the scores on individual items within the test). Each type of validity arose from a set of practices that the field developed to address a particular type of practical application of test use. Content validity was the means of validating tests used to sample a content domain and evaluate mastery within that domain. The *unified view of validity* initiated by Jane Loevinger and Lee Cronbach, and elaborated by Samuel Messick, sought to forge a single theory of test validation that subsumed these disparate practices.

The basic practical concern involved the fact that assessment of the representativeness of the content domain achieved by a set of items does not provide a sufficient basis to evaluate the soundness of inferences from scores on the test. For example, a student correctly answering arithmetic items at a level above chance offers stronger support for the conclusion that he or she can do the arithmetic involved than the same student failing to correctly answer the items offers for the conclusion that he or she cannot do the arithmetic. It may be that the student can correctly calculate 6 divided by 2 but has not been exposed to the 6/2 notation used in the test items. In another context, a conscientious employee might be rated low on a performance scale because the items involve tasks that are important to and representative of the domain of conscientious work behaviors, but opportunities for which come up extremely rarely in the course of routine work (e.g., reports defective equipment when encountered). Similarly, a test with highly representative items might have inadequate reliability or other deficiencies that reduce the validity of inferences from its scores. The traditional approach to dividing up types of validity and categorizing tests with respect to the **[p. 240 ↓]** appropriate type of validation tends in practice to encourage reliance on just one kind of validity evidence for a given test. Because just one type alone, including content-related validation evidence, does not suffice to underwrite the use of a test, the unified view sought to discourage such categorical typologies of either tests or validity types and replace these with validation methods that combined different forms of evidence for the validity of the same test.

As Stephen Sireci and others have argued, the problem with the unified approach with respect to content validation stems directly from this effort to improve on inadequate test validation practices. A central ethos of unified approaches involves the rejection of a simple checklist approach to validation in which completion of a fixed set of steps results in a permanently validated test that requires no further research or evaluation. As an antidote to this checklist conception, Michael Kane and others elaborated the concept of a *validity argument.* The basic idea was that test validation involves building an argument that combines multiple lines of evidence of the overall evaluation of a use or interpretation of scores derived from a test. To avoid a checklist, the argument approach leaves it open to the test validator to exercise judgment and select the lines of evidence that are most appropriate in a given instance. This generally involves selecting the premises of the validation argument that bear the most controversy and for which empirical support can be gathered within practical constraints on what amounts to a reasonable effort. One would not waste resources gathering empirical evidence for claims that no one would question. Similarly, one would not violate ethical standards in order to validate a test of neural functioning by damaging various portions of the cortex in order to experimentally manipulate the variable with random assignment. Nor would one waste resources on an enormous and costly effort to test one assumption if those resources could be better used to test several others in a less costly fashion. In short, the validity argument approach to test validation does not specify that any particular line of evidence is required of a validity argument. As a result, an effort to discourage reliance on content validation evidence alone may have swung the pendulum too far in the opposite direction by opening the door to validation efforts that exclude content validation where it could provide an important and perhaps necessary line of support. These considerations have led to proposals to modify the argument approach to validation in ways that make content-related evidence necessary or at least strongly recommended for tests based on sampling from a content domain.

Contemporary approaches to content validation typically distinguish various aspects of content validity. A clear *domain definition* is foundational for all the other aspects of content validity because without a clear definition of the domain, test developers, test users, or anyone attempting to do validation research has no basis for a clear assessment of the remaining aspects. This aspect of content validation closely relates to the emphasis in the *Standards for Educational and Psychological Testing* on clearly defining the purpose of a test as the first step in test validation.

A second aspect of content validity, *domain relevance*, draws a further connection between content validation and the intended purpose of the test. Once the domain has been defined, domain relevance describes the degree to which the defined domain bears importance to the purpose of the test. For example, one could imagine a test that does a very good job of sampling the skills required to greet visitors, identify whom they wish to see, schedule appointments, and otherwise exercise the judgment and complete the tasks required of an effective receptionist. However, if the test use involves selecting applicants for a back office secretarial position that does not involve serving as a receptionist, then the test would not have good domain relevance for the intended purpose. This aspect of content validation relates to a quality of the defined domain independent of how well the test taps that domain.

In contrast, *domain representation* does not evaluate the defined domain but rather evaluates the effectiveness with which the test samples that domain. Clearly, this aspect of content validation depends on the previous two. Strong content representation does not advance the quality of a test if the items represent a domain with low relevance. Furthermore, even if the items do represent a domain well, the test developer has no effective means of ascertaining that fact without a clear domain definition. Domain representation can **[p. 241 ↓]** suffer in two ways: Items on the test may fail to sample some portion of the test domain, in which case the validity of the test suffers as a result of construct underrepresentation. Alternatively, the test might contain items from outside the test domain, in which case these items introduce construct-irrelevant variance into the test total score. It is also possible that the test samples all and only the test domain but does so in a way that overemphasizes some areas of the domain while underemphasizing other areas. In such a case, the items sample the entire domain but in a nonrepresentative manner. An example would be an addition test where 75% of the items involved adding only even numbers and no odd numbers.

An additional aspect of content validation involves clear, detailed, and thorough documentation of the test construction procedures. This aspect of content validation reflects the epistemic aspect of modern test validity theory: Even if a test provides an excellent measure of its intended construct, test users cannot justify the use of the test unless they know that the test provides an excellent measure. Test validation involves justifying an interpretation or use of a test, and content validation involves justifying the test domain and the effectiveness with which the test samples that domain. Documentation of the process leading to the domain definition and generation of the item pool provides a valuable source of content-related validity evidence. One primary element of such documentation, the *test blueprint*, specifies the various areas of the test domain and the number of items from each of those areas. Documentation of the process used to construct the test in keeping with the specified test blueprint thereby plays a central role in evaluating the congruency between the test domain and the items on the test.

The earlier passages of this entry have left open the question of whether content validation refers only to the items on the test or also to the processes involved in answering those items. Construct validation has its origins in a time when tests as the object of validation were not yet clearly distinguished from test scores or test score interpretations. As such, most early accounts focused on the items rather than the processes involved in answering them. Understood this way, content validation focuses on qualities of the test rather than qualities of test scores or interpretations. However, as noted above, even this test-centered approach to content validity remains relative to the purpose for which one uses the test. Domain relevance depends on this purpose, and the purpose of the test should ideally shape the conceptualization of the test domain. However, focus on just the content of the items allows for a broadening of content validation beyond the conception of a test as measuring a construct conceptualized as a latent variable representing a single dimension of variation. It allows, for instance, for a test domain that spans a set of tasks linked another way but heterogeneous in the cognitive processes involved in completing them. An example might be the domain of tasks associated with troubleshooting a complex piece of technology such as a computer network. No one algorithm or process might serve to troubleshoot every problem in the domain, but content validation held separate from response processes can nonetheless apply to such a test.

In contrast, the idea that content validity applies to response processes existed as a minority position for most of the history of content validation, but has close affinities to both the unified notion of validation as an overall evaluation based on the sum of the available evidence and also with cognitive approaches to test development and validation. Whereas representativeness of the item content bears more on a quality of the stimulus materials, representativeness of the response processes bears more on an underlying individual differences variable as a property of the person tested. Susan Embretson has distinguished *construct representation*, involving the extent to which items require the cognitive processes that the test is supposed to measure, from *nomothetic span*, which is the extent to which the test bears the expected patterns of association with other variables (what Cronbach and Paul Meehl called *nomological network*). The former involves content validation applied to processes whereas the latter involves methods more closely associated with criterion-related validation and construct validation methods.

Content Validation Methodology

Content-related validity evidence draws heavily from the test development process. The content domain should be clearly defined at the start of **[p. 242 ↓]** this process, *item specifications* should be justified in terms of this domain definition, *item construction* should be guided and justified by the item specifications, and the overall test blueprint that assembles the test from the item pool should also be grounded in and justified by the domain definition. Careful documentation of each of these processes provides a key source of validity evidence.

A standard method for assessing content validity involves judgments by *subject matter experts* (SMEs) with expertise in the content of the test. Two or more SMEs rate each item, although large or diverse tests may require different SMEs for different items. Ratings typically involve domain relevance or importance of the content in individual test items. Good items have high means and low standard deviations, indicating high agreement among raters. John Flanagan introduced a *critical incident technique* for generating and evaluating performance-based items. C. H. Lawshe, Lewis Aiken, and Ronald Hambleton each introduced quantitative measures of agreement for use with criterion-related validation research. Victor Martuza introduced a *content validity* *index*, which has generated a body of research in the nursing literature. A number of authors have also explored multivariate methods for investigating and summarizing SME ratings, including factor analysis and multidimensional scaling methods. Perhaps not surprisingly, the results can be sensitive to the approach taken to structuring the judgment task.

Statistical analysis of item scores can also be used to evaluate content validity by showing that the content domain theory is consistent with the clustering of items into related sets of items by some statistical criteria. These methods include factor analysis, multidimensional scaling methods, and cluster analysis. Applied to content validation, these methods overlap to some degree with construct validation methods directed toward the internal structure of a test. Test developers most often combine such methods with methods based on SME ratings to lessen interpretational ambiguity of the statistical results.

A growing area of test validation related to content involves cognitive approaches to modeling the processes involved in answering specific item types. Work by Embretson and Robert Mislevy exemplifies this approach, and such approaches focus on the construct representation aspect of test validity described above. This methodology relies on a strong cognitive theory of how test takers process test items and thus applies best when item response strategies are relatively well understood and homogeneous across items. The approach sometimes bears a strong relation to the facet analysis methods of Louis Guttman in that item specifications describe and quantify a variety of item attributes, and these can be used to predict features of item response patterns such as item difficulty. This approach bears directly on content validity because it requires a detailed theory relating how items are answered to what the items measure. Response process information can also be useful in extrapolating from the measured content domain to broader inferences in applied testing, as described in the next section.

Role in Validity Arguments

At one time, the dominant approach was to identify certain tests as the type of test to which content validation applies and rely on content validity evidence for the evaluation of such tests. Currently, few if any scholars would advocate sole reliance on content validity evidence for any test. Instead, content-related evidence joins with other evidence to support key inferences and assumptions in a validity argument that combines various sources of evidence to support an overall assessment of the test score interpretation and use.

Kane has suggested a two-step approach in which one first constructs an argument for test score interpretations and then evaluates that argument with a test validity argument. Kane has suggested a general structure involving four key inferences to which content validity evidence can contribute support. First, the prescribed scoring method involves an inference from observed test-taking behaviors to a specific quantification intended to contribute to measurement through an overall quantitative summary of the test takers' responses. Second, test score interpretation involves generalization from the observed test score to the defined content domain sampled by the test items. Third, applied testing often involves a further inference that extrapolates from the measured content domain to a broader domain of inference that the test does not fully sample. Finally, most applied testing involves a final set of **[p. 243 ↓]** inferences from the extrapolated level of performance to implications for actions and decisions applied to a particular test taker who earns a particular test score.

Interpretation of statistical models used to provide criterion- and construct-related validity evidence would generally remain indeterminate were it not for the grounding of test score interpretations provided by content-related evidence. While not a fixed foundation for inference, content-related evidence provides a strong basis for taking one interpretation of a nomothetic structure as more plausible than various rival hypotheses. As such, content-related validity evidence continues to play an important role in test development and complements other forms of validity evidence in validity arguments.

Unresolved Issues

As validity theory continues to evolve, a number of issues in content validation remain unresolved. For instance, the relative merits of restricting content validation to test content or expanding it to involve item response processes warrant further attention. A variety of aspects of content validity have been identified, suggesting a multidimensional attribute of tests, but quantitative assessments of content validity generally emphasize single-number summaries. Finally, the ability to evaluate content validity in real time with computer-adaptive testing remains an active area of research.

Keith A. Markus and Kellie M. Smith

http://dx.doi.org/10.4135/9781412961288.n74
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.

Crocker, L. M.,, Miller, D., &, and Franks, E. A. Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education, (1989). vol. 2, pp. 179–194.

Embretson, S. E. Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, (1983). vol. 93, pp. 179–197.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing, ed. & T. M. Haladyna (Eds.), Handbook of test development (pp. pp. 131–153). Mahwah, NJ: Lawrence Erlbaum.

McKenzie, J. F.,, Wood, M. L.,, Kotecki, J. E.,, Clark, J. K., &, and Brey, R. A. Research notes establishing content validity: Using qualitative and quantitative steps. American Journal of Health Behavior, (1999). vol. 23, pp. 311–318.

Popham, W. J. Appropriate expectations for content judgments regarding teacher licensure tests. Applied Measurement in Education, (1992). vol. 5, pp. 285–301.

Sireci, S. The construct of content validity. Social Indicators Research, (1998). vol. 45, pp. 83–117.