Encyclopedia of Measurement and Statistics

Validity Theory

Contributors: Neil J. Salkind & Kristin Rasmussen
Print Pub. Date: 2007
Online Pub. Date:
Print ISBN: 9781412916110
Online ISBN: 9781412952644
DOI: 10.4135/9781412952644
Print pages: 1033-1036

10.4135/9781412952644.n471

The concept of *validity* is one of the most influential concepts in science because considerations about its nature and scope influence everything from the design to the implementation and application of scientific research. Validity is not an abstract property of any observable, unobservable, or conceptual phenomenon, such as a measurement instrument, a personality trait, or a study design. Rather, validity is a characteristic of the inferences that are drawn about phenomena by human agents and the actions that result from these inferences. Specifically, validity is always a matter of degree and not absolutes. This stems partly from the fact that validity is not an observable characteristic of inferences and actions but something that has to be inferred also.

The evaluation of the degree to which inferences are valid and resulting actions are justifiable is, therefore, necessarily embedded in a social discourse whose participants typically bring to the table diverse frameworks, assumptions, beliefs, and values about what constitutes credible evidence. Specifically, modern frameworks for validity typically list both rational and empirical pieces of evidence as necessary, but in each individual context, what these pieces should look like is open to debate. Put differently, a coherent statement about the validity of inferences and actions requires negotiation as well as consensus and places**[p. 1033 ↓]** multiple responsibilities on the stakeholders who develop such a statement.

Negotiating Validity

A metaphor may illustrate complications that can arise in a discourse about validity. If an educational assessment is viewed as the construction of a house, inferences are markers of the utility of the house. In this sense, an evaluation of the validity of inferences can be viewed as an evaluation of the degree to which the house provides structural support for the purposes that are envisioned for it. Obviously, the parties who are envisioning a certain use of the house are not necessarily the same as the designers or builders of the house, and so discrepancies can arise easily. Of course, other reasons for a mismatch are possible and could stem from a miscommunication between the designers of the house and the users of the house or from a faulty implementation of the design plans for the house. In a sense, the search for inferences

that can be supported can be viewed as the search for how a house can be transformed into a home.

In general, the stakeholders in an assessment can be coarsely viewed as belonging to four complementary groups. First, there are the test developers, who create a research program, a framework, or an instrument under multiple considerations, such as theoretical adequacy and feasibility for practical implementation. Second are the examinees, whose needs in the process are typically more practical and may differ quite substantially from those of the other stakeholders involved. Third are the test users, or the decision makers who utilize the scores and diagnostic information from the assessment to make decisions about the examinees; only rarely are the examinees the only decision makers involved. Fourth are the larger scientific and nonscientific communities to which the results of an assessment program are to be communicated and whose needs are a mélange of those of the test developers, the test users, and the examinees. Therefore, determining the degree to which inferences and actions are justifiable is situated in the communicative space among these different stakeholders.

Not surprisingly, examples of problems in determining the validity of inferences abound. For example, the inferences that test users may want to draw from a certain assessment administered to a certain population may be more commensurate with an alternative assessment for a slightly different population. However, that is not a faulty characteristic of the assessment itself. Rather, it highlights the difference between the agents who make inferences and the agents who provide a foundation for a certain set of inferences, of which the desired inferences may not be a member.

Historical Developments of Validity Theories

Until well into the 1970s, validity theory presented itself as the coexistent, though largely unrelated, trinity of *criterion*-based, *content*-based, and *construct*-based conceptions of validity. According to the criterion-based approach, the validity of an assessment could be evaluated in terms of the accuracy with which a test score could predict or estimate the value of a defined criterion measure, usually an observable performance

measure. The criterion-based model, notably introduced by Edward L. Thorndike at the beginning of the 20th century, owed much of its lingering popularity to an undisputable usefulness in many applied contexts that involve selection decisions or prognostic judgments, such as hiring and placement decisions in the workplace or medical and forensic prognoses. Depending on whether the criterion is assessed at the same time as the test or at a subsequent time, one can distinguish between *concurrent* and *predictive* validity, respectively. Though a number of sophisticated analytical and statistical techniques have been developed to evaluate the criterion validity of test scores, the standard methods applied were simple regression and correlation analyses. The resulting coefficient was labeled *validity coefficient*. Occasionally, these procedures were supplemented by the *known-groups method*. This approach bases validity statements on a comparison of mean test scores between groups with hypothesized extreme values (e.g., devout churchgoers and members of sex-chat **[p. 1034 ↓]** forums on the Internet on a newly developed sexual permissiveness scale).

The content-based model of validity comes into play when a well-defined and undisputed criterion measure is not readily available, especially when the prediction is targeted at a broader and multifaceted criterion (e.g., achievement in a content area like mathematics). An argument for content validity is usually established through a panel of experts, who evaluate the test content in terms of (a) relevance and (b) representativeness for the content area under scrutiny. Not surprisingly, the vagueness and subjectivity of the evaluation process has led many psychometricians to discount the content-based model as satisfying *facevalidity* requirements at best. However, modern proponents of the content-based model have applied a wealth of sophisticated quantitative procedures to ensure and evaluate interrater agreement, thereby trying to lend credibility to otherwise qualitative and judgment-based validity evidence.

Shortcomings of the criterion-based and the content-based models of validity incited the American Psychological Association to set forth technical recommendations for justifying interpretations of psychological tests. As a result of this endeavor, the term *construct validity* was coined and later elaborated by Lee J. Cronbach and Paul Meehl. In the beginning, they tied their validity theory closely to a more general and abstract nomological network, which was described in 1952 by Carl G. Hempel in his classic essay *Fundamentals of Concept Formation in Empirical Science*. Metaphorically and graphically, the constructs are represented by knots, and the threads connecting

these knots represent the definitions and hypotheses included in the theory. The whole system, figuratively speaking, “floats” above the plane of observation and is connected to it by “strings,” or rules of interpretation. The complex system of theoretical definitions can be used to formulate theoretical hypotheses, which can, in turn, be used to formulate empirical hypotheses about relationships among observable variables. In this framework, validity is not a characteristic of a construct or its observed counterpart but of the interpretation of defined logical relations of a causal nature that function to semantically circumscribe a theoretical network of constructs and construct relations.

An obvious epistemological problem arises, however, when the observed relationships are inconsistent with theory, which is exacerbated by the dearth of developed formal theories in many psychological and social science domains. This lack of strong theory led Cronbach to coin the phrases “weak program” and “strong program” of construct validity. He cautions that, without solid theory (i.e., with only a weak program of construct validity), every correlation of the construct under development with any other observed attribute or variable could be accepted as validity evidence. Consequently, in the absence of any coordinated argument, validation research would then resemble more an empirical shotgun procedure than a scientific program.

Such problems notwithstanding, by the 1980s, the notion of construct validity became accepted as the basis for a new framework of validity assessment that is characterized by its unifying nature. The unifying aspect stems primarily from the acknowledgment that interpretive elements like assumptions and value judgments are pervasive when measuring psychological entities and, thus, are unavoidable in any discourse about any aspect of validity. As Samuel Messick, the most prominent proponent of a unified theory of validity, has framed it, “The validation process is scientific as well as rhetorical and requires both evidence and argument.”

The most controversial aspect of the unified concept of validity as developed by Messick pertains to the role of consequences in the validation process. In this view, a validity argument must specifically consider and evaluate the social consequences of test interpretation and test use, which are describable only on the basis of social values. Importantly, his notion of social consequences does not refer merely to test misuse but, specifically, to the unanticipated consequences of legitimate test score interpretation and use. A number of critics reject his idea that evidential and consequential aspects of

construct validity cannot be separated, but despite this debate and recent clarifications on the meaning of the consequential aspects, the question of value justification within a unified validity approach persists.

[p. 1035 ↓]

Philosophical Challenges

The unification of validity theory under a constructivist paradigm has challenged the prevailing implicit and explicit philosophical realism that many applied social scientists had hitherto followed in their practical measurement endeavors. In philosophical realism, a test's task was to accurately measure an existing entity and not to question whether such an entity existed in the first place (an ontological question) or whether it could be assessed at all (an epistemological question). In the constructivist view, it is not a test that is validated but its interpretation (i.e., the inferences that are drawn from a test score). Therefore, it is insufficient to operationalize validity through a single validity coefficient. Rather, validation takes the form of an open-ended argument that evaluates the overall plausibility of the proposed test score interpretations from multiple facets. Currently, the strengthening of cognitive psychology principles in construct validation as described by Susan Embretson and Joanna Gorin, for example, appears to be one of the most promising avenues for developing validity theory toward a more substantive theory that can truly blend theoretical models with empirical observations. Models with genesis in cognitive psychology enable one to disentangle and understand the processes that respondents engage in when they react to test items and to highlight the test instrument as an intervention that can be used to search for causal explanations, an argument that was developed recently in detail by Borsboom, Mellenbergh, and van Heerden.

Perspectives for the Future

To comprehensively develop a unified theory of validity in the social sciences, a lot more must be accomplished besides a synthesis of the evidential and consequential bases of test interpretation and use. In particular, a truly unified theory of validity would

be one that crosses methodological boundaries and builds on the foundations that exist in other disciplines and subdisciplines. Most prominently, consider the *threats-to-validity* approach for generalized causal inferences from experimental and quasi-experimental designs, the closely related *validity generalization* approach by virtue of meta-analytical techniques, and the long tradition of validity concepts in qualitative research. In the end, it may be best to acknowledge that validity itself is a complex construct that also needs to be validated every once in a while.

André A. Rupp and Hans Anand Pant

10.4135/9781412952644.n471

See also

Further Reading

Borsboom, D. , Mellenbergh, G. J. , and van Heerden, J. The concept of validity. Psychological Review vol. 111 no. (4) pp. 1061 (2004).

Cureton, E. E., Cronbach, L. J., Meehl, P. E., Ebel, R. L., Ward, A. W., & Stoker, H. W. (1996). Validity. In A. W. Ward, ed. , H. W. Stoker,, ed. & M. Murray-Ward (Eds.), Educational measurement: Origins, theories, and explications: Vol. 1. Basic concepts and theories. Lanham, MD: University Press of America.

Embretson, S. and Gorin, J. Improving construct validity with cognitive psychology principles. Journal of Educational Measurement vol. 38 no. (4) pp. 343 (2001).

Hempel, C. G. (1952). Fundamentals of concept formation in empirical science. Chicago: University of Chicago Press.

Kane, M. T. Current concerns in validity theory. Journal of Educational Measurement vol. 38 no. (4) pp. 319–342 (2001).

Markus, K. A. Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? Social Indicators Research vol. 45 pp. 7–34 (1998).

Murphy, K. R. (2003). Validity generalization: A critical review. Mahwah, NJ: Erlbaum.

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

Whittemore, R. , Chase, S. K. , and Mandle, C. L. Validity in qualitative research. Qualitative Health Research vol. 11 no. (4) pp. 522–537 (2001).