URN: urn:nbn:de:bvb:473-irb-472429
DOI: https://doi.org/10.20378/irb-47242

Clarifying the concept of validity – From measurement to everyday language

Matthias Borgstede (formerly Buntins), Katja Buntins, Frank Eggert

Abstract

Test validity is widely understood as the *degree to which a test measures what it should measure* (cf. Cattell, 1946). We argue that this conceptualization does not refer to a psychometric problem but to the correspondence between scientific language and everyday language.

Following Steven's (1946), test results give an operational definition of attributes, qualifying any test as valid by definition. Following the representational theory of measurement (Krantz, Luce, Suppes & Tversky, 1971), an attribute is defined by an empirical relational structure and a corresponding measurement model. Since measurement depends on the specified empirical structure, if a test measures anything, it must be valid.

However, the question of validity can be asked in a meaningful way, if one interprets test results in the context of everyday language. We conclude that validity can be understood as the *degree to which the variable measured by a test corresponds to concepts of everyday language*.

Clarifying the concept of validity – From measurement to everyday language

Introduction

Test validity is one of the key concepts of psychometric theory. Traditionally, it refers to *the degree to which a test measures what it should measure* (Buckingham, 1921; Cattell, 1946; Courtis, 1921; Kelley, 1927). Although validity seems to be a very basic and simple concept, there has been a considerable amount of debate about its exact meaning. When Cronbach and Meehl (1955) introduced the idea of construct validity, they tried to explicate the concept in relation to a nomological network. Construct validity thus understood refers to the question whether test score interpretations fit a certain theoretical background. This idea was further developed by Loevinger (1957) and later Messick (1989) who broadened the concept to include social consequences of psychological test use. In accordance with this conception, Kane (1992, 1994) highlights the argumentative character of test validity. Thus, the concept of validity drifted from a measurement specific test characteristic to an evaluative statement about the interpretation of test scores (cf. Shear & Zumbo, 2014). In more recent years, some new approaches emerged (e.g. Embretson, 2007); Lissitz & Samuelsen, 2007; Zumbo, 2007, 2009), which all depart from the original meaning.

However, the theoretical work on validity theory remained largely academic, bearing little or no significance to the fact that most psychologists still think of validity in the traditional sense (Borsboom, Mellenbergh & van Heerden, 2004). As a consequence, Borsboom (2005) revived the original conception of test validity by introducing a very straightforward explication of validity: A test is valid if the attribute to be measured exists and causes variation in test scores.

Like Borsboom et al. (2004) the paper at hand is concerned with the traditional view of validity, because it is still the most widely accepted account of the concept when it comes to

actual test use. However, we arrive at a completely different conclusion: If one explicates test validity from a measurement theoretical view, it turns out that the degree to which a test measures what it should measure is an obsolete concept. It is further argued that the original meaning of test validity relies on a naïve conception of measurement, which has its origins in everyday language and common-sense psychology. Following this line of reasoning, we propose that validity does not pose a psychometric problem whatsoever but rather a problem of language use.

Validity and psychological measurement

In order to speak meaningfully of the question whether a test measures what it should measure, one needs to know what the test should measure. However trivial this insight may seem, it has been overseen in validity theory so far, resulting in an incomplete analysis of the concept. Following this insight, we conclude that a statement about a test's validity can only be meaningful, if the attribute to be measured is *well-defined*[1].

The second step is to realize that what counts as a definition for an attribute depends on the underlying theory of measurement. Whereas this question is uncontroversial within natural science, the meaning of the word 'measurement' has been a point of debate in psychology ever since (cf. Michell, 1999). In psychology, there are two main positions about the nature of measurement: Stevens (1946) argues that measurement is the *assignment of numerals to objects according to a specified rule*. On the other hand, there are advocates of the so-called representational theory of measurement (e.g. Krantz et al., 1971), who define measurement as a *homomorphism from an empirical relational system to a numerical relational system*.

[1] We do not agree with the much stronger claim that the attribute must *exist*, however (cf. Borsboom, 2005).

Validity from an operationalist point of view

Stevens‘ (1946) conception of measurement as the assignment of numerals to objects according to rule follows an operationalist view of measurement (Borsboom, 2005). Operationalism understands measurements not as referring to real entities nor as an abstraction of empirical relations, but as the result of an *arbitrary* operation. The only constraints to this operation are the use of numerals and of a rule. Because the rule of assignment may be arbitrary, it is not justified to interpret test scores in a sense, which surpasses the actual measurement operation. Following Stevens, the attribute to be measured is entirely defined by the specified measurement procedure. A psychological attribute like e.g. ‘intelligence’ is thus nothing but a numeral that is read from a table which compares the number of correct answers in an intelligence test with a specified reference population. A theoretical interpretation in the sense of a cognitive ability would go beyond the measurement result and therefore be speculative (or metaphysical).

Since the attribute, which should be measured by a test, is defined as the result of the measurement procedure, test scores are identical with the level of the attribute by definition. Hence, it is impossible to measure anything else but the attribute, which should be measured, because the attribute is defined as the result of the test. Consequently, every test is valid by definition. The question whether a test *actually* measures what it should measure is – following Sevens’ definition of measurement – obsolete.

Validity from a representationalist point of view

The representational theory of measurement takes a fundamentally different position (e.g. Krantz et al., 1971). The representationalist sees measurement as a mapping of empirical relations onto numerical relations, such that each level of the attribute to be measured is assigned exactly one number and the relations between attribute levels are preserved in the

corresponding numerical relations (mathematically speaking, the mapping forms a *homomorphism*). The central concept is the preservation of empirical structures when numbers are assigned to objects. Whereas Stevens' account qualifies any (arbitrary) rule of assignment as measurement, the representational theory of measurement restricts measurement to those cases where numbers actually express empirical facts about the attribute which should be measured. It is important to notice the fact that within this framework, measurable attributes are *entirely* characterized by the empirical relational structure, which is attempted to be modelled.

We shall give a simple example in order to illustrate this idea: Imagine a collection of stones, which are put on a beam scale in order to compare their weight. The empirical relation in this scenario would consist in the position of the beam scale when stones are put into the pans. Formally speaking, we deal with a dominance relation ($\succ$) over the cross product of this set of stones with itself. We might call this relation: 'weighs more than'. If we were to measure the weight of these stones, we would have to assign numbers to the stones such that every time a stone weighs more than another does, its assigned number is indeed bigger than the number of the other stone. These numbers could then be called 'weight', expressing an ordered sequence of objects, which are compared by means of a beam scale[2].

How then is an attribute defined within the framework of the representational theory of measurement? Since measurement is understood to be a mapping of empirical relations onto numerical relations, two steps are involved in the definition of an attribute: Firstly, one must select an empirical relational system, and secondly, a suitable measurement model must be applied to this structure. If such a model is successfully applied to the empirical structure under consideration, we may interpret the resulting numbers as measurements. In contrast to

[2] Usually, when we weigh objects another relation is involved, which consists in empirical concatenation of different objects (like putting more than one object into the pans at the same time).

the operationalist view, the actual act of measuring (the *operation* involved) is merely a technical detail, which does not affect the meaning of the results.

In the context of psychological measurement, we mostly deal with empirical structures involving people's responses to a set of test items. 'Intelligence' for example, could be characterized by the empirical relation of a person giving a correct answer to a test item. This would correspond to a dominance relation on the cross product of test items and persons. Suitable measurement models for this kind of structure could be the Guttman-Model (Guttman, 1950), or its probabilistic version, the Rasch-Model (Rasch, 1960). To evaluate the question whether measurement was successful in this context, one would simply conduct a series of empirical model tests. If the model fits the empirical structure, one may speak of measuring intelligence with this test. 'Intelligence', however, is now defined within the context of the measurement model and refers to the ability of a person to solve this kind of test item.

The definition of an attribute not only depends on the involved empirical structure, but is *determined* by the successful application of a measurement model to this structure. The question whether measurement results actually refer to the attribute, which is intended to be measured, is therefore identical to the question whether the measurement model fits the underlying empirical structure[3]. If it does not, there is no measurement whatsoever. The attribute that should be measured is not defined, if the model does not fit. Thus, it is not possible to have a test that measures something, but does not measure the attribute it should measure. Consequently, in the context of representational measurement, the concept of validity is obsolete, too.

[3] This was anticipated by Borsboom et al. (2004) when they state that '…little remains of the validity problem, which is virtually reduced to the question whether this theory of response behavior is true'.

Conclusion

It has been shown that both operationalist and representationalist accounts of measurement render the question of test validity obsolete. Following operationalism, every measurement is valid by definition. Representationalism, on the other hand, does not allow to talk about measurement without specifying what is measured – hence if a procedure counts as measurement, it is necessarily valid, too.

The definition of what a test should measure seems to be incompatible with a coherent use of the word 'validity'. Since the concept of validity cannot be applied without such a definition, we arrive at a logical dilemma: Whether you define what a test should measure or not – the concept of validity cannot be applied in a meaningful way within the framework of measurement theory.

Validity and the nature of scientific language

The above sections argued that from the viewpoint of measurement theory, there is no validity problem. Why then is the traditional concept of test validity so widespread? What do we mean when we ask the question whether a test is valid? We shall try to give an answer to these questions by taking a closer look at how scientific concepts develop from everyday language.

From the perspective of measurement theory, the definition of an attribute is arbitrary – for example, there is no formal criterion to decide which items should be included in an intelligence test. However, from a practical point of view, the choice of test items in an intelligence test is by no means arbitrary – depending on the theoretical concepts of the scientific community, some items clearly resemble the intended concept and others do not.

In the natural sciences, the meaning of theoretical concepts does not pose any problem, because science usually relies on explicit definitions within a formal theory. In the social

sciences, however, the meaning of many theoretical terms is not unambiguous. Especially in psychology, we face the problem that many concepts are indeed only vaguely defined (Blumer, 1940; Buntins, 2014; Buntins, Buntins & Eggert, 2015). The key problem is that the question which test items are instances of the concept underlying a measurement procedure ultimately is a matter of how the underlying concept is used in everyday language. According to the vague boundaries of many psychological terms in everyday language, the collection of test items used to formally define the concept may vary from test to test. This is indeed the case, resulting in a great variety of psychological tests seemingly aiming at measuring the same thing.

Presuming that concepts of common-sense psychology are useful at least to a certain degree, adapting these concepts is a reasonable starting point for psychological measurement. However, common sense terms need to be clarified and revised based on accumulating empirical evidence, in order to arrive at a formal definition in the context of a scientific theory. The gradual advance from concepts of everyday language to theoretical terms seems to be a natural step in scientific progress. Moreover, forming unambiguous definitions of the terms used in scientific language is a necessary precondition to the empirical evaluation of theories. This process of clarifying the meaning of theoretical terms thus serves an essential part of scientific theory formation (cf. Hempel, 1965).

However, at the same time the scientific vocabulary eventually drifts away from the original meaning of the terms in use, as found in everyday language. Scientific concepts are no longer comprehensible without profound knowledge of the underlying theories. The person parameter estimates of a Rasch-scaled intelligence test, for example, are semantically similar to what is meant with ‘intelligence’ in everyday language, nevertheless they are not identical in meaning. They are part of a scientific language, which can only be understood if one is familiar with the assumptions of the Rasch model and the underlying theory of

intelligence. We find the same phenomenon in the natural sciences: The concepts of physical force, work or energy, for example, are closely related to the corresponding terms in everyday language. However, no one would try to apply these scientific terms in their original common-sense meaning – like, for example, applying the concept of 'work' in Newtonian mechanics to the context of 'work-life-balance'.

Curiously, this is exactly what happens in modern day psychology. Psychological theories naturally develop from common-sense conceptions and, consequently, use their vocabulary. However, as science evolves towards formal theories (which are a necessary precondition of measurement), terms like 'intelligence' undergo a considerable change in meaning. But instead of accepting this as a sign of scientific progress, psychologists seem to be desperate to re-establish the original meaning by asking, whether the terms of their theories resemble the corresponding common concepts. It is this line of reasoning which motivates questions like 'do the person parameters of this Rasch-scaled intelligence test *actually* refer to his/her intelligence?'.

Having this pointed out, we propose that the question what a test should measure does not refer to the actual attribute underlying test scores, but to the meaning of the term used to label this attribute as it is understood in everyday language. Thus, the question of test validity does not pose a psychometric problem at all. Rather, it refers to the correspondence between scientific vocabulary and everyday language.

Discussion

This paper dealt with the concept of test validity. A careful analysis of its meaning within the context of measurement theory yielded the result that validity is not a question of psychometrics but deals with the correspondence between scientific language and common psychological concepts that are used in everyday language.

What are the implications of this result? First of all, from a theoretical point of view, the explication of a meta-theoretical term like test validity is vitally important. Conceptual clarity is one of the most important characteristics of the scientific method. This is especially the case when dealing with methodological concepts. Secondly, the above analysis turns out to be extremely relevant for the field of psycho-diagnostics. The use of psychological tests plays an important role in modern western society. Psychological tests often form the basis of decision processes, both on the individual level (e.g. in the context of employee selection), and on the level of whole populations (e.g. large-scale assessments in educational science). Within this context, the concept of validity is a crucial factor when it comes to give a social justification for decisions made on the ground of psychological test scores. Our analysis, however, shows that validity is not a scientific concept at all. Instead, the concept seems to serve a social purpose – that is to justify selection processes and political decisions on scientific grounds. This result seems to be anticipated by Messick (1989) when he states: 'Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores' (p. 13, italics in the original). However, Messick does not seem to recognize the theoretical implications of this conception, especially the fact that it obscures the boundaries between basic science and its application to solve social problems.

We think that meshing up scientific language with common-sense psychology is a bad idea. Not only does it tempt to draw dubious or even wrong conclusions from psychological test scores, it actually hinders scientific progress in the field of psychology. The development of a theoretical language is a natural process as science proceeds. It is not to be expected that the terms used in the scientific community have direct counterparts in everyday language. If scientific language did not differ from everyday language, what would be the point of theory formation? The aim of science should be to find out *new* things about the world we live in –

not to refrain our common-sense beliefs. Therefore, psychometricians should not try to model their tests after common-sense psychology, but after substantial scientific theory.

References

Blumer, H. (1940). The problem of the concept in social psychology. *American Journal of Sociology 45*, 707-719.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics*. Camebridge: Camebridge University Press.

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061-1071.

Buntins, M. (2014). *Psychologische Tests und mehrwertige Logik – Ein alternativer Ansatz zur Quantifizierung psychologischer Konstrukte.* Wiesbaden: Springer VS.

Buntins, M., Buntins, K. & Eggert, F. (2015). Psychological tests from a (fuzzy-)logical point of view. *Quality and Quantity* (online first).

Cattell, R. B. (1946). *Description and measurement of personality.* New York: World Book Company.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52* (4), 281-302.

Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*(1), 78-90.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researhcer, 36*(8), 449-455.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction: Studies in social psychology in World War II* (Vol. 4, pp. 60-90). New York: Princeton University Press.

Hempel, C. G. (1965). Fundamentals of Taxonomy. In C. G. Hempel, *Aspects of scientific explanation: And other essays in the Philosophy of Science*. New York: The Free Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (1994). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspective, 2*(3), 135-170.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmilllan.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement, vol. 1: Additive and polynomial representations.* New York: Academic Press.

Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Loevinger, (1957)

Messick, S. (1989). Validity. *Educational Measurement, 3* (1), 13-103.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept.* Cambridge: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Kopenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Shear, B. R. & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in Educational and Psychological Measurement. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral and health sciences* (pp. 91-111). Heidelberg: Springer.

Stevens, S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 45-79). Amsterdam: Elsevier.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity:*

Revisions, new directions and applications (pp. 65-82). Charlotte: Information Age Publishing Inc.