PSYCHOMETRIKA—VOL. 71, NO. 3, 441–445
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1491-2

IN PRAISE OF PLURALISM. A COMMENT ON BORSBOOM

MICHAEL KANE

NATIONAL CONFERENCE OF BAR EXAMINERS, MADISON, WI

I tend to agree with Professor Borsboom that psychology, and more generally the social sciences, could benefit from better psychometric modeling. However, if psychometric developments are to have more effect on everyday practice in psychology, psychometricians probably need to pay more attention to the substantive and methodological problems in various areas of psychology. For example, Professor Borsboom is critical of the *Standards for educational and psychological testing* (AERA, APA, NCME, 1999) for suggesting that group differences in test-criterion relationships are relevant to test bias. He bases his criticism on the finding that predictive invariance is not the same as measurement invariance. However, he fails to acknowledge the social, political, and ethical problems associated with failures of predictive invariance in high-stakes contexts (e.g., employment and admissions testing). The *Standards* are designed to provide test publishers and test users with guidance on a range of practical measurement problems. In this context, predictive invariance is a major issue in itself. If we want psychologists to pay more attention to psychometric analyses, these analyses need to recognize the psychologists' problems and goals.

Before getting into a discussion of Professor Borsboom's analyses, it is probably useful to consider his desired state of affairs. He mentions several options for the interpretation of attributes, but seems to prefer what he calls a "reflective latent variable modeling scheme," in which a latent attribute is assumed to cause the test behavior, and the psychometric model reflects this causal relationship. In an earlier paper, Professor Borsboom and his colleagues argued that a test is valid as a measure of an attribute:

> if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. (Borsboom, Mellenbergh, & Van Heerden, 2004, p. 1061.)

Under the reflective model, the latent attribute is the variable of interest, and the test is developed to reflect the attribute. I am comfortable with this kind of interpretation. It is certainly a common model for the interpretation of measurements in the social sciences. However, as illustrated below, it is not the only viable kind of score interpretation, and it is not the best interpretation in many cases.

Operational Definitions, Classical Test Theory, and Construct Validity

Professor Borsboom considers three theoretical obstacles to the integration of psychometrics and psychology: operational definitions, classical test theory, and construct validity. My remarks will focus on these three basic concerns.

Requests for reprints should be sent to mkane@ncbex.org.

Operational definitions were introduced into physics in reaction to the overthrow of traditional, common-sense assumptions about space and time by Einstein's theory of relativity. Bridgeman (1927) and the logical positivists sought to eliminate this kind of upheaval by eliminating implicit assumptions in science. They did not so much equate theoretical constructs with observable attributes, as strive to eliminate theoretical assumptions from their descriptions of observations. They tried to be absolutely clear about what they were doing (generally a good habit), but they also tended to downplay or eliminate theory altogether (not generally a good strategy).

Following this lead, some psychologists decided to define some theoretical attributes (e.g., intelligence) in terms of specific measures. Ironically, this simple replacement of theoretical attributes by observable attributes, called "operationism," had effects diametrically opposed to Bridgeman's goal (Ennis, 1973). Instead of eliminating unwarranted theoretical assumptions, "operationism" assigned all of the assumptions associated with a theoretical construct to the scores on a particular test, thus importing unwarranted assumptions by the carload. To define a theoretical term like intelligence narrowly in terms of a specific measure, while interpreting it broadly in terms of the traditional notion of intelligence, is clearly unwarranted.

However the operational specification of measurement procedures is certainly legitimate, if not essential. The operations used to collect data and to generate scores should be clearly described. Measurement procedures should be operationally defined, but theoretical attributes cannot be operationally defined.

Professor Borsboom sees the "true scores" of classical test theory as reinforcing operationist tendencies in psychology. The *true* score, which is defined as the expected score over replications of the measurement procedure, is clearly dependent on the operational definition of this procedure. However, true scores are used mainly as a basis for analyzing the precision, or reliability, of measurements, and in classical test theory, reliability is paired with validity, which examines the relationship between the true scores and the variable of ultimate interest. By focusing on the distinction between the true score and the variable of interest, validity theory tends to run counter to operationism.

The theory of validity has a long and checkered history, but by the 1980s, a general conception of construct validity provided a unified framework for validity (Messick, 1989). In the original formulation of construct validity (Cronbach & Meehl, 1955), substantive theory was assumed to provide a "nomological" network of relationships among theoretical constructs and observable attributes, and the meanings of the constructs were determined by their roles in this network. The validity of a measure of a theoretical construct would be evaluated in terms of how well its scores satisfied the relationships in the network.

Initially, the nomological networks were conceived of as formal theories (e.g., Newton's laws), but because such theories are rare to nonexistent in psychology, the requirement was relaxed to include open-ended collections of relationships involving the construct of interest. There was a shift from what Cronbach (1989) called the "strong form" of construct validity to what he called the "weak form" of construct validity.

Under the weak form of construct validity, the tight networks envisioned by Cronbach and Meehl (1955) were replaced by collections of relationships involving the construct. For constructs of any generality, such collections could be both vast and ill-defined, making it very difficult to evaluate the measure's fit to the network. Professor Borsboom's conclusion that construct validity functions as "a black hole from which nothing can escape" overstates the case, but by rolling all of the issues inherent in justifying a proposed interpretation into one big ball, many discussions of construct validity have tended to discourage would-be validators.

Nevertheless, the basic question addressed by validity theory, how to justify claims based on test scores, is of fundamental importance. I have suggested that validation can be simplified without being trivialized by requiring that the inferences and decisions to be derived from test

scores be spelled out and evaluated (Kane, in press). This approach allows for a variety of possible interpretations and uses for test scores, with the caveat that any proposed interpretation or use be justified by appropriate evidence. So, operationally defined variables are fine as long as we recognize them for what they are, and do not slide any theoretical claims in under the radar. A claim that the score resulting from a measure can be interpreted as an estimate of a latent attribute that causes the observed performances is also acceptable as long as the claim can be justified. A theory-based interpretation is admissible as long as the theory is specified and the measure's fit to the theory is established.

Professor Borsboom argues that construct validity "must be fundamentally ill-conceived for the simple reason that no physicists are currently involved in the 'neverending process' of figuring out whether meter sticks really measure length" (Borsboom, 2006, p. 431). Of course, it is also hard to find physiometric models (corresponding to our psychometric models) that specify a causal relationship between the latent attribute of length and the observed extension of objects in space. Length once provided a classic example of an operationally defined attribute (remember the platinum–iridium bar in a temperature-controlled chamber in Paris—the standard meter). Now, it can be considered a theoretical attribute within the special theory of relativity. The operational definition was adequate at one time and is still adequate in many contexts. The newer, theory-based definition is used when it is needed. Having methodologists tell scientists what they can and cannot do would limit progress, if the scientists paid any attention to this advice; luckily, they generally don't pay much attention.

The Role of Theory in the Development and Validation of Measures

Professor Borsboom suggests that a psychologist who proposes a measure for a theoretical attribute should spell out the relationship between the attribute and the proposed measure and that this would lead the researcher, "to start the whole process of research by constructing a psychometric model" (Borsboom, 2006, p. 429). This is all well and good, but how does the researcher go about defining the attribute, the measure, and the relationship between the two? Presumably, the process is not arbitrary. We would generally not propose a measure of intelligence that was based on speed in running the mile, just as we would not propose a measure of physical fitness based on completing verbal analogies. However, on a finer level, what tasks should be included in a measure of intelligence (or fitness) and what should be left out? And, how do we evaluate how well the measure is working?

One important determiner of how this process is likely to work is the state of theory development in the area under investigation, with little or no substantive theory at one end of the continuum and a formal theory that can guide test development at the other end.

Development and Validation of Measures without Theory

Assuming that no formal theory exists (the usual case), the test-development process is necessarily ad hoc, guided by general conceptions of the attribute of interest. For example, intelligence is assumed to promote success on a wide range of cognitive tasks, and measures of intelligence generally consist of samples of such tasks. The model is causal but not detailed or formal.

In the absence of a formal theory that specifies a particular form for the psychometric model, the researcher who follows Professor Borsboom's advice is likely to adopt some standard unidimensional IRT model. We have a test consisting of a set of tasks that are thought to reflect the latent attribute, and the IRT model is applied to responses to these tasks. The estimation

of the model parameters requires large sample sizes, which are not available in many areas of psychology, but if the model can be applied, it yields estimates of a latent ability for each person and of one or more parameters for each task.

The resulting latent ability scale derives most, if not all, of its substantive meaning from the sample of tasks on which it is defined. A formal IRT model can be applied to different kinds of tasks to define different scales; like all formal systems, it does not, in itself, contribute substantive content. The population on which the scaling is conducted can also influence the proposed interpretation, but in most applications of IRT, it is assumed that the scale is invariant across persons and groups. So, the scale derived from a standard IRT analysis is largely determined by the observations used to estimate the model parameters. The latent scale is operationally defined in terms of the domain of tasks on which it is based (not that there is anything wrong with that). The fact that the responses have been scaled using a psychometric model does not turn the scale into a theoretical construct. Bridgeman's (1927) conception of operational definitions allowed for the use of sophisticated mathematical models (Benjamin, 1955).

The choice of which of the currently available IRT models to use is often dictated by the preferences of the modeler; some like fewer parameters and some like more parameters. The choice may also be influenced by fit statistics, with preference going to the model that provides the best fit to the data. In the absence of strong substantive theory, this is a reasonable basis for evaluating models, but it also reflects the dependence of the scale on the observations generated by the measurement procedure and not on an a priori conceptualization of the theoretical attribute.

In the absence of theory, we do not know how the latent attribute has its effect (although we may have some hypotheses), and we do not know how this attribute is related to other variables (although we may have some hypotheses). In order to develop our understanding of these issues, we will need to do some empirical research and some theorizing. To draw conclusions about the attribute (e.g., that test results can be generalized to new contexts) simply because we scaled the responses and assigned a trait label to the scale (e.g., "intelligence") would be unwarranted.

Assuming that we want to use our scale scores to make some predictions about future performance on nontest tasks in nontest contexts, it would be prudent to examine how well these predictions turn out. Assuming that we want to make causal claims about how the attribute affects performance on the tasks included in the measure or on other tasks, we would need to develop support for these inferences, and the support for such causal inferences generally involves both empirical research and theory development. The procedures used to develop support for such inferences have been discussed under the heading of construct validity (Messick, 1989).

It is tempting to interpret the latent ability estimates generated by the IRT model as a real, causal attribute, and if no claim is to be made beyond this immediate causal claim (i.e., that some otherwise unspecified latent attribute causes the observed performances), the causal attribution does not make much difference. However, if the hypothesized causal claim is used to justify other inferences (e.g., predictions about future performance on other tasks or in other contexts), then these additional claims need to be examined.

Development and Validation of Measures Based on Theory

If solid theory exists, it can be used to guide test development in a way that builds support for a reflective interpretation in terms of a causal, latent attribute. In particular, if the theory provides a causal explanation of the relationship between the latent attribute and performance on some set of tasks, performance on the tasks can be used to draw conclusions about the causal attribute. This approach works well in areas with highly developed quantitative theories, but it cannot be implemented otherwise. A specification of how a theoretical attribute produces certain effects

is possible only after the theory is in place; it is not the first step, but one of the last steps in a research program.

An alternative approach that has been applied to several kinds of test performance is to take a well-established measure of some attribute (e.g., intelligence) and develop causal models for the performances included in the measure (Embretson, 1998). This research tradition takes performance on the measure as an observable variable of interest and seeks to explain the performance in terms of latent abilities and task characteristics. The dependent variables in these analyses, performances on specific tasks, are operationally defined. The independent variables used to specify task characteristics are also operationally defined. Latent ability, which is determined by overall performance on the test as a whole, is also initially operationally defined, but can be interpreted as a latent, causal attribute after the causal model of performance has been developed.

The Role of Psychometricians

At the end of his paper, Professor Borsboom suggests that psychometricians read widely and that they get involved in the development of substantive theory. This is a great suggestion. It can foster the development of both substantive areas and psychometric theory. I think that it would be especially useful for psychometricians to join research groups as full participants who are actively engaged in all aspects of research projects. I am not talking about a consulting gig or an elegant reanalysis of some existing data set (not that there is anything wrong with either of these activities), but rather about participating in the preliminary analysis, hypothesis development, study design, and data collection, as well as the analysis and interpretation of the results.

Psychologists are likely to make use of psychometric models if they perceive these models to be helpful to them in achieving their goals. Getting them to make greater use of psychometric models is partly a function of making the models accessible (through education, better textbooks, computer programs), as Professor Borsboom suggests, but it is also a function of getting their attention, and a great way to get their attention is to show them what you can do for them.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Benjamin, A. (1955). *Operationism*. Springfield, IL: Charles C. Thomas.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440

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

Bridgman, P.W. (1927). *The logic of modern physics*. New York: Macmillan.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

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

Embretson, S.E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Ennis, R. (1973). Operational definitions. In H. Brody, R. Ennis, & L. Krimerman (Eds.), *Philosophy of educational research* (pp. 650–669). New York: Wiley.

Kane, M. (in press). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Manuscript received 15 MAR 2006
Final version received 20 MAY 2006
Published Online Date: 23 SEP 2006