PSYCHOMETRIKA—VOL. 71, NO. 3, 451–455
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1497-9

PSYCHOMETRICS IN PSYCHOLOGICAL RESEARCH: ROLE MODEL OR PARTNER IN SCIENCE?

KLAAS SIJTSMA

TILBURG UNIVERSITY

This is a reaction to Borsboom's (2006) discussion paper on the issue that psychology takes so little notice of the modern developments in psychometrics, in particular, latent variable methods. Contrary to Borsboom, it is argued that latent variables are summaries of interesting data properties, that construct validation should involve studying nomological networks, that psychological research slowly but definitely will incorporate latent variable methods, and that the role of psychometrics in psychology is that of partner, not role model.

Key words: construct validity, latent variable models, role of psychometrics in psychology, status of latent variables.

Introduction

Borsboom (2006) addresses a topic that probably has occupied psychometricians' minds for quite some time: Why is it that psychology takes so little notice of the modern developments in psychometrics? According to Borsboom, these developments are, in particular, latent variable models such as item response models, confirmatory factor models, and latent class models. Borsboom suggests three causes for this lack of interest.

First, their operationalistic orientation leads psychologists to ignore multidimensional attribute structures or nominal attributes as possible causes of test scores. Instead, psychological attributes are equated with test scores such as the number-correct score. This results in assigning, more or less habitually, the same properties to attributes than to test scores. An example is the linear ordering of individuals on a single dimension.

Second, their reliance on classical test theory prevents psychologists from seeing the distinction between observable variables and psychological attributes. This is due to the definition of the true score as the expectation of an observable variable, the test score. Thus, research is led in the direction of investigating reliability and validity of test scores and away from studying relationships between psychological attributes and behavior elicited by the test.

Third, psychologists have embraced the Popperian idea that theory is continuously under construction, thus accepting beforehand that it cannot ever be fully determined what a test measures. Thus, the process of construct validation (i.e., finding the meaning of a test score) becomes an enterprise that is always underway, and the discouraging thought of endlessly investigating the meaning of measurement is often taken as an excuse for not trying at all.

Borsboom's way out of this dead-end street is to be found in latent variable modeling. Although I think that the propagation of latent variable methods will stimulate the construction of better instruments, I do have four critical comments. They concern the status of statistical latent variables relative to psychological attributes, construct validation, the state of psychological research, and the role of psychometrics in psychology.

Requests for reprints should be sent to Klaas Sijtsma, Department of Methodology and Statistics, FSW, Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: k.sijtsma@uvt.nl.

Latent Variables and Psychological Attributes

Like Borsboom, I believe that the widespread use of latent variable models will stimulate researchers to think about:

(1) the possibility that their attributes may be multidimensional rather than unidimensional, and that an attribute may be represented by nominal categories rather than a continuum;
(2) the distinction between observable item scores and latent traits, factors, and latent classes so that more thought is given to the causes of item responses; and
(3) using item response models with restrictions on the item parameters, cognitive diagnosis models that combine features of multidimensional item response models with such restrictions, and latent class regression models and other latent class structures to better understand what tests measure.

Unlike Borsboom, however, I think that latent variables—latent traits, factors, and latent classes—are summaries of the data and nothing more, and that, compared to Borsboom's ambitions, this seriously limits the possibilities of latent variable models. Let me explain my point of view. Suppose a psychologist uses 20 items for measuring the attribute of inductive reasoning, an attribute that is at the heart of the intelligence construct. He hypothesizes that the items elicit responses based on the same cognitive mechanism; that a higher level of inductive reasoning (or, similarly, a better degree of functioning of the cognitive mechanism that is labeled inductive reasoning) increases the likelihood that a respondent solves the items correctly; and that the attempts to solve an item do not in any way affect the probability that subsequent items are solved correctly. In Borsboom's and my perfect worlds, the researcher puts his hypotheses to the test by means of a unidimensional, monotone, locally independent item response model.

This is where the strength of latent variable models resides, as far as I am concerned: in the possibility to test the assumptions of the models through their observable consequences. Classical test theory does not provide the researcher with the means to do this and leaves him with "measurement by fiat." However, the possibility of testing of a model for the data—indeed a mighty weapon—is also where the strength of these models ends. The latent trait in an item response model is a mathematical summary of variation in the data between cases. Its direct meaning derives from a mathematical model, not a psychological theory, and it is estimated from the data, not the cognitive processes behind the data. The best that can happen, indeed, is that the item response model fits the data, which may then be taken as support for the hypothesis that the test is driven by inductive reasoning.

I use the word "support," not "proof," because there always remains a "gap" between fitting a model to data and the final piece of evidence that the test indeed is driven by the hypothesized attribute. This gap can only be "crossed" by inference or human judgment, and the hypothesis or the theory becomes more likely the more evidence is collected in their support. However, there is no way in which it can be decided that at a certain point the evidence is complete, other than that different researchers of good reputation agree that it does (Hofstee, 1980). How might additional evidence look in our inductive reasoning example?

It might have the form of other tests for inductive reasoning that use a different kind of item, yet be hypothesized to elicit the same or nearly the same cognitive processes as the previous test. Different cognitive skills or item properties might be distinguished using the new test(s), and a componential item response model (e.g., De Boeck & Wilson, 2004)—actually, akin to a nonlinear regression model—might be fitted to new data. Both the fit and the misfit of such models can contribute valuable knowledge to theory formation for inductive reasoning. A new set of items may give rise to another—that is, not exactly the same as the first—latent trait or even a set of latent (sub)traits, which is not at all unlikely in an item set designed to elicit different

skills or to let different item properties exercise their influence on cognitive processes. This need not worry anyone, as long as one sees latent variables as tools for summarizing data, not entities independent of the data on which they are fitted. My conclusion is that statistical latent variables help describe variation in data that is consistent with a putative psychological attribute; but, in isolation, goodness of fit of a latent variable model to data does not illuminate the existence or functioning of the attribute.

Construct Validation

Based on the assumptions that psychological attributes exist and exercise a causal influence on item responses and that latent variables represent psychological attributes, Borsboom proposes to limit the process of construct validation to latent variable modeling of item response data alone and discard studying relationships with other variables in a nomological network. I just explained that I disagree with the second assumption; below I will comment on the first.

The first assumption boils down to a conception of construct validation that entails the use of a substantive theory about the attribute of interest for predicting the pattern of responses to a set of items and, reversely, using latent variable modeling of these responses for establishing construct validity as a property of a test (Borsboom, 2006; Borsboom, Mellenbergh, & Van Heerden, 2004). This conception excludes tests for the vast majority of psychological attributes that are not supported by the kind of detailed and established theory that Borsboom seems to have in mind. For these attributes the "theory" will make inaccurate predictions and, as a result, the latent variable model will not fit. But what will one do next?

Taking substantive theory as a starting point for test construction is an excellent idea that has existed for a long time but is not widely practiced. The reason probably is that much theory is still in its puberty, infancy, or even at the fetal stage. Given this state of affairs one often has no other choice than to cling onto about every piece of evidence available in learning about test validity, including relationships with other interesting variables. There is no reason to exclude well-developed theories about attributes and their tests. For example, transitive reasoning is an example of a theoretically well-developed attribute (Bouwmeester, 2005), but its relationship with several verbal abilities may be interesting in its own right. Such studies are justified when different researchers of good reputation disagree about this relationship, and may shed more light on transitive reasoning but also, perhaps unexpectedly, on verbal intelligence.

Borsboom's assumption about the ontology and causality of psychological attributes seems to lead to a very restrictive conception of the process of construct validation: Elegant in its rigor but impractical for psychology (and many others areas). It seems to me that we still know so little about the functioning of the human brain in general and cognitive processes including those underlying personality traits and attitudes in particular, that it is difficult even to say what an "attribute" is. In the absence of such knowledge, I prefer to consider psychological attributes as organizational principles with respect to behavior. Thus, my point of view is that psychological attributes define which behaviors hang together well and are useful to the degree in which tests sampling these behaviors play a role in predicting interesting psychological phenomena.

The State of Psychological Research: A Case Study

In his own words, Borsboom sketches a grim picture of psychological academic research. I agree that occasionally psychologists are capable of wild adventures but not unlike any breed of academicians—including those involved in psychometrics, I would like to add. However, I believe psychology is in a better state than Borsboom suggests. I also think that psychometrics

has much help to offer, but perhaps less spectacularly than Borsboom would hope for. Here is what I see in present-day test and questionnaire construction.

At the time of writing this reaction, I was involved in several projects together with researchers from education, psychology, marketing, and medicine. Each of them uses questionnaires to measure an attribute: attitude toward homework and study (education), self-concealment (i.e., keeping things secret from others; psychology), service-quality (of computer helpdesks and restaurants; marketing), and perceived educational climate in Dutch hospitals (medicine). Each of the researchers is trying hard to work with a good definition of the attribute of interest that is well founded in the relevant literature; to find a useful operationalization of the attribute into a set of items; to be aware of item wording, use of both positive and negative item phrasing, and the threat of response sets; and to think about the composition of the sample and the way in which the data should be collected. They all use item response models or other modern statistical methods, such as latent class models and multilevel models. Of course, they do many of these things acting on my advice but what counts is that they are motivated and will carry on their knowledge to others. Thus there is progress which, however, proceeds slowly.

I have to admit that it is difficult to explain to my colleagues why latent variable models are better methods than Cronbach's alpha, the item-rest correlation, and principal components analysis. After all, what item response theory does is model the dimensionality of one's data and represent persons and items on a scale, but isn't this exactly what principal components analysis and classical test theory also do? A psychometrician shakes his head in disbelief about so much naivety but a psychological researcher who has not been trained to see the difference thinks this is "much ado about nothing." An effective recipe to make people see the—admittedly often subtle—differences is to do the classical and modern analyses next to one another and report both. For example, what convinced my fellow researchers to use Mokken scale analysis was that it allows the investigation of dimensionality by means of user-friendly software and without the artifacts of principal components analysis and factor analysis caused by discrepancies in the frequency distributions of the item scores. Notice these are practical arguments. Given these experiences, I think that researchers from substantive disciplines will accept modern psychometric methods if they are convinced of their practical advantages over classical methods and if results can be obtained without much trouble (which could mean including a psychometrician in the project).

The Role of Psychometrics in Psychological Research

Borsboom spurs his fellow psychometricians to take the lead in psychological research and use their latent variable models as blueprints for psychological measurement devices. This is motivated by the lack of fine-grained psychological theories that define exactly what a particular attribute stands for in terms of cognitive processes and functions, and how it should be operationalized in terms of items. The question then is whether in the absence of substantive theory an "empty" statistical model can fill the void and determine how attributes are measured. For example, unidimensionality, monotonicity, and local independence are necessary to have at least an ordinal scale, but they do not imply the kinds of tasks and data psychologists may find ideal for the assessment of a particular attribute.

As I see it psychometrics cannot replace substantive theorizing about intelligence and personality for designing good measurement instruments; it can only provide support. To learn about intelligence and personality, more and more research has to be done in the best traditions of these fields. Good substantive theories are the basis for good operationalizations and measurement procedures. The role of psychometricians is to make researchers more aware of the importance of sound theory and its operationalization, an appropriate research design, a correct definition

of the population and corresponding stratification of the sample, and the pitfalls in designing a test or questionnaire. Important additional questions, several mentioned by Borsboom, are: Do I expect a unidimensional or a multidimensional latent structure underlying the data? Are dimensions continuous are categorical? Should the items be questions, statements, tasks, games, or assignments? Should responses be oral, in writing, or sensomotoric? Should the data be correct/incorrect scores, ordered rating scale scores, category membership scores, or response times? Such choices determine which method should be used for data analysis.

Borsboom's approach is different in that he would take a psychometric model as point of departure and say: For a measurement instrument to have these particular properties, this is how your test should look like and these are the kinds of data you have to collect. It looks as though this view is somewhat at odds with Borsboom et al. (2004) who posit a substantive theory as point of departure. However, given that they assume an ontological status for the attribute and assume that a latent variable represents an attribute (also, Borsboom, 2006), it follows that latent variable models indeed provide blueprints for theory about attributes.

Instead of blueprints for theory, I see latent variable models as tools for analyzing data. They perform best, like any statistical method, when the data result from a well-established substantive theory. Nothing beats a good theory: if one knows which strings to pull, the expected data structure will stand out clearly and statistical analysis will be simple. Test construction should always be based on substantive theory, no matter how primitive, because only then does one know what one is looking for by means of statistical analysis, and only then can expectations be refuted or supported. Absence of theory leads to data beset with many weak signals and an overdose of noise, and the outcome of data analysis depends to a high degree on the statistical model used instead of substantive theory. Alternatively, running many models will usually not contribute greatly to theory formation other than, for example: "It looks like your data are primarily unidimensional but this may depend largely on the items you used." Thus, the role of psychometricians in psychological research is to be found in propagating the formation of theory, the operationalization of the attribute, the construction of the test, and the choice of the appropriate psychometric methods for analyzing the data, in that order.

References

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.

Bouwmeester, S. (2005). *Latent variable modeling of cognitive processes in transitive reasoning*. Tilburg, Tilburg University (PhD dissertation).

De Boeck, P., & Wilson, M. (2004) (Eds.). *Explanatory item response models. A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Hofstee, W.K.B. (1980). *De empirische discussie. Theorie van het sociaal-wetenschappelijk onderzoek. (The empirical discussion. Theory of social science research*). Meppel, The Netherlands: Boom.

Manuscript received 7 MAY 2006
Final version received 19 MAY 2006
Published Online Date: 23 SEP 2006