PSYCHOMETRIKA—VOL. 71, NO. 3, 447–450
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1500-5

WHEN A PSYCHOMETRIC ADVANCE FALLS IN THE FOREST

LEE ANNA CLARK

UNIVERSITY OF IOWA

Borsboom (2006) attacks psychologists for failing to incorporate psychometric advances in their work, discusses factors that contribute to this regrettable situation, and offers suggestions for ameliorating it. This commentary applauds Borsboom for calling the field to task on this issue and notes additional problems in the field regarding measurement that he could add to his critique. It also chastises Borsboom for occasionally being unnecessarily perjorative in his critique, noting that negative rhetoric is unlikely to make converts of offenders. Finally, it exhorts psychometricians to make their work more accessible and points to Borsboom, Mellenbergh, and Van Heerden (2003) as an excellent example of how this can be done.

Key words: psychometrics, psychological measurement, construct validity, critique

Borsboom (2006, p. 435), in his own words, "has sketched a rather grim picture of the role of psychometrics in psychology." He deplores the fact that advances in psychometric modeling have had little impact on psychological testing, and illustrates his point with two examples, taking the field to task for treating factors that have been identified via principal components analysis (PCA) as latent variables and for ignoring measurement invariance when interpreting group differences on psychological tests. He then discusses three domains in which major obstacles exist: theory/philosophy of science, pragmatic factors, and substantive theory.

As one who considers measurement of fundamental importance in psychology and who has developed several psychological instruments, I found myself alternatively cheering, indignant, and concerned as I read Borsboom's paper: Cheering, when his critique reflected my own dismay at the lack of respect and attention that many psychologists afford measurement; indignant, when I felt that he belittled—whether fairly or unfairly—substantive measurement psychologists, a group in which I include myself; and concerned, when he criticized practices that I feared that, in my own ignorance, I myself might have engaged in. The last I felt strongly enough that I read half-a-dozen of the papers Borsboom cites before beginning this commentary, in which I elaborate on the basis for each of these reactions.

I begin by cheering Borsboom's critique. There is no question that most psychologists receive too little training in mathematics and thus lack the skills needed to understand fully various complex psychometric issues and even to use advanced psychometric techniques in their work. Some recognize and compensate for this deficiency by collaborating with those who have the needed skills, but many too often choose the easier path of analyzing their data using only what they know already—and that is if they even conduct research; to wit, most psychologists are not researchers at all. There also is no question that many studies have too few participants and that the field would be well-served by reviewers and editors insisting that, to use Borsboom's language, researchers "buy" rather than "steal" the assumption that one can "directly estimate population differences in the theoretical attribute on the basis of observed scores" (Borsboom, 2006, p. 433). Additionally, the lack of substantive theory to guide psychological measurement is also beyond doubt. Psychology is still a young science and I am optimistic that appropriately explicit, testable

I wish to thank Frank Schmidt for his help in preparing this paper.
Requests for reprints should be sent to la-clark@uiowa.edu.

theories will be developed eventually, but it is disappointing that the field generally does not seem eager to embrace that future.

In some ways, Borsboom does not go far enough in his critique. Except as implied by his critiques of the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998) and the common practice of interpreting scores on an ad hoc paper-and-pencil test of X "as if they automatically applied to the *attribute*" of X (emphasis in original, p. 428), he does not criticize what likely are thousands of published studies in which the outcome of an experimental manipulation or the difference between two naturally occurring groups is assessed with an instrument or procedure developed for that particular study, with the resulting scores treated as a psychological construct (i.e., attribute), with no apparent thought given to the measurement issues involved. When others later review the literature, such studies are categorized by whatever label the researchers selected for their purported construct, be that comparative optimism, social attraction, or self control. At least those who use PCA know they should—and make an explicit attempt to—address measurement issues.

A related pet peeve of mine that he also could have mentioned is the far-too-common justification for the selection of study measures amounting to little more than a mantra: "According to the literature, measure X's reliability is good and it has been shown to be valid (citation)," with no numbers for reliability, nor even a "valid for Y" nod to criterion validity (about which I share Borsboom's concerns). As a reviewer, I often request that the authors either take reliability and validity at least somewhat seriously and provide some relevant data, or simply eliminate the mantra, which has no added value. Perhaps this sad practice reflects helplessness in the face of the "aura of intractability that surrounds the problem of 'construct validity' " (2006, p. 430), but I concur that this is "merely a poor excuse for not taking the measurement problem seriously" (2006, p. 431).

Nor does Borsboom take on projective testing with its theoretically rich but empirically questionable relation between responses and attributes. Given the recent official statement by the Board of Trustees of the Society for Personality Assessment (BTSPA) that "the Rorschach possesses reliability and validity similar to that of other generally accepted personality assessment instruments, and its responsible use in personality assessment is appropriate and justified" (BTSPA, 2005) including it in his critique would have been timely, but perhaps he judged projective testing not to be worth the time and journal space.

On the other hand, portions of Borsboom's critique aroused my indignation, and illustrated another reason, which he did not discuss, that also contributes to the general failure of psychologists "to incorporate psychometric modeling techniques in their methodological inventory" (2006, p. 425), specifically, a tendency of the mathematically inclined to denigrate the efforts of those who lack their measurement skills. At times this emerges as sarcasm: "Once a question gets labeled as a problem of construct validity, its difficulty is considered superhuman and its solution beyond a mortal's ken" (2006, p. 431). He goes on to note that physicists aren't "involved in the 'never-ending process' of figuring out whether meter sticks really measure length" and wonders, "So why should construct validity be such an enormous problem in psychology"? (2006, p. 431).

I have acknowledged that this point has some validity. Nevertheless, those who take construct validity seriously and also recognize the difficulty of fully understanding psychological constructs, don't deserve to be mocked, any more than we might mock theoretical physicists who acknowledge that the Big Bang Theory or General Relativity are not finished products and that their validation is similarly a "never-ending process" (Messick, 1988, cited in Borboom, 2006, p. 431), even duly noting the difference between validating a theory and a measure. Perhaps I should fight fire with fire and deplore that there apparently are psychologists who are not able to distinguish the qualitative difference between validating a measure of intelligence and a measure of length?

At other times this "superior" attitude takes the form of simplifying complex issues and then using this simplification to make those who do not attend to the complex issue look foolish. For example, Borsboom uses the argument developed by Millsap (1997) to illustrate how psychologists routinely ignore important measurement issues. He summarizes Millsap's argument, "if two groups differ in their latent means, and a test has prediction invariance across the levels of the grouping variable, it must have measurement bias with regard to group membership" (2006, p. 427). He goes on to say that if psychologists were good psychometricians, then "to put forward invariant regression parameters as evidence for measurement invariance would be out of the question in every professional and scientific work that appeared after 1997" (2006, p. 427). Finally, he documents that this has not happened, citing both official publications and Hunter & Schmidt (2000), concluding that "test bias is in acute need of scientific scrutiny (Borsboom, 2006, p. 427)."

When someone takes issue with Frank Schmidt, my antennae go up, so I took a close look (admittedly for the first time) at Millsap's argument, which made the narrower point that in measuring two groups on a test and a criterion that share a single common factor, it is not possible for there to be simultaneously "prediction invariance" (i.e., the same regression slope and intercept), "measurement invariance" (i.e., the same factor structure, which in the case of a single factor means equivalent factor loadings), and different observed-score distributions (specifically, differences in the variance of the common factor). So why would Hunter and Schmidt (2000, p. 151) declare that "the issue of test bias is scientifically dead"? From the arguments that Hunter and Schimdt make in their paper, I suspect that the answer lies in the fact that most psychological constructs used in applied settings are broadly specified rather than narrowly precise, so that in cases relevant to Millsap's argument—and with factor invariance evaluated via significance tests—trivial differences in factor variation emerge as significant. In other words, whereas Millsap's point, strictly speaking, may be true, it makes no practical difference when applied in real world settings and thus has been largely ignored.

To his credit, Borsboom (2006, p. 428) acknowledges "that psychometricians insufficiently communicate their advances to psychologists." Ironically, Millsap (1997) is an excellent case in point. Borsboom implies that the paper is "clearly written" (p. 427) and he may find it so, but I did not, which directly relates to his point that psychology journals may reject papers for being too difficult (i.e., containing mathematics). This problem is only partly with journal standards—it is not unreasonable for journals to want their contents accessible to their readers—but also that mathematical psychologists typically do not write in a way that is accessible to "mainstream" journal readership.

If Millsap (1997) were written so I could assign it in a first-year graduate class (and I believe it could be), it then would have a better chance to advance the field. Perhaps the currently popular push to "translate" research findings into practically usable forms will spill over into measurement, and mathematically inclined psychologists will begin to make their advances readily available to those working in more applied areas such as personality assessment and clinical practice. This might not hasten psychology becoming a mathematically grounded science, it might even delay it by at least partially obviating the need for mainstream psychologists to learn the relevant mathematics, but it also might fill the gap in the meanwhile.

Finally, Borsboom's paper engendered my concern that I was missing something important in my own work, so I read a number of the paper he cites in his criticism of the current status of psychometrics in psychology. I was relieved to find that—whereas the papers were quite useful for sharpening my thinking and learning some psychometric terminology—I did not need to totally revamp my research program. As I hope I made clear in my cheering section, I do not dispute that many of Borsboom's critiques are justified, but at the same time, some of them are exaggerated (to make a point?) in ways that ultimately distract from his message. First, his critique appears to be addressed to all psychologists who are not psychometricians, whereas actually it is aimed at only

a subset (though admittedly a large subset). Although, to his credit, he admits that his "extreme examples cannot simply be generalized to the field of psychology as a whole" (2006, p. 428), this point runs counter to the general tenor of his paper. Regrettably, his critique is most directly aimed not at those who will read his paper, but at those who should, but won't. My point here is that important psychometric advances must be communicated to mainstream psychologists, so the question is how to preach to the congregation and not just the choir? To use another analogy, if an important psychometric advance falls in a forest where there are no psychologists to hear it, does it make a sound? Apparently not.

Second, I believe Borsboom exaggerates the gulf between schools of thought. He has few kind words for classical test theory and yet cites Mellenbergh (1994) who states that "a Rasch-type model can be formulated as a special case of classical test theory" (p. 300, citing Moosbrugger & Müller, 1982). When you are trying to make a persuasive argument, it is more helpful to bridge from what one's audience understands or believes already to the point that one wants them to embrace. Alienating one's audience by simply rejecting key elements of their current understanding as fallacious is more likely to meet with resistance than open arms.

Let me end with more cheering. As mentioned, I read several papers Borsboom cites, and the most valuable paper from my perspective was one of his own (Borsboom et al., 2003, p. 205) which laid out quite clearly (indeed, although there are a few path diagrams with Greek letters, the most complex equation is "$2 + 2 = \ldots$"), the problem with what they call a "uniformity-of-nature assumption," that "the relation between mechanisms that operate at the level of the individual and models that explain variation between individuals is often taken for granted, rather than investigated" (p. 215). The paper served as an important reminder—and provided a well-reasoned argument to support it—that a complete science of psychology necessarily will require understanding of both processes and structures as well as their interrelations. For this, we will need psychologists across our broad discipline to bring their special expertise to bear on common problems, so that together we can find uncommon solutions.

References

Board of Trustees of the Society for Personality Assessment (BTSPA) (2005). The status of the Rorschach in clinical and forensic practice: An official statement. *Journal of Personality Assessment*, *85*, 219–237.

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

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219.

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Hunter, J.E., & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*, 151–158.

Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

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