PSYCHOMETRIKA—VOL. 71, NO. 3, 425–440
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1447-6

THE ATTACK OF THE PSYCHOMETRICIANS

DENNY BORSBOOM

UNIVERSITY OF AMSTERDAM

This paper analyzes the theoretical, pragmatic, and substantive factors that have hampered the integration between psychology and psychometrics. Theoretical factors include the operationalist mode of thinking which is common throughout psychology, the dominance of classical test theory, and the use of "construct validity" as a catch-all category for a range of challenging psychometric problems. Pragmatic factors include the lack of interest in mathematically precise thinking in psychology, inadequate representation of psychometric modeling in major statistics programs, and insufficient mathematical training in the psychological curriculum. Substantive factors relate to the absence of psychological theories that are sufficiently strong to motivate the structure of psychometric models. Following the identification of these problems, a number of promising recent developments are discussed, and suggestions are made to further the integration of psychology and psychometrics.

Key words: Psychometrics, modern test theory, classical test theory, construct validity, psychological measurement

In a recent overview of the psychometric developments of the past century, Embretson (2004, p. 8), noted that "[. . .] at the end of the 20th century, the impact of IRT on ability testing was still limited" and that "[t]he majority of psychological tests still were based on classical test theory." This conclusion applies with equal force to many other areas of mainstream experimental and quasi-experimental research, such as research on personality, attitudes, cognitive development, and intelligence. In fact, throughout psychology, one rarely encounters serious psychometric modeling endeavors.

Thus, even though psychometric modeling has seen rapid and substantial developments in the past century, psychometrics, as a discipline, has not succeeded in penetrating mainstream psychological testing to an appreciable degree. This is striking. Measurement problems abound in psychology, as is evident from the literature on validity (Cronbach & Meehl, 1955; Messick, 1989; Borsboom, Mellenbergh, & Van Heerden, 2004), and it would seem that the formalization of psychological theory in psychometric models offers great potential in elucidating, if not actually solving, such problems. Yet, in this regard, the potential of psychometrics has hardly been realized. In fact, the psychometric routines commonly followed by psychologists working in 2006 do not differ all that much from those of the previous generations. These consist mainly of computing internal consistency coefficients, executing principal components analyses, and eyeballing correlation matrices. As such, contemporary test analysis bears an uncanny resemblance to the psychometric state of the art as it existed in the 1950s.

The question that arises is why psychologists have been so reluctant to incorporate psychometric modeling techniques in their methodological inventory. The goal of the present paper is to answer this question by identifying and analyzing the factors that have hindered the incorporation of psychometric modeling into the standard toolkit of psychologists. A second goal is to offer some suggestions for improving the integration of psychological theory and psychometric techniques. However, to communicate the urgency of this situation, I will first consider some examples where things have gone seriously wrong.

This research was sponsored by NWO Innovational Research grant no. 451-03-068. I would like to thank Don Mellenbergh and Conor Dolan for their comments on an earlier version of this manuscript.

Requests for reprints should be sent to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: d.borsboom@uva.nl

Crisis? What Crisis?

There might be people who are wondering whether the incorporation of psychometric theory is really all that important for psychology or, perhaps, there are even some who are of the opinion that things are actually going rather well in psychological measurement. This is not the case. The daily practice of psychological measurement is plagued by highly questionable interpretations of psychological test scores, which are directly related to the lack of integration between psychometrics and psychology. The following examples serve to substantiate this.

Principal Components and Latent Variables

Many investigations into the structure of individual differences theorize in terms of latent variables, but rely on Principal Components Analyses (PCA) when it comes to the analysis of empirical data. However, the extraction of a principal components structure, by itself, will not ordinarily shed much light on the correspondence with a putative latent variable structure. The reason is that PCA is not a latent variable model but a data reduction technique (e.g., Bartholomew, 2004). This is no problem as long as one does not go beyond the obvious interpretation of a principal component, which is that it is a conveniently weighted sumscore. Unfortunately, however, this is not the preferred interpretation among the enthusiastic users of principal components analysis.

Consider, for instance, the personality literature, where people have discovered that executing a PCA of large numbers of personality subtest scores, and selecting components by the usual selection criteria, often returns five principal components. What is the interpretation of these components? They are "biologically based psychological tendencies," and as such are endowed with causal forces (McCrae et al., 2000, p. 173). This interpretation cannot be justified solely on the basis of a PCA, if only because PCA is a formative model and not a reflective one (Bollen & Lennox, 1991; Borsboom, Mellenbergh, & Van Heerden, 2003). As such, it conceptualizes constructs as causally determined by the observations, rather than the other way around (Edwards & Bagozzi, 2000). In the case of PCA, the causal relation is moreover rather uninteresting; principal component scores are "caused" by their indicators in much the same way that sumscores are "caused" by item scores. Clearly, there is no conceivable way in which the Big Five could cause subtest scores on personality tests (or anything else, for that matter), unless they were in fact not principal components, but belonged to a more interesting species of theoretical entities; for instance, latent variables. Testing the hypothesis that the personality traits in question are causal determinants of personality test scores thus, at a minimum, requires the specification of a reflective latent variable model (Edwards & Bagozzi, 2000). A good example would be a Confirmatory Factor Analysis (CFA) model.

Now it turns out that, with respect to the Big Five, CFA gives Big Problems. For instance, McCrae, Zonderman, Costa, Bond, & Paunonen (1996) found that a five factor model is not supported by the data, even though the tests involved in the analysis were specifically designed on the basis of the PCA solution. What does one conclude from this? Well, obviously, because the Big Five exist, but CFA cannot find them, CFA is wrong. "In actual analyses of personality data [. . .] structures that are known to be reliable [from principal components analyses] showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure" (McCrae et al., 1996, p. 563).

I believe this rather points to serious problems in psychologists' interpretation of principal components; for it appears that, in the minds of leading scholars in personality research, extracting a set of principal components equals fitting a reflective measurement model (or something even better). The problem persists even though the difference between these courses of action has been

clearly explicated in accessible papers published in general journals like *Psychological Bulletin* (Bollen & Lennox, 1991) and *Psychological Methods* (Edwards & Bagozzi, 2000). Apparently, psychometric insights do not catch on easily.

Group Comparisons

The interpretation of group differences on observed scores, in terms of psychological attributes, depends on the invariance of measurement models across the groups that figure in the comparison. In psychometrics, a significant array of theoretical models and associated techniques has been developed to get some grip on this problem (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993). In practice, however, group differences are often simply evaluated through the examination of observed scores—without testing the invariance of measurement models that relate these scores to psychological attributes.

Tests of measurement invariance are conspicuously lacking, for instance, in some of the most influential studies on group differences in intelligence. Consider the controversial work of Herrnstein and Murray (1994) and Lynn and Vanhanen (2002). These researchers infer latent intelligence differences between groups from observed differences in IQ (across race and nationality, respectively) without having done a single test for measurement invariance. (It is also illustrative, in this context, that their many critics rarely note this omission.) What these researchers do instead is check whether correlations between test scores and criterion variables are comparable (e.g., Lynn & Vanhanen, 1994, pp. 66–71), or whether regressing some criterion on the observed test scores gives comparable regression parameters in the different groups (e.g., Herrnstein & Murray, 2002, p. 627). This is called prediction invariance. Prediction invariance is then interpreted as evidence for the hypothesis that the tests in question are unbiased.

In 1997 Millsap published an important paper in *Psychological Methods* on the relation between prediction invariance and measurement invariance. The paper showed that, under realistic conditions, prediction invariance does not support measurement invariance. In fact, prediction invariance is generally indicative of *violations* of measurement invariance: 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. Conversely, when a test is measurement invariant, it will generally show differences in predictive regression parameters. One would expect a clearly written paper that reports a result, which is so central to group comparisons, to make a splash in psychology. If the relations between psychometrics and psychology were in good shape, 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.

So what happens in psychology? In 1999, two years after Millsap's paper appeared, the American Psychological Association is involved in the publication of the 1999 *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). With regard to the problem of test bias, we read: "[u]nder one broadly accepted definition, no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question" (AERA, APA, & NCME, 1999, p. 79). Another influential source, the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial Organizational Psychology, 2003), quite explicitly favors predictive invariance over measurement invariance as a method for investigating test bias (pp. 31–34). Perhaps, then, it should not come as a surprise that the fact, that prediction invariance is hardly ever violated, leads Hunter and Schmidt (2000, p. 151) to conclude that "the issue of test bias is scientifically dead." Unfortunately, on the basis of Millsap's work, one would rather say that, in the absence of violations of prediction invariance, the issue of test bias is in acute need of scientific scrutiny.

The Psychometric Connection

Of course, the impression resulting from these admittedly extreme examples cannot simply be generalized to the field of psychology as a whole. On the other hand, the involvement of influential theorists in psychology, as well as some of our most important professional organizations, indicates that we are not merely dealing with a few exceptions that prove the rule. In fact, it suggests that these examples are symptomatic of a serious problem in psychological measurement. One side of the problem is that psychologists have a tendency to endow obsolete techniques with obscure interpretations. The other side is that psychometricians insufficiently communicate their advances to psychologists, and when they do they meet with limited success. The result is a disconnect between psychometric theory on the one hand, and psychological research on the other. As a consequence, scientific progress in psychology is slowed. The questions that now arise are: (1) Why does this situation exist; and (2) What can we do about it? I address these questions in turn.

Obstacles to a Psychometric Revolution

The obstacles to a successful integration of psychometrics and psychology fall into three broad classes. The first class is theoretical in nature, and has to do with the dominant philosophical views on psychological measurement. The second concerns a loosely connected set of practical, social, and pragmatic factors that limit the amount of methodological innovation that the researcher in psychology can afford. The third relates to a shortage of substantive theory that is sufficiently detailed to drive informed psychometric modeling.

Theoretical Factors

Operationalism Rules

One of the main breakthroughs of the past century in psychometric thinking about measurement consists in the realization that measurement does not consist of finding the right observed score to *substitute* for a theoretical attribute, but of devising a model structure to *relate* an observable to a theoretical attribute. An essential precondition for this realization to occur is that, either intuitively or explicitly, one already holds the philosophical idea that theoretical attributes are, in fact, distinct from a set of observations, i.e., that one rejects the operationalist thesis that theoretical attributes are synonymous with the way they are measured (Bridgman, 1927).

Although one would expect most psychologists to subscribe to the thesis that theoretical attributes and measures thereof are distinct—after all, the rejection of operationalism was one of the driving forces behind the cognitive revolution—the standard research procedure in psychology is, ironically, to pretend that it is not true. Both in textbooks on psychological methods and in actual research, the dominant idea is that one has to find an "operationalization" (read: observed score) for a construct, after which one carries out all statistical analyses under the false pretense that this observed score is actually identical to the attribute itself. In this manner, it becomes defensible to construct a test for, say, self-efficacy, sum up the item scores on this test, subsequently submit these scores to analysis of variance and related techniques, and finally interpret the results as if they automatically applied to the *attribute* of self-efficacy because they apply to the *sumscore* that was constructed from the item responses.

This would be a relatively minor problem if psychologists widely realized that such an interpretation is crucially dependent on the assumption that the summed item scores can serve as adequate proxies for the actual attribute, and, perhaps more importantly, that violations of this

strong assumption present a threat to the validity of the conclusions reached. But this realization seems to be largely absent. The procedure mentioned is so common that it must be considered paradigmatic for psychological research. It brings with it the idea that properties that pertain to the sumscore somehow must also pertain to the attribute under study, so that, for instance, attributes are presumed to induce a linear ordering of people because sumscores do. But of course the assumption that an attribute induces a linear ordering cannot be derived from the fact that sumscores have this property; for the latter are linearly ordered by definition, while the former are not. Moreover, for many psychological attributes, alternative structures (like latent class structures or multidimensional structures) are no less plausible. However, the default strategy in psychological research precludes the consideration of such alternatives. And nobody knows how often these alternatives may actually be more accurate and truthful.

Classical Test Theory

It is an unfortunate fact of psychometric life that every introductory textbook on psychological research methods starts and ends its section on measurement with ideas and concepts that are based on classical test theory. Classical test theory may be considered the statistical handmaiden of the philosophy of operationalism that, at least as far as actual research practice is concerned, dominates psychology.

The reason for this is that the central concept of classical test theory—the true score—is exhaustively defined in terms of a series of observations; namely, as the expectation of a test score over a long run of replicated administrations of the test with intermediate brainwashing (Lord & Novick, 1968; Borsboom & Mellenbergh, 2002; Borsboom, 2005). Thus, the connection between the theoretical attribute (the true score) and the observation (the test score) is, in classical test theory, fixed axiomatically (Lord & Novick, 1968, Chap. 2). Therefore, within classical test theory, this relation is not open to theoretical or empirical research. This is in stark contrast with modern test theory models, in which the relation between test scores and attributes (conceptualized as latent variables) can take many forms (Mellenbergh, 1994).

Instead, classical test theory draws the researcher's attention to concepts such as reliability and criterion validity. The latter concept is especially important because it suggests that what is important about psychological tests is not how they work, but how strongly they are correlated with something else. This shift of attention is subtle but consequential. In an alternative world, where classical test theory never was invented, the first thing a psychologist, who has proposed a measure for a theoretical attribute, would do is to spell out the nature and form of the relationship between the attribute and its putative measures. The researcher would, for instance, posit a hypothesis on the structure (e.g., continuous or categorical) of the attribute, on its dimensionality, on the link between that structure and scores on the proposed measurement instruments (e.g., parametric or nonparametric), and offer an explanation of the actual workings of the instrument. In such a world, the immediately relevant question would be: How do we formalize such a chain of hypotheses? This would lead the researcher to *start* the whole process of research by constructing a psychometric model. After this, the question would arise which parts of the model structure can be tested empirically, and how this can best be done.

Currently, however, this rarely happens. In fact, the procedure often runs in reverse. To illustrate this point, one may consider the popular Implicit Association Test (IAT), which was developed by Greenwald, McGhee, and Schwartz (1998). This test is thought to measure so-called implicit preferences. A typical IAT application involves the measurement of implicit racial preferences. Subjects are presented with images of black and white faces and with positive and negative words on a computer screen. They are instructed to categorize these stimuli as quickly as possible according the following categories: A: "either a white face or a positive word," B: "either a black face or a negative word," C: "either a white face or a negative word," and D: "either a

black face or a positive word." The idea is that people who have an implicit preference for Whites over Blacks will be faster on tasks A and B but slower on C and D; the reverse is the case for those who have an implicit preference for Blacks over Whites. The IAT-score is computed by subtracting the log-transformed average response latency over compatible trials (A and B) from that over incompatible trials (C and D). Higher values on the resulting difference score are then considered to indicate implicit preference for Whites over Blacks.

Note the following facts about the psychometric work under consideration. First, the original paper puts forward no psychometric model for the dynamics underlying the test whatsoever. Second, even though the test is described as a measure of individual differences, the main evidence for its validity is a set of mean differences over experimental conditions, and no formal model explicating the link between these two domains is offered. Third, a Web of Science search reveals that the paper has been cited in over 420 papers. Some of these publications are critical, but most involve extensions of the test in various directions as well as substantive applications to different topics, which indicates that the IAT is a popular measurement procedure despite these points. Fourth, it took no less than eight years for a detailed psychometric modeling analysis of the proposed measure to see the light (Blanton, Jaccard, Gonzales, & Christie, 2006); and that analysis suggests that the scoring procedures used are actually quite problematic, because the various possible psychometric models on which they could be predicated are not supported by the data.

This illustrates a typical feature of the construction of measurement instruments in psychology. Let us say that in the ideal psychometric world, nobody could publish a test without at least a rudimentary idea of how scores are related to attributes, i.e., the outline of a psychometric model, and an attempt to substantiate that idea empirically. From the IAT example it is obvious that our world differs in two respects. The first concerns theory formation: the construction of a formal model that relates the attribute to its indicators is not necessary for a measurement procedure to be published and gain substantial following. The second concerns data analysis: Psychologists do not see psychometric modeling as a necessary tool to handle data gathered with a newly proposed testing procedure; running observed scores through ANOVA machinery, and computing correlations with external variables is perceived as adequate.

It is important to consider how these aspects are conceptually related, because quite often psychometricians try to sell the data analytic machinery to psychologists who have never asked themselves what the relation between the attribute and the test scores might be in the first place. It is obvious that such psychologists have no use for these modeling techniques; they may even perceive them as a silly mathematical circus. Psychometric modeling is only relevant and interesting to those who ask the questions that it may help answer. And because classical test theory axiomatically equates theoretical attributes with expected test scores, it has no room for the important and challenging psychometric question of how theoretical attributes are related to observations. Therefore, researchers who think along the lines of classical test theory simply do not see the need to ask such questions.

The Catch-All of Construct Validity

Operationalist thinking and classical test theory are mutually supportive systems of thought, which sustain a situation in which researchers habitually equate theoretical attributes with observational ones. However, although such practices may conceal the measurement problem, they do not make it go away; and many researchers are, at some level, aware of the fact that, with respect to psychological measurement, there is something rotten in the state of Denmark.

Now, psychologists can do a fascinating sort of Orwellian double-think with respect to the measurement problem: They can ask good psychometric questions, but then relegate them to a special theoretical compartment, namely that of "construct validity," instead of trying to answer

them. Relevant questions that are routinely dropped in the catch-all of construct validity are: What is it that the test measures? What are the psychological processes that the test items evoke? How do these processes culminate in behaviors, like marking the correct box on an IQ-item? How do such behaviors relate to individual differences? What is the structure of individual differences themselves? What is the relation between such structures and the test scores? In fact, looking at this list, it would seem that a question is considered to concern construct validity at the very instance that it becomes psychometrically challenging.

Construct validity functions as a black hole from which nothing can escape: Once a question gets labeled as a problem of construct validity, its difficulty is considered superhuman and its solution beyond a mortal's ken. Validity theorists have themselves contributed to this situation by stating that validation research is a "never-ending process" (e.g., Messick, 1988), which, at most, returns a "degree of validity" (Cronbach & Meehl, 1955; Messick, 1989), but can by its very nature never yield a definitive answer to the question whether a test measures a certain attribute or not. This effectively amounts to a mystification of the problem, and discourages researchers to address it. In addition, this stance must be fundamentally ill-conceived for the simple reason that no physicists are currently involved in the "never-ending process" of figuring out whether meter sticks really measure length, or are trying to estimate their "degree of validity"; nevertheless, meter sticks are doing fine. So why should "construct validity" be such an enormous problem in psychology?

The general idea seems to be based on the conviction (taken from the philosophy of science, and especially the work of Popper, 1959) that all scientific theories are by their nature "conjectures that have not yet been refuted"; i.e., tentative and provisionally accepted working hypotheses. Whether one subscribes to this idea or not, it is evident that it cannot be specifically relevant for the problem of validity, because this view concerns not just validity, but every scientific hypothesis, and, by implication, applies to every psychometric hypothesis. Thus, if validity is problematic for this particular reason, then so are reliability, unidimensionality, internal consistency, continuity, measurement invariance, and all other properties of tests, test scores, and theoretical attributes, as well as all the relations between these properties that one could possibly imagine. But this is thoroughly uninformative; it merely teaches us that scientific research is difficult, and that we hardly ever know anything for sure. While this may be an important fact of life, it has no special bearing on the problem of test validity and most certainly cannot be used to justify the aura of intractability that surrounds the problem of "construct validity."

It can be argued that, if the construction and analysis of measurement instruments were done thoroughly, this process would by its very nature force the researcher to address the central questions of construct validity before or during test construction (Borsboom et al., 2004). Not being able to do so, in turn, would preclude the construction of a measurement instrument. Thus, the fact that basic questions such as "What am I measuring?" and "How does this test work?" remain unanswered with respect to an instrument, which is considered fully developed, implies that we cannot actually take such an instrument seriously. In fact, a discipline that respects its scientific basis should hesitate to send tests, for which such basic problems have not been solved, out for use in the real world. A reference to the effect that such problems concern construct validity, and therefore their solution is impossible, cannot be taken as an adequate justification of such practice. So used, construct validity is merely a poor excuse for not taking the measurement problem seriously.

Pragmatic Factors

Although the ideological trinity of operationalism, classical test theory, and construct validity forms an important obstacle to the incorporation of psychometric modeling into the standard

methodology of psychology, there are also more mundane factors at work. These concern a hodgepodge of factors relating to the sociology of science, the research culture in psychology, practical problems in psychometric modeling, and the poor representation of psychometric techniques in widely used computer software. We will consider these factors in turn.

Psychometrics Is Risky

Suppose that an unconventional thinker in psychology were to stumble across a psychometric model, and recognize its potential. Suppose also that the psychologist were to use the model to analyze data that were gathered using the average psychological test. The researcher would quickly encounter a problem. Namely, psychometric models have a tendency to disprove commonly held assumptions, like unidimensionality and measurement invariance. The researcher then gets involved in fundamental problems concerning the structure of the attributes under investigation and the relation that they bear to the observations. Such questions are among the most fascinating and important ones in any science, but they are not popular in psychology. So, even if the psychologist has some success in answering these questions and coming up with a reasonable model for the observations, it will turn out difficult to get these results published, because many editors and reviewers of scientific journals are not overly familiar with psychometric models, and will often suggest that these results be published in a psychometric journal rather than a psychological one. This, of course, is not what the researcher necessarily wants; moreover, psychometric journals may refuse the work for the reason that it is not sufficiently psychometrically oriented, so that the researcher gets stuck between a rock and a hard place. Thus, career-wise, turning to psychometric modeling techniques is risky.

It Shouldn't Be Too Difficult

This problem is compounded by the research standards that are currently accepted in psychology. Even though the research topic of psychology—human behavior and the mental processes that underlie it—is perhaps the most complicated ever faced by a science, the contents of scientific papers that deal with it are required to be below a certain standard of difficulty. I have seen at least one case where a manuscript that used psychometric modeling was rejected by a major journal because, according to the editor, it was too difficult for the journal's audience since it contained some basic matrix algebra (i.e., addition and multiplication). That a scientific journal should reject a paper for being difficult is almost surrealistic; yet, the use of equations, in general, is discouraged in many psychological journals. This is detrimental to the development of psychology. If physics journals had existed in the seventeenth century and had adhered to this policy, it would have been impossible to achieve the break with Aristotelian theory that is now known as the Scientific Revolution. The current research culture in psychology, however, actively works against the formalization of theories and the use of mathematical modeling techniques, which include psychometric models.

Educational Issues

Psychologists typically do not receive a substantial amount of mathematical training. In this respect it is illustrative to compare the average educational program of psychologists with that of, say, economists. Every trained economist understands basic calculus, for instance, while trained psychologists often do not know what calculus is in the first place. The reason for this difference is clear. It is simply impossible to read advanced economics texts without substantial mathematical knowledge, and hence mathematical training is a bare necessity. Evidently, no such baggage is required in psychology. As Lykken (1991, p. 7.) stated, "there are many courses in the psychology curriculum, but few have real prerequisites. One can read most psychology texts

without first even taking an introductory course." It is not strictly necessary for a student to have even a rudimentary understanding of mathematics in order to complete a degree in psychology; and neither is such understanding required on the part of the researcher who studies psychology's advanced texts. The consequence of this situation in the present context is that psychologists often lack the necessary skills to understand what psychometric models do or what they can be used for, which hampers the dissemination of advances in psychometrics.

But It's Not in SPSS!

The reason that, say, Cronbach's alpha and principal components analysis are so popular in psychology is not that these techniques are appropriate to answer psychological research questions, or that they represent an optimal way to conduct analyses of measurement instruments. The reason for their popularity is that they are default options in certain mouse-click sequences of certain popular statistics programs. Since psychologists are monogamous in their use of such software (most in my department are wedded to SPSS) there is little chance of convincing them to use a model—any model—that is not "clickable" in the menus of major statistical programs. For reasons that defy my understanding, psychometric models are not well represented in such software. In fact, for a long time the psychometric modeler was typing in obscure commands next to a command prompt on a black and white screen. Considerable improvement on this point has been made (e.g., Muthén & Muthén, 2001), but this improvement is of a relatively recent date.

Thou Shalt Not. . .

Psychometric models are often considered to have a normative component. People who subscribe to this point of view, see psychometrics as a set of golden rules that the researcher should live by. I am thinking of the "no Rasch, no good" philosophy and associated doctrines. The presentation of psychometrics in terms of strictures (instead of opportunities, for instance) is damaging to its public image; for it is a law of human psychology that people whose behavioral repertoire is limited to "you are not allowed to do that" do not get invited to parties. Thus, it is important to get psychologists to see that psychometric modeling gives them new possibilities, instead of presenting them with strictures and limitations; good examples of such a positive strategy can be found in De Boeck and Wilson (2004).

Sample Size Issues

The estimation and testing of psychometric models is not always feasible due to the fact that one often needs large data sets for this purpose. In experimentally oriented research, for instance, sample sizes typically involve dozens rather than hundreds of subjects, and in such cases the use of psychometric models with latent variables is often hard to justify. In contrast, treating variables as "observed," i.e., as recorded without measurement error, returns the possibility of doing science with 15 to 30 subjects per cell in a standard factorial design. So why would one then even consider the use of psychometric techniques in psychological research? Simply adding up some scores and calling the sumscore "self-efficacy" does the job just as well, but with much fewer subjects. The question is, of course, whether this is true.

Coombs (1964) has said that we buy information by assumption. In many cases, however, one instead needs information to buy assumptions. For instance, if one knows one is using a good measurement instrument, one can use this information to "buy" the very useful assumption—common to all observed score regression techniques, including the various species of ANOVA—that one models "observed variables." This effectively means that one can drop the measurement problem and directly estimate population differences in the theoretical attribute on the basis of observed scores. However, there is no uncertainty concerning the question whether psychologists

have the knowledge needed to buy such assumptions: they do not. Hence, any argument against psychometric modeling on the basis of the larger sample sizes needed, when compared to analysis of variance and related observed score techniques, is in fact based on a wager. This wager involves the question whether observed score techniques are still trustworthy, if one does not buy the assumptions needed, but steals them. That is, does the run-of-the-mill research design plus analysis still work, if one pretends to have solved the measurement problem, while one has in fact ignored it? I do not know the answer to this question, but given the small and notoriously unstable effects that psychologists usually find, I would not like to bet on psychology's chances in this gamble.

Substantive Factors

It will be obvious by now that the integration of psychology and psychometrics faces significant obstacles of various natures. When ideological, theoretical, practical, and sociological factors conspire against the incorporation of a method, it is no surprise that such a method has trouble getting off the ground. However, we have not yet devoted attention to what may be the single most important problem that faces psychometric modeling. This is the almost complete absence of strong psychological theory.

The problem is best illustrated with an example. Suppose we are interested in personality and want to construct a measurement instrument for, say, conscientiousness. Like most people, we have a common-sense idea about the characteristics of conscientious people. For instance, they tend to be in time for appointments, do their best to succeed on a job, feel guilty when they fail to meet obligations, etc. Suppose that we assess these characteristics through a self-report questionnaire; for ease of exposition, assume we construct a set of items in the spirit of "I feel guilty when I fail to meet obligations," and score them dichotomously in a yes/no format. How do we then relate the item responses to the attribute of conscientiousness?

Consider the following list of options. We could view the items as a sample from a domain of behaviors, and define the attribute as the proportion of the behaviors in that domain that any given person exhibits, which would lead us toward generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). We could also view conscientiousness as a causal function of the behaviors assessed in the items (people are conscientious because they meet appointments). This would lead us toward a formative model, like a PCA model (Bollen & Lennox, 1991). Alternatively, we could reverse the causal relation (people meet appointments because they are conscientious), which would lead us toward a reflective latent variable modeling scheme (Edwards & Bagozzi, 2000). We could further envision the causal relation to govern person-specific changes or as a relation between sets of individual differences (Borsboom et al., 2003; Hamaker, Dolan, & Molenaar, 2005; Molenaar, 2004).

Within any of these schemes of thinking, we still have many choices to make (see also Mellenbergh, 2001). Suppose, for instance, that we opt for the reflective latent variable approach. Should we then conceive of conscientiousness as a continuous variable (pushing us toward IRT) or as a categorical variable (pushing us toward a latent class approach)? If we suppose 0conscientiousness is continuous, what is the form of the relation between the item responses and the trait? Is it a monotonic function or not (Stark, Chernyshenko, Drasgow, & Williams, 2006)? If it is monotonic, is the function smooth, like a logistic function (which would suggest a one- or two-parameter logistic model; Rasch, 1960; Birnbaum, 1968), or erratic (which would suggest a nonparametric alternative; Mokken, 1970)? If it is smooth, can we then assume that zero and one are the right asymptotes for the Item Response Function or are different asymptotes more realistic (Hessen, 2004)? Is a logistic function at all appropriate?

This is a psychometric embarrassment of riches. Even within this concise set of questions to be answered we encounter no less than four radically different conceptualizations of the relation between conscientiousness and conscientious behaviors: a universe-sampling relation (generalizability theory), a formative causal relation (formative model), a reflective causal relation with the latent variable categorical (latent class model), and a reflective causal relation with the latent variable continuous (IRT). Moreover, as the IRT example shows, within each of these conceptualizations there are many more fine-grained choices to be made before we truly have a candidate model. Literally none of these choices are dictated by substantive theory.

Of course, *researchers* make such choices all the time—otherwise they could do nothing with their data. For instance, personality traits are usually taken to be continuously structured and conceived of as reflective latent variables (even though the techniques used do not sit well with this interpretation). The point, however, is that there is nothing in personality *theory* that motivates such a choice, and the same holds for the majority of the subdisciplines in psychology. Thus, the crucial decisions in psychological measurement are made on the basis of pragmatic or conventional grounds, instead of on substantive considerations.

This may be the central problem of psychometrics: psychological theory does not motivate specific psychometric models. It does not say how theoretical attributes are structured, how observables are related to them, or what the functional form of that relation is. It is often silent even on whether that relation is directional and, if so, what its direction is. It only says that certain attributes and certain observables have something to do with each other. But that is simply not enough to build a measurement model.

The Light at the End of the Tunnel

This paper has sketched a rather grim picture of the role of psychometrics in psychology. Fortunately, however, several positive developments have also taken place in the last decade or so. Three developments are especially noteworthy.

The first concerns the increasing number of conceptually and practically oriented introductions to psychometric modeling that have appeared since the late 1980s. Important examples, among others, are the books by Bollen (1989), Frederiksen, Mislevy, and Bejar (1993), Embretson and Reise (2000), Embretson and Hershberger (1999), Hagenaars (1993), Heinen (1996), Kaplan (2000), and Sijtsma and Molenaar (2002). These works present the subject matter in a relatively accessible way, thereby facilitating a transition to psychometric modeling in psychology.

A second promising development is the increase of user-friendly software for psychometric modeling. Of course, the one program based on psychometric ideas that has, in the past decades, made something of a breakthrough is the LISREL program by Jöreskog and Sörbom (1996). This prepared the road for various more recent latent variable modeling computer programs including the versatile Mplus program by Muthén and Muthén (2001) and Vermunt's and Magidson's Latent Gold (2000). The increasing popularity of freeware statistical computing programs like R (Venables, Smith, & The R Development Core Team, 2005) and Mx (Neale, Boker, Xie, & Maes, 2003) is also promising. Finally, the group of Paul de Boeck (e.g., De Boeck & Wilson, 2004) has worked out effective ways to do IRT modeling through the program SAS by specifying IRT models as mixed regression models. One hopes that such developments will necessitate the most widely used program in psychology, SPSS, to incorporate latent variable modeling options in its basic test analysis section. In all, these developments are certainly going in the right direction, and hopefully will result in a situation where psychometric modeling is a realistic option for the average researcher in psychology within a decade or so.

A third positive development is the increasing number of psychometrically informed research papers that have been appearing in the past decade. The recently introduced section

on applied psychometrics in *Psychometrika* presented some good examples of such papers (e.g., Bouwmeester & Sijtsma, 2004; Van Breukelen, 2005). Substantial psychometric literatures are further building up on process-based psychometric modeling of intelligence (sub)tests (Embretson, 1998; Mislevy & Verhelst, 1990; Süss, Oberauer, Wittmann, Wilhelm, & Schulze, 2002) and cognitive development (Jansen & Van der Maas, 1997, 2002; Dolan, Jansen, & Van der Maas, 2004). Formal modeling approaches are also gaining momentum in personality theory (Fraley & Roberts, 2005), emotion research (Ferrer & Nesselroade, 2003), and social psychology (Blanton et al., 2006). In a more general sense, the framework of explanatory IRT highlights the potential of psychometrics in substantive contexts (De Boeck & Wilson, 2004).

These developments are creating a set of background conditions against which major changes in psychometric practice become possible. The development of accessible introductory works on psychometrics and the development of user friendly computer programs remove some significant practical obstacles; and the fact that substantively driven psychometric modeling endeavors are published in both technically oriented and substantive journals may create the momentum that is needed to establish the breakthrough of psychometric modeling. Hopefully, this will lead to a change in the psychological research culture and ultimately further progress in psychology. It is important that psychometricians support and, where possible, accelerate this process. There is room for improvement on several fronts.

Write Good Introductory Textbooks

Psychologists' first, and at some places only, contact with the theory of psychological measurement occurs through introductions to psychological research methods. Such books usually contain a section on psychological measurement. This section is always outdated and often flawed. The student is mesmerized through the formula $X = T + E$ but not given the tools to understand what it means. The caricature of classical test theory so induced is invariably accompanied by the misinterpretation of "true scores" as "construct scores" (Borsboom & Mellenbergh, 2002; Borsboom, 2005), which sets the stage for the operationalist mode of thought described earlier in this paper. Also, there is usually an explanation of reliability (as test–retest reliability or internal consistency) and a section that emphasizes, but does not elucidate, the difficulties of validity. If one is lucky, there is a treatment of the so-called convergent and divergent validity coefficients. This teaches one how to eyeball correlation matrices, which cannot be considered an optimal research procedure but is better than nothing. That is where it stops. No attention is given to the crucial questions how psychological attributes are structured, how they are related to observed scores, how one can utilize substantive theory in test construction, or how one can test one's ideas through the construction and evaluation of measurement models. It would be a significant advance if these books were to update their sections on measurement, so that a psychologist no longer has to unlearn earlier ideas when she/he decides to take the measurement problem seriously. It seems psychometricians are the right candidates for this job.

Read and Publish Widely

The founding fathers of the Psychometric Society—scholars such as Thurstone, Thorndike, Guilford, and Kelley—were substantive psychologists as much as they were psychometricians. Contemporary psychometricians do not always display a comparable interest with respect to the substantive field that lends them their credibility. It is perhaps worthwhile to emphasize that, even though psychometrics has benefited greatly from the input of mathematicians, psychometrics is not a pure mathematical discipline but an applied one. If one strips the application from an applied science one is not left with very much that is interesting; and psychometrics without the "psycho" is not, in my view, an overly exciting discipline. It is therefore essential that a psychometrician keeps up to date with the developments in one or more subdisciplines of psychology. This

is not to say that the purely conceptual and mathematical study of psychometric models is unimportant. On the contrary. However, as a discipline, psychometrics should consciously and actively avoid a state of splendid isolation. This requires regular reading of psychological journals and visits to substantively oriented conferences. Psychometricians should, in my view, also actively promote such behavior in their (PhD) students, who often cannot see the substantive forest for the mathematical trees. Substantive involvement ideally leads to a greater number of psychologically oriented publications by psychometricians, and hence to a more prominent presence of psychometrics in psychological research; this, in turn, may facilitate the acceptance of the ideas of modern psychometric theory in psychological circles.

Psychometrics and Theory Formation

Traditionally, psychometricians develop mathematical models, but leave the development of the substantive theory that is supposed to motivate these models to psychologists. As such, psychometrics takes the attitude of an ancillary discipline, which helps psychology with the formalization of theory into statistical models, and with the analysis of psychological data. I think that, with respect to the measurement problem, a century of experience teaches us that this procedure does not work very well. As has been discussed above, psychological theories are often simply too vague to motivate psychometric models. Most psychologists appear neither capable of, nor interested in, constructing more precise theories. I suggest that the more adventurous psychometricians would do well to take matters into their own hands, and start developing psychometric theories (as opposed to psychometric models) with a substantive component.

As it happens, there is another, quite similar discipline that is also waiting for a revolution that never happened (Cliff, 1992), namely, mathematical psychology, where a good deal of experience and know-how on the development of formal psychological theories is available. The question of how attributes may be structured, for instance, has received ample attention in the work on fundamental measurement theory (Krantz, Luce, Suppes, & Tverky, 1971). However, for reasons that elude me, and that are probably historical in nature, there is very little communication and collaboration between the fields of psychometrics and mathematical psychology, even though they manifestly have so much in common. Much could be gained by a further integration of these disciplines. Some psychometricians and mathematical psychologists (e.g., Scheiblechner, 1995; Tuerlinckx & De Boeck, 2005; Falmagne, 1989; Doignon & Falmagne, 1999) have already explored some of the common ground with promising results. That common ground may harbor significant further opportunities to promote the development of formal psychological theorizing.

Discussion

This paper has traced the lack of successful integration between psychometrics and psychology to a number of theoretical, pragmatic, and substantive factors that obstruct necessary changes in the research practices of psychologists. These factors are wide-ranging and distinct in nature, and thus render the task of breaking the resistance of psychology to psychometric modeling a formidable one. However, the incorporation of psychometrically sensible thinking in psychological research is important, not just for the progress of psychology as a science, but also for society as a whole. For, insofar as psychological research remains purely scientific, the lack of psychometrically defensible analyses may obstruct progress; but apart from that it is mostly harmless. However, psychological testing also has a significant and direct impact on people's lives—for instance, through the use of tests in psychiatric diagnoses or for the selection of employees—and at present such applications do not always stand on firm grounds, to say the

least. Thus, we face a pressing obligation to improve the practice of psychological research, and this obligation is not merely of a scientific nature.

There is no question that there is ample room for such improvement. The current practice of psychological measurement is largely based on outdated psychometric techniques. We should not sit around while the psychometric procedures of our fellow psychologists slip into obsolescence. Psychometricians may prevent this through active participation in the education of psychologists, the dissemination of psychometric insights among researchers, but also through the development of formalized psychological theory. At the very least, the psychometricians of the twenty-first century should strive to play a more pronounced role in substantive psychological research than is currently the case.

Max Planck stated that it hardly ever happens that scientists radically change their established ideas: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." Perhaps this holds to an even stronger degree for methodological innovations, because these necessitate not just the revision of theoretical ideas, but also require researchers to learn new skills. Several promising processes, like the development of versatile computer programs and the increasing number of successful psychometric investigations in substantive psychology, suggest that the tide may be turning. I suggest we work as hard as possible to facilitate the emergence of a new generation of researchers who are not afraid to confront the measurement problem in psychology.

References

AERA, APA, & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) Joint Committee on Standards for Educational and Psychological Testing (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bartholomew, D.J. (2004). *Measuring intelligence: Facts and fallacies*. Cambridge: Cambridge University Press.

Blanton, H., Jaccard, J., Gonzales, P.M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*, 192–212.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K.A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314.

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

Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*, 505–514.

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

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

Bouwmeester, S., & Sijtsma, K. (2004). Measuring the ability of transitive reasoning, using product and strategy information. *Psychometrika, 69*, 123–146.

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

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186–190.

Coombs, C. (1964). *A theory of data*. New York: Wiley.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

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

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

Doignon, J.P., & Falmagne, J.C. (1999). *Knowledge spaces*. New York: Springer-Verlag.

Dolan, C.V., Jansen, B.R.J., & Van der Maas, H.L.J. (2004). Constrained and unconstrained normal finite mixture modeling of multivariate conservation data. *Multivariate Behavioral Research, 39*, 69–98.

Dolan, C.V., Roorda, W., & Wicherts, J.M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32*, 155–173.

Edwards, J.R., & Bagozzi, R.P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155–174.

Embretson, S.E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods, 3*, 300–396.
Embretson, S.E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement, 2*, 1–32.
Embretson, S.E., & Hershberger, S.L., Eds. (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Erlbaum.
Embretson, S.E., & Reise, S., Eds. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Falmagne, J.C. (1989). A latent trait theory via stochastic learning theory for a knowledge space. *Psychometrika, 54*, 283–303.
Ferrer, E., & Nesselroade, J.R. (2003). Modeling affective processes in dyadic relations via dynamic factor analyses. *Emotion, 3*, 344–360
Fraley, R.C., & Roberts, B.W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*, 60–74.
Frederiksen, N., Mislevy, R.J., & Bejar, I.I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.
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.
Hagenaars, J.A. (1993). *Loglinear models with latent variables*. Newbury Park: Sage.
Hamaker, E.L., Dolan, C.V., & Molenaar, P.C.M. (2005). Statistical modeling of the individual: Rationale and application of multivariate time series analysis. *Multivariate Behavior Research, 40*, 207–233.
Heinen, T. (1996) *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks: Sage.
Herrnstein, R.J., & Murray, C. (1994). *The Bell curve*. New York: The Free Press.
Hessen, D.J. (2004). A new class of parametric IRT models for dichotomous item scores. *Journal of Applied Measurement, 5*, 385–397.
Hunter, J.E., & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests. *Psychology, Public Policy & Law, 6*, 151–158.
Jansen, B.R.J., & Van der Maas, H.L.J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.
Jansen, B.R.J., & Van der Maas, H.L.J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383–416.
Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8 User's reference guide* (2nd ed). Chicago: Scientific Software International.
Kaplan, D. (2000). *Structural equation modeling. Foundations and extensions*. Thousand Oaks, CA: Sage.
Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*, Vol. I. New York: Academic Press.
Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Lykken, D.T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W.M. Grove (Eds.), *Thinking clearly about psychology, Vol. 1*. Minneapolis, MN: University of Minnesota Press, pp. 3–39.
Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.
McCrae, R.R., Costa, P.T., Jr., Ostendorf, F., Angleitner, A., Hrebickova, M., & Avia, M.D., et al. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology, 78*, 173–186.
McCrae, R.R., Zonderman, A.B., Costa, P.T., Jr., Bond, M.H., & Paunonen (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566.
Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.
Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.
Mellenbergh, G.J. (2001). Outline of a faceted theory of item response data. In: A. Boomsma, M.A.J. Van Duijn, & T.A.B. Snijders (Eds.), *Essays in item response theory*. New York: Springer-Verlag.
Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–543.
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequence of measurement. In H. Wainer, & H.I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248–260.
Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17*, 297–334.
Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.
Mokken, R.J. (1970). *A theory and procedure of scale analysis*. The Hague: Mouton.
Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.
Muthén, L.K., & Muthén, B.O. (2001). *Mplus user's guide* (2nd ed.). Los Angeles, CA: Muthén & Muthén.
Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2003). *Mx: Statistical modeling* (6th ed.). Box 980126 MCV, Richmond, VA 23298, USA.
Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281–304.
Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
Society for Industrial Organizational Psychology (2003). Principles for the application and use of personnel selection procedures. Bowling Green, OH: Society for Industrial Organizational Psychology.
Stark, S., Chernyshenko, O.S., Drasgow, F., & Williams, B.A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25–39.
Süss, H., Oberauer, K., Wittmann, W.W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence, 30*, 261–288.
Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*, 629–650.
Van Breukelen, G.J.P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359–376.
Venables, W.N., Smith, D.M., and The R Development Core Team (2005). *An introduction to R, Version 2.2.0*. R-Project, 2005. URL: http://CRAN.R-project.org
Vermunt, J.K., & Magidson, J. (2000). *Latent GOLD user's manual*. Boston, MA: Statistical Innovations Inc.

Manuscript received 6 JAN 2006
Final version received 20 APR 2006
Published Online Date: 23 SEP 2006

Editor's Note

Denny Borsboom was the 2004 winner of the Psychometric Society Dissertation Prize. This essay and the subsequent commentary grew out of conversations following Borsboom's presentation of his work at the International Meeting of the Psychometric Society 2005, Tilburg, The Netherlands.