Item Response Models, Pathological Science and the Shape of Error

Reply to Borsboom and Mellenbergh

Joel Michell
UNIVERSITY OF SYDNEY

ABSTRACT. There is nothing in Borsboom and Mellenbergh's (2004) response that refutes my thesis that psychometrics is a pathology of science. They seek to defend item response models from my charge of pathological science without apparently realizing that my charge relates to psychometricians, not to models. They appeal to the Quine–Duhem thesis in an attempt to argue that item response models do not allow the hypothesis that psychological attributes are quantitative to be tested in isolation, but their argument is based upon a misinterpretation of Duhem. In any experiment, what is being tested depends on what the experimenter already takes to be true, and it is possible that a psychometrician could be testing just one of the hypotheses constituting an item response model. Furthermore, using the theory of conjoint measurement, it is possible to isolate predictions that depend upon psychological attributes being quantitative, as opposed to merely ordinal. Despite this, Borsboom and Mellenbergh agree with the first part of my thesis. They do not discuss the second part, but an examination of textbooks on item response models shows that psychometricians disguise their failure to test the hypothesis that psychological attributes are quantitative by simply declining to mention that this hypothesis is presumed in their models. Claims to measure psychological attributes based upon these models depend exclusively upon the weakest part of these models: the hypothesis that the distribution of 'errors' takes a specific form.

KEY WORDS: item response models, pathological science, psychological measurement, psychometrics, quantification

Borsboom and Mellenbergh (2004) note that I did not discuss item response models in my assessment of psychometrics as pathological science (Michell, 2000). This was because my critique of psychometrics (namely that mainstream psychometricians have neglected to test the hypothesis that psychological attributes are quantitative, and that they attempt to disguise this fact)

THEORY & PSYCHOLOGY Copyright © 2004 SAGE Publications. VOL. 14(1): 121–129
DOI: 10.1177/0959354304040201 www.sagepublications.com

applies to all mainstream psychometricians, including advocates of these models. I disagree with Borsboom and Mellenbergh's conclusions, but for reasons of space, I confine my reply to the main question: do they provide scientific reasons to revise my critique?

That they do not is because, in part, their arguments are premised upon a misreading. They would exclude *item response models* from the category of pathological science, whereas this category was never of models, but always *only* of scientists. This was explicit in my statement: 'My thesis is that *psychometricians* are not only uncritical of an issue basic to their discipline but that, in addition, they have constructed a conception of quantification that disguises this' (Michell, 2000, p. 639, emphasis added) and my description of pathological science as a two-level breakdown in processes of critical inquiry wherein 'not only is some hypothesis accepted within the mainstream of a discipline without a serious attempt to test it, but that fact is not acknowledged or, in extreme cases, is disguised' (p. 641). The *mainstream* was obviously of scientists because it is scientists who *accept*, *fail to test* and *disguise*.

The hypothesis at the root of this pathology is that psychological attributes (such as the various intellectual abilities, personality traits and social attitudes) are quantitative. My thesis is that mainstream psychometricians have never seriously attempted to test this hypothesis and have not only declined to acknowledge this failure, but have also attempted to disguise it. (By *mainstream psychometricians* I mean those who attempt to measure psychological attributes or who propose theories intended to sustain such attempts, excluding [Michell, 1997, 1999] the tradition stemming from the writings of Luce and Suppes [e.g. Krantz, Luce, Suppes, & Tversky, 1971] on the foundations of measurement, this tradition still remaining outside the mainstream [Cliff, 1992].)

Borsboom and Mellenbergh agree that psychometricians have not tested the hypothesis that psychological attributes are quantitative, but attribute this to the 'impossibility' of testing hypotheses in isolation. They claim that:

> The fact that we cannot test a hypothesis in isolation, however, is characteristic not of pathological science, but of science in general. From a philosophy of science viewpoint, it seems a particular instance of the Quine–Duhem thesis (which states that no hypothesis is ever tested in isolation), rather than of pathological science. (p. 114)

Regarding this *Quine–Duhem thesis*, Duhem (1914/1954) wrote that the scientist

> . . . can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (p. 187)

Duhem thought that hypotheses could not be tested in isolation, in the sense that testing always requires conjoining hypotheses with other propositions to deduce predictions, but he did not mean that hypotheses could not be tested in isolation in the sense that an experimental test can never be just of a single hypothesis. He noted that a scientist may identify the cause of failed predictions as residing in a

> . . . proposition he wishes to refute, but is he sure it is not in another proposition? If he is, he accepts implicitly the accuracy of all the other propositions he has used, and the validity of his conclusion is as great as the validity of his confidence. (p. 185)

That is, an experiment *taken on its own* teaches less than an experiment *combined with background beliefs*. Hypotheses are only testable in the light of *things already taken to be true*. Otherwise, nothing could ever be tested via data because data, like relevant background beliefs, must be assumed true for a test to proceed, and data are as fallible as such beliefs. Were psychometricians to accept certain parts of item response models along with the data involved, the hypothesis that the relevant attribute is quantitative could be tested in isolation, in the sense that were predictions to fail, it and it alone would be rejected.

So my thesis is unaffected by the Quine–Duhem thesis. Refuting my thesis would require showing that psychometricians, in testing item response models, are as prepared to reject the hypothesis that psychological attributes are quantitative as to reject any other implicit in these models. One way to assess this would be to look at the interpretations psychometricians offer when their models fail.

An examination of some relevant textbooks (Bond & Fox, 2001; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; Suen, 1990) reveals a consistent pattern: the issue of whether the relevant psychological attribute is quantitative is *never* raised as a source of model misfit. Other issues, such as the unidimensionality of the underlying attributes, item-discrimination parameters and local independence, are raised, but item response modellers appear never to question that their attributes are quantitative. How could this striking omission slip Borsboom and Mellenbergh's attention?

Borsboom and Mellenbergh also neglect the second part of my thesis, namely the claim that psychometricians have constructed a conception of quantification that ignores or disguises their failure to investigate the question of whether the relevant attributes are quantitative. I maintained that psychometricians who accept Stevens' (1946) definition of measurement are deflected from investigating this question. However, item response modellers often disregard Stevens' definition. Some (e.g. Bond & Fox, 2001), indeed, are allies in opposing it. Despite this, pathology is not avoided; one way of maintaining it is merely abandoned. Item response modellers hold

that when their models fit, estimates obtained from item responses are measurements of underlying traits. But one will search in vain for acknowledgement of the presumption that the relevant attributes are quantitative. For example, Hambleton et al. (1991) list the assumptions of item response models as unidimensionality, local independence and the various features of the item characteristic functions (e.g. one-parameter logistic, etc.). Item characteristic functions describe how the hypothesized trait is thought to relate to the probability of an item response of a specific kind (say, of a correct response) and are always taken to be continuous. This presumes that the trait is a continuous quantity. I know of no publications in this area where this assumption is mentioned, nor any in which it is recognized that being quantitative is an empirical condition.

Not only do item response modellers fail to consider whether their attributes are quantitative, but also their failure is masked by their declining to list this hypothesis as an assumption. They take it for granted that the attributes to which they apply their models are quantitative and, silently presuming this, select sets of items that produce data fitted by their models, regarding this as sufficient to claim measurement. Nonetheless, could not such instances of fit be interpreted as confirming the hypothesis that the relevant attribute is quantitative, even if this hypothesis is unacknowledged?

Superficially, such instances of model fit might look like tests of this hypothesis, but they are not *serious* attempts, for two reasons. First, in experimental science, in order to test any hypothesis, the relevant theory must be sufficiently detailed to specify the kinds of situations relative to which specific predictions can be deduced. In psychometrics, these situations are, in part, the items constituting the relevant psychological test. At present, psychometrics does not possess theories of sufficient richness. As Goldstein and Wood (1989) noted and, indeed, as Borsboom and Mellenbergh mention in another publication (Borsboom, Mellenbergh, & van Heerden, 2003), item response models contain nothing about the character of the attributes that psychometricians aspire to measure, other than requiring that they be quantitative. (Why psychometricians should require this when they are able to say so little about these attributes is no mystery; see Michell, 2003a.) In the typical psychometric model fitting exercise, items are never deduced from a theory of the attribute. Rather, items are constructed using informal procedures, and items thought to contribute to model misfit are culled. Given that there is no scientific reason to believe that the relevant attributes are quantitative, this way of proceeding can mislead because data can fit these models when no underlying quantitative attribute exists (e.g. Wood, 1978).

Consider the data I have reported elsewhere (Michell, 1994). By imposing an ordinal, semantic structure upon a set of attitude statements, a body of data was produced fitting, almost perfectly, a deterministic item response model, namely Coombs' (1964) unfolding theory, a result replicated by

Sherman (1994) and Johnson (2001). Probabilistic models are similar to deterministic ones except that probabilistic models relax the conditions necessary for fit. So, if a body of data almost fits a deterministic model in a case where the relevant attribute is only ordinal, then it is likely to fit the less demanding requirements of a probabilistic model. Bond and Fox (2001) present other cases where responses to items with a predetermined ordinal structure (derived from Piaget's non-quantitative theory) fit the Rasch model. Even if the underlying psychological attribute was no more than a partial order, a strictly ordered subset of items could be selected, and item response models could well fit ensuing data. That is, these models may fit even when the relevant attributes are non-quantitative.

Second, a genuine test of the hypothesis that an attribute is quantitative requires more than a theory from which test items can be deduced; it requires distinguishing predictions that depend upon an attribute's merely being ordinal from those that depend upon quantitative structure over and above mere order. The theory of conjoint measurement shows that one-parameter item response models, such as the Rasch model, predict a hierarchy of cancellation conditions capable of distinguishing ordinal from quantitative structure (Michell, 1990, 2003b). Attention still needs to be given to the problem of identifying cancellation tests logically independent of mere order in the attribute. While solved for double (Michell, 1988) and triple cancellation tests (Richards, 2002), the more general problem is not. Despite this, enough is known for this issue to have been addressed across the range of relevant attributes. (The work of Karabatsos—e.g. his 1999 paper—is also relevant here.)

Borsboom and Mellenbergh distract the reader from these facts by attempting to make a mountain of the distinction between deterministic and probabilistic models. Currently, probabilistic models are in vogue, but it is disingenuous to think that because they are probabilistic they are thereby superior to deterministic alternatives, such as the Guttman scale model (Guttman, 1944). It is fashionable to dismiss Guttman's model as 'very restrictive' (Borsboom & Mellenbergh, 2004, p. 108), but the fact is that any data fitting a probabilistic item response model will fit a deterministic model, albeit, perhaps, a multidimensional one (Coombs, 1964). Which model is correct could only be discovered if the issue of dimensionality could be resolved in advance, say, via an explicit theory of the attributes, and, as noted, psychometricians studiously avoid that kind of theorizing.

Apart from the fact that most probabilistic item response models are unidimensional, probabilistic models promise measurement, while deterministic item response models, such as Guttman's, always deliver less. Mainstream psychometricians want to be able to promote psychological tests as instruments of measurement (Michell, 1999), and, so, probabilistic models are preferred. Yet the locus of quantitivity in probabilistic models

derives not from anything known directly about the relevant attribute, but from the postulated shape of 'error'.

Probabilistic item response models are usually expressed in terms of the probability of a person making a response of a certain kind to a test item, given the relevant person and item parameters. Considering one-parameter ability test models, Sutcliffe (1986, p. 91) suggests thinking of a person's correct or incorrect item performance on any occasion as due to a person parameter he calls *capability*, which is ability, as normally understood in these models, plus a random *error* component, notionally drawn on each occasion from a probability distribution of possible 'errors', normal, for example, in normal ogive item response models (Lord, 1952) and of an inverse hyperbolic tangent form for logistic ogive models (Rasch, 1960). When item difficulty exceeds capability, the response is incorrect, otherwise it is correct. Now, if a person's correct response to an item depended solely on ability, with no random 'error' component involved, one would only learn the ordinal fact that that person's ability at least matches the difficulty level of the item. Item response modellers derive all quantitative information (as distinct from merely ordinal) from the distributional properties of the random 'error' component. If the model is true, the shape of the 'error' distribution reflects the quantitative structure of the attribute, but if the attribute is not quantitative, the supposed shape of 'error' only projects the image of a fictitious quantitivity. Here, as elsewhere, psychometricians derive what they want most (measures) from what they know least (the shape of 'error') by presuming to already know it.

If the random 'error' concept is retained, but it is admitted that the shape of these 'errors' is unknown, then at best only ordinal relationships between people (or items) follow from test performances (Grayson, 1988) unless the cancellation conditions alluded to above (namely double cancellation, triple cancellation, etc.) obtain. By any fair-minded reckoning, our state of knowledge at present does not support claims to be able to measure psychological attributes using item response models. A small number of modellers (see Sijtsma & Molenaar, 2002), following Mokken's (1971) work, have embraced this weaker admission, developing *non-parametric* item response models, but, interestingly, Mokken's name does not rate a mention in the textbooks surveyed above.

Indeed, a realistic assessment of item response models is not to be found in any of these. Instead, the tempo regarding the prospects of measurement is consistently upbeat and issues relating to the ordinal versus quantitative distinction are not discussed. With the best will in the world, I find no scientific reason in the response by Borsboom and Mellenbergh to revise my assessment of psychometrics. The fact remains that psychometrics has existed for a century and item response models have been proposed for over half a century. In this time, psychometricians have consistently failed to address the issue of whether the attributes they aspire to measure are

quantitative and have masked this fact by colluding in ignoring it. If their aim is the scientific one of attempting to discover the character of the attributes underlying differences in performance on psychological tests, then this is a pathological way of going about it.

References

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Borsboom, D., & Mellenbergh, G.J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology*, *14*, 105–120.

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

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

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

Duhem, P. (1954). *The aim and structure of physical theory* (P.P. Wiener, Trans.). Princeton, NJ: Princeton University Press. (Original work published 1914.)

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139–167.

Grayson, D.A. (1988). Two-group classification and latent trait theory: scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Johnson, T. (2001). Controlling the effect of stimulus context change on attitude statements using Michell's binary tree procedure. *Australian Journal of Psychology*, *53*, 23–28.

Karabatsos, G. (1999). *Axiomatic measurement theory as a basis for model selection in item-response theory*. Paper presented at the 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA.

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

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7. 17 (4, pt. 2).

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Michell, J. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*, 466–473.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology*, *38*, 244–273.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88*, 355–383.
Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.
Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*, 639–667.
Michell, J. (2003a). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology, 13*, 5–31.
Michell, J. (2003b, May). *The theory of conjoint measurement and the Rasch model*. Paper presented at the conference on Epistémologie de la mesure dans sciences sociales: perspectives contemporaines, Institut d'Histoire et Philosophie des Sciences et des Techniques CNRS – Université Paris 1.
Mokken, R.J. (1971). *A theory and procedure of scale analysis with applications in political research.* The Hague: Mouton.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
Richards, B. (2002). *Unidimensional unfolding theory and quantitative differences between attitudes*. Empirical thesis submitted in partial fulfilment of the requirements for the BSc (Honours) degree in Psychology, School of Psychology, University of Sydney.
Sherman, K. (1994). The effect of change of context in Coombs' unfolding theory. *Australian Journal of Psychology, 46*, 41–47.
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680
Suen, H.K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.
Sutcliffe, J.P. (1986). On untrue scores and rash applications. In R.A. Heath (Ed.), *Current issues in cognitive development and mathematical psychology: John A. Keats Festschrift Conference* (pp.87–99). Newcastle, NSW: Department of Psychology, University of Newcastle.
Wood, R. (1978). Fitting the Rasch model—a heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27–32.

ACKNOWLEDGEMENTS. My thanks to the editor of *Theory & Psychology* for his invitation to write a response to the paper by Denny Borsboom and Gideon Mellenbergh. I also thank my colleague David Grayson and research assistant Sharon Medlow for their comments.

JOEL MICHELL teaches psychometrics and the history and philosophy of psychology at the University of Sydney. His main research interest is the history and philosophy of quantification in psychology. He has written two books on this subject: *An Introduction to the Logic of Psychological Measurement* (Erlbaum, 1990); and *Measurement in Psychology: A Critical History of a Methodological Concept* (Cambridge University Press, 1999). He recently collaborated with Philip Bell and Phillip Staines on *Logical*

Psych: Reasoning, Explanation and Writing in Psychology (University of New South Wales Press, 2001) and is currently working on a book with Fiona Hibberd, Agnes Petocz and David Grayson about the relationship between quantitative and qualitative research methods in psychology. ADDRESS: Department of Psychology, University of Sydney, Sydney, 2006, NSW, Australia. [email: joelm@psych.usyd.edu.au]