Measurement, 6: 78–123, 2008

ISSN 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366360802035521

Psychology Press
Taylor & Francis Group

COMMENTARIES

The Consequence of Sustaining a Pathology: Scientific Stagnation—a Commentary on the Target Article "Is Psychometrics a Pathological Science?" by Joel Michell

Paul Barrett
Hogan Assessment Systems, Inc.

For more than 20 years now, Joel Michell has carefully explicated the properties and requirements of quantitative measurement, in books, papers, conference papers, and workshops. In 1997, he published what was essentially one of the most important papers in the whole of psychology for a generation, titled "Quantitative Science and the Definition of Measurement in Psychology."

As Michell notes in his target article, the response from psychologists and psychometricians alike has been largely silence. From my own attempts to relay his thesis in papers, conferences, and academic seminars, I have been met with a mixture of disbelief, disinterest, anger, accusations of intellectual nihilism, ridicule, and generally a steadfast refusal to seriously question the facts. And, make no mistake, Michell deals with facts about measurement. The only coherent responses I have read where the author made a passable attempt at really engaging with Michell's work was in the book *Measurement: Theory and Practice* by Hand (2004), a mathematical statistician and machine learning expert, and Borsboom and Mellenbergh (2004) in a direct response to the paper published earlier by Michell (2000).

Correspondence should be addressed to Paul Barrett, Chief Research Scientist, Hogan Assessment Systems, Inc., 2622 E 21st Street, Tulsa OK 74114. E-mail: pbarrett@hoganassessments.com

Although many who read the current exposition by Michell might again choose to "look the other way" or otherwise downplay the significance of his statements, for those who actually engage in empirical psychological science, rather than as Kline (1998, p. 197) put it "atheoretical response counting, the province of clerks," Michell's observations will hit hard. There are real-world consequences to his thesis, and we are living through them right now. The most important of these is that while psychometricians have advanced their thinking and technical sophistication in leaps and bounds over the past 40 years or so, the practical consequences have been almost nonexistent except in the domain of educational testing and various examination scenarios.

The whole point of establishing a metric for a variable is presumably so we can more precisely explain variations in the magnitudes of phenomena we observe—relating orders or perhaps even standard units of magnitude of the hypothesized variable to observed empirical outcomes. As we move from classes, to simple orders, to a standard unit around which ratios of magnitudes of quantity might be expressed, so will the precision with which we describe variation increase. But, paradoxically, the accuracy of our description will actually decrease if we assume we are measuring a quantitative variable when in fact our empirical observations may support only simple orders.

For example, what if psychological variables are not quantitatively structured, but we proceed anyway to treat them as if they were? What are the likely consequences when we come to use these variable magnitudes to explain or predict the phenomena we consider important for their validity? Well, just think what happens if we make measurement of length with a ruler whose units vary in size over the length of its scale. Then, using this ruler, we seek to measure the various heights of individuals. What happens is that we would observe an approximate order between our measures and the empirical observations of these individuals. Some individuals would be observed to be taller than their measures seem to indicate, some shorter. Some investigators might rightly conclude that the empirical observations are veridical and conclude that something is wrong with the measuring process. Psychometricians and IRT/latent variable theorists instead seem to find ways to avoid explaining the lack of concordance of observations with their measures by treating the error as "necessary" (as in IRT), or develop assumption-laden methods of "correcting" data for "artifacts which cause attenuated relationships" (as in meta analysis), or begin to exaggerate the verbal description of what are essentially inconsequential analysis results, or as in interpretation of statistical parameters such as estimates of correlation, the link between the magnitude of a parameter to its accuracy of explanation is subverted (Hemphill, 2003).

The real-world consequences of this systematic aversion to properly considering the presumed status of a psychological variable is that our journals are now filled with studies that are largely trivial exemplars of mostly inaccurate

explanations of phenomena. Nothing seems to have changed since Lykken's similar observations of this phenomenon back in 1991. Sophisticated statistical models are now used to produce results that seem to have little real-world practical or even scientific consequence. For those who might doubt this, try and name just one result from the application of psychometrics in the last 50 years which has yielded a finding so important that it has changed the course of investigation and understanding of a phenomenon. There are a few of these within the recent empirical science of psychology and neuroscience, such as Gigerenzer and colleagues' work on fast and frugal heuristics (Gigerenzer & Todd, 1999) and Gould and colleagues' work on neurogenesis (Gould, Tanapat, McEwen, Flugges, & Fuchs, 1998), neither of which rely on the kinds of statistical machinations employed by those who seek to create "measures" from what are mostly questionnaire responses.

For example, Freedman (2004, p. 200) states the following at the end of his book *Statistical Models, Theory and Practice*:

> In the social and behavioral sciences, far-reaching claims are often made for the superiority of advanced quantitative methods—by those who manage to ignore the far-reaching assumptions behind the models. ... The goal of empirical research is—or should be—to increase our understanding of the phenomena, rather than displaying our mastery of technique.

And, I am not alone in pointing out that psychometric test theory has yielded little of scientific value. Blinkhorn (1997, p. 175), in a review of 50 years of test theory, states the following in the abstract to his paper:

> Contemporary test theory, with its emphasis on statistical rather than psychological models, has become inaccessible to the majority of test users, and predominantly reflects educational rather than psychological concerns. Real progress may depend on the emergence of a new and radical reconceptualization.

What kind of science promulgates the fundamental requirement for IRT that, as Michell puts it in is target article, "improving the precision of our observational conditions decreases the precision of our observations"? Andrich (2003, p. 27) has tried to provide an answer to this "paradox" in a recent paper, stating that

> the need for probabilistic models in the social sciences does not result from lack of precision in the data, but the contrary; it implies that social scientists who work in the units of their measuring devices may need probabilistic models just because they are working at precise levels relative to the variation of the values of the traits they are measuring. To support this perspective, it was observed that, where measuring devices in the physical sciences have become very precise, that is at the quantum level, probabilistic formulations have been invoked.

This is a peculiar line of reasoning for two reasons. First, what are these "precise levels" at which psychologists work? The precision is accorded by the quite arbitrary use of a real-valued, additively structured, numerical relational system with arbitrary units, which forms the imposed quantitative structure of the latent variable. Precise it may be, but the assumed one-to-one relationship between the numerical relational system and the observed empirical relational system is rarely tested, as Michell (this journal) notes. Second, the argument that physicists have to use probabilistic models because of the increased precision of their measuring devices is somewhat inappropriate. For example, Penrose (1997) notes that the accuracy of deterministic Newtonian mechanics is about 10^{-7}, that of Einstein's relativity is about 10^{-14}, and that of quantum mechanics is about 10^{-11}. Measurement at the deterministic level in physics is so far beyond the precision of any measurement we can make of any psychological variable that Andrich's arguments are rendered irrelevant for that reason alone. However, the implied inference—that because measurement made at the quantum level of the physical world requires a probabilistic formulation of events, so therefore do psychometric models—ignores the reasons *why* a probabilistic formulation was required in quantum mechanics. These involve a change in perspective of the very nature of causality (nonlocal quantum entanglement) and Heisenberg's uncertainty principle (Giancoli, 1988), which states that we cannot simultaneously measure both the position and the momentum of an object precisely. However, if Andrich is indeed recommending that the worldview of quantum physics should be adopted by psychologists, well, that is something else entirely.

In the end, none of what Michell states is actually that difficult or complex to understand. As he notes, we can barely specify the relations between variables and their measures beyond a simple order. We have no body of sufficiently detailed theory as to why we can even observe such orders for the vast majority of these variables. Instead of simply "dealing" with this state of affairs as scientists might, which is to think about why it is proving so difficult and perhaps concentrate more on methods for establishing some decent levels of predictive accuracy of whatever we hold as "important," psychometricians have instead created a self-sustaining illusion that data-model-driven statistical complexity equates to more accurate science. Where is the evidence for this proposition?

For example, look at the revolution in the area of risk prediction of recidivism in forensic psychology. Society and the courts required more from forensic psychologists than "clinical hand-waving" when it came to predicting risk of recidivism and dangerousness (Ziskin, Faust, & Dawes, 1995). The best tool for predicting violent recidivism, the Violence Risk Assessment Guide (Webster, Harris, Rice, Cormier, & Quinsey, 1994), turns out to be a straightforward behavioral checklist, using a simple integer importance-weighting scheme for its items, which, when summed, produces a classification accuracy of 72%, and

has been replicated in many countries since—no structural equation modeling, no data-model-driven regressions, no "made-up" latent variables, and, above all, no assumptions made as to the "quantitative" nature of the scale of risk propensity. Clearly, the understanding behind why it works is limited, and there is no obvious set of sufficiently detailed theoretical propositions that could be tested to explain the many behavioral constituents of this risk scale; this will come over time. But it works, to a quantified degree of predictive accuracy, and has changed international forensic-legal practices.

Given the work by Fan (1998) and the recent PhD thesis by Courville (2004), showing that IRT latent variable yields "scores" and parameter estimates which are not appreciably different from simple sum-score test theory, clearly, the debate over the superiority of modern "latent variable" psychometrics as part of an investigative science over "back of the matchbox" use of sum-scores is pointless. The fact of the matter is that both seem to work equally as well in producing what are essentially the same scores, when properly representative samples of questionnaire data are analyzed.

The obvious conclusion is that if we are to generate progress in our science, increasingly sophisticated methods of statistical data-model analysis are not going to help. Something about the way we go about construing psychological variables, their measurement, and their presumed causal relations is wrong. That's where our efforts should be concentrated as scientists, not on yet more questionnaire item test theory and assumption-laden structural latent variable models that ignore the fundamental issues of quantitative measurement. We might seriously consider whether psychology might be, for all intents and purposes, a nonquantitative science best approached with variables that vary in simple orders and even just classes, and where equifinality (many causes of the same observed outcome) is the order of the day (Richters, 1997).

Adherence to modern psychometric dogma, inasmuch as I agree with Michell that it is a pathology of science, is strangling innovation and creativity in areas concerning the investigation of psychological phenomena. In 1997, Blinkhorn concluded his own review of 50 years' of test theory with the following:

> A more radical view is that current styles of test theory simply have no more of practical value to offer, that the implicit assumptions which have guided research for nearly a hundred years place constraints on what can be achieved which have no more elasticity, and that a new start is needed. It is not a story of failure but of pragmatic success that has reached its limits, and indeed reached them some time ago.
>
> The challenge is to recover the element of surprise—astonishment even—which greeted the developments of the first decade of this century when effective psychological measurement was first demonstrated. For too long test theory has concerned itself with ever cleverer accounts of unimproved practical effectiveness and not with recommendations as to how to devise more effective tests. (p. 183)

Eleven years later, nothing has changed. If ever anyone needed convincing that sustaining a pathology has consequences, this is it.

REFERENCES

Andrich, D. (2003). On the distribution of measurements in units that are not arbitrary. *Epistemology of Measurement*, *42*(4), 557–589.

Blinkhorn, S. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology*, *50*(2), 175–186.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology*, *14*(1), 105–120.

Courville, T.G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*. PhD thesis, Texas A & M University. Retrieved January 28, 2008, from http://txspace.tamu.edu/handle/1969.1/1064

Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381.

Freedman, D. A. (2004). *Statistical models: Theory and practice*. New York: Cambridge University Press.

Giancoli, D. C. (1988). *Physics for scientists and engineers with modern physics* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall International.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gould, E., Tanapat, P., McEwen, B. S., Flugges, G., & Fuchs, E. (1998). Proliferation of granule cell precursors in the dentate gyrus of adult monkeys is diminished by stress. *Proceedings of the National Academy of Sciences*, *95*, 3168–3171.

Hand, D. J. (2004). *Measurement: Theory and practice: The world through quantification*. London: Arnold Publishers.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. New York: Routledge.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove. (Eds.), *Thinking clearly about psychology. Volume 1: Matters of public interest* (pp. 3–39). Minneapolis, MN: University of Minnesota Press.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Michell, J. (2000). Normal science, pathological science, and psychometrics. *Theory & Psychology*, *10*(5), 639–667.

Penrose, R. (1997). The mysteries of quantum physics. In R. Penrose, A. Shimony, N. Cartwright, & S. Hawking (Eds.), *The large, the small, and the human mind* (pp. 50–92). Cambridge, UK: Cambridge University Press.

Richters, J. E. (1997). The Hubble hypothesis and the developmentalist's dilemma. *Development and Psychopathology*, *9*(2), 193–229.

Webster, C. D., Harris, G. T., Rice, M. E., Cormier, C., & Quinsey, V. L. (1994). *The violence prediction scheme: Assessing dangerousness in high risk men*. Toronto, Canada: University of Toronto, Centre of Criminology.

Ziskin, J., Faust, D., & Dawes, R. (1995). *Coping with psychiatric and psychological testimony* (5th ed.). Los Angeles, CA: Law and Psychology Press.