Educational Measurement (4th ed.)

Article *in* Structural Equation Modeling A Multidisciplinary Journal · October 2009

DOI: 10.1080/10705510903206097

CITATIONS
8

READS
92

1 author:

Denny Borsboom
University of Amsterdam
191 PUBLICATIONS **7,282** CITATIONS

SEE PROFILE

Structural Equation Modeling, 16:702–711, 2009

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510903206097

Psychology Press
Taylor & Francis Group

BOOK REVIEW

Educational Measurement (4th ed.). R. L. Brennan (Ed.). Westport, CT: Praeger, 2006, 779 pages, $125.00.

Reviewed by Denny Borsboom
University of Amsterdam

Not counting its editor, R. L. Brennan, who has earned my eternal admiration for successfully managing what must have been a monstrous project, I am probably the fourth person alive—with Wainer (2007), Green (2008), and Cizek (2008)—to have read the fourth edition of *Educational Measurement* from cover to cover. Given the late date of this book review, I expect this to remain the case, as this book is unlikely to be read in its entirety except by those intent on reviewing it. On the back flap, *Educational Measurement* is marketed as "the bible in its field," and as far as readability goes, the comparison to the Holy Scripture is quite apt. *Educational Measurement* is not a page turner.

Its limited readability need not be an insurmountable problem for the book, however, because over the years the editions of *Educational Measurement* have grown to become reference texts more than anything else. In that respect, the book definitely stands its ground. In fact, there is nothing that compares to it. If *Educational Measurement* contains a chapter on what you are looking for, call it *A,* then you can be sure (with one or two exceptions) that the relevant chapter will (a) provide an exhaustive list of the many ways in which *A* has been done in the past, (b) discuss the manifold of problems people encountered when doing *A,* (c) review a dozen or more empirical studies designed to figure out the best way of doing *A* (usually inconclusive), and (d) urge researchers to do more in the way of validation research with respect to *A.* Although, at times, the uniform adoption of the laundry list model makes reading this book feel like a mild form of torture, it works well in terms of organization, and I have to say that I did learn very much about educational testing, as most chapters are of high quality in terms of scholarship and content.

However, despite the many topics that receive detailed attention in the book and its considerable size (779 pages; Wainer, 2007, estimates its weight at 4.5 pounds), some of the most interesting facts about this monumental volume concern the things it does not contain. There are some glaring omissions that, as an outsider (not an American, not an educational tester, not a construct validity enthusiast), stared me quite directly in the face, but appear to have gone

Correspondence should be addressed to Denny Borsboom, Department of Psychology, University of Amsterdam, Roeterstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: d.borsboom@uva.nl

entirely unnoticed among the contributors to the book. These issues are interesting because, in my view, they concern the core of the educational testing adventure. I will get to them shortly; however, first an overview of the book's contents is provided.

WHAT IS IN THE BOOK

Apart from an introductory chapter by Brennan, which deals with the history and probable future of educational testing, the fourth edition of *Educational Measurement* consists of three parts: *"Theory and General Principles,"* which addresses basic psychometric theory; "Construction, Administration, and Scoring," which mainly deals with practical issues that arise in the implementation of testing programs; and *"Applications,"* which reviews specific examples of testing programs as well as legal and ethical issues that have emerged from those.

The first part of the book kicks off with a chapter on validation by Kane, who aims to give a blueprint for constructing decent arguments that ought to support test score interpretations. Compared to Messick's (1989) chapter in the previous volume, this is validity-lite, as Kane is concise and does not immerse himself in much philosophy. The most controversial aspect of the previous edition of *Educational Measurement,* Messick's consequential basis of construct validity, appears to be on its way out. Both Kane and Brennan discuss the consequences of test use, as does Shepard, but their discussions are sufficiently polite to convince me that we can safely say goodbye to the consequential part of construct validity. Kane's chapter further contains little that one could reasonably object to, but that is partly because it does not take a stance regarding issues one would really care to object to (more on that later).

Haertel follows with a chapter on reliability that mainly covers the concept as it is used in true score theory and its big cousin, generalizability theory. The modern test theory concept of information is not covered in Haertel's chapter, but in the next chapter on item response theory (IRT), by Yen and Fitzpatrick, which I found infelicitous as information is clearly more basic than classical reliability (Mellenbergh, 1996) and could have been used to counter many mistaken but widely prevalent interpretations of reliability quite effectively. Yen and Fitzpatrick's chapter gives a relatively accessible overview of the many models that are currently available to the test developer and user. The chapter does not suffer from the often seen IRT tunnel vision by which all latent variables are continuous (a dogma that still plagues many of its applications, and some of the other chapters in the book as well) and gives some attention to alternative latent structures.

Kolen's chapter on scaling and norming offers an extensive discussion of what to the naive reader appears a bewildering variety of scoring functions that are in use in educational testing. Holland and Dorans's chapter on linking and equating gives a fundamental, and theoretically integrative, overview of the many ways in which scores on one test can be projected onto scores on another test—they also give due attention to the extremely strong assumptions that lie behind the practice of equating. Camilli's chapter on test fairness reviews the conventional wisdom on the topic, including, unfortunately, some of its mistakes—like the idea that "[i]f a test item is fair, it is ... equivalent across groups with respect to measurement and prediction" (p. 226). This situation is, in general, unlikely to obtain and in many cases impossible, as invariance in measurement and prediction are often mutually inconsistent properties—see Millsap (1997, 2007). Mislevy closes the first part of the book with a sparkling chapter on cognitive psychology

and educational assessment, which is a creative and eclectic synthesis of ideas taken from general philosophy, the philosophy of science, linguistics, cognitive psychology, connectionism, and psychometrics. I personally doubt whether the grand scheme that Mislevy sketches is necessary to motivate the type of modeling strategies that it ultimately converges on. However, taken as a whole, the chapter is original and thought provoking, and in this sense it stands out from the rest of the volume.

The second part of the volume is opened by a discussion on test development by Schmeiser and Welch, which provides a huge (not to say daunting) list of things to take care of while designing a test. They open their chapter with the question of whether test design is an art or a science, and end it by saying it is both; however, in my mind the construction of current educational tests resembles industrial production more than anything else. That image is reinforced by Cohen and Wollack's chapter on administration, security, scoring, and reporting, which, given the topic, provides a surprisingly interesting (at times even mildly funny) description of the many practical problems involved in educational measurement, ranging from handling special procedures for the disabled to automated essay grading.

Lane and Stone's chapter on performance assessment provides a mixed image of the opportunities, limitations, and benefits of performance assessments (for which examinees are required to show their skills directly; e.g., when music students are required to play a piece of music). Although popular in the public eye for its transparency, a lengthy review of research results gives no clear picture of whether performance assessments are, in general, worth their money. Also, modeling performance assessments with standard measurement models feels like squaring the circle, and it is not clear that more appropriate measurement models are forthcoming. Hambleton and Pitoniak give an overview of methods and problems involved in the construction of categories like *proficient* in the chapter "Setting Performance Standards." They review a large number of methods to accomplish such standards with the use of panelists, which tend to give varied and often conflicting results. The second part of the volume closes with Drasgow, Luecht, and Bennett's chapter on technology and testing, where they sketch a vivid, at times almost futuristic, image of what testing programs are on the verge of becoming. The authors do an excellent job of explaining how drastically technology is changing the nature of what educational testers are doing.

The third part of *Educational Measurement* opens with a discussion of second language testing by Chalboub-Deville and Deville. The chapter offers a historical overview of the practice of second language testing, and is the first chapter in the volume to use examples from outside of the United States. The rest of the chapter showcases abstract theorizing, sometimes reminiscent of postmodernism, that I found quite hard to follow and of questionable relevance for testing practices.

The next two chapters, by Koretz and Hamilton on testing for accountability in K–12, and by Ferrara and DeMauro on standardized assessment of individual achievement in K–12, offer a critical discussion of the psychometric practices in educational testing in the United States system (for non-Americans, K–12 means kindergarten to Grade 12 and spans roughly the period between ages 5 and 18). Both chapters give due attention to assessment problems that have arisen as a result of the No Child Left Behind Act, which mandates that states and schools set targets for educational improvement and show that they meet these targets using standardized tests (which nevertheless vary by state; the United States does not have a centrally organized examination system that regulates admittance to higher education as, for

instance, many European countries do). In this way, tests and test scores have become political instruments-with all the dangers inherent to those, as the chapters illustrate vividly.

The critical spirit is taken forward by Shepard, in her chapter on classroom assessment. Shepard reviews literature on the effects of grading, advocates the use of formative assessment (assessment that serves an explicit instructional goal), and in fact suggests that a "complete paradigm and cultural shift" (p. 641) is necessary. That shift should apparently be based on sociocognitive theory. Historically speaking, complete paradigm shifts based on the shaky grounds of social science have proven even more dangerous than politicians wielding tests, especially if implemented by an elite of believers, so I would say that care should be taken.

Zwick, in her chapter on higher education admission testing, provides an overview of the use of tests like the SAT and ACT. She provides a balanced discussion of the predictive utility of such tests; the overall image is that the correlations with study success are smaller for standardized test scores than for high school grade-point average (GPA), and that the incremental predictive utility of test scores over GPA is small (typically on the order of .1 or less); moreover, test scores are of highly limited predictive value with respect to educational attainment within a given college or graduate school (likely due to self-selection and restriction of range). She also offers an extensive discussion on fairness that complements Camilli's chapter well.

"Monitoring Educational Progress With Group-Score Assessments," by Mazzeo, Lazer, and Zieky, covers formal and practical aspects of the use of large-scale assessments to make inferences about groups of people (e.g., districts, states, nations, etc.), like the National Assessment of Educational Progress (NAEP) and the Program for International Student Assessment (PISA). Clauser, Margolis, and Case follow with a chapter that reviews testing for licensure and certification in the professions. The chapter mainly focuses on methods for the estimation of classification accuracy and provides a discussion of validity threats that are typical for this form of testing.

The final chapter of the book is by Philips and Camara, and reviews legal and ethical issues. The chapter is built around the discussion of lawsuits relating to test fairness, the release of test items to the public after administration, and various issues relating to special test conditions (e.g., for persons with disabilities). The chapter offers a very interesting peek into the legal forces that shape American testing practices as well as the conceptual framework of educational testing. I ended up thinking that every non-American psychometrician should be required to study this chapter, for it is much easier to understand how the conceptual framework of educational testing has become what it is today, if one recognizes the profoundly important role the framework plays in the United States court of law.

WHAT IS NOT IN THE BOOK

Michell (1997, 2000, 2008) has criticized social scientists for relying on inadequate definitions of measurement, most notably that of Stevens (1946) as the assignment of numerals according to rule; in a literature search, he found that 39 out of 44 books that were concerned with psychometrics or measurement in the social sciences used a variant of this definition (Michell, 1997). Should he choose to include *Educational Measurement* in his sample, he will have to

add a missing value to his data file, for despite its title, this book contains no attempt to define measurement.

In fact, although the word *measurement* figures as prominently in the book as the title suggests, there is no discussion of what it might mean; no discussion of the extant philosophy of science literature on the topic; no discussion of formal measurement theory; no discussion of how the activity of measurement relates to the activity of testing; and no discussion of the relevance of all these issues to the question of validity and the process of validation. With the exception of Mislevy, who in my opinion entertains a dubious interpretation of measurement models as reasoning devices in a narrative space, but nevertheless touched on the issue long enough to convince me that he is not making such identifications, the authors of the book chapters appear to consistently conflate the terms *testing* and *measurement, test scores* and *measurements,* and *test theory* and *measurement theory,* as if any test score is automatically a measure, any testing procedure is automatically a measurement procedure, and test theory is more or less the same thing as measurement theory.

Now, this is clearly mistaken. One can concoct all kinds of tests that should not be considered measures. I could register your birth year, multiply it by your length, and divide the resulting number by the square root of the number in your postal code. The resulting test scores could surely be used for various purposes, some more defensible than others, and I am certain that Kane's argument-based approach to validity could be used to back up any interpretative inferences you might be prepared to make. However, it would be far-fetched to say that the created test scores should be considered measurements of an underlying attribute. Test scores can sometimes be interpreted as measurements, that is, as bearing a systematic connection to a measured attribute so that testees' positions on this attribute can be inferred from their test scores, but not always and not necessarily so. The question of whether educational test scores can be so interpreted, and under which conditions, receives no attention at all in *Educational Measurement*-which must be considered remarkable, especially in view of the attention that this question has received in the recent literature on psychological measurement (e.g., Michell, 1997, 1999, 2008).

A related omission that would appear to be relevant is the distinction, probably familiar to most readers of this journal, between formative and reflective models (e.g., Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). This distinction could have been used, in my view, to furnish a distinction between test scores that can be considered measures (namely, those that are correctly described by reflective models) and test scores that should not be considered measures but rather composite variables (those correctly described by formative models). However, the *Educational Measurement* volume provides no references to the literature on formative versus reflective models, nor a discussion of the relevance of this distinction. This is unfortunate, as it is plausible that many of the test scores used in educational testing are plausibly constructed as being governed by a formative rather than a reflective model—and for formative models, the rules of the game are considerably different (Bollen & Lennox, 1991).

None of the authors in this volume show awareness of the fact that there might be a relevant choice to make here. However, by equating the notions of test score and measure, one is unwittingly importing the notions associated with reflective measurement models (i.e., latent variables, reliability, information, etc.), even though these might be inappropriate. If one has test scores that are composites made out of measures of (sets of) distinct attributes, and treats them as reflective measures of a single attribute, the concepts imported from standard psychometric

theory might lead one to routinely require high internal consistency, for instance, although this is inappropriate (Bollen & Lennox, 1991); also, one might get completely stuck when one tries to answer the question of which construct the test scores really measure, because there is no single attribute that underlies each and every one of the items, so that in all honesty we should say that *no* construct is being measured (or, perhaps, that *very many* constructs are measured—I am uncertain what constructs are so I leave the answer to this question to the construct validity theorists).

This brings us to another notable omission in the volume, which concerns the lack of a definition of validity. Although the authors of the chapters cite previous definitions of validity quite frequently, even Kane (whose chapter, I must admit, is appropriately entitled "Validation" rather than "Validity") does not explicitly commit himself to one of the previous definitions—in fact, he gives no definition at all.[1] Neither does he (or anybody else in the volume) make an attempt at specifying what the necessary or sufficient conditions for validity are, or could be, or should be. In fact, Brennan, in his opening chapter, and Kane, in his chapter on validation, both make the appearance that they do not find it necessary to specify such conditions because they are not attainable anyway (the current edition, like previous ones, frequently addresses the never-ending character of validation research, rehearses at several places the all-models-are-approximations mantra, refers to the open-ended nature of validity evidence, etc.). I find this to be a serious mistake, for several reasons.

First, the fact that something is not attainable (an issue of which, in the case of validity, I am not convinced, by the way) does not excuse one from the duty of giving a definition or explication of what that something is—especially if one is writing "the bible in its field" and the field is educational measurement. If one writes a book that mentions the word validity at every other page, one is more or less obliged to give an idea of what one means with that concept. However, there is no flesh on the bones of validity in this volume of *Educational Measurement.* Apparently the authors think that one can discuss validation research without having a definition of validity in hand.

I doubt whether this should be taken seriously. Whether one understands validation as the process of finding out whether a test is valid (as I do) or as the process of finding out whether an interpretative inference is valid (as Kane apparently does), one should better have an idea of what one is trying to find out. Otherwise, one is indeed bound to get stuck in a never-ending research process—not because it is impossible to validate a test or test score interpretation, but because one has no idea of what one is trying to achieve.

Second, one simply cannot know that something is unattainable if one does not have a clear idea of what that something is. For instance, if one wants to claim that, say, truth is unattainable, such a statement is meaningless without a clear semantics for the word *truth* (such semantics are not obvious, as the philosophical literature testifies). The same holds for validity. I cannot see why validity should be considered unattainable if nobody gives a clear definition of what it is. It seems to me, in fact, that the authors of *Educational Measurement* conflate the task

[1] It is moreover somewhat unclear what Kane's chapter has to do with measurement in particular, as it discusses ways suited to back up any kind of test score interpretation with any kind evidential or theoretical warrant, whereas validity should, in my view (Borsboom, Mellenbergh, & Van Heerden, 2004), be concerned with a very specific kind of interpretation (e.g., this test measures verbal intelligence) and very specific kind of evidence (namely, evidence that bears on the question whether the test is indeed sensitive to variation in the attribute of interest, that is, whether variation in the attribute of interest causes variation in the test scores).

of giving a clear definition of validity (easy) with the task of saying which tests are plausibly valid and which are not (more difficult) or the project of specifying which kinds of evidence would be proof of validity in general (hard to impossible).

Clearly, one does not have to make the latter kind of specification to define validity. Just as one can say that truth consists of the correspondence of propositions to states of affairs in the world, without specifying evidential conditions under which one is licensed to conclude that any given proposition is true, one can say that validity consists in, say, a test's sensitivity to variation in the intended attribute without specifying the evidential conditions that license this conclusion. A clear semantics, in other words, does not require or presuppose standards of empirical evidence or justification.

Semantics is, in my view, by far the weakest point of psychometric theory, and authors on educational measurement are not champions of clarity either. It is therefore a major omission of the literature in this field in general, and the volume under review in particular, that the semantics of crucial concepts like measurement and validity receive so little attention. It is hard to escape the impression that some particularly difficult issues are being circumvented, although the fact that almost every author in this volume appears so completely oblivious to the issues at hand suggests that this is unintentional. Could this be a paradigmatic blind spot? I don't know, but I do know that it is often nearly impossible to gain a real understanding of what the authors are trying to say, because the meaning of their central concepts is never explicated. As usual, there is a lot of construct-validity-speak in the book (e.g., about latent constructs being tapped, test score interpretations being governed by construct theory, and test scores being screened for construct-irrelevant variance), but if one is not baptized in the construct validity church it is completely unclear what is meant by such talk (how on earth does one *tap* a *construct?*). As a result, as one who does not master the grammar of this language, I remained unsure that educational testing is not a case of shadowboxing; certainly impressive moves are made, but one wonders whether anything is ever really being hit at all.

Other notable omissions, from my perspective, are the following. First, causal reasoning and modeling have clearly become very important in educational testing. For instance, the use of value-added accountability systems requires schools or states to show, as Koretz and Hamilton put it, "how much schools or teachers contribute to student achievement growth" (p. 541). It is difficult not to construe such issues in terms of causality, and therefore a chapter on approaches to causal inference (which can no longer be called recent; e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 2000) could have provided a better insight with respect to issues like confounding, measurement invariance, and Simpson's paradox, all of which have become extremely important in educational measurement.[2] I could find no references to, or discussions of, any of the hundreds of papers and books that have appeared on this topic in the past, say, 20 years.

[2]As a side note, such investigations would probably reveal that consistent interpretations of procedures commonly followed in educational testing require latent variables to function as common causes of the item scores. For instance, latent variables should mediate effects of relevant background variables for the test scores to have measurement invariance. It is doubtful whether instrumentalist conceptions of latent variables, that figure prominently in this book (i.e., latent variables are just overall summaries of performance, narrative concepts, scaling devices, etc.) can sustain such interpretations, as overall summaries of performance are post-hoc constructions of the researcher that one assumes are causally inert with respect to the processes that lead to item responses (Borsboom, 2008; Borsboom, Mellenbergh, & Van Heerden, 2003, 2004).

Second, *Educational Measurement* is largely written from a highly specific psychometric perspective. Apart from a handful of authors taking their lead from generalizability theory, most authors either explicitly or implicitly reason from a two- or three-parameter logistic IRT model; that is, they assume unidimensionality, continuity, local independence, smooth item response curves, normal latent distributions, and so on. It would surprise me if such assumptions were indeed satisfied in typical applications of educational testing. A more thorough treatment of models with weaker assumptions (e.g., Junker & Sijtsma, 2001) and the inferences that they support would have been desirable given the topic. Also, various insights that are common ground in the structural equation modeling and factor analytic traditions appear to have been missed entirely by the educational testing field (e.g., the issue of prediction vs. measurement invariance as discussed in Millsap, 1997, 2007; the distinction between reflective and formative models as discussed in Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Howell, Breivik, & Wilcox, 2007). Finally, apart from a short discussion by Mislevy, there is little space for categorical latent structures as, for instance, used in cognitive diagnostic models (Junker & Sijtsma, 2001; Leighton & Gierl, 2007). In short, much of *Educational Measurement* rests on a highly specific and limited psychometric basis. Given the rapid developments in generalized latent variable modeling of the past 20 years (e.g., Muthén, 2002) a wider scope would have been both possible and desirable.

Third, *Educational Measurement* is a book written on Americans by Americans for Americans. Excluding parts of the chapters by Chelboub-Deville and Deville, and Mazzeo, Laker, and Zieky, the focus is entirely on the American system. Now, I understand very well that the volume is intended to function in the context of assessment in the United States, and I have no quarrels with that at all. However, by reviewing approaches taken in other parts of the world, and at appropriate points contrasting the U.S. system with other systems, the authors could have achieved greater clarity in their expositions—greater contrast always improves one's vision. One also assumes that, at least now and then, one could learn something by looking at one's neighbors. For instance, to a certain degree Shepard's paradigm shift toward teaching methods based on sociocognitive theory (methods that supposedly align with psychological development, and such) has been implemented in my own country of residence, The Netherlands, in a series of major educational reforms. The definitive evaluation of the results is a matter of contention, but the common sentiment in the Dutch media and among politicians is that the program has failed, with a resulting push back to traditional methods. One thing that the authors of this volume could certainly learn from the Dutch reforms is how not to implement them. In general, however, I think that an overview and comparison of the U.S. assessment system with other systems that exist in the world would have improved the volume considerably.

CONCLUSION

Perhaps the most interesting thing about large-scale testing and the psychometric technology it invokes is not how it works but that it exists at all. For what struck me most about *Educational Measurement,* by far, is the magnitude of the assessment project as currently implemented in the United States. The use of computers and continuous testing programs has led to an enormous industry that continuously updates its large item banks; all of these items have to be written by item writers (who have to be trained), reviewed, pretested, analyzed, evaluated,

put into an administration scheme, administrated, scored (sometimes judged by multiple raters, who have to be trained), and reported. At the same time, the security of the item bank has to be monitored, item parameters have to be updated, testing centers reserved, and computer systems maintained, while lawyers are defending the validity, reliability, and fairness of the tests in court. In short, educational testing in the United States has become a billion-dollar industry complete with everything that billion-dollar industries include.

Also, it is remarkable, not to say staggering, how quickly technology has changed the testing process to its core in the mere 20 years that separate the previous edition of *Educational Measurement* from the current one. For instance, to address security problems that arise in continuous testing programs, as well as court rulings that require disclosure after administration, item pools have to be refreshed continuously. Therefore, the very idea of a test (as a relatively fixed collection of items) has become diluted to the point that one wonders whether traditional psychometric ways of thinking about tests still make sense in this context. (Also, one wonders to what one should refer if one intends to criticize the validity of the tests in question: a set of items, a set of test scores, or a production system?) In fact, changes are happening so rapidly that this edition of *Educational Measurement* might not be an up-to-date source for a very long time.

In closing, *Educational Measurement* will be a significant influence on people's thinking about psychometrics, assessment systems, and educational testing. I disagree with Wainer (2007), who asks whether "we really need a cicada-like book that reappears every 18 years with 80% to 90% of the same material" (p. 485). *Educational Measurement* is more than a book; it is an anchor point in people's conceptual horizon. Certainly, the book has shortcomings, and considerable ones at that. Despite this, however, *Educational Measurement* provides a unique window onto current testing practices and sketches a breathtaking view of what must surely be one of mankind's more ambitious undertakings. Whether it all makes sense I do not know-but it certainly is impressive.

REFERENCES

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

Borsboom, D. (2008). Latent variable theory. *Measurement, 6,* 25–53.

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.

Cizek, G. J. (2008). Assessing *Educational Measurement*: Ovations, omissions, opportunities. *Educational Researcher, 37,* 96–100.

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

Green, B. F. (2008). Book review: Educational measurement (4th ed.). *Journal of Educational Measurement, 45,* 195–200.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12,* 201–218.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

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.

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.* New York: Cambridge University Press.

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

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6,* 7–24.

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

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 463–473.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, UK: Cambridge University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search.* New York: Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 667–680.

Wainer, H. (2007). A psychometric cicada: Educational Measurement returns. *Educational Researcher, 36,* 485–486.