How Cognitive Science Challenges the Educational Measurement Tradition

Robert J. Mislevy
University of Maryland

January 30, 2008

This work was supported by the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Introduction

The procedures through which measurement specialists investigate validity, establish reliability, and ensure fairness are enmeshed with the language and worldview of trait and behavioral psychology. The resulting narrative space articulates poorly with an emerging integration of individual, situative, and social perspectives on cognition—a "sociocognitive" perspective, in the terminology of Atkinson, et al. (2007). Cognitive science challenges the educational measurement tradition to bridge this widening chasm. Doing so entails a broader conceptions of the nature of proficiency and ways it is evidenced. It means rethinking what we are actually doing when we use measurement models. It requires for an articulation between, on the one hand, coarser-grained, between-persons measurement models for studying patterns in behaviors that are at the right level for many practical educational problems, and on the other hand, finer-grained, within-persons models for studying the genesis of those behaviors.

In this comment, I sketch key elements of a sociocognitive perspective, note some of its implications for assessment, describe a compatible view of measurement modeling, and call attention to issues that require immediate attention. We see that Borsboom (in press), Markus (in press), and Michell (in press) wrestle with problems that touch on this fundamental challenge, but staying within the measurement tradition limits their progress.

Snow and Lohman's Audacious Claim

In a publication no less foundational than the 3rd edition of *Educational Measurement* (Linn, 1989), Messick (1989) defines a trait as "a relatively stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (p. 15). What I will call "the educational measurement tradition" is a particular extension of this definition: Traits correspond to real properties of people, their character can be inferred by patterns of performances between people on tasks of different types, they are quantitative in nature, and inferences about their values can be drawn through models under which observable variables such as item responses or test

scores depend stochastically on traits. Snow and Lohman's chapter on cognitive psychology in the same volume questions its very foundations:

> Summary test scores, and factors based on them, have often been though of as "signs" indicating the presence of underlying, latent traits. … An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance. The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts. *Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs.* (Snow & Lohman, 1989, p. 317) [emphasis added]

The Sociocognitive Perspective

Snow and Lohman grounded their claim in the "cognitive revolution" of the 1960s and 1970s, exemplified by Newell and Simon's (1972) *Human Information Processing*, which probed beyond the trait metaphor to study the nature of knowledge and how people might acquire, store, and retrieve it. The so-called first generation cognitive science drew on the metaphor of analytic computation, in the form of rules, production systems, task decompositions, and means–ends analyses.

Intervening research enriches the meaning of Snow and Lohman's claim. The view that these information processing structures directly reflect the implementation of cognition has been largely supplanted by a connectionist metaphor that brings together results from psychology on learning, perception, and memory *within* individuals (e.g., Hawkins & Blakeslee, 2004) and fields such as linguistics and anthropology on the shared patterns of meaning and interaction *between* people (e.g., Gee, 1992; Strauss & Quinn, 1998). Kintsch's (1998) construction-integration (CI) model for comprehension describes how interactions with situations activate myriad associations built up from past experiences, and understanding arises from those which cohere with one another and the situation of the moment: situation models, in Kintsch's terms; mental spaces, in Fauconnier and Turner's (2002). They are blends of particular circumstances and more

general patterns that are partly personal, due to our unique experiences, and partly shared with others, both because they tap cultural models and because they build up as extensions of universal human experiences such as putting things into containers and making objects move by bodily action (Lakoff & Johnson, 1999).

Model-Based Reasoning

Experts in semantically rich domains, such as the ones we study in school and work, organize their knowledge around fundamental principles and conceptual frames, for effective perception, understanding, and action (Ericsson, et al., 2006). Kintsch and Greeno (1985) showed how CI theory can be applied to model-based reasoning in science, both for experiential and immediate thinking about situations in terms of mental models and for the reflective thinking and public activity that characterizes research and applied work. The situated nature of cognition under the connectionist metaphor explains why a student who can solve physics problems on a test falls back on naïve explanations of formally identical situations on the playground: "learning" kinematics in isolated instructional settings does not lay down associations that are likely to be activated in the real world. Proficient model-based reasoning in science is thus more than just reasoning within the abstract spaces and symbol systems of models. It is ultimately about creating apt blends of models and real-world situations for reasoning about those situations within the metaphor, through the structures of the model (Stewart & Hafner, 1994): what to include in the model and what to leave out, and searching for patterns in the data that don't accord with the patterns the model can accommodate. Model-fit can't tell you if the metaphor you're reasoning through is the best way to think about a situation, but it can tell how well the patterns it can express match the patterns in your observations, and if they match so poorly that you need a different model, different data, or a different metaphor.

Although the terms "metaphor" and "model" are used in many ways, we can distinguish senses that are useful for discussing educational measurement. The trait, information-processing, and sociocognitive metaphors lay out basic concepts and relationships for thinking about the nature of knowledge, and how it is acquired and used. They are simplified, human-scale, narrative frames for thinking about an exceedingly

complex subject (Sfard, 1998), an example of what has been called an embodied reasoning view of cognition (Varela, Thompson, & Rosch, 1992). Mathematical/statistical models support reasoning within metaphors. The formal elements in models, such as variables and relationships among them, acquire meaning through the metaphors, then further as situated in particular circumstances (which corresponds roughly to Markus's notion of constructs defined without and within specific populations in mind). Models for within-person processing under the sociocognitive metaphor include neural network models for perception and Kintsch's (1998) latent semantic index (LSI) illustrations of CI theory. Production systems such as Newell's (1998) SOAR and cognitively diagnostic psychometric models (Leighton & Gierl, 2007; Nichols, Brennan, & Chipman, 1995) are examples under the information-processing metaphor. Classical test theory, factor analysis, generalizability theory, and item response theory are models that developed under the trait metaphor. Snow and Lohman suggest, however—and this is the point of the present commentary—that models which evolved under one metaphor can be gainfully employed, if suitably re-interpreted, under another.

In addition to models for aspects of knowledge and performance, educational measurement uses models for probability. The grounding metaphor for probability is our experience with tangible, replicable situations like Markus's card experiments (Shafer, 1976). The set theoretic basis of probability models tells us much about their properties as reasoning tools, and the better we understand them the more effectively we can reason within the model space. In contrast with the embodied-reasoning interpretation of models, a Fregian or correspondence interpretation of models maintains that the properties of the entities in the model are reflect, with the same crispness, independently existing properties of their real-world counterparts. In this view, as Markus describes, latent variables are determined as sets of individuals who share properties *as a presumed feature of the world*, rather than as a part of the frame that an analyst posits to reason about a situation in the world. The reification of traits that troubles Snow and Lohman is one pitfall of this interpretation.

Another warning signal in Markus's article is the difficulty of determining which population(s) an individual belongs to: "Given that the population level nomic relations

define the population, and that these need not manifest themselves in the individuals who compose the population, it remains unclear how one assigns individual cases to populations" (p. xx). This is a problem if the locus of variables and probabilities is taken to be the world. It is not a problem under the alternative subjective Bayesian interpretation of probability proposed by de Finetti (1974). The same set-theory axioms apply to ground reasoning in the model space, but now probability is about an analyst's degree of belief about an approximated real-world situation, through a model, in light of the analyst's state of knowledge. The reasoning that applies to random sampling from tangible populations is extended to situations in which an analyst holds individuals as *exchangeable* for a given inferential problem—a property of the analyst's use of the model to support reasoning about a situation in the world, as to what aspects of the situation to not model explicitly, rather than a property of the world itself. Thus an analyst can use different models to discern different aspects of the same situation, at different levels of analysis or for different purposes, and individuals will be exchangeable under some but not under others (Lindley & Novick, 1981). This embodied-cognition approach addresses the challenge Markus correctly notes, of "understanding how researchers might project the same concept into different modal spaces for different problems," but in a more satisfying way than refined definitions of concepts.

Some Implications for Educational Assessment

A sociocognitive perspective on knowledge holds profound implications for educational assessment. The first concerns what we want to know about students, i.e., the targets of inference, and what kinds of things we need to see them do in what kinds of situations—taken together, the elements of assessment arguments. From this perspective, the target of learning in the semantically rich domains of school and work is becoming attuned to the ways people see, talk, and do things; the ways they represent information, solve problems, and communicate with others; the ways they use the tools and principles of the domain productively and interactively in unique real-world situations—in short, the epistemic frame of the domain (Shaffer, 2007). Cognition, in this view, is not just something that happens inside individuals' heads, but a coordinated interplay of actions within and among people in a socially-structured space.

Questions about a student's capabilities that come to mind from this perspective are qualitative and contextual: What associations does a student activate in what conditions, to lead to effective perception, comprehension, actions, and interactions? Direct evidence for answering these kinds of questions requires observing the student acting in relevant situations. It is a commonplace of industrial/organizational psychology that the best predictors of job performance are samples of situations that are as much like the job as possible. Simulation-based (Baker, Dickieson, Wulfeck, & O'Neil, 2007) and game-based assessments (Gee, in press) are of increasing interest for the same reason. Making sense of examinees' performances in these kinds of assessments involves understanding the interlinked system of interfaces, background materials, task design, scoring rules, and statistical models, all tuned to the purposes of the assessment (Bennett & Bejar, 1998). These are contextual considerations the sociocognitive perspective makes us aware of in regard to claims about students and evidence to support those claims. They do not in and of themselves determine what models ought to be used in this pursuit, nor the meanings that formal elements of models should accrue.

Measurement Models in Educational Assessment

Model-based reasoning takes place in conceptual spaces that blend particular real-world situations and the more general structures that models afford, to support more precise understanding and reasoning. Models address only certain abstracted entities, relationships, and processes in a general form in order to support reasoning within that structure. They provide a frame, or narrative space, for thinking about aspects of a situation, from some particular perspective. The educational measurement tradition provides a particular narrative space to organize reasoning about students' capabilities and performances. At its heart is the measurement metaphor. Both Michell Borsboom (in press) and Snow and Cronbach (1989) question the scientific value of work carried out under its aegis, Michell from inside the metaphor and Snow and Lohman from without.

Are they really "measurement" models?

Michell characterizes psychometrics as a science in which the central hypothesis is that psychological attributes are quantitative. The formal requirements for quantitative measurement (Michell cites Hölder, 1901) extend the measurement metaphor from the tangible, replicable experience of physically comparing objects to a standard, to inferred quantities such as force. Luce and Tukey's (1964) theory of conjoint measurement furthers extends the metaphor with a model to simultaneous measures of two quantities inferable from regularities in the interactions between collections of entities such as responses of persons to tasks. But the typical practice in educational and psychological testing of adding up task scores and calling the totals "measures," to imply quantitative variables akin to length and force, does not verify the relationships among the elemental observations that would satisfy the axioms (specifically, cancellation conditions). In other words, most applications of "measurement" models in educational and psychology measurement reason through a blend of the trait metaphor and the measurement metaphor, without having checked the aptness of the measurement portion. The quantitative nature of scores is simply presumed in analyses of their relationships with other variables in order to evaluate policies or guide instruction using models such as regression, growth curves, and value-added analysis.

Accepting Michell's observation, one can respond in different ways that vary as to their view of the measurement metaphor. The most constrained response would confine psychometrics to the traditional quantitative framework for traits, restrict models to those of fundamental measurement which have no stochastic layer for observations, and limit application to situations with data that satisfy axiomatic requirements. The resulting package wouldn't contain much, but seems to be Michell's preference. Psychometricians such as Bond (2001) and Borsboom and Mellenbergh (2004) argue for a further extension of the measurement metaphor with latent variable models such at IRT that propose their own requirements for model fit at the level of individual observations. The family of Rasch models possesses particularly strong argument in its own right in terms of locally-objective quantitative comparisons of response probabilities (Rasch, 1977), as well as in terms of providing a probabilistic overlay for conjoint measurement. Statistical tests of model fit are readily available. But applied work exhibits a strong confirmation bias in the use of such models, and it is easy to find violations of model fit in educational

assessments with respect to students' demographic backgrounds, instructional histories, and solution strategies—phenomena that are troublesome under the measurement metaphor but wholly unexpected from a sociocognitive perspective.

A pragmatic alternative still within the traditional measurement metaphor is to view models such as IRT as narrative frames to organize thinking about masses of observations. For dichotomous responses, for example, capturing the patterns that some people tend to make more correct responses than others (metaphorically, they can exert greater force; Rasch, 1960/1980), some tasks tend to be harder than others (they offer greater resistance), and the tasks pretty much line up the same way for everyone. Under the Rasch model, the crisp modeled facsimile of the data gives probabilities that accord with the axioms of conjoint measurement; this is the basis of interpretation through the model, but the analyst realizes this is a property of the model and not necessarily of the real-world situation. The practical question is whether the model fit is adequate to support reasoning for the job at hand. Paradigms of data-gathering and modeling that work together well and consistently in different situations are indicative of pervasive patterns, the stuff of which science is made; a kind of correspondence with reality is evidenced, though not necessarily at the grainsize, the degree of simplicity, or the crispness that the model would suggest. This interpretation of latent variables in measurement models entails a stronger ontological stance than instrumentalism, but weaker than correspondence realism (see Borsboom, 2005, in press).

Model fit is a far less stringent requirement than establishing the quantitative nature of a variable in Michell's sense, but it does entail an oft-neglected responsibility for model fit. Subsequent analyses of the resulting scores that presume quantitative measures need to be scrutinized for their robustness to departures from this assumption. Some, such as inferences based on ordering, will be very robust, especially if an IRT model is found to fit fairly well. Others, such as analyses of gain scores and estimates in value-added models, will be more precarious.

Viewing measurement models from the sociocognitive perspective

Snow and Lohman's critique comes into play even if every care is taken within the measurement metaphor, however, in two ways. The first is the highlighted conclusion of

the quotation, that the tradition's interpretations do not suffice as scientific explanations of aptitude and achievement. Snow and Lohman could make this claim in 1989 as finer-grainer, within-person alternative metaphors were becoming available under the information-processing and sociocultural perspectives. I interpret them to mean that the acknowledged utility of the traditional measurement approach can, and should, be understood as emergent from processes that can now be studied with metaphors and models at lower levels, much as chemical phenomena can be understood in terms of quantum mechanics. This view is consistent with Borsboom's observation that it is possible to entertain distinct within- and between-persons latent variable models, and the distinct models can be consistent within the systems and situations for which they are posited. These observations do not obviate the usefulness of scores in traditional measurement models for work at the levels for which they evolved

It does, however, transform the meanings we impart to the elements of measurement models, and causes us to rethink their role in instruction and policy. Thus the second way that the measurement metaphor is too limiting is a pragmatic as well as scientific sense: issues of learning and performance that are central to the practical applications lie outside its scope as a narrative frame. Neither the genesis of performance nor the nature of the processes producing it are addressed in the metaphor or the attendant probability models of classical test theory or IRT (although cognitively diagnostic models, about which I will say more below, step in this direction). Therefore, instructional guidance and policy decisions based solely on the measurement metaphor, because they do not connect with research on the nature of learning and cognition, can be inefficient or counterproductive; the measurement models yield inferences, within the metaphor, that are at odds with what we are learning about the nature of cognition and performance. The measurement metaphor can help us see if desired results are occurring at the level at which it is meant to operate, but it can't tell us what to do to achieve them.

Construing Measurement Models through the Sociocognitive Metaphor

In chemistry, the quantum foundation both constrains the patterns that are possible to model at a coarser grainsize in terms of chemistry, and motivates productive lines of research in chemistry. A similar articulation is needed between the phenomena that can

be studied through the sociocognitive metaphor and the between-person patterns studied in educational measurement and used at larger scales. An understanding of the elements of measurement models from a sociocognitive metaphor will improve applied work carried out within the measurement metaphor.

Movements in this direction in the are proceeding along several fronts educational measurement community, and actually have been for some time in the guise of informal practice. I will mention some examples shortly. But because the current rapid progress in learning and cognition can be understood only from the sociocognitive perspective, because of growing demands on assessment such as the No Child Left Behind act, and because failing to do so can work against of larger educative goals, it is essential to recognize working out the articulation as an explicit and urgent priority.

One path of development has been in the way that measurement models have been used informally, namely in using traditional measurement models but in ways that conform to sociocognitive interpretations—simply because that is what was needed for successful applications of an assessment in its context and for its purpose. College Board's Advanced Placement Studio Art portfolio assessment (Mitchell, 1992) is a case in point. The program employs classical test theory to monitor broad patterns and call attention to outliers in a social system of tens of thousands of readers' evaluations. Scores are not interpreted as measures of artistic ability, but summaries of evaluations of particular work in accordance with criteria that reflect learning goals. The models are not about measurement per se, but about narrow-channel communication among raters, teachers, and students as part of a broader framework that concerns the qualities that are valued, how you recognize them when you see them, and how the things you do and the things you think when you create your works show them—in short, the epistemic frame of being an artist. Note that both the student variables and the observations in the model are interpreted through a sociocultural frame; in Borsboom's sense, all are latent variables, and all are jointly conceived as syntheses of real-world phenomena and the analyst's metaphor, model, and state of information.

AP Studio is by no means unique. Countless applications of educational measurement have used and interpreted assessment results in ways that are consistent

with sociocognitive ideas for pragmatic reasons, without having done so explicitly and therefore without the benefit of the more recent conceptualizations and research base. An assessment argument cast in sociocognitive terms might use the same probability model as a trait argument, but the situated meanings of the elements, the interpretation of observations, and nature of the target inferences would be different, with regard to (1) alternative interpretations of performance, (2) generalizations beyond the observational context, and (3) the role of information about the examinee/situation relationships.

More explicit development of these ideas is beginning to appear in learning domains. Redish (2003), for example, describes key ideas of a sociocognitive perspective to physics educators, and shows them how to leverage the ideas to improve teaching and assessment. Language testing at forefront of work along these lines, because language use is by nature social and interactive, and requires students to develop, through experience, the capabilities to carry out these actions through shared patterns of language and culture. The ascendant communicative emphasis in language testing was strongly influenced by the work of sociolinguist/anthropologist Dell Hymes (1972). Current practice is advanced with respect to targets of inference, task models, and evaluation procedures, but is struggling with the issues of interpretation and modeling. Chalhoub-Deville (2003) describes the central challenges of language testing as "amending the construct of individual ability to accommodate [how] language use in a communicative event reflects dynamic discourse, co-constructed among participants; and … reconciling [this local nature of language ability] with the need for assessments to yield scores to generalize across contextual boundaries" (p. 373). "The way forward," she says, "is to recognize that, while some contexts activate stable ability features, others produce more variable performance from learners" (p. 372).

From among the armamentarium of models developed under the trait metaphor, generalizability theory (g-theory; Cronbach, et al., 1972; Brennan, 2001) offers some machinery to serve these ends. Beyond "estimating scores" and "calculating reliabilities," g-theory has from the start been about understanding how peoples' performance varies under different conditions, what observations from students in one set of situations with such and such characteristics might tell you about what they are likely to do in others. The g-theory terminology of "universe scores" and "errors" from the trait

metaphor clashes with the sociocognitive metaphor. A re-interpretation of g-theory in terms of the percepts, concepts, and actions of the targeted learning models (e.g., from science, language, computer networks, studio art) can be envisaged, with student variables viewed as attunements; task situations modeled in terms of features as seen through the cultural models; and performances characterized in terms of appropriateness and effectiveness through the lens of the same targeted model. A theory of situations can be developed in terms of particular targeted models, to characterize both assessment situations and criterion situations. Within an extended network of variables, they can be used to condition inference about performance in the latter from observations in the former, incorporating ideas from structural equations models and Bayesian inference networks. In language testing, Bachman and Palmer (1996) develop the idea of creating performance tasks around key features of target language use situations; Messick, 1994, expands on these ideas in connection with validity in performance assessment, using the language of the trait metaphor. The definition of a trait we cited from Messick earlier still holds under a sociocognitive metaphor, but with an interpretation quite different than that of the strict measurement metaphor that Michell advocates or the Fregian definition in terms of the set of people who have that trait that Markus draws upon.

Another line of methodological developments in the psychometric literature is also pertinent, namely that of cognitive diagnosis (see, for example, the Winter 2007 special issue of the *Journal of Educational Measurement* devoted to this topic). Cognitive diagnosis models use Bayesian machinery to build models and draw inferences about students under a information-processing metaphor (Rupp & Mislevy, 2007). Student models concern production rules, task models concern relevant features of situations, and observable variables concern qualities of action. Especially given that situated versions of production rule systems approximate some targets of learning, cognitive diagnosis models provide additional tools for drawing inferences about students' capabilities from a sociocultural perspective. Models adapting features of generalizability theory, cognitive diagnosis, and standard measurement models would seem to be a suitable starting point for a psychometrics to support assessment under the sociocognitive metaphor. It bears emphasizing, however, that the issue should not be thought of primarily as one of building more complex models. Even classical test theory models, as models, can

support inference under the sociocognitive perspective with appropriately conceived tasks, evaluations, and interpretations of behavioral tendencies. More complex models can serve important roles, and we probably will use more complex models with more complex tasks, but it is the interpretive framework rather than the models per se that constitutes the needed shift.

Conclusion

This is a time of great change in the psychology of education, and, for the better or worse, it will be a time of great change in education. The better the changes in educational practices and policies accord with what we are learning about how people learn, think, and act, the better the outcome will be. Assessment is the link, because assessment shapes the way we think about our options for instruction and policy, and dominates how we gather information and make sense of it, to see how well we are doing and figure out how to do better. Many needs to gather information and make decisions will continue to reside at higher levels and coarser grainsizes than the methodologies associated with the sociocognitive metaphor address. We will still need to make decisions at the grainsize at which the traditional measurement metaphor evolved.

I suggest that many of the models and much of the formal machinery developed there will remain helpful, if properly reconceived. This means seeing them in a way which is itself motivated by the view of model-based reasoning emerging from sociocognitive research: Models as narrative frames, to guide perception, understanding, and action within a metaphor, using probability-based tools to support reasoning. It is clearly good thing to have a better understanding of the formal properties of measurement-models as tools, as in the articles of Borsboom, Markus, and Michell, and it is may also good that some researchers continue to wrestle with the problems that defined psychometrics a century ago. But this is not the direction in which psychometrics can most contribute to educational assessment. Rather, it is experience with modeling and probability tools to reason about students' capabilities from limited information, which can be applied to assessment arguments cast in any psychological perspective.

References

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Atkinson, D., Churchill, E., Nishino, T., and Okada, H. (2007). Alignment and interaction in a sociocognitive approach to second language acquisition. *Modern Language Journal, 91,*169-188.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, E. L., Dickieson, J., Wulfeck, W., & O'Neil, H. F. (Eds.) (2007) *Assessment of problem solving using simulations.* Mahwah, NJ: Erlbaum.

Bennett, R. E., & Bejar, I .I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bond, T.G. (2001) Book Review 'Measurement in psychology: A critical history of a methodological concept'. *Journal of Applied Measurement, 2, 1,* 96-100.

Borsboom, D. (in press). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives.*

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

Brennan, R. L. (2001). *Generalizability theory* (2nd ed.). New York: Springer.

Chalhoub-Deville, M. (2003) Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369-383.

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.

de Finetti, B. (1974). *Theory of probability (*Volume 1). London: Wiley.

Ericsson, K.A., Charness, N., Feltovich, P., Hoffman, R.R. (Eds) (2006) *Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press.

Fauconnier, G., & Turner, M. (2002). *The way we think.* New York: Basic Books.

Gee, J. P. (1992). *The social mind: Language, ideology, and social practice.* New York: Bergin & Garvey.

Gee, J. P. (in press). Game-like learning: An example of situated learning and implications for opportunity to learn. In P. Moss, D. Pullin, J. P. Gee, & E. Haertel (Eds.), *Assessment and opportunity to learn: New voices, new views*. Cambridge: Cambridge University Press.

Hawkins, J, & Blakeslee, S. (2004). *On intelligence.* New York: Times Books.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse, 53: 1-46.

Hymes, D.H. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguins Books.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W., & Greeno, J.G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109-129.

Lakoff, G., & Johnson, M. (1999) Philosophy in the flesh: The embodied mind and its challenge to western thought. New York: Basic Books.

Leighton, J.P. & Gierl, M. J. (Eds.) (2007). *Cognitive Diagnostic Assessment: Theories and Applications.* Cambridge: Cambridge University Press.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics, 9,* 45-58.

Linn, R.L. (Ed.) (1989), *Educational measurement* (3rd Ed.) New York: American Council on Education/Macmillan.

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Markus, K.A. (in press). Constructs, concepts and the worlds of possibility: Connecting the measurement, manipulation, and meaning of variables. *Measurement: Interdisciplinary Research and Perspectives*.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Michell, J. (in press). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives.*

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools*. New York: The Free Press.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Newell, A. (1990) *Unified theories of cognition.* Cambridge: Harvard University Press

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press (reprint).

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Redish, E. F. (2003) *Teaching Physics with the Physics Suite*, John Wiley & Sons, Inc.

Rupp, A.A., & Mislevy, R.J. (2007). Cognitive foundations of structured item response models. In J.P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment: Theories and Applications.* Cambridge: Cambridge University Press.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher* 27, 4–13.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: Princeton University Press.

Shaffer, D.W. (2007). *How computer games help children learn.* New York: Palgrave Macmillan.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 263-331). New York: American Council on Education/Macmillan.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp 284-300). New York: Macmillan.

Strauss, C., & Quinn, N. (1998). *A cognitive theory of cultural meaning*. New York: Cambridge University Press.

Varela, F. J., Thompson, E.T., & Rosch, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press.