MEASUREMENT, *2*(1), 33–54

COMMENTARIES

The Future of Ability Testing: What Happened to the Science of Mental Ability?

Paul Barrett
Department of Psychology
University of Auckland, New Zealand

In an article dedicated to "ability testing," it is strange to find no technical definition of what exactly is said to constitute "an ability." Further, the word *intelligence* is mentioned in passing within the first section of the article entitled "Precursors" and appears in various other places interspersed with the term *abilities.* Yet, the author ignores dealing with either terminology. Instead, the reader is assumed to know the definition of *ability* to the degree that measures of it can be discussed and projected into the future. Jensen (1998) dispensed with the term intelligence altogether on the basis of it being indefinable or at best, arbitrarily defined. In fact, it is what might be called, using Maraun's (1998) terminology, a "common or garden psychological concept." But, in discarding the concept of "intelligence" in this way, it now requires that "ability" is defined as a technical construct in order that measurement might be made of this "ability," and so that we can subsequently recognize that such measurement is yielding valid estimates of magnitude of an "ability." For example, is the observed behavior that characterizes the solving of matrix-type problems the result of a single "ability" to do so, or the result of multiple abilities being applied to a problem solution to achieve a final unitary item response? This question is important when it comes to constructing measurement.

Requests for reprints should be sent to Paul Barrett, Department of Psychology, University of Auckland, Human Sciences Building, Symonds Street, Auckland, New Zealand. E-mail: paul.barrett@auckland.ac.nz

That is, what is to be measured, the outcome behavior or the constituent "abilities" that will be modeled as causal and predictive of that outcome behavior?

The author makes a strong case that the second century of ability testing will be built largely on the groundwork of probabilistic Item Response Theory (IRT), computer technology, and item-generation algorithms, not on any kind of reconceptualization of what actually might constitute an ability, or intelligence. I think this is far too restricted a view of the impact that fundamental scientific investigation of "mental abilities" might have on ability testing in the future. Consider the impact of evolutionary psychological theory on the nature of what might be construed as a mental ability. The work by Gigerenzer and Goldstein (1996) and Gigerenzer and Todd (1999) on "fast and frugal" algorithms and the modularity hypothesis of Tooby and Cosmides (2000), even allowing for the counterarguments of Fodor (2000) and Karmiloff–Smith (2000), should perhaps cause the most ardent test technologist some disquiet about the very nature of testing for "an ability." Further, the whole approach to computational intelligence (Poole, Mackworth, & Goebel, 1998) seems to offer completely novel ways of construing and modeling a "mental ability" and "intelligence."

QUANTITATIVE MEASUREMENT

Following Michell's (1997, 1999) unambiguous definition of quantitative measurement, and his recent article (Michell, 2001) on the failure of current psychometrics to consider precisely what is to be meant by the measurement of a magnitude, I find it puzzling that Embretson seems to consider the impact of such work as irrelevant to the future of ability testing in the next century. I have written about the consequences of such neglect elsewhere (Barrett, in press), but, for the purposes of this commentary, let me quote from Michell:

> Measurement, as a scientific method, is a way of finding out (more or less reliably) what level of an attribute is possessed by the object or objects under investigation. However, because measurement is the assessment of the magnitude of a level of an attribute via its numerical relation (ratio) to another level of the same attribute (the unit selected), and because only quantitative attributes sustain ratios of this sort, measurement applies only to quantitative variables. Psychometrics concerns the measurement of psychological attributes using the range of procedures collectively known as psychological tests. As a precondition of psychometric measurement, these attributes must be quantitative. (2001, p. 212)

There is a telling quote from Embretson (this issue) in the section of the article entitled "Measurement of Qualitative Aspects of Individual Differences:" "In the first century of ability testing, a single aspect of ability was measured: namely, its

level. However, it was often acknowledged that examinees also differ qualitatively so that the meaning of their ability scores differs … ."

Embretson then proceeds to outline how such qualitative variants in item score patterns might be used to augment interpretation of test scores. This is no longer the quantitative measurement of a single "ability." The process of characterizing item-response behavior with a series of operationally defined "latent variables" followed by post hoc explanations of "discovered" latent classes among individuals is an excellent approach to accounting for observed phenomena with a few broad descriptive latent variables. However, it is not an excellent approach for the better understanding and quantitative measurement of what may be universal properties of a human cognitive system that are subsequently causal for the manifest observations of these "mental abilities."

PSYCHOMETRIC TEST THEORY AS A FORM OF "SCIENTIFIC RETARDANT?"

I read this article with perhaps a different viewpoint from many who work in educational assessment. That is, I purposely took the term ability to refer to a broad psychological construct that is definitional for a range of complex observed phenomena associated with a human cognitive system. With this view, and its greater reliance on the relevance of constitutive theory of what might be referred to as a human "ability," and the very nature of what is to be said to constitute "measurement," the highly focused technological and educational measurement related predictions of the author seem rather limited in scope.

I am also reminded of Blinkhorn's (1997) review of 50 years of preceding work on test theory, and his conclusion:

> 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. (p. 175)

My own view is that the article reflects exactly that insular preoccupation with ever more sophisticated statistical modeling of "educationally-relevant" data. The recent article by Fan (1998) demonstrated that there are no significant differences between item and person statistics and invariance properties using sophisticated IRT or very simple Classical Test Theory scoring with large samples of respondents. For me, this demonstrates all too clearly that an overriding preoccupation with ever more "advanced" test theory might paradoxically serve to retard substantive progress in assessment in the future, rather than accelerate it.

REFERENCES

Barrett, P. T. (2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology, 3,* 421–439.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357–381.

Fodor, J. (2000). *The mind doesn't work that way.* Cambridge, MA: MIT Press.

Gigerenzer , G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

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

Karmiloff–Smith, A. (2000). Why babies' brains are not like Swiss army knives. In H. Rose & S. Rose (Eds.), *Alas, poor Darwin: Arguments against evolutionary psychology* (pp. 144–156). London: Jonathan Cape.

Jensen, A. R. (1998). *The 'g' factor: The science of mental ability.* Westport, CT: Praeger.

Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8,* 435–461.

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: Critical history of a methodological concept.* Cambridge, England: Cambridge University Press.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36,* 211–217.

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational intelligence: A logical approach.* Oxford, England: Oxford University Press.

Tooby, J., & Cosmides, L. (2000). Toward mapping the evolved functional organization of mind and brain. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 1167–1178). Cambridge, MA: MIT Press.

Technology and Theory as Drivers for Deeper-Than-Predicted Changes

C. Victor Bunderson
EduMetrics Institute

Embretson's predictions of the future of ability testing include 1 general and 11 specific predictions. In this commentary, I endorse the 11, split off another category, and seek to understand how these are being driven by technology and theory changes. The general prediction, that the second century will mirror the last one in rapid pace of change for a few decades, slowing down during the last two thirds, seems poorly motivated. Other analyses that probe historical, philosophical, technological, and related developments in a different light are called for. These should include asking why so little seems to have changed in the face of opportunities to do so.

Seen as an insular activity of a specialized in-group of psychologists and educators, progress in ability testing may have slowed down after 1930. But seen in a broader social context, this slow-down is anomalous. Advances in information and communication technology (ICT) are propelling rapid changes in every field, including automated test and item development, and the development and testing of new tools and methods (Irvine & Kyllonen, 2002). These are important and promising, but do not depend entirely on artificial intelligence (AI). Other industrial strength techniques not requiring AI will also be used. As for the predicted great broadening of tasks used in measuring abilities, ICT, with emphasis on multimedia display and diverse response entry options, is the main driver. The proliferation of richer tasks has already begun and will accelerate in the near future. Fast computation is opening new doors for presenting adaptive, simulative performance situations and using complex scoring algorithms. But ICT broadens the base of users and methods of production and delivery well beyond the psychology-related disciplines. This may create a continuing driving force to overturn the prediction that the second century will mimic the first in tapering off in innovation after 30 years.

Why has so little change occurred since 1930? Why has Item Response Theory taken hold so slowly? Why haven't some of the innovations possible for some time with technology taken root sooner? As asides, Embretson gives two possible reasons:

Requests for reprints should be sent to Victor Bunderson, 560 S. State, Orem, UT 84058. E-mail: cvicb@attbi.com

1. "The testing industry became large and lucrative."
2. Revisions required 18 months to 2 years, and often effort over 5 years.

Her three predictions about distributed and automated test development procedures, if valid, will change the revision time and cost dramatically. Then why should the last two thirds of the 21st century repeat the slow progress in the 20th? Consider also that the "large and lucrative" testing industry is also changing rapidly. Attend one of the Association of Test Publishers conferences soon. Observe the vista of a whole set of new players and new applications of testing in industry and education. To stay preeminent in this industry, the large legacy testing companies will have to be nimble and creative. Whoever succeeds, the stability and lack of incentive for change that marked the second third of the last century has vanished forever. Here is an alternate prediction: Competitive forces and new players with different professional backgrounds will work against inertia, math avoidance, and other possible forces retarding change, and instead, will accelerate change.

Advances in theory is another force behind several of the 11 predictions. It is introduced in the section on item development by cognitive design principles. Embretson and a growing cadre of others are building items from theoretical propositions about invisible thinking processes—unthinkable by earlier ability theorists using earlier science philosophies. Fundamentally, measurement in every science depends on theory, and theory and new measurement instruments accelerate one another.

Although endorsing the 11 predictions, I would stress the importance of the work of Mislevy and his colleagues (Mislevy, Steinberg, & Almond, 2002, 2003; Almond, Steinberg, & Mislevy, 2002) on Evidence Centered Design (ECD), beyond its utility in providing "flexible mixtures of evidence for ability" (Embretson, this issue, p. 15). Mislevy has tried to dig down to foundations, then build up from those foundations, and in the process, reinvent assessment. As this happens, we see true reinvention, not just continuous elaboration from foundations laid early in the last century. The addition of the term *design* into the ECD concept is of great significance. As principle-based designers, we are using cognitive and sociocultural theories to design and construct both better theories freed from logical–empiricist strictures, and new technologies based on them. ECD offers a substantial, disciplined design process that includes domain analysis, domain modeling, and a four-part conceptual assessment framework. It offers an assessment delivery system that can be conceived of as having four main processes. ECD thus offers a breakdown of assessment design, development, and delivery into levels, and within each level, a language is offered so that communication and later specialization can be achieved. In any engineering and construction-like discipline (including instruction and assessment—both are design disciplines, not objective sciences), layers and languages must develop to create a mature industry (Gibbons, 2003). It would be helpful in the creation of this industry for many in the measurement community to adopt Mislevy

and his colleagues' copiously documented concepts and languages when discussing the design of assessments. The end result may or may not be called ECD, but it will help create collaboratively this new technology and science-based industry.

Embretson made three important predictions in her final set. I heartily endorse these predictions and her list of requirements for *objective* dynamic assessment. However, has she gone far enough in her predictions? Will domain-based interpretations produce larger changes in how abilities are understood, measured, and used than she implies? Is modifiability of abilities so great that the opportunity of improving abilities will have a larger market and greater impact on education and society than the measurement of abilities for selection, placement, and the like has ever had? She predicts that dynamic assessment will increasingly become the mainstay in measuring ability, but implies that ability *testing* will continue to be the main concern. She alludes to the possibility that Bennett (1998) is right, and that testing will indeed be reinvented when instruction and testing merge fully, but confines her predictions to ways of measuring ability. Both Bennett and earlier writers (Bunderson, Inouye, & Olsen, 1989) predict a third generation of computerized testing, where measurement is continuous, and seamlessly integrated with instruction. In Bennett's words: "… nothing short of reinvention will prepare it (assessment) to meet the dramatically different demands it will soon face" (p. iii).

These differing predictions mix the aptitude and achievement distinction. The third-generation predictions feature achievement–learning progress, with abilities in a supporting role. A goal that honors objective dynamic assessment, but includes and transcends it, is to combine the assessment both of achievement and of relevant abilities into live learning settings using continuous measurement technology. Theory is developing to provide the basis for the continuous measurement of learning progress toward high achievement. Theory is developing for aptitude as well. Space does not permit the review of many important aptitude theorists, but it is important to mention the posthumous book edited by Lee Cronbach, *Remaking the Concept of Aptitude* (Corno et al., 2002). This volume brings some closure to Richard Snow's many decades of research on aptitudes and their role in learning. This work makes it clear that we must go beyond the cognitive processing paradigm in our theories of aptitude in instruction. We must give an account of affect and conation as well, and include an account of situations and their affordances. In Snow's view, "Abilities are reflected in the person's tuning to the demands and opportunities of each situation, and thus reside in the union of person in situation, not in the mind alone" (p 163).

The prediction "that domain-referenced interpretations of ability will become prevalent" combined with the prediction about dynamic testing, could be expanded into a fourth category. This category would involve the seamless integration of learning with measurement for continuing growth of both abilities and achievement. The domain-referenced prediction paves the way for building interpretive frameworks to track learning progress in achievement as well as aptitude domains. It includes the important idea that common-scale measurement should be obtained conjointly not

only for items and persons, but also for cognitive processes. By so constructing interpretive frameworks of learning paths or scales, this conjoint theory gives an account of both task difficulty and processing complexity. Although past research has used the term *item domain,* that idea is too limited. What is needed is a theory of learning and growth in an achievement domain and a related theory of abilities that cut across achievement domains. To work together with instruction, relevant abilities may be selected based on the situations and affordances found in the achievement domain. Messick (1995) introduced the term *domain theory* as defining the boundaries of a domain for measurement purposes: "A major goal of domain theory is to understand the construct-relevant sources of task difficulty" (p. 745). His ideas can be applied to ability domains that cut across achievement domains as well. Elsewhere, he talks of the related substantive processes that provide interpretation of the constructs. I have cited the work of Snow and his successors to include affective and conative, as well as cognitive processes of substance. A group of my colleagues are continuing to develop theories of learning progress in domains, or domain theories, and how to use design processes to create them (Bunderson, 2002). This concept is highly compatible with the three final predictions of Embretson. However, it leads to more aggressive predictions about both adapting to and learning to improve abilities continuously, not merely measuring them occasionally.

REFERENCES

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retreived from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Bennett, R.E. (1998) *Reinventing Assessment.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Bunderson, C. V. (2002, April). *How to build a domain theory: On the validity-centered design of construct-linked scales of learning and growth.* Paper presented at the Institute of Objective Measurement Workshops, New Orleans, LA.

Bunderson, C. V., Inouye, D. I., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367–407). New York: ACE/Macmillan.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., et al. (2002). *Remaking the concept of aptitude.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gibbons, A. S. (2003). What and how do designers design? A theory of design layers. *Tech Trends, 47*(5), 2–27.

Irvine, S. H., Kyllonen, P. (2002), *Item Generation for Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741–749.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2002). Design and analysis in task-based language assessment. *Language Assessment, 19,* 477–496.

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

History, the Future, and Developments in Ability Testing

William P. Fisher, Jr.
MetaMetrics, Inc.
Durham, North Carolina

Embretson presents an even-handed, systematic, and comprehensive description of English-language developments in ability testing over the last century, and projections for further developments in the coming years. Two observations in particular emerge in the reading of the article.

First, as Michell (1999) pointed out, similar previous historical reviews of ability testing have failed to acknowledge the relevant contributions of fundamental measurement theory and practice. Embretson's review continues in this tradition, despite the fact that Rasch's probabilistic conjoint measurement models have been situated in the context of fundamental measurement theory for some years, as also is pointed out by Michell.

Second, work in the role of historical narrative in identity development (Venema, 2000) provides a framework for selecting and organizing related events en route to recording a synthesis of the past that projects a trajectory of achievable, significant developments for the future. Unifying past and future in a coherent narrative is vital to the development of a sense of identity both for individuals and for communities. The field of ability testing is a community of inquirers that, like many similar communities in the human sciences, is in particular need of revitalizing the tradition of shared narratives that tell its story.

This commentary aims only to sketch an alternative approach to thinking about the past and future of ability testing to provide a point of contrast with Embretson's presentation. Of particular interest is the possibility of a convergence between the kind of story that makes for a good history and the kind of story that makes for good measurement.

Requests for reprints should be sent to William P. Fisher, Jr., MetaMetrics, Inc., 2327 Englert Drive, Suite 300, Durham, NC 27713. E-mail: wfisher@lexile.com

Wright (1997, p. 34) offers a promising point of entry into the construal of measurement as storytelling in the context of fundamental measurement theory, saying that the "purpose of inference is to estimate [from historical data] what future data might look like before we encounter them." The point of mathematical modeling as far as measurement is concerned is less one of describing every kind of variation in a set of historical observations that will never occur again in exactly the same way, than it is one of prescribing the structural consistency that must be obtained for inferential stability.

Historians of science recognize that, to prescribe the data structures associated with particular variables, researchers require advance knowledge of what numbers will be inscribed as measures in association with what observations. As Kuhn (1977, p. 219; original emphasis) put it: "*The road from scientific law to scientific measurement can rarely be traveled in the reverse direction.* To discover quantitative regularity one must normally know what regularity one is seeking and one's instruments must be designed accordingly." To design instruments according to a known quantitative regularity, researchers must know how to make stable inferences from past experience. Fundamental measurement models project the structure of scientific laws and so enable researchers to determine whether abilities and attitudes exhibit stable regularities. Given that the hypothesis of a kind of regularity is not falsified, it must then be possible to calibrate all tools measuring the same variable so they express amounts in a common quantitative language interpreted in a common framework. Designing instruments in accord with known quantitative regularities is the means through which "the true union of mathematics and measurement" (Roche, 1998, p. 145) is achieved.

In the context of the full union of mathematics and measurement, Kuhn (1977, p. 221) ventured "the following paradox: The full and intimate quantification of any science is a consummation devoutly to be wished. Nevertheless, it is not a consummation that can effectively be sought by measuring." The basic point, then, is that the human sciences lack fully quantified variables because, rather than calibrate the coordinated and distributed sign systems constituting mathematical languages, researchers prefer to focus on assigning numbers to observations, mistaking that activity for the means by which generalized quantification will be achieved. Kuhn (1977, p. 221) further suggested that "maturity comes most surely to those who know how to wait," implying that, in the manner of the emergence of professional metrology in the 19th century, full quantification may have to happen by itself in its own time, if it will happen at all.

This would seem to be the expectation not only of Embretson in her sense of the past and future of ability testing, but of the mainstream of researchers and practitioners in this field. Virtually none of the ability testing literature mentions or sets goals relative to the invariance of constructs, reference standard metrics, metrological traceability, or the work in fundamental measurement theory and practice that has been underway over the last 75 years (since Thurstone's classic

work in the 1920s), laying the foundations for the full union of mathematics and measurement in the human sciences. Be that as it may, the current plethora of scale- and sample-dependent ordinal numbers produced by ability tests nonetheless can be called measures only if we are willing to devalue completely the meaning of the word. The truth of this assertion becomes evident as soon as it is recognized just how very effective most alleged ability measures are in preventing the practical quantification of growth or change, as would be achieved should researchers succeed in coherently unifying past performances and future possibilities into a single narrative for each ability measured.

That is, if real measurement were in hand, changes in ability would be routinely expressed in universally available uniform metrics. Teachers could have research-supported expectations for performance as a basis for individualized instruction. Schools could be held accountable to standards that are meaningfully integrated with the curriculum. Parents and students could interpret feedback on performance in the same metric across the entire lifespan. Employers could connect school outcomes with workplace ability requirements. Researchers could have firm criteria for recognizing generalizable accumulations of new learning about effective ways of supporting and promoting the natural course of ability development. And, finally, purchasers of educational services could know how much change in ability to expect per dollar invested, a development that, if broadly realized across the human sciences, could transform the economies of human and social capital.

Our ability to tell not just true but meaningful stories about ourselves is limited only by our willingness to follow through on our scientific culture's mathematical metaphysics to a full integration of qualitative richness and quantitative rigor in measurement (Fisher, 2003a, 2003b). Thus, Wright, in language similar to Ricoeur's in its emphasis on prescribed value over described fact, opens the door to a path that leads toward mature quantification and opportunities for telling the story of development as it occurs in the professions, and in the lives of professionals and students (Fisher, 2003c). Some of the developments in ability testing reviewed or projected by Embretson are playing important roles in effecting the transition from the statistical modeling of whatever historically factual observations happen to be recorded to the measurement modeling of inferentially valuable data structures. Unfortunately, it is difficult to tell from the microscopically detailed text which developments are facilitating, and which are hindering, that process. Perhaps, as the structures of historical narrative, the necessity and sufficiency of fundamental measurement theory for quantification, and the metaphysics of meaning and its place in method and measurement are more fully elaborated, historical accounts of ability testing will be written that prioritize the capacity to share more effectively substantive ability narratives over time, space, and individuals.

Kuhn (1977) held that the full mathematization of the natural sciences that occurred around 1840 contributed to the emergence of what he called a second scien-

tific revolution. The full mathematization of the human sciences may well provoke a similar revolutionary burst of new insights and productivity. Whether it will, we can discover only by trying. But we can try only if we recount the story of our past in a way that opens the doors of the future.

REFERENCES

Fisher, W. P., Jr. (2003a). Consequences of mathematical metaphysics for metrology in the human sciences. In A. Morales (Ed.), *Renascent pragmatism: Studies in law and social science* (pp. 118–153). Brookfield, VT: Ashgate Publishing Co.

Fisher, W. P., Jr. (2003b). Mathematics, measurement, metaphor, and metaphysics: Part I. Implications for method in postmodern science. *Theory & Psychology, 13,* 753–790.

Fisher, W. P., Jr. (2003c, April). Provoking professional identity development: The postmodern legacy of Benjamin Drake Wright. Paper presented at the conference, "A Celebration of the Career and Contributions of Benjamin D. Wright," held at the Rehabilitation Institute of Chicago.

Kuhn, T. S. (Ed.). (1977). *The function of measurement in modern physical science.* Chicago: University of Chicago Press.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge, England: Cambridge University Press.

Roche, J. (1998). *The mathematics of measurement: A critical history.* London: The Athlone Press.

Venema, H. I. (2000). *Identifying selfhood: Imagination, narrative, and hermeneutics in the thought of Paul Ricoeur.* Albany: State University of New York Press.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45, 52.

Automatized Reading Comprehension Ability Assessment: How Much or What?

Richard M. Golden
Behavioral and Brain Sciences
University of Texas at Dallas

Susan R. Goldman
Psychology Department
University of Illinois at Chicago

This commentary considers Embretson's contention that "domain-referenced interpretations of ability require both a psychometric and cognitive foundation" from the perspective of "complex" reading comprehension assessment. Research in both experimental cognitive psychology (see Kintsch, 1998, for a review) and educational practice (see Gunning, 1998, for a review) indicates the importance of revealing a reader's mental model of the text.

In complex comprehension tasks, the reader's mental model is likely to include inferences drawn from the text that are not necessarily implied with high certainty by the text. Complex comprehension focuses on thinking and reasoning that integrates text information with information in other texts and with prior knowledge. The integration process involves reasoning with and about the presented information and is the means by which readers construct interpretations or mental models of what the text is about (Goldman & Varma, 1995; Goldman, Varma, & Coté, 1996; Kintsch, 1998; Johnson–Laird, 1983).

CLASSICAL PSYCHOMETRIC METHODOLOGIES FOR READING COMPREHENSION ASSESSMENT

A strength of standardized reading comprehension assessment tests is that they provide reliable psychometric assessments of reading achievement. They typically

Requests for reprints should be sent to Richard M. Golden, Behavioral and Brain Sciences, GR4.1, University of Texas at Dallas, Box 830688, Richardson, TX 75080. E-mail: golden@utdallas.edu

consist of a series of short reading passages followed by a series of multiple-choice questions and can be administered and graded automatically using computerized adaptive testing methodology (e.g., Lord, 1980; see Wainer, 2000, for a recent review).

A difficulty with classical standardized assessment tests, however, is that such constrained response methodologies are effective only for evaluating inferences that are implied with high likelihood from the text. That is, they are more useful for the assessment of "basic comprehension" as opposed to complex comprehension. On the other hand, free response data can be especially revealing of the cognitive structures underlying a reader's mental model and thus seems especially likely to play an increasing role in complex comprehension assessment methodologies.

Unfortunately, as noted later, few existing psychometric-based measurement technologies are capable of automatically processing free response essay data while maintaining reasonably high construct validity from a cognitive measurement perspective. Note that the term *psychometric-based* assessment is used here to refer to a theory that guarantees uniqueness of parameter estimates, supports hypothesis testing, test equating, bias detection, and so on.

AUTOMATED READING COMPREHENSION ASSESSMENT USING FREE RESPONSE DATA

As reviewed by Embretson in the target article, current state-of-the-art methods for automatized essay grading tools are based on analyzing surface structure properties of the student's response. Information such as the relative frequency of specific syntactical constructions, key word searches, and "overall semantic similarity" between a student-produced essay and an ideal "correct" essay using techniques such as Latent Semantic Analysis (LSA), syntax, vocabulary, and word meaning are exploited (Burstein & Chodorow, 1999; Foltz, Kintsch, & Landauer, 1998; Landauer, Foltz, & Laham, 1998; Rehder et al., 1998; Wolfe et al., 1998).

However, such syntax and LSA-based essay grading methods have several inherent problems as measurement tools. First, although such methods have been shown to be reasonably correlated with the ratings assigned by human scorers, these surface structure methods are at best highly indirect methods of assessing comprehension because they are based on word co-occurence and thus have poor construct validity (Bennett & Bejar, 1998). Second, as mentioned in Embretson's target article, these surface structure methods often do not employ psychometric methods such as those described in the target article for the purposes of statistically controlling assessment reliability across families of testing materials and identifying latent classes of examinees in a principled manner. And third, E-rater and other LSA-type auomated graders are "task-scoring" methodologies whose parameters are task-bound.

For the purposes of complex comprehension assessment, methodologies whose parameters are bound more directly to a theoretical model of the reader's understanding of the text are more appealing. In particular, methodologies such as Tatsuoka's (1985) rule theory, Embretson's (1991) multidimensional cognitive Item Response Theory, and the more recent probabilistic graphical modeling methods (Mislevy et al., in press) mentioned by Embretson as well as recent work by Golden, Goldman, Oney, Thomas, & Macleod (2003; see also Durbin, Earwood, & Golden, 2000; Golden, 1998), appear to offer promising alternative tools for modeling and identifying the products of more complex comprehension processes.

AUTOMATIZED COMPLEX READING COMPREHENSION ASSESSMENT: THE NEXT TWO DECADES

Within the next two decades, we envision automated diagnostic tools will be developed for analyzing student free response data (e.g., "essay" data written in response to specific probe questions, recall data, summary data) for the purpose of identifying a psychometric model of the reader's mental representation of the text. Such a tool, used for instructional purposes, would be consistent with Embretson's description of the "third generation" of computerized testing where testing reinvents itself and merges with instruction. Finally, we expect that in the course of the next two decades, the development of such a measurement technology will be essential for the future growth of educational and cognitive measurement technologies.

REFERENCES

Bennett, R. H., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 7*(4), 9–17.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonative English speakers. *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, USA.*

Durbin, M. A., Earwood, J., & Golden, R. M. (2000). Hidden Markov models for coding story recall data. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the 22nd Annual Cognitive Science Society Conference* (pp. 113–118). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 64,* 407–433.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25,* 205–308.

Golden, R. M. (1998). Knowledge digraph contribution analysis of protocol data. *Discourse Processes, 25,* 179–210.

Golden, R. M., Goldman, S. R., Oney, B., Thomas, P. R., & Macleod, S. (2003, June). *Modeling text understanding: Applications for diagnostic assessment of reading.* Talk presented at the meeting of the International Hanse Conference for "Higher Level Language Processes in the Brain: Inference and Comprehension Processes," Delmenhorst, Germany.

Goldman, S., & Varma, S. (1995). Capping the construction-integration model of discourse comprehension. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 337–358). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Goldman, S., Varma, S., & Coté, N. (1996). Extending capacity-constrained construction integration: Towards "smarter" and flexible models of text comprehension. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 73–113).

Gunning, T. G. (1998). *Assessing and correcting reading and writing difficulties.* Boston: Allyn & Bacon.

Johnson–Laird, P. (1983). *Mental models.* Cambridge, MA: MIT Press.

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

Landauer, T., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259–284.

Lord, F. M. (1980). *Applications of item response to practical testing problems.* Hillsdale: NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., et al. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Tech. Rep. No. 580). Los Angeles: Center for Studies in Evaluation, University of California Los Angeles.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25,* 337–355.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12,* 55–73.

Wainer, H. (2000). *Computerized adaptive testing: A primer.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., et al. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25,* 309–336.

Extending the Conversation: Comments on Embretson's "The Second Century of Ability Testing: Some Predictions and Speculations"

Robert J. Mislevy
School of Education
University of Maryland

Professor Embretson's article, "The Second Century of Ability Testing: Some Predictions and Speculations," does a fine job of reviewing key developments in ability testing over the past century, and extrapolating from them to the next century. My comments consist of questions that came to me as I considered whether this extrapolation is sufficient. They are questions that might not have been posed at the beginning of the 20th century, maybe could not have been. At the beginning of the 21st century, we must ask them. I believe the advances Embretson describes as having been developed to address the assessment questions of the past century will be instrumental to addressing those of the next, even if their surface forms or the purposes they serve are unfamiliar (Mislevy, 1994, 1996; Mislevy, Steinberg, & Almond, 2003).

Embretson's Abstract poses an interesting premise: that the key ideas from psychometrics and testing were in place by 1930, and most of the progress of the remainder of the century was in applying them. Whether the future will be further extensions along the same lines remains to be seen. But isn't psychology vastly different than it was in 1930? Can a view of assessment that arose from the psychology of nearly a century ago, which circumscribed not only questions but potential answers to them, be up to the challenges of a more contemporary understanding of the ways people acquire and use knowledge? Yes, the "Stanford–Binet IV. … remains remarkably similar to the early test." Is this good? Is the continued interpretation of test results from a trait perspective, often through unidimensional models, useful, even if one uses results from cognitive psychology to construct the tasks?

Requests for reprints should be sent to Robert J. Mislevy, EDMS, Benjamin 1230–C, University of Maryland, College Park, MD 20742. E-mail: mislevy@umd.edu

How far can the introduction of qualitative differences contribute, before we no longer have simple measures for comparing examinees?

The focus of the article is "ability testing," in contrast to "achievement testing." One of the main areas of progress in cognitive psychology since the 1970s has concerned the nature and acquisition of knowledge. Does the growing understanding of achievement (for example, the importance of knowledge structures that must be learned, and through which reasoning proceeds, and the social context of learning) leave ability testing behind? Tests that are integrated with instruction move assessment inside the content and social context, rather than attempting to remain outside it. To what extent are the advances noted here then applicable to achievement testing? If ability testing is measuring capabilities that are relatively robust across particular knowledge and social contexts, will it be diminishing in importance when we turn our attention to knowledge and skill that we wish to change as much as we can?

A related question on this issue is that of the invariance of Item Response Theory (IRT) estimates, mentioned in connection with prospects for continuous test renewal. Invariance is a property of estimates with respect to collections of items and people—IF the model is correct. This is not to be presumed as a state of nature, but rather a falsifiable proposition. Results from cognitive and sociocultural psychology give us insights into the limits of such a practically useful working assumption, as well as pointing us to potential violations that should be continually monitored (as Embretson notes in the text). Both the utility and the limitations of the principled approach to test design are reflected in experience with assessing document literacy. Mosenthal and Kirsch (1991) have provided an information-processing based model for the ways people find and use information in documents. They account for a great deal of the variance in item difficulties with a model using features of documents (such as their structure, organizing categories, and so on), the demands of the directive, or task to be carried out (e.g., locating versus integrating, number of features that need to be matched), and correspondence between the document and the directive (e.g., occurrence of near-matches to needed information). But variation from one person to the next depends on such factors as context of use and familiarity with the document form. (I once found I could run an IRT computer program with documentation in German—not because I knew the language of Germany, but because I knew the languages of IRT and FORTRAN.) The Mosenthal and Kirsch frameworks have proven useful for constructing and analyzing literacy tasks in large-scale surveys. But don't instructional approaches such as working with documents with content familiar to students presume a lack of item parameter invariance, to exploit it to maximize learning in strategy more in line with the work of Vygotsky or Piaget than that of Spearman or Gulliksen? And might not teaching students about particular classes of documents and strategies for working with them decrease rather than increase invariance, in an outcome that nevertheless pleases us?

Measurement at the beginning of the 20th century had a fair amount of "one size fits all" character: the same observational settings for everyone; the same work products required; the same evaluation procedures, mapping work into common observable variables; the same method of synthesizing this information into variables that characterized examinees; and the same interpretation of the results in terms of constructs. A methodology arose that provided a sound conceptual basis for measurement under these circumstances. Many of the advances Embretson describes have been relaxing various of these constraints without abandoning the underlying logic. These are advances in a dimension that is orthogonal to the technology of testing and the efficiency of administration. A relatively new front on which we may anticipate further extensions, as required in the integration of assessment and instruction, will allow for different forms of data or tailored evaluation rules in light of students' current knowledge and contextual factors.

REFERENCES

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14,*147–180.

What Skills Should Be Measured in the Second Century of Ability Testing?

Robert J. Sternberg
PACE Center
Yale University

Embretson's tour de force focuses on developments in the means of testing that are likely to arise during the second century of ability testing. I focus here on the skills that should be tested during this century.

Requests for reprints should be sent to Robert J. Sternberg, PACE Center, Yale University, Box 208358, New Haven, CT 06520–8358. E-mail: robertsternberg@sbcglobal.net

Embretson (this issue) has written a tour de force regarding future possible and even likely developments in the means by which testing will be done during the second century of ability testing. The article is forward-looking, comprehensive, and masterful. I found nothing with which to disagree. In writing this commentary, I merely attempt to complement Embretson's piece by discussing not the means of ability testing—which she handles in detail—but the content.

Ability testing has been dominated by what I view as a rather narrow approach to content (Sternberg, 1985, 1997, 1999). There have been exceptions, most notably, Guilford (1967), who seriously considered the importance of assessing creative and practical skills in his theory and his tests. For the most part, however, the tests have measured memory and analytical skills. These are important skills, but they are by no means the only ones that matter for understanding the structure of abilities or success in life.

There are many reasons that we have focused on the traditional skill set in psychometric testing of abilities. First, these abilities have been measured before, and tradition is powerful in the ability-testing field. Second, the abilities are relatively easy to measure using so-called objective formats. Third, they can be measured with good reliability and validity. We are nowhere near as far along in the development of tests of other kinds of abilities, but then, we have not yet had a century to develop them. Fourth, people who go into the ability-testing field are probably people who, on average, have done well on traditional ability tests, and thus capitalize on their own strengths in studying this area. But as tends to be true in any area, people often value what they do best, judging others by the skills in which they, themselves, excel (see Okagaki & Sternberg, 1993). Fifth, systems tend to be self-propagating (Sternberg, 1997), and once a system of testing is in place, it tends to select those who will later make selections, and so forth, so that whatever system is in place continues to be in place. Finally, no one can argue with the financial success that the testing industry has experienced, and money goes a long way in determining what research is done and what products are invented.

BEYOND THE TRADITIONAL SKILL SET

In my own work (Sternberg, 1997), I have proposed adding creative and practical abilities. Some believe that these abilities are *g*-based, but in maintaining this belief, they ignore statistical evidence but more importantly the evidence of their own eyes—namely, people who are analytically strong but weak in either creative or practical skills. Indeed, one might argue that countless studies arguing for the importance of analytical abilities over creative abilities make the point—that they are analytical but, in their endless repetition of what has been shown before—not terribly creative. Psychologists may be slow to come around. No one else is. Hiring

continues to be done in ways that look beyond *g* to other kinds of important skills. One might argue that, in this respect, psychologists have been blinded by their own expertise (Frensch & Sternberg, 1989; Sternberg & Lubart, 1995).

Creative and practical abilities can be measured in a variety of ways. In our Rainbow Project (Sternberg & The Rainbow Collaborators, in press), we have used performance-based measures as well as more traditional ones. The former include writing short stories, telling short stories, and captioning cartoons to measure creative abilities, and situational-judgment tests of responses to school situations and practical situations presented either in written or movie format to measure practical abilities. We have found that such tests increase prediction of college grades and also reduce score differences among ethnic groups.

Other intellectual abilities may come to be measured as well. Although Howard Gardner (1983, 1999) has not been particularly supportive of traditional cognitive tests, abilities in his theory of multiple intelligences that have not been traditionally measured, such as bodily-kinesthetic, musical, interpersonal, intrapersonal, and naturalistic, may come to be measured in the tests of the second century.

I believe that the most important skill set is not measured at all, and that is the skill set involved in wisdom—using one's intelligence and experience for a common good (Sternberg, 1998, 2002a). People can be smart but not wise. When they are, they are susceptible to committing four fallacies:

1. The egocentrism fallacy, whereby they come to believe that the world revolves, or, at least, should revolve, around them. They then act in ways that benefit them, regardless of what the effects may be on other people.
2. The omniscience fallacy, whereby they come to believe that they know all there is to know, and therefore do not have to listen to the advice and counsel of others.
3. The omnipotence fallacy, whereby they come to believe that their brains and education somehow make them all-powerful.
4. The invulnerability fallacy, whereby they come to believe that not only can they do what they want, but that others will never be clever enough to figure out what they have done or, even if others do figure it out, to get back at them.

Can we devise ways of measuring the extent to which smart people think in stupid ways (Sternberg, 2002b)? Given the fiascoes we have observed in politics (e.g., during the Nixon and Clinton administrations) and in business (e.g., Enron, WorldCom, Arthur Andersen), perhaps no kinds of measurements are more important.

We can continue to devise new ways of measuring the same things we have measured before. Such efforts may be useful. But efforts to develop ways of measuring new constructs may prove to be even more useful.

ACKNOWLEDGMENTS

Preparation of this article was supported by Grant REC–9979843 from the National Science Foundation and by a government grant under the Javits Act Program (Grant No. R206R000001) as administered by the Institute of Educational Sciences, U. S. Department of Education. Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the positions or the policies of the U.S. government, and no official endorsement should be inferred.

REFERENCES

Frensch, P. A., & Sternberg, R. J. (1989). Expertise and intelligent thinking: When is it worse to know better? In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 157–188). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century.* New York: Basic Books.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw–Hill.

Okagaki, L., & Sternberg, R. J. (1993). Putting the distance into students' hands: Practical intelligence for school. In R. R. Cocking & K. A. Renninger (Eds.), *The development and meaning of psychological distance* (pp. 237–253). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (1997). *Successful intelligence.* New York: Plume.

Sternberg, R. J. (1998). A balance theory of wisdom. *Review of General Psychology, 2,* 347–365.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology, 3,* 292–316.

Sternberg, R. J. (2002a). It's not just what you know, but how you use it: Teaching for wisdom in our schools. *Education Week, 22,* 42, 53.

Sternberg, R. J. (2002b). Smart people are not stupid, but they sure can be foolish: The imbalance theory of foolishness. In R. J. Sternberg (Ed.), *Why smart people can be so stupid* (pp. 232–242). New Haven, CT: Yale University Press.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity.* New York: Free Press.

Sternberg, R. J., & The Rainbow Collaborators. (in press). Augmenting the SAT through assessments of analytic, practical, and creative skills. In W. J. Camara & E. W. Kimmel (Eds.), *Choosing students: Higher education tools for the 21st century.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.