Research Memorandum

Improving Assessment: The Intersection of Psychology and Psychometrics

Ida M. Lawrence

Edward C. Shea

November 2008

ETS RM-08-15

ETS®

Listening. Learning. Leading.®

Improving Assessment: The Intersection of Psychology and Psychometrics

Ida M. Lawrence and Edward C. Shea
ETS, Princeton, NJ

November 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2008 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

Abstract

Psychometricians can and must learn from psychologists, and for several important reasons. One is that psychologists can enhance the ability of psychometricians to design assessments in substantively principled ways so that what test takers know is measured. A second is that psychologists can identify how test takers reason and how their knowledge and reasoning is influenced by situational factors and their background knowledge. In addition, psychologists are very aware of the diverse characteristics of the test takers who take assessments, and those characteristics have implications for assessment design and score interpretation. Finally, psychologists can help psychometricians provide useful and comprehensible information results from testing.

The article begins by making some points about educational assessments and the work of psychometricians. It then describes why and how psychometricians can learn from psychologists. The article acknowledges that the recognition that psychological findings and perspectives are critical for a well-informed psychometrics has many precedents. It then identifies some limitations of current assessments before turning to a vision for improved assessments, including examples from work being done in psychology.

Key words: Psychometrics and psychology, improving educational measurement

Acknowledgments

This paper is revised version of an invited address given to members of Division 5 (Evaluation, Measurement, and Statistics) at the August 2007 meeting of the American Psychological Association.

From its beginnings, psychometrics has had a split personality. On the one hand, it has been concerned with practical means of measurement and prediction, including not only the construction of instruments but also the mathematical and statistical basis for obtaining reliable and valid measurements—or what is commonly called "test theory." On the other hand, the very notion of validity—particularly the notion of "construct validity"—implies that one be at least somewhat bothered by the problem of what a test measures (Carroll, 1976, p. 27).

Carroll's suggestion that psychometrics can be too narrowly focused on "test theory" introduces our fundamental assertion, that psychometricians need help from psychologists in order to move assessment forward as a way to improve education.

This article begins by making some points about educational assessments and the work of psychometricians. We will draw upon our first-hand experience from the work we do in a research and development division at a large educational research organization to support ongoing and new testing programs. Then we will describe why we believe psychometricians can learn from psychologists. In doing so we will acknowledge that the recognition that psychological findings and perspectives are critical for a well-informed psychometrics has many precedents. We will then identify some limitations of current assessments before turning to a vision for improved assessments, including examples from work being done in psychology that we believe can lead to improvement.

Assessments and the Work of Psychometricians

The best educational assessments do two things. First, a high-quality educational assessment provides data that are reliable. Second, a high-quality assessment measures constructs in ways that can impact learning, sometimes directly, sometimes indirectly, in meaningful ways. Testing experts work from the assumption that scores reported to users must have meaning, which requires that certain technical and construct assumptions be met. These include the following:

- The scores generated by the assessment are an accurate reflection of the domain in question.
- Scores from the same testing program can be sensibly compared.
- The amount of noise in any given score is relatively small and can be described to score users.

- The relationships of scores to other indicators are sensible given their intended meaning and use.
- Similarly, the absence of relationships with variables that are not pertinent to the intended use of the test is an assumption that has to be tested.
- Finally, the intended inferences from scores can be supported through research (validity evidence).

In our organization, the work involved in building assessment fits into these four categories: (a) designing assessments, (b) developing items and tests, (c) scoring assessments, which includes equating, and (d) conducting research to ensure that these assessments are performing as intended; that they are reliable, fair, and valid; and that they are free from influences that detract from the meaning of the scores.

All of this research and development work involves a variety of experts. These experts include assessment developers—specialists with deep subject matter expertise in a variety of content fields as well as knowledge of good item writing and test development practice. The assessment developers work alongside psychometricians. These experts have graduate training in scaling, scoring, equating, and analyzing the quality of assessments. Close collaboration between assessment developers and psychometricians is a cornerstone of ensuring high-quality assessment.

So what do psychometricians actually do, day-to-day? The psychometricians working on testing programs in our organization perform a number of roles. Here are some of them:

- They contribute to the design of assessments—in other words, they specify the detailed set of features needed to create fair and valid assessments.
- They carry out statistical analysis procedures required for accurate score reporting and interpretation of testing program results.
- They prepare statistical reports documenting the performance of assessments.
- They design, develop, and document best practice technical and operating procedures and statistical guidelines for testing programs.

- They develop score interpretation materials so clients and score users can understand the meaning of assessment results, and they respond to special data requests from clients.
- They serve as liaisons to client advisory boards and committees.
- They identify ways to gather increased validity evidence for the meaning of test scores.
- They identify research and measurement-related problems and their psychometric implications and design studies to identify solutions.
- They contribute to the development and evaluation of new constructs, testing modes, item types, and psychometric methods.
- They are part of teams charged with development of new assessments and provide guidance for "pushing the envelope," so to speak, so that new assessments reflect what we have learned from other assessments.

Camara (2007) had a nice description of the characteristics of good psychometricians that seems exactly right. He said:

> Psychometricians…[combine] superb technical skills and organizational abilities, amazing amounts of drive and personal motivation, and an ability to communicate with nonresearchers and explain technical and complex issues in clear and simple language. In addition they enjoy solving complex problems in a practical and applied environment. (p. 352)

The point Camara made about liking to solve problems is an important one. Good psychometricians like to experiment: They like to identify a problem, hypothesize how to solve it, and design a study to capture data to test their hypothesis. It is worth noting that a robust interest in experimentation is a characteristic of many kinds of psychologists.

We know that psychometricians are trained to use sophisticated statistical, quantitative, and analytical methods and tools to design, develop, interpret, and use assessments. And it hardly our intention to deny that expertise in using these methods and tools is essential to good educational measurement. But it is also the case that many psychometricians lack extended training in psychology (Boorsboom, 2006). For that reason, we believe the best psychometric

work increasingly is the result of the collaboration of teams that bring unique expertise and perspectives to bear on the major categories of psychometric work described above. One key perspective and expertise is that of the psychologist.

How Can Psychology Inform Psychometrics?

Psychometrics needs to learn from psychology for several important reasons. One is that psychologists can help enhance the ability of psychometricians to design assessments in substantively principled ways so that what test takers know is measured. A second reason is that psychologists can identify how test takers reason and how their knowledge and reasoning is influenced by situational factors and their background knowledge. In addition, psychologists are very aware of the diverse characteristics of the test takers who take assessments and those characteristics have implications for assessment design and score interpretation. Finally, psychologists can help psychometricians provide useful and comprehensible information results from testing.

The foregoing inputs from psychology to psychometrics are natural if one considers that psychometrics shares the paradigm that underlies much of psychology, namely the study of person-by-situation interactions. In its most abstract form, psychometrics is the modeling of test-taker responses (behavior) in response to items (situations).

In effect, psychometric models are about formulating approaches to mathematically model the encounter of persons with items. The predominant family of such models is labeled collectively *item response theory* (e.g., Lord, 1980). The *theory* in this case is not, or at least it has not been, a psychological theory, as pointed out by Goldstein and Wood (1989). However, over the last 20 years or so, it has become increasingly clear that psychological theorizing can be fruitfully incorporated into psychometrics models. By augmenting psychometric models with psychological theorizing, it becomes feasible to infer from a series of those person-by-situation encounters the standing of persons on well-defined psychological constructs rather than simply estimating a score that remains to be validated. Moreover, from the analysis of those encounters, we can learn much about the items (situations) that elicit responses or reactions and refine the psychological theory that was the basis for characterizing persons and situations. This last point is important. A test designed with psychological principles in mind embodies a set of hypotheses about what is expected to transpire in a series of person-by-situation interactions. This means

that truly useful information can be obtained, often with extraordinarily large samples, about psychological hypotheses.

Thus, the value of closer collaboration between psychology and psychometrics is twofold. On the one hand, the collaboration can provide assessment developers with the most current knowledge about the psychology of the domain. On the other hand, as the test is administered, valuable information is generated about the psychological theory employed in the design of the test.

The importance of psychological findings and perspectives for psychometrics is a discussion that has been going on for a long time. Cronbach, for example, made a similar argument in his presidential address to the American Psychological Association (APA; Cronbach, 1957). As we will discuss in more detail below, two presidential addresses to APA's Division of Evaluation and Measurement, as it was known then (Anastasi, 1967; Glaser, 1981), emphasized the need for a broader perspective. One thing is clear, however: There has never been enough collaboration, and today this collaboration is more important than ever. Clearly, the use of testing has greatly expanded in the last several years, in large part because of No Child Left Behind (NCLB) legislation. An article in *The New York Times* (Herszenhorn, 2006) recently estimated that 45 million educational tests are administered a year in the United States. Many measurement organizations are building new assessments for populations they have not typically assessed before, sometimes measuring new kinds of skills. So, more than ever, it is critical that the fields of educational measurement and psychometrics be informed by insights from a number of psychological fields.

It is useful to take a look through the literature in the measurement field and to see that this issue of the need for closer collaboration between psychometricians and psychologists has been an ongoing theme. An historical perspective shows that the need for different kinds of collaborations between the disciplines has evolved alongside the prevailing criticisms of assessments.

We can begin with Thorndike in 1910. From his perspective, there was a great optimism about what psychology was capable of doing for education. He claimed, perhaps optimistically, "A complete science of psychology would tell every fact about everyone's intellect and character and behavior, would tell the cause of every change in human nature, would tell the result which every educational force—every act of every person that changed any other or the agent

himself—would have" (Thorndike, 1910, p. 6). Thorndike believed in the potential for psychology to measure change and to study the learning process. While his approach is clearly more behaviorist and less cognitively based than that of researchers writing in the later 20th century, his emphasis on the need for psychologists to enrich the field of educational measurement is an early example of our theme.

In fact, the field of mental measurement began in psychology. Thorndike studied learning extensively, as well as methods for measuring it and for evaluating the quality of those measures. (Thorndike conducted the first College Board predictive validity study and created achievement tests for the College Board in the 1920s.) Another example of an early pioneer in mental measurement who was a psychologist is Carl Brigham, who applied psychological methods to the study of how students solved test items long before the field of cognitive psychology became prominent (Brigham, 1932).

Another earlier argument for closer collaboration between psychology and psychometrics can be found in Loevinger's landmark paper, *Objective Tests as Instruments of Psychological Theory* (Loevinger, 1957). Loevinger raised concerns about contemporary concepts of test theory. She was highly critical of the field's inattention to construct issues and pointed out that certain classical notions of validity, such as predictive, concurrent, or content validity, were quite inadequate by themselves. Instead, she called for a more comprehensive notion of construct validity, one informed by psychological theory. According to Loevinger, "The argument against classical criterion-related psychometrics is thus two fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful" (p. 641).

Although no specific field in psychology was identified by Loevinger, the thrust of her paper was to "celebrate the extension of the concept of validity as an indication that psychometrics is recognized as truly the handmaiden of psychology rather than merely of psychotechnology" (Loevinger, 1957, p. 636). She wanted to develop a coherent framework of psychometrics that melded test theory and test construction, and she looked to psychology to help do this. Much of her paper on construct-oriented psychometrics was based on the earlier contributions of Cronbach and Meehl (1955).

Ten years later, in 1967, in an article that began as a presidential address to APA's Division of Evaluation and Measurement, Anastasi (1967) took test makers to task and asserted that they were, to some extent, responsible for the increasing anti-testing environment prevalent at the time. She voiced serious concerns about the nature of tests (in particular their content), the meaning of scores from these tests, and how tests were used.

Anastasi (1967) pointed to inadequate collaboration between psychometricians and psychologists as a factor responsible for the prevalent criticisms of testing. She argued,

> Testing today is not adequately assimilating relevant developments from the science of behavior…. Psychometricians appear to shed much of their psychological knowledge as they concentrate upon the minutiae of elegant statistical techniques…. [The psychometrician] may become so deeply engrossed in the technical refinements of his specialty that he loses touch with relevant developments in other psychological specialties. Yet these developments may basically alter the meaning of the very tests he is busy elaborating and refining. (pp. 300, 302)

Anastasi suggested some possible solutions. In particular, she recommended that the assessment field take advantage of modern psychological knowledge (1967, pp. 302-305). This knowledge included the importance of *change* and the *nature of intelligence.* For example, she argued that in assessing individual differences, it is important to focus on change in someone's status because tests can be used as instruments to facilitate change. (This is relevant to contemporary emphases today on assessment *for* learning, an idea to which we will return.) She pointed out the fact that the nature of intelligence varies by time and situation, noting how structural analyses (Anastasi used the narrower term of factor analyses) and other kinds of analyses can provide information about how abilities develop over time and become organized. This, of course, is highly relevant to the interest today in cognitive psychology to understand how someone moves from novice to expert in a particular domain. Insight about the role of time and situation may also help to increase awareness of the conditions that bring about change, for instance, in a classroom.

Finally, Anastasi (1967) urged her readers to understand that test performance is influenced by personality factors such as self-concept, persistence, and goal orientation. To assess abilities without considering the impact of these other factors is likely to be misleading.

In another presidential address to APA's Division of Evaluation and Measurement at the beginning of the 1980s, Glaser (1981) also argued for the salutary influence of psychology, in this case both cognitive and developmental psychology, on psychometrics. As Anastasi had, Glaser acknowledged a prevalent anti-testing bias and pointed out that, in important ways, skepticism about the "psychometric enterprise" had "accelerated technical and scientific examination of the fundamental assumptions and basic knowledge that underlie test design and various assessment procedures" (pp. 923-924). He suggested that diagnosis of performance regularities at different levels of learning, analysis of the nature of competence, and investigation of certain measures of aptitude (such as self-regulatory and metacognitive skills) would be part of a fruitful research agenda for psychology that ought to lead to improvements in educational assessment.

The promise of cognitive psychology to improve assessment was also a theme in another paper from this same period by Sternberg (1981). Sternberg asserted that cognitive testing procedures currently in use should be supplemented by information-processing testing procedures. He indicated that that these procedures would deliver assessments of the information-processing components, mental representations, and strategies that people use in performing actual tasks. And he pointed to a particular field in psychology where the potential for help was apparent to him: "…it would seem reasonable to believe that contemporary developments in cognitive psychology should have implications for the psychometric testing of mental abilities" (Sternberg, p. 1181).

Similarly emphasizing the importance of studying how people perform actual tasks, Embretson (1983) discussed a view of construct representation that identifies the theoretical mechanisms that underlie task performance. This point had been made much earlier (Glaser, 1963) but at an inauspicious time when behaviorism dominated psychology. By 1983, cognitive psychology had gained dominance over behaviorism, which made it feasible for Embretson and others to suggest the need to carry out task decomposition as part of assessment design and the validation of those assessments. Embretson emphasized that construct validation research needed to be concerned with the qualities that are reflected in the test score, and that this reflection can take on two interpretations. She distinguished between the construct that directly impacts performance (either successful or unsuccessful) on a test item and the possibility that the construct is correlated with other constructs that are directly related to performance. She gave an

example of a personality trait that could be reflected in an aptitude test score because of environmental conditions or developmental conditions (Embretson, 1983, p. 195). Clearly, collaboration between psychologists and psychometricians would be critical to carrying out this kind of validation research.

As a final example, the theme of interaction between psychology and psychometrics was reflected in Bejar's "Speculations on the Future of Test Design" (Bejar, 1985). Here Bejar offered explicit advice about how to improve test design and argued that fundamental change is likely to come about through an integration of cognitive psychology and psychometric theory. He argued for the need to "cognitivize psychometrics" (p. 285) in two major ways: by understanding test performance in terms of cognitive constructs to help in validation of test scores and by improving the design of current and new tests through better understanding sources of the difference among items with respect to their characteristics, especially difficulty. Bejar's speculations were the closing chapter in a book edited by Embretson (1985), which, though out of print today, contains many examples of psychology informing test design.

To sum up, there is an interesting history of arguments, spanning the last century, for the closer collaboration of psychometrics and psychology, with the motivation for these arguments often having to do with the limitations of assessment perceived at a particular moment in time.

Assessment Today—What Are the Limitations? What Are Some Visions for the Future?

The need for collaboration between psychometrics and psychology is stronger today than ever before, and current assessments continue to be limited in several ways. For instance, the 2001 National Research Council (NRC) report *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) stated that current assessments suffer from four major limitations:

- Important aspects of cognition and learning are not addressed in these assessments. These aspects include things like students' organization of knowledge, problem representations, use of strategies, self-monitoring skills, and individual contributions to group problem-solving.
- Current assessments provide limited information that teachers and administrators can use to identify both why students do not perform well and what instructional interventions are likely to improve achievement.

- Most assessments are static instruments. They do not capture the progression of students' conceptual understanding over time—they do not capture the heart of learning. This is because most assessments lack a theoretical model for how student understanding in a particular domain develops.
- There are ongoing concerns about fairness and equity of assessments, particularly in the use of assessments for high-stakes decisions and about whether those assessments are up to the job.

The NRC report made a detailed and compelling case for how advances in the sciences of thinking and learning—cognitive science, which includes cognitive psychology—should inform the building of good assessments. Among many useful prescriptions, the report urged those who design and build assessments to be very clear about what test takers know, how they know it, and how they are able to use that knowledge to answer questions, solve problems, and engage in additional learning.

More recently, Ferrara (2006) called for a "psychology of …educational achievement testing" (pp. 3-4). He suggested six very useful and concrete capabilities for testing programs: (a) defining achievement constructs as models of cognition and learning, (b) modeling the relationship between assessment tasks and students responses, (c) understanding affective processes and their influence on proficiency, (d) understanding the role of language and culture in the development of proficiency, (e) developing items that really do assess the targeted constructs, and (f) communicating performance information in clear and helpful ways. This list provides a good summary of some of the arguments by Loevinger (1957), Anastasi (1967), Glaser (1981), Sternberg (1981), Embretson (1983), and Bejar (1985), and the capabilities in the list address the limitations of assessments outlined in the NRC report (Pellegrino et al., 2001).

We have described two fairly comprehensive visions for better assessment—that contained in the NRC report (Pellegrino et al., 2001) and that contained in Ferrara's paper (2006). Let us briefly mention three more we believe are very suggestive. Each has implications for the design, development, scoring, and interpretation of assessments, and draws on work in psychology.

One of these visions has predicted that, among many changes in the field of assessment, test development procedures will be characterized by continuous test revision, automated validity

studies, and item development by artificial intelligence (Embretson, 2004). A testing system that involves all three of the characteristics will be possible only if design principles for items and tests are based on a framework rooted in principles from cognitive science. In this way, it will be possible to calibrate new test questions that have not been previously field-tested and can be assumed to behave similarly to other items that have been previously calibrated. In this system, each item type will require a distinct cognitive design system that has been verified by a research effort based on a theory of how items (and persons) behave. The idea is to understand what processes, strategies, and knowledge are required to solve items and what features of the items call upon those processes, strategies, and knowledge.

A recent article by Fadel, Honey, and Pasnik (2007) offered another comprehensive view of improved assessment. The authors argued that learning in the 21st century is about the process of integrating and using knowledge, not just about the acquisition of facts and procedures. To accomplish this new role, assessments need to have the following features: (a) be performance-based so how students apply content knowledge to problem-solving and critical thinking situations can be identified, (b) reveal the strategies a student uses to solve a problem, (c) produce data that is actionable, and (d) provide frequent opportunities for feedback and revision during assessment so that assessment itself is a learning event. Assessments with these characteristics can be called assessments *for* learning rather than simply *of* learning.

A final vision comes from Stiggins (2007), who argued that the assessment *for* learning paradigm has largely not been implemented. He identified two things that must be done. First, it needs to be recognized that assessment is about more than technical quality; assessment is also about the effect of the score on the student. What good is the assessment if it causes the student to give up? Second, it needs to be understood that students' reactions to the results of an assessment are as important as those of adults. A student's emotional reaction to the results of an assessment will determine what he or she does in response to the results. In short, Stiggins correctly emphasized an affective dimension of assessment results.

These are some excellent recent suggestions about the future of assessments and how assessments can benefit from a wider perspective on learning, on what should be measured and how it should be measured, and on the psychological consequences of assessment results.

Some Examples of Work in Psychology That Can Advance Assessment

We now want to explore some areas in psychology in which we believe critical work has been done in the recent past or is being done today from which people who make assessments can learn. This work can provide the kind of fertilization we believe is essential to advance assessment.

As suggested earlier, there has been a great deal of research in the last quarter century to reconceptualize our understanding of how people learn. Bransford and his colleagues (Bransford, Brown, & Cocking, 1999) identified several areas in which there is much new information about the nature of learning. These areas include memory and the structure of knowledge, analysis of problem-solving and reasoning, the early foundations of learning (the learning of infants and young children), metacognition processes and self-regulatory capabilities, and cultural experience and community participation.

The work in these areas is clearly very important. For example, understanding how expert learners and novice learners go about solving problems is obviously crucial in helping to improve instructional methods in schools. But the work is clearly also of critical value to psychometricians in their role as designers of assessments. Gorin (2006) summarized various strands of work that point to the centrality of cognitive models in helping assessment designers more precisely specify the constructs being measured in an assessment. Cognitive models are used to specify the "learners' representations of a domain in terms of requisite *knowledge, skills and abilities*" (Gorin, p. 22) and indicate the relationships between these.

Another area of critical work being done in psychology is accounting for the difficulty of a problem, or the probability of a correct response to one. Why are some problems more difficult than others? Although this might seem a simple question to answer, psychology is required to answer it fully, thereby offering the possibility for fruitful collaboration between psychometricians and psychologists. Here is an example. Cognitive psychologists (Kotovsky, Hayes, & Simon, 1985; Kotovsky & Simon, 1990) have addressed the issue of problem difficulty explicitly. They wonder, for example, why problems that are structurally identical differ in difficulty. One possibility, of course, is that the structural attributes of a problem, as postulated by a theory, may or may not correspond to how a student actually approaches the problem. That is, the student provides his/her own structural attributes, and, unless there is a match between the attributes postulated by a psychologist and the attributes that actually underlie how the student

approaches the problem, there is likely to be a mismatch in the observed and predicted or theoretical difficulty.

A second possibility is that the surface attributes of a problem, such as its wording, have more than a passing influence on difficulty. For clusters of test takers, the specific content and wording of a problem may or may not trigger schemas that are relevant to the solution of a problem. Such clusters could be based on any number of person attributes. These attributes include gender (Cole & Willingham, 1997), prior educational experiences (Berk, 1980; Lehman, Lempert, & Nisbett, 1988), and language background (Abedi & Gandara, 2006; Abedi & Lord, 2001).

Of course, both possibilities for why problems that are structurally identical differ in difficulty could be operating simultaneously, leading to situations where the difficulty prediction works in some cases but not in others. So, different groups of students could differ in the approach they deploy to classes of problems, depending perhaps on their educational experiences or factors specific to the class of problems. When such interactions are present, the interpretation of the main effects—persons and items in the present case—is not straightforward. The interaction could be artifactual or could be the essence of the phenomenon of interest. A psychological analysis can be used to inform a dimensionality analysis when the hypothesized item structure within the assessment is formulated from a theory based on psychological principles. Going further, the theory could be used to manipulate items experimentally and then subject the data to a dimensionality analysis in a confirmatory study.

The work of educational psychologists to more fully understand individual differences, such as gender and ethnic differences, has clear implications for the kinds of tests that psychometricians build. A worthwhile goal for test makers is to use this information about individual differences to design assessments that measure the construct of interest but that do so in ways that do not introduce task types that produce construct-irrelevant variance. An awareness of these issues is also very relevant to the work psychometricians do to design studies to ensure that assessments are valid for all examinee subgroups. To control for construct irrelevant variance, Haladyna and Downing (2004) suggested that assessment designers need to consider the extent to which deficits in reading interfere with performance on an achievement test, as well as the role of test anxiety, motivation, and engagement. The work of language and reading experts and social and personality psychologists to study the impact of these kinds of factors can be used to improve assessment designs.

One example of an improvement might be in the design of item formats. It may be possible to design innovative item formats that would, for example, reduce the impact of test anxiety. For instance, traditional assessments do not usually provide immediate feedback about the correctness of an answer to an open-ended question. Recent experimental work by Attali (Attali & Powers, 2008; Attali, Powers, & Hawthorn, 2008) showed that the reliability of an assessment can be improved by providing test takers with immediate feedback on the correctness of an answer and the opportunity to revise the answer. The study found that test takers were able to improve their performance on the test significantly and that the scores were more reliable. More research, based in psychology and psychometrics, is needed to understand the validity of this kind of format.

The work we have been sketching out here, then, has relevance for the important work psychometricians must do to improve assessments. Another strand of work relevant to improving assessments is being carried out in social and personality psychology. This is work to identify and reliably measure key noncognitive traits and skills: qualities such as persistence, tenacity, collegiality, communication, and enthusiasm. Expanding *what* is measured is critical as psychometricians consider how assessments might be improved. It may be that valid and reliable assessments of noncognitive constructs can provide information about training, school, or job success beyond what cognitive ability measures can provide. A growing body of work being done here to identify coherent and measurable non-traditional constructs can provide the impetus to broaden what is being measured. (Kyllonen, 2005)

Finally, we should briefly mention the contributions of developmental psychologists. A developmental perspective can provide a framework for the trajectory within which people develop certain proficiencies, such as language or certain understandings of math and science concepts at different ages. These trajectories can influence performance on assessments and need to be accounted for in test design.

A key developmental concern is how proficiency changes over time. An example of how developmental psychology principles can inform assessment is Wilson's learning progressions work (Wilson, 2004). Wilson has shown that things known about people need to be combined with things known about test questions. Some type of interactions need to be isolated either to utilize the factors or to engineer the influence of the factors out of the system. What a question is

designed to measure needs to be considered as well as what a test taker has to do to demonstrate proficiency and at what levels of development can certain levels of mastery be expected.

Test makers need to consider other developmental issues such as that the timing of a test should be age appropriate, that the pacing of a test should be facilitated, and that reading materials should be age appropriate and engaging. Related questions test makers should ask include whether younger students take large-scale assessments more seriously than older students and whether there are age-related ways to engage students better to increase motivation levels to get more valid scores.

The recent decision in New York City to study the impact of providing cash incentives for performance on NCLB tests is an example of where a theory based on principles of developmental psychology can form a framework for validating this kind of assessment intervention. For example, one needs to understand what tactics work best to motivate performance and then ensure the tactics are used uniformly so that the scores are fair. Before moving forward with such an intervention, it would be necessary to distinguish between real learning differences and differences in motivation to perform well. Of course, other policy implications would also need to be well understood.

Summary

As have others before us, we have argued in this paper that psychometricians and psychologists need to collaborate closely in all phases of assessment work, from design to development, from scoring to interpretation. We believe the opportunity for psychology to fertilize the field of psychometrics is enormous and should be exploited. With the increased use of tests in the classroom and for high-stakes decisions, the need to have a complete understanding of score meaning has become paramount. The world is looking to the field of assessment to solve many educational problems.

Interestingly, this need on the part of psychometrics to draw from psychology comes at a time when psychologists are worried about communicating the value of the field to their students and other disciplines. While we were completing this paper, we came across an article in a recent issue of the Association for Psychological Science journal *Observer* titled “Exporting Psychology: Communicating the Value of Our Field to Students” (Baron, 2007). One area that could be emphasized by psychology faculty to their students at both the undergraduate and graduate level is that psychology as a field has considerable insight to contribute to educational

measurement and psychometrics and that, through close collaboration between psychology and psychometrics, we can improve educational measurement by providing scores that are meaningful and that can have a positive impact on education.

References

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36-46.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22*, 297-306.

Attali, Y., & Powers, D. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended GRE Subject Test items* (GRE Board Research Rep. No. 04-02). Princeton, NJ: ETS.

Attali, Y., Powers, D., & Hawthorn, J. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended sentence-completion items* (GRE Board Research Rep. No. 03-15). Princeton, NJ: ETS.

Baron, R. A. (2007). Exporting psychology: Communicating the value of our field to students. *Observer 20*(6), 27-30.

Bejar, I. I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (p. 279-294). New York: Academic Press.

Berk, R. A. (1980). *Criterion-referenced measurement: The state of the art.* Baltimore: Johns Hopkins University Press.

Boorsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Brigham, C. C. (1932). *A study of error: A summary and evaluation of methods used in six years of study of the Scholastic Aptitude Test of the College Entrance Examination Board*. New York: College Entrance Examination Board.

Camara, W. (2007). Improving test development, use, and research: Psychologists in educational and psychological testing organizations. In R. Sternberg (Ed.), *Career paths in psychology: Where your degree can take you* (2nd ed.). Washington, DC: American Psychological Association.

Carroll. J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence*. Hillside, NJ: Lawrence Erlbaum Associates.

Cole, N. S., & Willingham, W. W. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 11*, 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement, 2,* 1-32.

Fadel, C., Honey, M., & Pasnik, S. (2007, May). Assessment in the age of innovation. *Education Week, 26*(38), 40.

Ferrara, S. (2006). Toward a psychology of large-scale educational achievement testing: Some features and capabilities. *Educational Measurement: Issues and Practice 25,* 2-5.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36,* 923-936.

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

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25,* 21-35.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice 23*(1), 17-27.

Herszenhorn, D. M. (2006, May 5). As test-taking grows, test-makers grow rarer. *The New York Times.* Retrieved October 31, 2008, from http://www.nytimes.com/2006/05/05/education/05testers.html

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard?: Evidence from Tower of Hanoi. *Cognitive Psychology, 17*, 248-294.

Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology, 22*, 143-183.

Kyllonen, P. (2005). *The case for noncognitive assessments* (R&D Connections No. 3). Princeton, NJ: ETS.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43*(6), 431-442.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Sternberg, R. J. (1981).Testing and cognitive psychology. *American Psychologist, 36*, 1181-1189.

Stiggins, R. (2007, May). Assessment through the student's eyes. *Educational Leadership: Educating the Whole Child, 64*(8), 22-26.

Thorndike, E. L. (1910). The contribution of psychology to education. *The Journal of Educational Psychology, 1,* 5-12.

Wilson, M. (2004). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.