Large-Scale Assessments That Support Learning: What Will It Take?

Naomi Chudowsky, James W. Pellegrino

Theory Into Practice, Volume 42, Number 1, Winter 2003, pp. 75-83
(Article)

Published by Ohio State University College of Education

Naomi Chudowsky
James W. Pellegrino

Large-Scale Assessments That Support Learning: What Will It Take?

Large-scale assessments can and should support learning. But for that to happen, greater clarity is needed about the underlying constructs, or aspects of thinking and learning, that are the most important targets for assessment. This article describes the construct problem that has long existed in achievement testing, and argues that current research in the cognitive sciences, measurement, and technology make this an opportune time to make a significant leap forward in assessing critical aspects of learning. However, designing new kinds of situations for capturing the complexity of learning requires breaking out of the current paradigm of drop-in-from-the-sky standardized testing. It also requires a sustained, collaborative effort among specialists in academic content, learning, and assessment.

> Needed are classroom and large-scale assessments that help all students learn and succeed in school by making as clear as possible to them, their teachers, and other education stakeholders the nature of their accomplishments and the progress of their learning. (National Research Council, Committee on the Foundations of Assessment, 2001, pp. 1-2)

Naomi Chudowsky is an educational testing consultant in Washington, DC, and James W. Pellegrino is Distinguished Professor of Cognitive Psychology and Education at the University of Illinois at Chicago.

IN THIS ERA OF EDUCATIONAL ACCOUNTABILITY it has become increasingly common to hear calls, like the one just mentioned, for better large-scale assessments that not only measure but also improve student learning (Commission on Instructionally Supportive Assessment, 2001; Shepard, 2000; Wiggins, 1998). Policy makers, educators, and the public are looking to large-scale assessments to serve a variety of purposes including gauging student learning, holding education systems accountable, signaling worthy goals for students and teachers to work toward, and providing useful feedback for instructional decision making.

Can large-scale assessments live up to the demands being placed on them? Can assessments be developed that can both measure and support student learning? And if so, why has so little progress been made in this regard? We set forth the proposition that large-scale assessments can and should do a much better job of supporting learning. But for that to happen, education leaders will need to rethink some of the fundamental assumptions, values, and beliefs that currently drive large-scale assessment practices in the United States. The knowledge base to support change is available but has yet to be harnessed.

In this article, we focus on the assessment of school learning or *achievement*. However, later we discuss how achievement testing has been influenced by past practices designed for a different purpose, assessing individuals' *aptitude* for learning

THEORY INTO PRACTICE, Volume 42, Number 1, Winter 2003

for predictive purposes. In keeping with the theme of this issue, we also mainly focus on assessments used in large-scale contexts—that is, assessments that are administered at the direction of users external to the classroom, such as policy makers—as opposed to assessments used by teachers in their own classrooms. Note that many of the issues raised in this article also apply to classroom assessment, which we believe has an even more central role to play in supporting instruction and learning, as described more fully elsewhere (NRC, 2001).

We begin by setting forth what we believe is the most significant impediment to improving large-scale achievement testing—the lack of clarity about the underlying constructs to be assessed. The term *construct* refers to the competency, or aspect of thinking or learning, that one is aiming to assess (e.g., an area of knowledge, skill, or ability). Contemporary research on learning and thinking illuminates critical constructs that should be targets of assessment, but many of them remain untapped by most large-scale tests, including how students organize knowledge, the representations they generate to solve problems, their use of strategies and self-monitoring skills, and their individual contributions to group problem solving (National Academy of Education, 1997). Uncovering the constructs that are most central to learning in a particular subject domain and developing assessments that focus on them requires a sustained, collaborative effort among specialists in academic content, learning, and assessment. Such multidisciplinary efforts are needed to bring to bear the best available research to answer fundamental questions pertinent to the design of assessments of school achievement, such as the following:

1. What are the most essential domains of knowledge and skill that students should master in school to be productive members of society?
2. What kinds of performances differentiate beginning, competent, and expert learners in each domain?
3. What are the central conceptual structures within each domain that students must grasp to successfully move on to higher levels of understanding?
4. What are typical difficulties or misconceptions that learners have in each domain that, if identified early, could be remediated with instruction?

After providing some brief background about the construct problem that has long existed in achievement testing, we will show that current research in the cognitive sciences, along with advances in measurement and technology, make this an opportune time to make a significant leap forward in defining and assessing critical academic achievement constructs. However, we will also argue that only limited improvements in large-scale assessment are possible so long as we remain wedded to current constraints and typical standardized testing scenarios. Designing new kinds of situations for capturing the complexity of cognition and learning requires breaking out of the current paradigm of drop-in-from-the-sky standardized testing and exploring alternative approaches.

The Construct Problem in Achievement Testing

The lack of clarity and theoretical basis for the constructs underlying achievement testing is a byproduct of earlier theories of measurement and learning. During the first few decades of the 20th century, techniques were developed for assessing general intelligence or specific aptitudes for learning, for the purpose of sorting an increasingly diverse student population into instructional tracks for more efficient instruction. The assessment techniques of that time were based on an underlying assumption that individuals have general and specific aptitudes to learn (some people having more and others less), which influences their performance across a broad range of situations and content areas. Although adequate theories of constructs such as aptitudes were lacking, assessments were nonetheless developed to differentiate among individuals with respect to these hypothesized underlying latent traits.

Recognizing that assessment requires a balance between defining constructs at a conceptual level and finding ways of operationalizing them, assessment validation nonetheless came to be dominated by pragmatic, as opposed to theoretical, approaches. That is, because construct validity (the assembly of evidence that a test measures the underlying construct it purports to measure) was difficult to demonstrate, validation tended to focus instead on criterion or predictive validity (evidence

that a test predicts some outcome further down the road). For instance, the SAT, formerly known as the Scholastic Aptitude Test, is widely used for college selection purposes because it has been shown to predict to some extent the likelihood of success in a typical college learning environment, not because it has been shown to measure competencies required to do college level work.

We now understand that operational and conceptual definitions can become redundant, as in the case of defining intelligence as that which intelligence tests measure (Mosher, in press; Pellegrino, 1992). Still today, the vast majority of aptitude tests, including the SAT, fit the theories of measurement and learning developed early in the 20th century. What is less obvious to many is that the measurement and test development model created for aptitude assessment has served as the prototype for the assessment of academic achievement. Rather than measuring verbal or mathematical aptitude, we measure reading or mathematics achievement, but many of the fundamental assumptions are the same—that students have general, unidimensional proficiencies, or tendencies to behave in certain ways across diverse situations. The adoption of a latent trait approach in achievement testing has been aided and abetted by earlier theories of learning that failed to deal with the true complexity entailed in the mastery of academic content domains. Achievement came to be represented as a step-by-step accumulation of facts, procedures, definitions, and other discrete bits of knowledge and skill, and measurement meant how much more or less one person knew relative to his or her peers.

Most large-scale tests continue to define the student achievement construct they purport to measure in a largely operational form (i.e., the construct is defined by the content and item types on the test). This is contrasted with clearly defining the construct beforehand in terms of the specific knowledge and skills that accompany competence and then selecting content and item types to assess the critical constructs. This remains true for many large-scale testing programs despite the fact that progress has been made in developing performance assessments that tap more complex forms of knowledge and skill. Many state achievement tests now use constructed response tasks that may require more complex skills than traditional multiple-choice items. Yet much of the potentially rich information about student learning is lost because such tests still lack explicit theories and models of the underlying achievement construct(s) to guide task design and selection, scoring of students' responses, and reporting of results.

Mosher (in press) argues that even the National Assessment of Educational Progress (NAEP), considered by many to be the gold standard of large-scale assessment of academic achievement, does not pay close enough attention to the structure of subject matter knowledge and skills. He concludes that the NAEP measures some confounded combination of aptitude and achievement constructs that have yet to be sorted out. Related arguments have been made by the NRC (1999a) and NAE (1997) committees that have evaluated NAEP.

Opportunities for a Significant Leap Forward

Over the last few decades there have been advances made in several areas. Taken together, they provide the foundations for developing achievement tests that have greater construct validity because they better reflect current understandings of the complexities of learning and the high standards of academic achievement we espouse for all students.

Content standards

Virtually every state now has standards in place that outline, in general terms, what all students should know and be able to do in core subjects. Standards represent an important start toward identifying key competencies that should be the focus of assessment; however, they generally fall short of their intentions. Most state standards are too vague to be useful blueprints for instruction or assessment. Some call on students to learn a broad range of content rather than focusing in-depth on the most central concepts and methods of a particular discipline; some standards are so voluminous that the big ideas are lost or buried (American Federation of Teachers, 1999; Finn, Petrilli, & Vanourek, 1998).

Recently, five national education associations convened the Commission on Instructionally Supportive Assessment, in which one of the authors

took part. Experts in assessment, curriculum, and instruction were brought together to recommend ways that state-administered achievement tests could not only satisfy public demands for accountability, but also improve the instruction of children. In the resulting report, *Building Tests to Support Instruction and Accountability*, the Commission's (2001) first two recommendations speak clearly to the limitations of standards for guiding instructionally supportive assessment. First, they recommend that, "A state's content standards must be prioritized to support effective instruction and assessment." Given limited instructional time, educators cannot adequately address the large number of content standards. In addition, state tests cannot adequately assess all of the content standards, and thus tend to focus on those that are easiest to test. Second, the Commission recommends that, "A state's high-priority content standards must be clearly and thoroughly described so that the knowledge and skills students need to demonstrate competence are evident." This reflects the concern that many content standards are not worded with sufficient clarity to support instructional planning and assessment design. Cognitive and learning research has much to offer in response to these kinds of concerns.

Advances in the sciences of thinking and learning

The work of another recent panel of which the authors were part, convened by the NRC (2001), focused on the implications of advances in the cognitive sciences and measurement for improving assessment. That committee's report, *Knowing What Students Know: The Science and Design of Educational Assessment*, demonstrates that we have accrued a substantial scientific knowledge base about how students think and learn in particular subject domains. However, most of what we know has yet to be applied to large-scale educational assessment. The report proposes that the constructs to be assessed should be largely determined by a model of cognition and learning that describes how people represent knowledge and develop competence in the domain. Such a model suggests the most important aspects of student achievement about which one would want to draw inferences, and provides clues about the types of assessment tasks that will elicit evidence to support those inferences (see Figure 1).

The model of learning that informs assessment design should include as many of the following key features as possible:

1. Be based on empirical studies of learners in the domain.
2. Identify performances that differentiate competent and less competent performance in the domain.
3. Provide a developmental perspective, laying out typical progressions from novice levels toward competence and then expertise, and noting landmark performances along the way.
4. Allow for a variety of typical ways that children come to understand the subject matter.
5. Capture some, but not all, aspects of what is known about how students think and learn in the domain. Starting with a theory of how people learn the subject matter, the designers of an assessment will need to select a slice or subset of the large theory as the targets of inference.
6. Lend itself to being aggregated at different grain sizes so it can be used for different assessment purposes (e.g., to provide fine-grained diagnostic information as well as coarser-grained summary information).

Figure 1. Features of a model of cognition and learning. (NRC, 2001)

Research on cognition and learning has produced a rich set of descriptions of domain-specific knowledge and performance that can serve as the basis for assessment design, particularly for certain areas of mathematics and science (e.g., American Association for the Advancement of Science, 2001; NRC, 2001). Yet much more research is needed. The current literature contains analyses of children's thinking conducted by various types of professionals, including teachers, curriculum developers, psychologists, and educational researchers, for a variety of purposes. Thus it should come as no surprise that existing descriptions of thinking differ on a number of dimensions: Some are highly detailed, whereas others are coarser-grained; some focus on procedures, whereas others emphasize conceptual understanding; and some focus on individual aspects of learning, whereas others emphasize the social nature of learning. Differing theoretical

descriptions of learning should not necessarily be viewed as competitive, even when they are focused on the same content. Rather, aspects of such descriptions can often be combined to create a more complete picture of student competence that better achieves the purposes of a particular assessment.

Advances in methods of measurement

Existing measurement methods (or psychometrics) are often blamed for constraining the types of large-scale tests that can be created, largely because such methods have long been associated with the types of tests aimed at ranking individuals and measuring general proficiencies. But many measurement experts point out that the problem lies not so much with the range of measurement models (e.g., the statistical methods for interpreting test data) available, but with the outdated conceptions of learning and observation that underlie the most widely used assessments. This takes us back again to the construct problem discussed earlier.

As laid out in the NRC (2001) report, a wide variety of measurement models are currently available to support the kinds of inferences about student knowledge that cognitive science suggests are important to pursue. It is now possible to (a) characterize student achievement in terms of multiple aspects of proficiency rather than relying on a single score; (b) chart students' progress over time instead of simply measuring performance or status at a particular point in time; (c) deal with multiple paths or alternative forms of valued performance; and (d) model performance not only at the level of students, but also at the levels of groups, classes, schools, and states.

Many of these methods are not yet widely used because they are not easily understood or packaged in accessible ways. Furthermore, having a broad array of models does not mean the measurement problem has been solved. Much hard work remains to focus psychometric model-building on the critical features of achievement that we desire to asses—but once again we note that those constructs must first be clearly delineated.

The role of technology

Information technologies, such as computers, are helping to remove some of the constraints that have limited assessment practice in the past, and technologies are expanding the types of constructs that can be tapped through assessment. By enriching assessment situations through the use of multimedia, interactivity, and control over the stimulus display, it is possible to assess a much wider array of constructs than was previously possible. New capabilities afforded by technology include directly assessing problem-solving skills, making visible sequences of actions taken by learners in solving problems, and modeling complex reasoning tasks (NRC, 2002). Technology also makes possible data collection on concept organization and other aspects of students' knowledge structures, as well as representations of their participation in socially interactive discussions and projects.

In sum, the advances in cognition, measurement, and technology just reviewed hold promise for creating assessments that are more useful and valid indicators of what students have learned (Figure 2). In fact, some promising examples of assessments that have capitalized on these advances (see NRC, 2001), especially in the form of classroom formative assessment tools, already exist. For instance, intelligent tutoring systems have been created that seamlessly integrate computerized, self-paced instruction with assessment of student correct responses and errors in areas of the curriculum such as algebra and physics. The computer program recognizes when a student is progressing in the desired learning direction and when a mistake has been made, and provides advice and remediation to keep the learning on track in the desired direction (Koedinger, Anderson, Hadley, & Mark, 1997; VanLehn & Martin, 1998). But to derive real benefits from the above advances for large-scale assessment requires finding ways to cover a broad range of competencies and to capture rich information about the nature of student understanding. A major problem is that only limited improvements in large-scale assessments are possible under current constraints and typical standardized testing scenarios.

Alternative Approaches to Large-Scale Assessment

Large-scale assessments are designed to meet certain purposes under constraints that often include (a) providing reliable and comparable scores for

individuals as well as groups; (b) sampling a broad set of curriculum standards within a limited testing time per student; and (c) offering cost-efficiency in terms of development, scoring, and administration. To meet such demands, designers typically create assessments that are given at a specified time, with all students taking the same tests under strictly standardized conditions (often referred to as "on-demand" assessment). Tasks are generally of the kind that can be presented in paper-and-pencil format, that students can respond to quickly, and that can be scored reliably and efficiently. In general, the competencies tapped are those that lend themselves to being assessed in these ways, while aspects of learning that cannot be observed under such constrained conditions are not addressed. To design new kinds of situations for capturing the complexity of cognition and learning will require breaking out of the current paradigm to explore alternative approaches to large-scale assessment.

Population sampling versus census testing

If the primary purpose of the assessment is program evaluation, the constraint of having to produce reliable individual student scores can be relaxed, and population sampling can be useful. Instead of all students taking the same test (census testing), a population sampling approach can be used whereby different students take different portions of a much larger assessment, and the results are combined to obtain an aggregate picture of student achievement. The best known example of this is the NAEP, which is a national survey intended to provide policy makers and the public with information about the academic achievement of students across the United States in nine subject areas at grades 4, 8, and 12; thus its coverage is broader than any particular curriculum. The challenge for the NAEP is to assess the breadth of learning goals that are valued across the nation. This is accomplished through a matrix sampling design where each student takes only a small portion of the entire assessment.

Maryland is one of the few states that decided to optimize the use of assessment for program evaluation, forgoing individual student scores. For nearly 10 years, Maryland has used a sampling approach where each student takes only one-third of the entire

Advances in the cognitive sciences, methods of measurement, and technology also have implications for improving classroom assessment practices. A summary of some key implications follows. For a more detailed discussion see NRC (2001).

1. Providing students with frequent information about particular qualities of their work and about what they can do to improve is crucial for maximizing learning.
2. In the classroom, theories of cognition and learning can be particularly helpful by providing teachers with a picture of intermediary states of student understanding on the pathway from novice to competent performance in a subject domain. By assessing a learner's current state of understanding, instruction can center on what is most important for the next stage of learning.
3. Students also have a critical role to play in making classroom assessment effective. They should be taught to ask questions about their own work and revise their learning as a result of reflection. Just as teachers should adopt models of cognition and learning to guide instruction, they should also convey a model of learning (perhaps a simplified version) to their students so the students can monitor their own learning. This can be done through techniques such as getting students to develop scoring rubrics or other criteria for evaluating their work.
4. The effectiveness of classroom assessment rests on a bedrock of informed professional practice. Preservice and professional development are needed to uncover teachers' existing understandings of how students learn, and to help them formulate models of learning so they can identify students' initial understandings and build on those to move students toward more sophisticated understandings.
5. Teachers should not be expected to design all of their own assessment tools themselves. Sophisticated cognitive theories and measurement models can be embedded in easy-to-use instruction and assessment materials for classroom use (e.g., in the form of computerized tutoring systems). A goal for the future is to provide tools that make high-quality classroom assessment more feasible for teachers.

Figure 2. Classroom assessments that support learning.

assessment, which includes a variety of performance tasks such as essays, science laboratory reports, and open-ended mathematical reasoning problems. This means an individual student's results do not give a complete picture of how that child is performing (although parents can obtain a copy of their child's results from the local school system). What is gained is a program evaluation instrument that covers a much more comprehensive range of learning goals than that addressed by a traditional standardized test. Unfortunately, state leaders just recently announced that they are abandoning the performance assessment program in favor of a more traditional, largely multiple-choice, census testing program. Much of the impetus for Maryland's change in course is due to new federal education legislation, which requires individual test scores, at more grade levels, and with speedy turnaround of test results (Hoff, 2002). We return later to discussion of the implications of federal education policies for the future of achievement testing.

Embedded assessment

If individual student scores are required, broader coverage of the domain can be achieved by extracting evidence of student performance from classroom work produced during the course of instruction (often referred to as "curriculum-embedded" assessment). This can be used to supplement the information collected from an on-demand assessment. Although rarely used today for large-scale assessment purposes, curriculum-embedded tasks can serve policy purposes of assessment if there is some standardization (e.g., tasks are centrally determined to some degree; consistent rules are used for scoring student work), with flexibility built in for schools, teachers, and students (e.g., some choices about which tasks from a set of possibilities to use, and when to have students respond to them).

Curriculum-embedded assessment approaches afford additional benefits. In on-demand testing situations, students are administered tasks that are targeted to their grade levels but not otherwise connected to their personal educational experiences. It is this relatively low degree of contextualization that renders these data good for some inferences, but not as good for others (Mislevy, 2000). If the purpose of assessment is to draw inferences about whether students can solve problems using knowledge and experiences they have acquired in school, an on-demand testing situation in which every student receives a test with no consideration of his or her personal instructional history can be unfair. In this case, to provide valuable evidence of learning, the assessment must tap what the student has had the opportunity to learn (NRC, 1999b). Curriculum-embedded assessment offers an alternative to on-demand testing for cases in which there is a strong desire to maintain a correspondence among curriculum, assessment, and actual instruction.

The Advanced Placement (AP) Studio Art portfolio assessment is an example of an assessment designed to certify individual student attainment over a broad range of competencies, and to be closely linked to the actual instruction students have experienced (College Board, 1994). Student work products are extracted during the course of instruction, collected into a portfolio, and then evaluated by a group of artists and teachers. Instructional goals and the criteria by which students' performance will be evaluated are made clear and explicit early on. Numerous readings go into the scoring of each portfolio, enhancing the fairness of the assessment process (Mislevy, 1996). Thus, by using a curriculum-embedded approach, the AP Studio Art program is able to collect rich and varied samples of student work that are tied to students' instructional experiences over the course of the year, but can also be evaluated in a standardized way for the purposes of summative assessment. It should be noted that some states attempting to implement large-scale portfolio assessment programs have encountered difficulties (Koretz & Barron, 1998). Therefore, while this is a good example of an alternative approach to on-demand testing, the authors recognize that there are many implementation challenges to be addressed.

Technology-supported assessment: A vision of the possible

What might happen if assessment of student learning became an integral part of instruction? It is both intriguing and useful to consider the possibilities that arise with intelligent uses of technology to support such an integration of instruction and assessment. One can imagine a future in which the

audit function of large-scale external assessments would be significantly reduced or even rendered unnecessary because the information needed to assess students, at the levels of description appropriate for various assessment purposes, could be derived from the data generated by students in and out of their classrooms. Technology could offer ways of creating over time a complex stream of data about how students think and reason while engaged in important learning activities. Information for assessment purposes could be extracted from this stream and used to serve both classroom and external assessment needs, including providing individual feedback to students for reflection about their states of knowledge and understanding.

A metaphor for this shift exists in the world of retail outlets, ranging from small businesses to supermarkets to department stores. No longer do these businesses have to close down once or twice a year to take inventory of their stock. Rather, with the advent of automated checkout and barcodes for all items, these businesses have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can business continue without interruption, but the information obtained is far richer, enabling stores to monitor trends and aggregate the data into various kinds of summaries. Similarly, with new assessment technologies, schools might no longer have to interrupt the normal instructional process at various times during the year to administer external tests to students, let alone spend large amounts of time preparing to take such tests.

While people are sure to be divided as to the practicality and advisability of pursuing the scenario just described, we offer it as food for thought. Clearly there are a number of associated pragmatic, equity, and privacy issues that would need to be worked through.

Conclusion

To develop large-scale assessments that give a valid picture of students' understanding of school subject matter, signal worthy educational goals to work toward, and provide instructionally useful feedback, we must reflect more deeply about what it is, exactly, that large-scale tests should be measuring. Opportunities exist for better defining the constructs of achievement testing. Advances in the cognitive sciences have illuminated qualities that differentiate beginning from more competent performers in particular subject domains, and advances in measurement and technology have expanded the capability to collect and interpret more complex forms of evidence about student performance.

The recently reauthorized ESEA legislation, with its emphasis on setting high academic standards and measuring students' attainment of those standards, reinforces the needs to clarify and focus educational goals. Unfortunately, the new requirements to quickly put in place census tests for multiple grades, such as yearly testing of all students in grades 3-8 in mathematics and reading, also present some real dangers (Popham, 2002). It is entirely possible that states will abandon efforts to develop new, improved forms of assessment of the type implied by our preceding discussion, and instead resort to the less costly alternative of traditional standardized tests. Such instruments have proven inadequate for supporting learning in the past and would fail to meet the intent of ESEA, which is to establish challenging academic standards and then measure student proficiency relative to those standards. Hopefully, education leaders will accept the conceptual as well as the operational challenge of ESEA. This includes recognizing the need for investing time and resources to pursue the improvement of large-scale assessment so it can provide the information needed to help all students learn and succeed in school.

References

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: Author.

American Federation of Teachers. (1999). *Making standards matter*. Washington, DC: Author.

College Board. (1994). *Evaluating the advanced placement portfolio in studio art*. New York: College Entrance Examination Board and Educational Testing Service.

Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability*. Retrieved April 30, 2002, from http://www.nea.org/issues/high-stakes/buildingtests.html

Finn, C.E., Jr., Petrilli, J.J., & Vanourek, G. (1998). The state of state standards. *Fordham Report* (Vol. 2). Washington, DC: The Thomas B. Fordham Foundation.

Hoff, D.J. (2002, April 3). Md. to phase out innovative testing program. *Education Week*.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30-43.

Koretz, D.M., & Barron, S.I. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS) (MR-1014-EDU). Santa Monica, CA: RAND.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Mislevy, R.J. (2000, December). *The challenge of context*. Presentation at the CRESST Conference, Los Angeles.

Mosher, F.A. (in press). What NAEP really could do. In L.V. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives*. Phi Delta Kappa and Washington, DC: American Educational Research Association and National Center for Education Statistics.

National Academy of Education. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: Author.

National Research Council. (1999a). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. J.W. Pellegrino, L.R. Jones, & K.J. Mitchell (Eds.). Washington, DC: National Academy Press.

National Research Council. (1999b). *High stakes: Testing for tracking, promotion, and graduation*. J.P. Heubert & R.M Hauser (Eds.). Washington, DC: National Academy Press.

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

National Research Council. (2002). *Technology and assessment: Thinking ahead—Proceedings of a workshop*. Washington, DC: National Academy Press.

Pellegrino, J.W. (1992). Understanding what we measure and measuring what we understand. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 275-300). Boston: Kluwer.

Popham, W.J. (2002). Implementing ESEA's testing provisions: Guidance from an independent commission's requirements. Retrieved April 30, 2002, from http://www.aasa.org/issues_and_insights/issues_dept/Commiss.n_Report_Book.pdf

Shepard, L.A. (2000, April). *The role of assessment in a learning culture*. Presidential address presented at the annual meeting of the American Educational Research Association, New Orleans.

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education, 8*, 179-221.

Wiggins, G. (1998). *Educative assessment*: *Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.