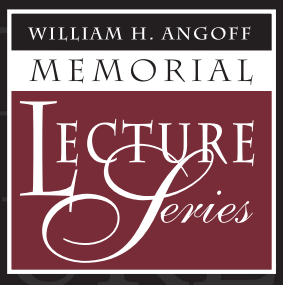
# THE EVOLUTION OF EDUCATIONAL ASSESSMENT: CONSIDERING THE PAST AND IMAGINING THE FUTURE

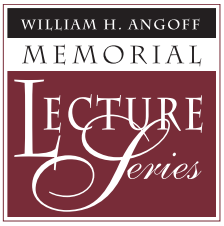
By James W. Pellegrino




Policy Evaluation and Research Center

Policy Information Center

# The Evolution of Educational Assessment:
## *Considering the Past and Imagining the Future*




*The sixth annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on November 17, 1999 This publication represents a modest revision and update of that lecture.*

James W. Pellegrino

University of Illinois at Chicago

Educational Testing Service

Policy Evaluation and Research Center

Policy Information Center

Princeton, NJ 08541-0001

# Preface

In the sixth annual William H. Angoff Memorial Lecture, Dr. James W. Pellegrino of the University of Illinois at Chicago reviews the issues of policy and practice that have had a significant impact on American educational assessment in the 20th century.

Using the past as a prologue for the future, Dr. Pellegrino looks at how current challenges facing educational assessment—particularly the high expectations for educational achievement engendered by standards being set at both state and national levels—are an important impetus for the evolution of the field. Arguing that the educational assessment community needs to substantially improve assessment design and implementation to meet those challenges, he offers insight into how the field can make that leap.

Dr. Pellegrino is a Distinguished Professor of Cognitive Psychology and Education and Co-Director of the Center for the Study of Learning, Instruction and Teacher Development at the University of Illinois, where he joined the faculty in fall 2001. At the time of this lecture, he was the Frank W. Mayborn Professor of Cognitive Studies at Vanderbilt University, where he also served as Dean of Vanderbilt's Peabody College of Education and Human Development (1992-1998) and as co-director of the Learning Technology Center (1989-1992).

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Bill made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Bill's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Drew Gitomer
Senior Vice President
ETS Research & Development
June 2004

---

# Acknowledgments

In addition to the lecturer's scholarship and commitment in the presentation of the annual William H. Angoff Memorial Lecture and the preparation of this publication, ETS Research & Development would like to acknowledge Madeline Moritz for the administrative arrangements, Kim Fryer and Loretta Casalaina for the editorial and layout work involved in this document, Joe Kolodey for his cover design, and, most importantly, Mrs. Eleanor Angoff for her continued support of the lecture series.

# Abstract

Multiple streams of influence, including social policy and societal goals, theories of the mind, and computational capacities, have affected the American educational assessment community over the past century and have prospects for continuing to do so well into the current century. The educational assessment community will have to face major challenges to improve approaches to educational assessment substantially. Solutions to current concerns, respectively denoted as top-down versus bottom-up approaches, address important issues in educational assessment, such as integrating assessment into the learning environment. If such solutions can be implemented, the landscape of educational assessment will be very different and much improved at the end of the current century.

# Introduction

This lecture considers a variety of issues that have impacted the past century of assessment in American education and examines their influence on both where the field has been as well as where it should be headed. The structure of this report consists of five major parts. The first part addresses general issues of policy and practice that have shaped the educational assessment community and that have prospects for continuing to do so well into the current century. The second part develops a case in support of the set of propositions outlined in part one and deliberates about whether the past serves as a prologue for the future. In part three, the argument shifts to imagining the future as it might be. This part looks at solutions to current concerns, respectively denoted as top-down versus bottom-up ways of addressing important issues in educational assessment. The fourth part of this report is devoted to major challenges that the educational assessment community will have to face to improve substantially approaches to educational assessment. The final section of this report imagines what the landscape might look like at the end of the current century, speculating on how things will have changed and the meaning of those changes.

# Part 1: Policy and Practice Issues

A variety of factors is shaping current policy with respect to educational assessment in the United States. American education has moved into a period with high achievement expectations for all children as evidenced by the promulgation of standards for educational attainment in multiple curricular areas. At the current time, virtually every state has developed specific standards with the goal of raising the bar for the education of all children. Simultaneously, Americans want to maintain high standards of excellence for the educational enterprise while arguing that those standards must be promoted for all children. Thus, equity and excellence are coincident and should be part of one educational agenda rather than considered to be separate pursuits.

A second factor shaping educational assessment is increased public demand for accountability, which can be observed every day and in multiple forms especially in the press and in public and political discussions about the need to improve the educational system. The focus on value-added approaches to the evaluation of programs, schools, and the quality of teachers in the assessment community is an obvious manifestation of this demand for accountability. Using a value-added approach as a way to assess how our system is doing and to hold various entities accountable has become part of the everyday discussions among citizens, politicians, and educational professionals. Not coincidentally, external assessments have become the instruments of the accountability movement. Almost every state has compulsory achievement tests at multiple grade levels in multiple subjects, and all are required to have such tests under current legislation (No Child Left Behind Act of 2001).

Assessment has become one of the most pervasive aspects of the American educational landscape. Many see this as a problem. Bob Stake captured the essence of the problem when he stated, "Assessing education well may depend on assessing it less" (Stake, 1999). Another take on the problem is the colloquial expression, "You don't fatten a pig by constantly weighing it." The message here is that weighing the pig won't cause it to grow—you still have to feed it. Many are concerned that the improvement of academic achievement won't come about by constantly assessing it—more is needed. Is this assertion true? Can academic achievement can be improved by constantly assessing it, and, if so, under what circumstances?

I propose that it is not just a matter of quantity or quality. Rather, we can improve educational outcomes through assessment but only when we have better assessment practices. Assessment that is external to an ongoing process of learning and teaching, which includes much of the formal educational assessment in this country, will not produce the desired outcomes by itself. When we combine it with other assessment practices and strategies, however, this kind of assessment can help, but it must be more valid and informative than is currently the case. Assessment that is integral to the process of learning and teaching can impact achievement significantly, but only if it becomes the focus of more efforts to develop academic programs. In other words, this kind of assessment must become an essential part of the design and enactment of contemporary learning environments.

If social and public goals regarding academic achievement are to be attained, then we must make more effort to improve assessment, especially assessment practices that can directly support enhanced outcomes for students. Thus assessment can become part of the solution rather than be part of the problem as many appear to believe at the current time.

# Part 2: Past as Prologue

To develop the case that improved assessment practices can enhance educational achievement, I will show how the past serves as prologue for the future. One can think of this as looking back to look ahead. It strikes me as especially appropriate to do this in the context of the "millennium madness" that occupied so much attention at the turn of the century. Multiple examples of this madness pervaded the press and airwaves. One of my favorites was "SportsCentury," an ESPN program reviewing the careers of the 50 greatest athletes of the 20th Century, as determined by a panel of sports journalists and observers. One can look at the previous century from a variety of perspectives, not simply the 50 greatest athletes or the 100 greatest events. For example, the past century was also the century of mental tests, when educational assessments came into widespread practice. Imagine a program that showcases the 50 greatest assessments of the 20th century. Think of this as ETS's "TestCentury." To determine which assessments merit inclusion, the program could collect votes through a toll-free number such as 1-800-TOP-TEST and or a Web site such as www.toptest.ets.org.

Some of the possible candidates that could be nominated include tests used for college admissions, such as:

- SAT®
- ACT
- Graduate Record Examinations® (GRE®)
- Law School Admission Test (LSAT)
- Graduate Management Admission Test® (GMAT®)

Another set of candidates might include tests to measure intelligence, such as:

- Stanford-Binet
- Wechsler Intelligence Scale for Children (WISC)
- Otis-Lennon
- Peabody Picture Vocabulary Test (PPVT)
- Wechsler Adult Intelligence Scale (WAIS)
- Primary Mental Abilities (PMA)
- Cognitive Abilities Tests
- Differential Aptitude Tests (DAT)
- Armed Service Vocational Aptitude Battery (ASVAB)

And a third candidate set could include assessments for students in kindergarten through twelfth grade, such as:

- National Assessment of Educational Progress (NAEP)
- Iowa Tests of Basic Skills (ITBS)
- Comprehensive Test of Basic Skills (CTBS)
- Stanford Achievement Test
- Trends in International Mathematics and Science Study (TIMSS)

Many of these candidates fit the prediction and selection model—that is, they constitute assessments of aptitude and intelligence. Their primary purpose is to predict performance in an educational environment or to select individuals for entry into those environments.

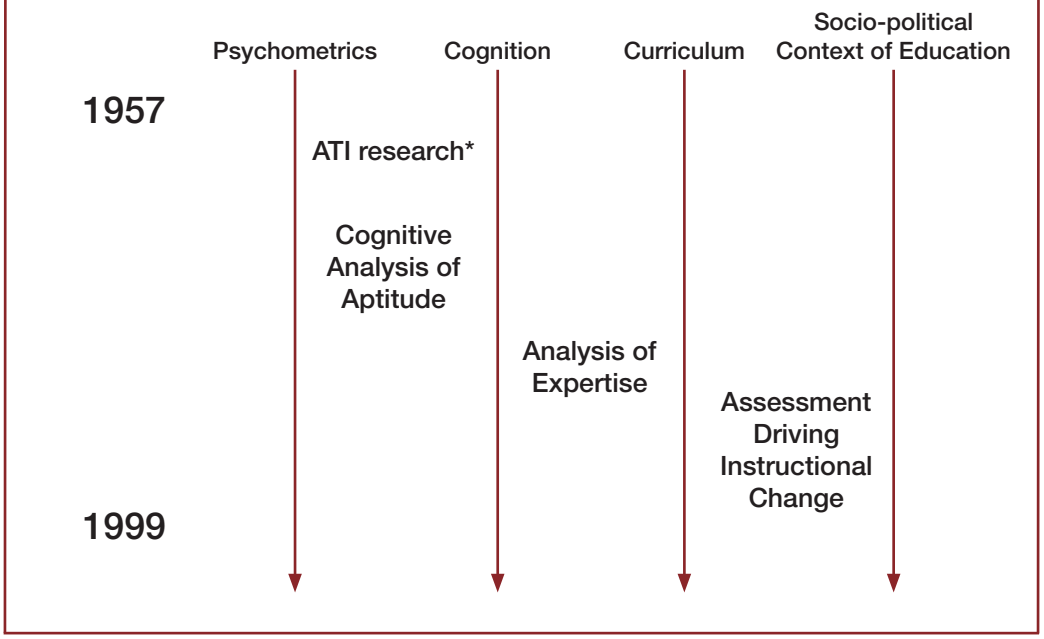
Some of these candidates fit the accountability and audit model. These are primarily assessments of achievement intended to determine how well students and educational systems are doing. Generally, all of the assessments in the top 50 serve the needs of audiences other than the examinee. In fact, they provide minimal direct and immediate feedback to the examinee. The information derived from such instruments is typically used by others who are relatively distal to the process of teaching and learning and who have purposes other than enhancement of the educational outcomes of individual examinees.

## *Major Factors Influencing Assessment*

Reviewing the past century of educational assessment gives rise to some critical questions. Where did the community of educational assessment stand at the close of the first century of mental tests? How did it get to such a place? Is this a comfortable and useful place? What needs to change and why? Meaningful answers to these questions can arise only from understanding that assessment practice is the product of multiple streams of influence, including social policy and societal goals, theories of the mind, and computational capacities. Interaction between these different forces takes many forms and changes over time.

Figure 1 shows four major components that influenced assessment practice—psychometrics, theories of cognition, the nature of curriculum, and the sociopolitical context of education—from 1957 through the present. I have a particularly salient reason for choosing 1957 as a starting point: That was when Lee Cronbach presented his visionary presidential address to the American Psychological Association (APA). Focusing on the two disciplines of scientific psychology, experimental and correlational (or differential) psychology, he described features of both and the benefits of their unification. He proposed linking theories and research on learning and instruction with the tradition of assessing individual differences in cognitive abilities (Cronbach, 1957).

**Figure 1. Four Major Forces That Have Influenced Educational Assessment Practice**




*Aptitude treatment interaction

Since then, individuals have tried bringing the two disciplines together to use assessment productively to support education. As shown in Figure 1, some of the early work involved research in aptitude treatment interactions (ATI) of the type Cronbach discussed in his 1957 presidential address. As cognitive theory developed through the 1960s, becoming richer in its constructs to deal with the nature of human knowledge, a transition occurred. The analysis of individual differences shifted from applying cognitive theory to intelligence and aptitude test performances to studying individual differences in the context of specific instructional or learning domains. This work took the form of the analysis of expertise and expert-novice differences. Thus, from 1957 to roughly 1990, educational assessment underwent an evolution in the attempt to join the study of individual differences with the study of the human mind and bring both fields closer to the domains of learning instruction known as the enterprise of schooling.

## *Using Assessment to Drive Academic Outcomes*

In 1991, Bob Glaser presented a very optimistic view on the state of knowledge as it existed at that point and its applicability to the issues of assessment:

> Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject matter competence can be assessed. As competence in a subject matter grows, evidence of a knowledge base that is increasingly *coherent, principled, useful,* and *goal-oriented* is displayed and test items can be designed to capture such evidence. (Glaser, 1991, p. 26)

Glaser's optimism as expressed in the foregoing quotation has yet to be fulfilled. There were, however, attempts to bring theories and analyses of cognition and expertise together with issues of curriculum and assessment during the 1990s. One major attempt used assessment as the vehicle to promote instructional change. This effort was rooted in the idea that changes in the tests would drive changes in instructional outcomes—in other words, what you test is what you get. The notion that testing limits the nature of teaching is pervasive, and thus the attempts to change tests were rooted in the theory that by testing more significant aspects of cognition, teachers will then focus more on these cognitive performances as part of their instructional repertoire. This

gave rise to the performance-assessment trend, which took hold in a variety of assessment programs such as California and Connecticut state assessments and NAEP. Examples can also be found in the efforts in mathematics and science curriculum reform such as the Quasar Project, as well as various hands-on science curricula.

It is beyond the scope of this lecture to try to review the work done in the area of performance assessment; however, suffice it to say that this work pointed out a number of problematic outcomes such as generalizability concerns. When various complex performances and tasks were embedded in high-stakes, large-scale assessments, the task and method variances were very high and thus the tasks had very little generalizability. In these situations, the design approach was task-centered rather than construct-centered. Another problematic outcome was a range of validity concerns. Baxter & Glaser (1998) provided a content-process space analysis of performance assessments that illustrated many were very weak with respect to the actual versus assumed cognitive demands. Their evidence suggested that many performance assessments, when tailored to fit the constraints of large-scale assessments, became knowledge-lean and process-constrained. Thus, rather than testing thinking and reasoning skills about science concepts and processes, the assessments were following sets of procedures that tapped little in the way of content understanding.

More generally, efforts at using assessment to drive instruction by integrating more complex tasks and performances into large-scale assessments, though well-meaning, revealed a fundamental problem on the assessment side. The problem is the absence of construct-centered design. As stated by Messick (1994),

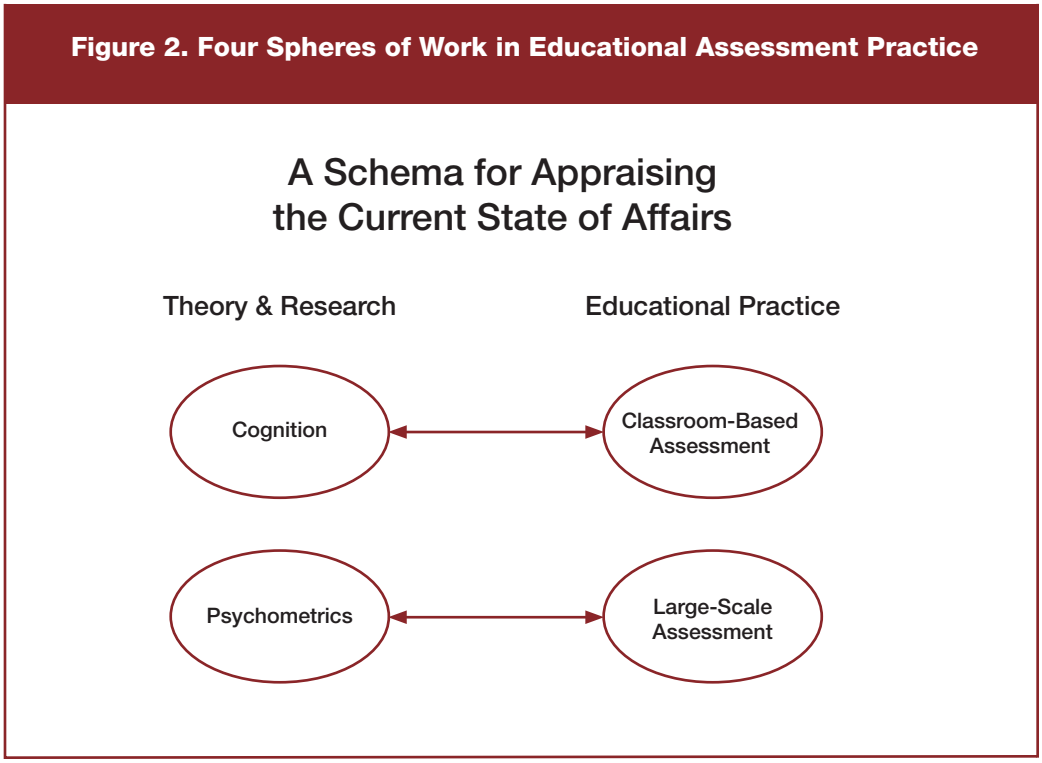
> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs and what tasks or situations should elicit those behaviors? Thus the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

Overall, efforts to use assessment to drive academic outcomes provide relatively little evidence that assessment external to an ongoing process of learning and teaching can in fact produce the desired educational outcomes. Rather, the educational assessment community is becoming increasing aware of the need to embed more valid and complex assessments into the fabric of instruction.

## *Theory and Practice in Educational Assessment*

Figure 2 provides a schema for appraising the current situation with respect to theory and practice in educational assessment. Four components in the figure denote the work in developing theories on the nature of human cognition and learning as well as theories and research on the nature of psychometric measurement. Under the heading of educational practice, the focus is on two levels of assessment: classroom-based assessment and large-scale external assessment. I deliberately depict these four spheres of work as separate entities in the figure. In fact, where links have formed between the spheres, I argue that they have been primarily between work in cognition and classroom-based assessment and work in psychometrics and large-scale assessment. A number of critical links that could connect the cognition and psychometrics spheres, however, are absent from the representation. The same separation exists between classroom-based and large-scale assessment within the educational practice domain. My summary appraisal of progress in the new millennium is that we have made progress in connecting cognitive theory with assessments designed to support learning and instruction at the classroom level. This effort, however, has involved little or no connection with formal psychometric criteria and practices or with large-scale assessment practices. This lack of connection is due in part to deep conceptual conflicts between cognitive theories and many of the psychometric models and practices that drive large-scale testing. After a century of mental tests, much still needs to be accomplished in linking assessment to the improvement of educational outcomes.

**Figure 2. Four Spheres of Work in Educational Assessment Practice**



# Part 3. Imagining the Future

Rather than dwell unnecessarily on the past and the limitations of what has been accomplished, imagine instead the future and consider the paths we need to follow in bringing together psychometrics, cognition, curriculum, and the socio-political context of education. Consideration of assessment-related "mega-trends," another term coined late in the last century, facilitates this imaginary journey. The first of these mega-trends concerns developments in the learning sciences that have major implications for curriculum and instructional practice as well as assessment.

## *Assessment-related Mega-trends*

Recently, the National Research Council (NRC) published two volumes on how people learn (Bransford, Brown, & Cocking, 2000; Donovan, Bransford, & Pellegrino, 2000). Three major findings in these volumes are of particular relevance for the future of assessment. The first is that students come to the classroom with preconceptions about how the world works. If their initial understanding is not engaged, they may fail to grasp the new concepts and information they are taught or they may learn them for purposes of a test but revert to their preconceptions outside the classroom. The implication is that teachers must draw out and work with the preexisting understanding of the subject matter that their students bring with them. Second is the idea that to develop competence in an area of inquiry, students must have a deep foundation of factual knowledge, understand facts and ideas in the context of a conceptual framework, and organize knowledge in ways that facilitate retrieval and application. The implication is that teachers must cover some subject matter in depth, providing many examples in which the same concept is at work, to give students a firm foundation of factual knowledge. Third, teachers need to use a meta-cognitive approach to instruction to help students learn to take control of their own learning by defining learning goals and monitoring their progress in achieving them. The implication is that the teaching of meta-cognitive skills should be integrated into the curriculum in a variety of subject areas.

While many other elements can be derived from the two NRC volumes on how people learn, one of the most critical things to arise from that work is the knowledge about understanding and designing effective learning environments and the components that are contained therein. Powerful learning environments are centered on four components:

- *Knowledge*. In a powerful learning environment, knowledge-centered elements include content organized around core concepts and big ideas. The focus is on the support of subsequent transfer.
- *Learners*. Attention to the knowledge, skills, and attitudes brought by each individual learner are part of the design of a powerful learning environment. A learner-centered approach recognizes the value and necessity of bringing to bear multiple perspectives and sharing them in the context of the learning activity.
- *Assessments*. Assessment-centered elements help make thinking visible to students, teachers, and others in the learning community. Furthermore, these elements support an ongoing process of work and revision that is focused on deepening understanding.
- *Community*. Community-centered elements create a sense of collaboration both within and beyond the immediate boundaries of the

classroom. They can serve to support shared and distributed expertise within the classroom and outside of it.

A second mega-trend to take into account when imagining the future of educational assessments is ubiquitous information technologies. These will have a major impact on the nature of available learning environments and the course of cognitive development. The most salient feature about technology is that it is a means to an end; it provides tools to support the creation and enactment of more powerful learning environments that contain the four components described above. Technology affords a variety of elements for enhancing learning and instruction. First, it supports multiple levels of complexity in the learning environment, and it can assist in the management of complexity. Second, it supports the production of new materials and resources. These include multimedia, simulations, and virtual reality environments, all of which can be focused on some element of knowledge and understanding. Third, it makes information accessible when needed. Fourth, it enables embedded assessment strategies that are integral to the enhancement to the process of learning. Fifth, it supports collaboration and communication and facilitates use of cooperative and distributive expertise in a learning environment.

In addition to developments in the learning sciences and ubiquitous information technologies, there are three other mega-trends: increases in computational power and statistical methods; the dynamics of population change, which will push us even greater in terms of the pursuit of equity and excellence; and, finally, the rhetoric and politics of accountability. These mega-trends form the foundation to derive a set of reasonable projections about the future. Foremost among these projections is that learning environments will change profoundly. Individuals will exert significant control over their learning environments, much more so than is the case presently. As a result, what students can come to know and understand will increase dramatically. Linguistic and cultural variation among learners will also be accommodated as a normal part of the teaching and learning setting rather than as something that is done through external assessments. And, finally, expectations regarding the outcomes from education will increase manyfold. The implications for assessment are profound. In 21st century learning environments, decontextualized, drop-in-from-the-sky assessments consisting of isolated tasks and performances will have zero validity as indices of educational attainments. High-stakes tests will fail badly on all five of Haertel's elements of a complex validity argument (Haertel, 1999). In essence, assessment will need to transform itself to remain relevant and useful.

### *Top-Down and Bottom-Up Solutions*

This transformation needs to come in two forms: top-down and bottom-up solutions to the assessment dilemma. Consider first what might be a top-down policy approach to improving educational outcomes through improved assessment. At the beginning of this lecture, I noted that assessment external to an ongoing process of learning and teaching will not produce the desired educational outcomes by itself. When assessment is combined with other strategies, however, it can have a substantial impact. But for that to happen, it must be more valid and informative than is currently the case. A major example of this argument is the development of a new paradigm for NAEP (Pellegrino, Jones, & Mitchell, 2000). NAEP has several functions in fulfilling its role of providing information to the nation regarding the academic achievement of America's students. Like many external assessments of a large-scale nature, NAEP functions as a social indicator. As such, it fulfills a variety of specific purposes

that include descriptive, evaluative, and interpretive functions. The descriptive function refers to providing detail about exact results in math or science or reading. The evaluative function denotes specifying how good the results are relative to particular sets of standards or goals for accomplishment at particular grades levels in specific subject matter areas. The interpretive function means answering the questions of what the results really mean, what they tell us, and why are they the way they are.

A recent review of the NAEP program argued that NAEP fails to meet the needs of the public policy makers and educators on two of the three important functions of a social indicator. At the interpretive level, it fails because no coherent picture of achievement is provided. Rather, the available information is sparse and generally is not very useful for informing policy in interesting and intelligent ways. At the evaluative level, NAEP also fails because the achievement levels that have been applied to the NAEP scales are largely post hoc attempts to apply meaning and set standards on a measurement instrument never designed for that purpose originally.

To address these problems, a construct-centered approach to NAEP needs to be adopted. Such an approach would view the assessment development process as defined in terms of a vision of student learning and by the inferences and conclusions about student performance that are actually desired in reports of NAEP results. This assessment development process transformation is one that extends from framework development through item development, field-testing, scoring, administration of national samples, and ultimately to reporting. If a construct-centered approach to NAEP is applied and assessments are developed that are guided by this strong vision of student learning and the desired outcome concerning student academic achievement, NAEP can fulfill its multiple functions as a social indicator.

There are, of course, several major issues for incorporating cognitive theory into a 21st century NAEP. One such issue is how to begin incorporating domain-based theories of performance into both the definition of the frameworks as well as into the design of the assessments and their scoring systems. NAEP also needs to move toward inclusion of a wider array of cognitive performances than those currently encompassed by the drop-in-from-the-sky test format that is characteristic of NAEP in its current form. Finally, assessment purpose must be in line with assessment method. Multiple methods including large-scale surveys as well as more complex extended performance tasks must become part of the NAEP portfolio of assessment approaches.

Certainly NAEP has taken some steps in the right direction. One of these is the incorporation of current knowledge into the material design and selection process, for example, in reading assessment, although much more could be done in that area. Another step, although it has yet to reach fruition, is the development of item sets and item families that are more informative about critical aspects of achievement. Examples of these can be found in certain blocks that have been designed for the math and science assessments, although much more needs to be done to pursue the analysis of the data resulting from these item sets and item families. Third, interpretive analyses need to be built in as a part of the initial analysis and reporting process. Thus, NAEP assessments need to be designed with interpretive analyses as a part of the ongoing analysis and reporting process so that interpretive data will actually be part of the item set and the analysis package. Adding on these analyses later introduces significant time delays in providing interpretive information. Finally, achievement levels must be conceptually defined and empirically validated as an integral part of the assessment development process. This

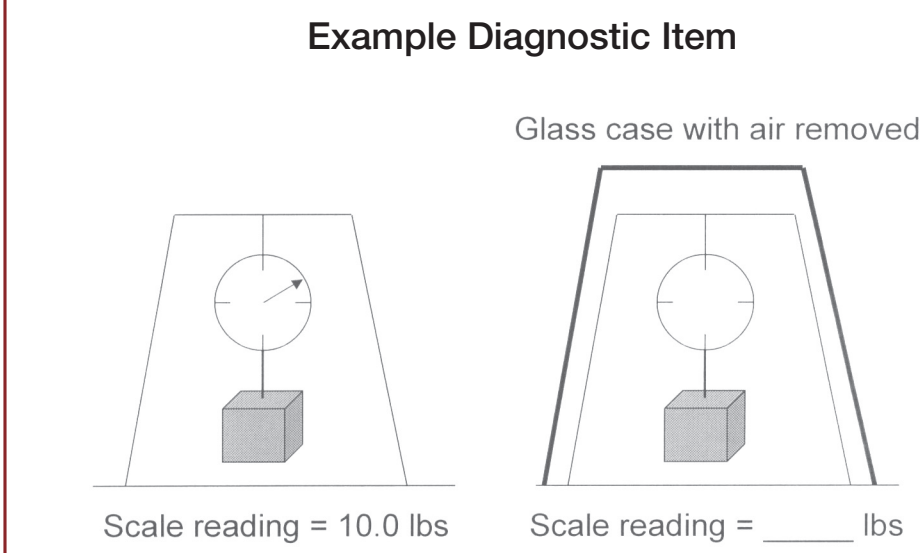
stands in stark contrast to the process that exists now in achievement level setting, although steps to a more conceptually driven achievement level setting design have been incorporated in more recent NAEP assessments.

As much as we need to change the forms of external assessments that provide information about student academic achievement, these changes alone will not be sufficient to improve educational outcomes. Rather, the top-down approach must be complemented by a bottom-up practice-oriented approach if we are to improve educational outcomes through improved assessment. As I argued at the beginning of this lecture, assessment integral to the process of learning and teaching can have a significant impact on achievement, but only if it becomes the focus of considerably more development efforts. In essence, it must become an essential part of the design and enactment of contemporary learning environments.

Fortunately, the decade of the 1990s provided glimpses of the development of theory-driven embedded assessments for classroom instructional practice. One such example comes from high school science and the teaching of physics. Jim Minstrell and colleagues at the University of Washington have conducted extensive research on mapping knowledge in multiple curricular areas. The focus of their work is on knowledge-rich domains such as high school physics with an emphasis on the development of students' conceptual understanding. In this work, assessment stems from deep analyses of how students conceptualize and explain situations. Furthermore, the assessment is integrated into recursive cycles of teaching, learning, and assessment and is aided through technology by using a program called "The Diagnoser" (Hunt & Minstrell, 1994).

Minstrell's work centers on conceptualizing the bits and pieces of knowledge that he refers to as facets. "A facet is a convenient unit of thought, an understanding or reasoning, a piece of content knowledge or strategy seemingly used by the student in making sense of a particular situation" (Minstrell, 1991). Figure 3 presents an example diagnostic item that Minstrell might use at the beginning of an instructional sequence (Minstrell, 2000). The item concentrates on separating the effects of a medium from gravitational effects.

**Figure 3. Example of a Diagnostic Item in Minstrell's System**



By examining the response that students give to such an item, Minstrell (2000) can map their understanding against a facet cluster such as that shown in Table 1. Facets in this cluster range from those that are a correct understanding, such as those at the top numbered 310 and 311, to those numbered 319, which are incorrect and contain significant misconceptions. Facets in the middle represent partial understandings in which the students display reasonable appreciation of some of the elements but not a correct or complete understanding. Such facets are often products of instruction rather than naïve misconceptions the students might bring with them. Also shown in the table is the percentage of responses representing each of the various facets that would occur on a pretest for the item shown in Figure 3.

**Table 1. A Facet Cluster for Separating Medium From Gravitational Effects**

| Facet | Student conceptions about gravitational effects | Frequency of response on pretest |
|---|---|---|
| 310 | Pushed from above and below by a surrounding fluid medium lend a slight support | 3% |
| 311 | A mathematical formulaic approach (e.g., rho x g x h1 – rho x g xh2 = net buoyant pressure) | |
| 314 | Surrounding fluids don't exert any forces or pushes on objects | |
| 315 | Surrounding fluids exert equal pushes all around an object | 35% |
| 316 | Whichever surface has greater amount of fluid above or below the object has the greater push by the fluid on the surface | |
| 317 | Fluid mediums exert an upward push only | 13% |
| 318 | Surrounding fluid mediums exert a net downward push | 29% |
| 319 | Weight of an object is directly proportional to medium pressure on it | 20% |

Note. From "Student Thinking and Related Assessment: Creating a Facet-based Learning Environment," by Jim Minstrell, 2000, in *Grading the Nation's Report Card: Research From the Evaluation of NAEP* (p. 52) edited by Nambury S. Raju, James W. Pellegrino, Meryl W. Bertenthal, Karen J. Mitchell, and Lee R. Jones, 2000, Washington, DC: National Academy Press. Copyright by National Academy of Sciences. Adapted with permission.

Minstrell uses such information to design an instructional strategy that includes the use of a series of diagnostic items that are part of the Diagnoser program. These items are constructed to provide information as to the exact nature of the students' understanding as mapped against a particular facet cluster. Students are asked in each problem to select a particular answer, and they are subsequently asked to justify the answer they chose. The program provides feedback to the student not about whether they are right or wrong but rather feedback that leads them to pursue further analyses of their own understanding. Minstrell's facet-based instruction has proven to be extremely successful in raising the scores of students in typical physics problem-solving. Evidence shows that facet-based instruction, when woven into an overall high school physics course, can significantly impact the performance of individuals all along the math aptitude range. Its value is not specific to individuals at the high or the low end of the aptitude distribution. In further work, Hunt and Minstrell (1994) have also shown that it is possible for other teachers to use the materials and the Diagnoser program to implement facet-based instruction in their own classroom. The results are the same for the other teachers, demonstrating that this is not something unique to the talents of Jim Minstrell as a teacher capable of weaving assessment into his learning environment.

Clearly, the Hunt and Minstrell (1994) work on facet-based instruction is but one example of the process of weaving assessment very carefully into the instructional process. It is a powerful demonstration of the role of careful conceptual and cognitive analysis in the design of effective diagnostic assessments that can become integral to the learning and teaching process. Other examples exist in the literature, although far more work needs to be done for this type of bottom-up approach to impact significantly the educational outcomes for children across all levels of the educational spectrum and in multiple curricular areas.

# Part 4: Challenges To Be Met

Without question, major challenges confront the process of reforming the assessment agenda—whether considering reform using either a top-down or bottom-up approach. One of these challenges is organizing the many disconnected pieces of the educational assessment puzzle into a coherent and more coordinated system. To do so requires the creation of a comprehensive design model that fits multiple purposes, uses, and grain sizes. A second challenge is the development of analytic methods and tools appropriate for the task. Many of the techniques currently used for the design and scoring of assessment items and tasks are insufficient for the kinds of analytic precision that will help reveal information of use to policy makers and practitioners. A third major challenge is communicating with multiple groups about the need for such a shift in thinking about testing and assessment. This is a significant problem.

Shifting the balance in terms of who is the primary assessment consumer is a fourth challenge. At the start of this lecture, I noted many examples of major assessments from the past (and current) century that have individuals other than the student or the learner as the primary consumer. We need to shift things so that the focus of assessment information is more on the level of the learner and the teacher. They should become the primary consumers and benefactors of the information derived from a more intelligent assessment processes. We also have to move from a model that is focused on the current process of disposable items and tests to one focused on the design and craftsmanship of high-quality, reusable assessments. This is perhaps one of the most difficult shifts to make since much of the testing regime is oriented toward the generation of thousands of items that can fulfill limited purposes and then are discarded.

Finally, sophisticated schemes will need to be developed to understand and analyze the cost-benefit consequences of the shift in assessment foci advocated in this report. Engaging in the kinds of assessment development and assessment utilization that I have advocated for programs such as NAEP as well as at the classroom level will involve considerable development and implementation costs. However, we need to consider the trade-off of those costs against the costs now allocated for the types of assessments that are minimally informative and that can have a very deleterious effect on classroom instructional process and overall academic outcomes. We do not yet have the conceptual schemes that will permit us to conduct the kind of cost-benefit analysis that is needed and to argue effectively for the long-term benefits of a richer and deeper model of assessment development and assessment implementation. If social and public goals regarding academic achievement are to be attained, then considerably more effort must be focused on improving assessment, especially assessment practices that can directly support enhanced outcomes for individual students.

# Part 5: Educational Assessment in 2099?

What might the world of educational assessment practice be like at the end of the current century? Consider a platonic ideal, a Unified Republic of Educational Assessment. In this world, theory and research on cognition and analytic methods are part of a larger connected universe, as are classroom and external assessment practices. Theory and research and the ways in which we measure what individuals know are incorporated into the design and the utilization of assessment in educational practice. Classroom-based and external assessments are seamless such that no discontinuities exist and both can be related to the same underlying theory of knowledge and measurement.

I have predictions about life within this platonic republic in the year 2099. First, the public will have a sophisticated understanding and appreciation of assessment that is an outgrowth of experiencing its direct value in their learning. As a consequence, drop-in-from-the-sky assessment designs will be seen as interesting and curious relics of the past. Second, technology-assisted dynamic learning environments will exist for multiple domains of knowledge and skill, with assessment as an integral component of the overall environmental architecture and design. Many of these environments will include intelligent tutoring systems built on powerful domain models and inference engines focused on what and how people learn. Third, the SAT, GRE, NAEP, and the other "top 50" assessments as we know them will be artifacts of history. Information about student competence and achievement will be captured as a part of the normal teaching and learning process. We will have ways to sample and aggregate data to address multiple needs, including the audit and accountability purposes that are so prevalent in driving the assessment agenda as we begin the 21st century. Such a world is worthy of striving for during this, the second century of mental testing.

# References


Baxter G. P., & Glaser R. (1998). The cognitive complexity of science performance assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Mind, brain, experience, & school.* Washington D.C.: National Academy Press.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (2000). *How people learn: Bridging research & practice.* Washington D.C.: National Academy Press.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice Hall.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly, (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice.* Cambridge, MA: MIT Press.

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

Minstrell, J. (1991, March). *Students' facets of physics knowledge and reasoning.* Paper presented at the International Workshop on Research in Physics Learning, University of Bremen.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.). *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44-73). Washington, D.C.: National Academy Press.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.*

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (2000). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, D.C.: National Academy Press.

Stake, R. (1999, May). The goods on American education. *Phi Delta Kappan, 80*(9), 668.

*Visit us on the Web at www.ets.org/research*

*Listening.*
*Learning.*
*Leading.*

08853-36320 • Y64E4 • Printed in U.S.A.

724932