SCIENCE ASSESSMENT MEETING PROCEEDINGS
JUNE 18-19
BOSTON

MEETING PROCEEDINGS

The following is a summary of the sessions from the June 18-19, 2004, meeting conducted by the Mathematics and Science Initiative, Office of the Secretary, U.S. Department of Education (ED). The subject of this meeting was science assessments, and it occurred in Boston in cooperation with the Council of Chief State School Officers Large Scale Assessment Conference. Additional information about the meeting includes an executive summary, an agenda, a list of participants, and background readings. To discuss this meeting in more detail, contact Scott May at scott.may@ed.gov, 202-441-7370 (cell), 202-401-6150 (ED) or Mike Snyder at mike.snyder@ed.gov, 202-401-0245 (ED).

A Call to Action – Susan Sclafani, Assistant Secretary for Vocational and Adult Education and Counselor to the Secretary, U.S. Department of Education

Dr. Sclafani opened the meeting by introducing the various issues to be discussed during the day. She explained the need for collaboration among the three participant groups: assessment directors, publishing companies, and researchers. Further, she stated that the joint efforts of these three groups can bring expertise and insight to bear in putting together effective science assessments for the states, consistent with the testing requirements of No Child Left Behind (NCLB).

Reports from State Assessment Directors

During this session, state assessment directors gave brief presentations regarding the progress of science assessment in their states as well as the major challenges they face. The progress that states have made varies a great deal. For example, some states have been conducting science assessments for years and are now working on improvements to their systems, while other states are just beginning to create assessments for the first time. The group described the following challenges:

- **Science Content**
  - NCLB requires that students be assessed three times in elementary and secondary education, once each in the following grade spans: K-4, 5-8, 9-12. There are important state policy decisions that arise: Should science assessments cover all science content knowledge learned over the grade span or just be a summary of the science learned during the year the test is given? How do states incorporate the scientific process and inquiry into the state assessment tests? How do states align tests to state content standards when there are so many standards to measure?

- In some states, content standards are organized by grade spans that do not coincide with the grades that are to be tested. It can also be difficult to determine what content needs to be tested since multi-year state standards often contain more content than that which can be reasonably tested. Further, test developers have to make decisions in terms of the depth at which key scientific ideas are tested.

- **Item Development**
  A common concern voiced by state assessment directors is how difficult it is to find and retain personnel who are qualified to create test items for the assessments. This is an issue whether the states or the developers are authoring the test items. Field-testing items proves to be a costly and lengthy process for many states. The challenge of using teachers to develop test items was widely discussed. Teachers are close to the actual curriculum taught, know what themes and concepts students understand, and provide credibility to other educators. However, they often do not have training in psychometrics and the science behind item development and use.

- **Test Design**
  There are pressures to create tests that serve multiple purposes, for instance, assessing an individual student's level of understanding and achievement on the one hand while simultaneously reflecting the quality of the school curriculum and instructional program, on the other. Hybrid assessments were suggested that test core content for all students and then use items with selected students that are developed for assessing the effectiveness of the curriculum and school program (matrix sampling).

- **Compatibility**
  Many states now require end-of-course tests at the high school level, and there are challenges in designing an NCLB-required test that is compatible with these other exams. Further, some states may experiment with computer-based delivery systems, and there may be challenges in aligning these with more traditional assessments.

Panel Presentation on What Science Is Important to Assess

Rodger Bybee, Executive Director, Biological Sciences Curriculum Study

Dr. Bybee articulated four key aspects of knowledge that need to be assessed to get a true measure of scientific literacy. He attributed the conceptualization of scientific knowledge to the work done by the American Association for the Advancement of Science (*Science for All Americans*, *Benchmarks for Science Literacy*) and the National Research Council (*National Science Education Standards*), and he urged the states to rely on these documents.

The four types of scientific knowledge that should be assessed are:

- *Content* – the core scientific ideas (i.e., physical, living, and earth systems) that all students should learn. An issue is how to reduce the amount of science that is taught to the most fundamental and essential ideas.
- *Process* – the key types of scientific processes, including inquiry, the nature of science, and the role of science and technology in society.
- *Contemporary issues* – science in a context that students will be able to apply to their lives. Examples of these contexts include health (e.g., obesity), population growth, resource use, and science literacy or how people use science in their role as citizens.
- *Cognitive abilities* – the student's ability to think in ways that are important in science such as engaging in scientific quests and identification, formulating explanations, and posing questions such as: Do students understand the essential role of evidence in science? Do students understand what it takes for the scientific community to use evidence, the rules of debate, and the process of skeptical review? Do students understand what counts as evidence in the scientific community?

Ioannis Miaoulis, President and Director, Boston Museum of Science

Dr. Miaoulis argued for including engineering and technology education along with science education throughout K-12 education. He explained that engineering and technology promote project-based instruction and problem solving, as science and mathematics are made relevant through the study of engineering. As described, technology and engineering are the basis for our way of life, and each student makes more decisions that are related to technology and engineering than any other discipline. Therefore, the incorporation of engineering with math and science is both vital and beneficial to the students.

Discussion

After the presentations of Dr. Rodger Bybee and Dr. Ioannis Miaoulis, participants discussed a variety of issues in science assessment:

- The need for teachers to become more knowledgeable about assessment and the positive impact that can have on assessment systems. More thought needs to be given to ways of helping teachers understand the results of state assessments and how that can impact their teaching.

- Assessment directors should clearly articulate, in advance of developing assessment systems, what type of data they want the system to generate so that the assessment can be designed to meet this goal and provide the results that are needed. A good design approach is to begin the assessment development process by describing what reports the assessment system needs to generate once it is in place.

- Assessment systems should measure a student's ability to apply key scientific ideas to new contexts or situations.

Panel Presentation on New Approaches to Conducting Assessments in Science

Eva Baker, Co-Director, National Center for Research on Evaluation, Standards, and Student Testing, University of California Los Angeles

Dr. Baker introduced the need for a comprehensive, coherent assessment system. This system should be focused on science learning and evidence-based frameworks. In addition, Dr. Baker argued for an assessment system that uses both individual student-level items and matrix-sampled assessments, resulting in hybrid tests that will meet the two key requirements of NCLB - understanding both how well individual students are doing while also reporting out results that can be aggregated on a classroom, school building, district, or state-level basis.

Richard Shavelson, Professor, School of Education, Stanford University

Dr. Shavelson explained three fundamental changes that can be made to assessment systems that will greatly increase their quality. These changes include:

- An assessment of what it means to achieve in science. Unfortunately, the majority of tests today rely almost entirely on testing declarative and procedural knowledge or the ability of students to recite back facts or perform routine procedures. One possible conception of science achievement includes the following five key types of knowledge: declarative (knowing that), procedural (knowing how to), schematic (knowing why), strategic (knowing when, where and how to apply knowledge), and epistemic (knowing how knowledge claims are warranted in science).
- An assessment system that links external science achievement assessment for accountability purposes ("summative assessment") with assessment for learning and teaching improvement ("formative assessment").
- A system that uses recent advances in technology and statistics. An important technological development is having the assessment system go beyond pencil-and-paper to include performance assessment and other dynamic assessments. This capacity, coupled with a system that relies on statistically sampling both students and assessment items, can be a powerful combination. In this way, the assessment system can include more complex and expensive performance-based assessments that are administered not to all, but to statistically representative samples of students. Conclusions can be made that allow us to infer science learning among larger populations of students.

  Dr. Shavelson also noted the importance of determining how one would overlay a developmental model on assessments so one can measure a student's movement from novice to expert in certain skills and understandings.

Robert Sternberg, IBM Professor of Psychology and Education, Yale University

Dr. Sternberg presented a model to assessment that not only measures learning but also can be used for diagnostic purposes. He contrasted the model with current assessment and instructional practices that generally fail students and educational systems. The CAP (creative, analytical, and practical) model employs the following components to teach and assess:

- Creative – the ability to create, invent, discover, and explore;
- Analytical – the skills to analyze, compare, and contrast; and
- Practical – the ability to apply, utilize, use, implement, and practice.

Dr. Sternberg highlighted the following areas where there are issues that need to be resolved in order to implement the CAP model:

- Whether performance assessments and simulations for assessment are at the point where they are practical to use;
- Addressing the "archer" question of what is the right level of abstraction to use; and
- Where the line should be drawn between assessment and instruction.

He presented to the group examples of how the CAP model is implemented.

Discussion

Following these presentations, a lively discussion ensued. Concerns about the validity, reliability, and expense of performance assessments were raised. In addition, a concern was raised about whether "scientific predictions" made during the course of a performance assessment are actually "unscientific guesses." Others held the view that assessment designers need to focus on the cognitive demands of the items. The differing viewpoints express the tension that exists in assessment development and related issues.

Presentations from Publishing Companies

Each publishing company representative spoke briefly about his or her current challenges when working with state education departments. Throughout the session, many of these representatives identified the following challenges and provided helpful suggestions:

- **Issues with the standards**
  Precise alignment of assessment systems with content standards in each state is very important, but it is difficult to achieve since the specificity and coherence of the standards vary widely from state to state. The publishers hoped that national groups could provide states with guidance in improving their standards.

- **Clear expectations**
  State officials need to specify what types of conclusions they need to be able to make from the assessment system, and they need to understand that increasing the types of data can increase the cost of the assessment.

- **Requests for Proposals (RFPs)**
  The RFPs need to be more clear and precise, and the necessary specificity should not be added to the RFPs after the bidding process has begun. Once the bidding process has begun, state officials need to share responses to clarification questions submitted by the bidders with all of the bidders. If states need help bringing specificity to their RFPs, then they should first release an RFI (Request for Information), followed later with an RFP.

- **Recruitment and training of staff**
  It is challenging for the testing companies to recruit and train staff who have both the necessary content knowledge and the essential item-writing skills.

- **Inquiry**
  The inquiry process should be assessed using the core scientific concepts, as opposed to being assessed separately.

- **Demands for individual scores**
  The publishing companies mentioned the difficulty that arises from the demand for individual scores. They explained that individual scores require many items, and that equating issues become a problem as items are released to the public. They know that states need reliable sub-scores in order to be diagnostic about instruction and sub-groups.

Presentation on Alignment of Science Frameworks and Assessments

Darvin Winick, Chairman, National Assessment Governing Board

Dr. Winick explained how the National Assessment Governing Board (NAGB) and the National Center for Education Statistics are developing a science framework and preparing to implement a National Assessment of Educational Progress (NAEP) in science. He cautioned state science assessment developers about the differences between the NAEP and the NCLB tests, but he also suggested areas where the states might be able to borrow from the NAEP effort. His major caution to the group was that the NCLB tests and the NAEP have different purposes, and constructing high quality tests that meet the requirements of NCLB presents a major challenge. This challenge is due primarily to the need to report results by school disaggregated by student sub-group. Given that, he suggested that the states could benefit from NAGB's work in the following areas:

- Quality items—all NAGB item development processes are being refined and even higher quality items are expected in the future.

- Science framework – a new science framework is being constructed, and it might be useful for states to refer to it as they develop their own standards.
- Motivational issues – a major issue that is currently being addressed by the NAGB is how to motivate students, especially those in high school, to take the tests and to try to do well on them.
- Technology and testing – significant effort is being made to continue to incorporate technology solutions into the NAEP testing system.
- Issues unique to science assessment- efforts are underway to address the inherent differences between assessing science and testing mathematics and reading. Plus, problems with students having the necessary reading and math skills to complete the science tests are being considered.

Breakout Sessions

After Dr. Winnick's session, participants separated into three groups: publishing companies, state assessment directors, and researchers. They were asked to consider key questions about how science assessments can be improved. Following the break-out session, each group reported its suggestions to the entire group. Below is a summary of the ideas that came out of each session, with no particular ranking of the ideas in terms of their merit or importance:

Publishing Companies

This group discussed strategies that states should use to improve their RFP processes and increase the quality of their assessment systems. Specifically, they suggest that states make the following improvements:

- **Increase the specificity of the RFP**
    - Describe what types of items should be in the assessment and how many of each type of item they desire.
    - Stipulate whether or not the assessment should include the use of manipulatives.
    - Notify developers in a timely manner if they need to use validity studies.
    - Define who will assume the key roles in the development process, including who will develop items and who will train the item developers.
    - Describe what the needs are for data from the assessment system, including how the data will be reported and to whom. If some types of data are not needed, then that should be included in the RFP as well.
    - Specify the level of cognitive demand and the dimensions of performance that are to be assessed.
    - Articulate the minimum needs and where there is room for fresh ideas or creative approaches.

- **Improve communication during the proposal submission period**
    - Host a forum where there are ongoing opportunities to clarify and ask questions about the RFP.

- Establish a process by which prospective bidders can access all of the questions that are asked by other bidders and that are answered by state assessment officials.
- Avoid making last minute changes to the RFP after it has been released, but if changes need to be made, do not make them close to the submission deadline and make sure all prospective bidders have access to them.

- **Stipulate the science content**
  - Be explicit as to what content standards need to be assessed, including which grade levels should be tested.
  - If the standards are going to be changed over the coming years, articulate how this will impact the assessment system and a desired timeline for how the assessment system will be modified.

- **Improve the timetable of the RFP and decision making process**
  - Lengthen the time between when the RFP is released and when the proposals need to be submitted to the state assessment officials.
  - Give more than one week between the end of the question and answer period and the final submission deadline.
  - Do not change the length of the proposal review period and final vendor selection deadline. This makes it difficult for the test development companies to plan and staff accordingly.

- **Consider using a two stage process**
  - If a state has a lot of uncertainty about its assessment system, then it should consider releasing first a request for information (RFI) that will help it later shape its RFP.
  - States should consider awarding a small contract to a developer to help them define components of their RFP/assessment system that are new or novel.

State Assessment Directors

This group discussed the issues in creating science assessments, and they acknowledged that many of the possible solutions are also applicable to other content areas. The directors discussed three important areas of concern: challenges with working with publishing companies, challenges involved with state consortiums, and the needs of state assessment directors. The following are the main points that emerged from these conversations.

- **Challenges stemming from working with publishing companies**
  - Publishing companies need to be realistic in their endeavors and express to the states the possibilities and impossibilities of the task.
  - The assessment development staff at publishing companies changes too frequently, and they often do not have the necessary content knowledge to create viable and appropriate items.
  - Staff burnout at testing companies seems to cause issues with continuity and progression from year to year in the development process.

- Publishing companies need to be more proactive in asking questions of the states and being "critical friends."
- As the assessment development task is a difficult and complicated one, it would be helpful if testing companies were more flexible and thought out-of-the-box to solve problems.
- State assessment directors would like more input from the testing companies on how to create more diagnostic scoring rubrics.

- **Challenges with consortia of state education agencies**
  - There is often difficulty reaching agreement on operational issues such as whether or not to release items and timing of the test.
  - State departments should provide guidance with their science content and performance standards, because it is hard to decide how much science to assess and at what points in an education science should be assessed.
  - State departments should strive to maintain comparability across classrooms, districts, and years – particularly at the high school level where students are in different courses.
  - States should develop assessment programs with universal design that can test all students (ESL and students with disabilities are particularly challenging).
  - There are almost always major budget issues for funding the assessment systems.
  - Political issues arise regarding content in areas such as evolution and dissection.
  - Security issues arise when multiple states share the same test.

- **Needs of state assessment directors**
  - Opportunities to collaborate with assessment directors from other states on assessment system development and assessment policy.
  - Sharing of information between experienced states and those states that are just beginning the assessment development process.
  - Addressing the problems of staffing at state offices.
  - Better communication with publishing companies.
  - A toolkit to help create RFPs.

Researchers

During the researchers' breakout session, the group discussed a wide variety of topics including the needs of states, aspects of RFPs, ways of gathering evidence and areas of needed research. They made recommendations to the state assessment directors and the test publishers in terms of what improvements can be made immediately to RFPs, what emerging areas they should be aware of, and what areas need additional research.

- **Ways to improve RFPs**
  The research group suggested several practical ways that states can improve the quality of their RFPs, including:

- Acknowledge that there is a lot of cognitive and assessment research that is not typically reflected in RFPs. State assessment people and contractors need to become more aware of this research and build it into their RFPs. The group also acknowledged that research results are often hard to comprehend so they suggested that syntheses of research should be commissioned to summarize what we know in layman's terms.
- Ask for hybrid-design assessment systems that utilize matrix sampling as well as a core set of items that all students are required to complete so that states can provide richer information.
- Require that peer review (including reviews from scientists/recognized science education experts and cognitive scientists) be a part of the item development process.
- Since science tests will be the most expensive to develop, accomplish economies of scale through shared resources such as RFPs, item banks, and networks. These would be both beneficial and cost effective for the states.
- Be explicit on what the test should do.
- Establish criteria in scoring and for developers and reviewers of the items.
- Conduct research on the science tests through validity studies and cognitive labs to examine what questions actually measure.
- Assess for transfer. The Program for International Student Assessment (PISA) may be of guidance here. There are classification schemes that can be of assistance.
- Focus on the assessment system as a whole.

- **Emerging areas**
  The researchers urged the group to follow closely these promising areas, as they can have important implications for improving science tests:
  - Apply lessons learned from integrated and project-based science, particularly with regard to assessing young children.
  - Use technology in areas such as simulations, test taking forums, as well as electronic grading of both multiple choice and open-ended items.
  - Build open-ended items and graphical items that allow multiple scores to be given to a single item.
  - Use the latest research on combining graphics and text.
  - Determine if matrix sampling will extend resources.

- **Needed research**
  The researchers also suggested two areas where much more research is needed, multiple choice item development as well as transfer. In the case of multiple-choice items, researchers are interested in ways that multiple-choice items can be constructed so that they are much more sophisticated. As for transfer, researchers noted the need to research whether students can transfer knowledge to new settings and how long information is remembered. They also noted the need to assess the effects of testing on teaching.

- **Additional areas where help is needed**
  The research group offered the following practical suggestions that can help improve the quality of science assessments:
    - The state officials requested that the research base in the area of performance standards of assessment tests (e.g., basic, proficient, advanced) be synthesized so that it is more easily understood by lay-people.
    - A research seminar should be developed that summarizes what is known in the areas of cognitive science, assessment research, and cognitive demands, and the seminar should be designed to make its content accessible to educators and assessment administrators.
    - Professional development programs should be developed for assessment administrators and teachers that help them understand quality assessment items and systems and learn to better use assessment data.

Group Discussion on Strategies for Helping States – Facilitated by Susan Sclafani

Once the breakout sessions had gathered back into the meeting room, the large group began to consolidate their ideas into possible solutions in helping the states. The researchers, publishing companies, and state assessment directors suggested that ED facilitate the creation of the following elements and activities:

Create and Support an Assessment Toolkit

The group concurred that ED should support the development of an assessment toolkit that is accessible via the Internet and other means. The toolkit can contain these and other elements:

- An easily understood synthesis of the latest research that is relevant to science assessments;
- An RFP guide and template to help states address key aspects early and to learn from exemplary RFP models;
- Widely-accepted testing standards;
- A catalog of best practices in science assessment;
- The Center for Research of Educational Evaluation and Student Testing (CREEST) tool;
- Advice on how to evaluate what you have done if you already have a science assessment;
- Database of exemplary assessment items;
- Model assessment systems;
- Guidance on evaluation of current science assessment; and
- Articulation of future research needs.

Professional Development for Assessment Staff

There was agreement that professional development programs about assessment for state and school district level personnel need to be developed and accessible nationwide. The

group suggested exploring online training. Key to this is that training needs to be centered on how to use data that come out of assessment systems. An example of an approach to this type of training is the University of Maryland assessment certification program.

Support Collaboration

There was a consensus that researchers, testing company experts, content specialists, and assessment directors need opportunities to collaborate. Suggested approaches include conferences, online forums, and just-in-time study groups.