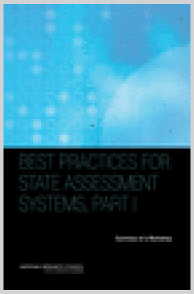


Best Practices for State Assessment Systems Part I: Summary of a Workshop

ISBN
978-0-309-15381-2

94 pages
6 x 9
PAPERBACK (2010)

Alexandra Beatty, Rapporteur; Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards; National Research Council

BEST PRACTICES FOR STATE ASSESSMENT SYSTEMS, PART I

Summary of a Workshop

Alexandra Beatty, *Rapporteur*

Committee on Best Practices for State Assessment Systems:
Improving Assessment While Revisiting Standards

Center for Education

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This project was sponsored by the James B. Hunt, Jr. Institute for Educational Leadership and Policy, with additional support from the Bill & Melinda Gates Foundation and the Stupski Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-15381-2
International Standard Book Number-10: 0-309-15381-6

Additional copies of this report are available from National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Research Council. (2010). *Best Practices for State Assessment Systems, Part 1: Summary of a Workshop.* A. Beatty, Rapporteur. Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON BEST PRACTICES FOR STATE ASSESSMENT SYSTEMS: IMPROVING ASSESSMENT WHILE REVISITING STANDARDS

Diana Pullin (*Chair*), Lynch School of Education, Boston College
Joan Herman, National Center for Research on Evaluation, Standards, and Student Testing
Scott Marion, Center for Assessment, National Center for the Improvement of Educational Assessment, Dover, New Hampshire
Dirk Mattson, Research and Assessment Division, Minnesota Department of Education
Rebecca Maynard, Graduate School of Education, University of Pennsylvania
Mark Wilson, Graduate School of Education, University of California, Berkeley

Judith A. Koenig, *Study Director*
Stuart Elliott, *Director, Board on Testing and Assessment*
Alexandra Beatty, *Senior Program Officer*
Kelly Duncan, *Senior Project Assistant*
Rose Neugroschel, *Research Assistant*

Preface

The idea that states might pool their resources and sign on to a common set of education standards has gone from a speculative concept to an emerging reality. Forty-eight states have now signed on to the "Common Core State Standards Initiative." States are also responding to the opportunity to compete for significant federal education funds through the Race to the Top Assessment Program, which calls on them to make improvements to their standards and assessments. Even before these recent developments, the context for decisions about assessment and accountability was shifting as federal and state policy makers began to take stock of the effects of the No Child Left Behind Act and to consider possibilities for changing it. At the same time, researchers—and a few states—have explored approaches to measuring student learning that are based on theoretical models distinctly different from those that have traditionally governed most state programs. Thus, states are reviewing their approaches to assessment and the role it can and should play in a standards-based accountability system in a complex environment of practical, political, theoretical, and technical questions.

With funding from the James B. Hunt, Jr. Institute for Educational Leadership and Policy, as well as additional support from the Bill & Melinda Gates Foundation and the Stupski Foundation, the National Research Council (NRC) planned two workshops designed to explore some of the possibilities for state assessment systems. The workshops were designed to pull together data and perspectives on the challenges in current assessment and accountability systems and on innovative assessment approaches that offer possibilities for reshaping the expectations educators and policy makers have for this educational tool. The Committee on Best Practices for State Assessment Systems was formed to plan the two workshops. The first workshop, held in December 2009, focused on lessons to be drawn from past experiments with innovative assessments, technical challenges, and the opportunities presented by the current common standards movement; this report describes the presentations and discussions. The second one was held in spring 2010 and provided a more detailed look at possibilities for developing coherent assessment systems that incorporate innovative approaches. This report describes the presentations and discussions from the first workshop; Part II will be available in early summer 2010.

These two workshops were designed to build on two previous ones that examined the possibilities for and questions about common standards for K-12 education. Although a separate committee was responsible for that series (*Common Standards for K-12 Education?: Considering the Evidence, Summary of a Workshop Series,* National Research Council, 2008), we hope that the body of information produced by all of these workshops will provide useful guidance for policy makers in a very fast-changing educational context. But the broader goal for this two-stage activity is to pull together research and perspectives from diverse sources that can contribute to thoughtful deliberation about longer-term questions and goals.

Many people contributed to the success of this first workshop. We would first like to thank the sponsors for their support of this work, particularly Judith Rizzo and Stephanie Dean with the James B. Hunt, Jr. Institute. The committee also recognizes the scholars who wrote papers and made presentations at the workshop, including Steve Ferrara, CTB McGraw-Hill; Margaret Goertz, University of Pennsylvania; Brian Gong, National Center for the Improvement of Educational Assessment; Ron Hambleton, University of Massachusetts-Amhert; Laura Hamilton, RAND; Steve Lazer, Educational Testing Service (ETS); Lorraine McDonnell, University of California, Santa Barbara; Lorrie Shepard, University of Colorado; Brian Stecher, RAND; Shawn Stevens, University of Michigan; Laurie Wise, HumRRO; and Rebecca Zwick, ETS and the University of California, Santa Barbara.

We are also grateful to senior staff members of the NRC's Division of Behavioral and Social Sciences and Education who helped to move this project forward. Michael Feuer, executive director, and Patricia Morison, associate executive director and acting director of the Center for Education, provided support and guidance at key stages in this project. Eugenia Grohman, associate executive director, and Kirsten Sampson Snyder, senior report review officer, offered their knowledge and experience with NRC procedures to guide the report through the NRC review process. Christine McShane, senior editor, provided expert editing assistance, and Yvonne Wise, production editor, adeptly moved this report through the publication process.

The committee also thanks the NRC staff who worked directly on this project. We thank Kelly Duncan, senior project assistant, and Rose Neugroschel, research assistant, both with the Board on Testing and Assessment, for their invaluable assistance in handling the logistical arrangements for the workshop. Their deft organizational skills and careful attention to detail helped to ensure the success of the workshop. We are especially grateful to Judy Koenig, staff director for this workshop project, who played an invaluable role in organizing an informative set of papers and presentations. Judy's wisdom and political skills made certain that the workshop would be a useful contribution to the public discourse on these important topics. We are also grateful to Stuart Elliott, director, and Alix Beatty, senior program officer, of the Board on Testing and Assessment for their contributions in formulating the workshop design and making it a reality. We particularly wish to recognize Alix Beatty for her superb writing skills and ability to translate workshop presentations and discussions into a coherent, readable report.

Finally, as chair of the committee, I wish to thank the committee members for their dedication and outstanding contributions to this project. They gave generously of their time in planning the workshop and actively participated in workshop presentations and discussions. Their varied experiences and perspectives contributed immeasurably to the success of the project and made them a delightful set of colleagues for this work.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the NRC. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We thank the following individuals for their review of this report: Suzanne Lane, Research Methodology Program, School of Education, University of Pittsburgh; Pat Roschewski, Statewide Assessment, Nebraska Department of Education, Lincoln; Theresa Siskind, Division of Accountability, State Department of Education, Columbia, SC; Thomas Toch, Executive Director's Office, Association of Independent Schools of Greater Washington; and Steven L. Wise, Research and Development, Northwest Evaluation Association, Lake Oswego, OR.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft of the report before its release. The review of this report was overseen by Edward H. Haertel of the School of Education of Stanford University. Appointed by the NRC, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author and the institution.

Diana Pullin, *Chair*
Committee on Best Practices for State Assessment Systems:
Improving Assessment While Revisiting Standards

Contents

APPENDIXES

1

Introduction

Educators and policy makers in the United States have relied on tests to measure educational progress for more than 150 years. During the 20th century, technical advances, such as machines for automatic scoring and computer-based scoring and reporting, have supported states in a growing reliance on standardized tests for statewide accountability. The history of state assessments has been eventful. Education officials have developed their own assessments, have purchased ready-made assessments produced by private companies and nonprofit organizations, and have collaborated to share the task of test development. They have tried minimum competency testing; portfolios; multiple-choice, brief and extended constructed-response items; and more. They have responded to calls for tests to answer many kinds of questions about public education and have encountered controversies related to literacy, student privacy, international comparisons, accountability, and even property values.

State assessment data have been cited as evidence for claims about many achievements of public education, and the tests have also been blamed for significant failings. Most recently, the implementation of the No Child Left Behind Act of 2001 has had major effects on public education: some of those effects were intended, and some were not; some have been positive, and some have been problematic.

States are now considering whether to adopt the "common core" academic standards that were developed under the leadership of the National Governors Association and the Council of Chief State School Officers, and they are competing for federal dollars from the Department of Education's Race to the

Top initiative.[1] Both of these efforts are intended to help make educational standards clearer and more concise and to set higher standards for students. As standards come under new scrutiny, so, too, do the assessments that measure their results—to be eligible for Race to the Top funds, a state must adopt internationally benchmarked standards and also "demonstrate a commitment to improving the quality of its assessments" (U.S. Department of Education, 2009).

The goal for this workshop, the first of two, was to collect information and perspectives on assessment that could be of use to state officials and others as they review current assessment practices and consider improvements, as Diana Pullin indicated in her opening remarks. In organizing the workshop, the Committee on Best Practices for State Assessment Systems identified four questions for consideration:

1. How do the different existing tests that have been or could be used to make comparisons across states—such as the National Assessment of Educational Progress (NAEP), the advanced placement (AP) tests, the SAT Reasoning Test (SAT, formerly, the Scholastic Aptitude Test and the Scholastic Assessment Test), ACT (formerly, American College Testing), and the Program for International Student Assessment (PISA)—compare to each other and to the existing state tests with their associated content and performance standards? What implications do the similarities and differences across these tests have for the state comparisons that they can be used to make?
2. How could current procedures for developing content and performance standards be changed to allow benchmarking to measures and predictions of college and career readiness and also promote the development of a small set of clear standards? What options are there for constructing tests that measure readiness with respect to academic skills? Are there options for assessing "21st century" or "soft" skills that could provide a more robust assessment of readiness than a focus on academic skills alone?

3, What does research suggest about best practices in running a state assessment system and using the assessment results from that system to improve instruction? How does this compare to current state capacity and practices? How might assessment in the context of revised standards be designed to move state practices to more closely resemble best practices?

[1] The Race to the Top initiative is a pool of federal money set aside for discretionary grants. States are competing to receive the grants on the basis of their success in four areas: standards and assessments, data systems, improving the teacher work force, and improving the lowest-achieving schools (see http://www2.ed.gov/programs/racetothetop/index.html [March 2010]).

4. How could assessments that are constructed for revised standards be used for accountability? Are there important differences in the use of assessments for accountability if those assessments are based on standards that are (1) shared in common across states, (2) designed to be fewer and clearer, or (3) focused on higher levels of performance?

For this workshop, held in December 2009, the committee focused on lessons to be drawn both from the current status of assessment and accountability programs and the results of past efforts to innovate, and this report describes the presentations and discussion.

Chapter 1 describes current approaches to assessment in the United States and some of the recent developments that have shaped them. Chapter 2 explores possibilities for changing the status quo, by changing both standards and assessments with the goal of improving instruction and learning, as well as some of the technical challenges of implementing innovative approaches on a large scale. Chapter 3 examines practical and political lessons from past and current efforts to implement innovative assessment approaches, and Chapter 4 focuses on the political considerations that have affected innovative assessment programs. Chapter 5 explores the possibilities presented by the current policy environment. The final chapter discusses the research needed to support states' efforts to make optimal use of assessments and to pursue innovation in assessment design and implementation.

The report of the second workshop, held in April 2010, is expected to be published in the summer of 2010.

CONTEXT

Standards-based accountability is a widely accepted framework for public education. Every state now has education standards, although the current array of accountability approaches is characterized by significant variation, as Diana Pullin noted. A previous set of workshops illuminated the extent to which content and performance standards vary in what they formally expect students to know and be able to do at each grade and the extent to which the implementation of state and district standards varies across schools and even classrooms (National Research Council, 2008). By design, assessments play a key role in standards-based accountability, but because standards are not working exactly as they were intended to, Pullin suggested, "assessments can be more powerful in driving what happens in schools than standards themselves." Recent research has indicated that the influence of assessments on curriculum and instruction has increased since 2002, when the No Child Left Behind (NCLB) Act was passed, and that tests themselves have changed in significant ways (see, e.g., Jennings and Rentner, 2006; McMurrer, 2007; Lai and Waltman, 2008;

Sunderman, 2008). Several presenters provided perspectives on the history and current status of assessment and accountability systems.

The idea that assessments should be used to evaluate not just individual students' progress, but also the quality of instruction and the performance of educators more generally, is one with longstanding roots, Joan Herman noted. Edward Thorndike, who published pioneering books on educational measurement in the first decades of the 20th century, viewed his work as useful in part because it would provide principals and teachers with a tool for improving student learning. Ralph Tyler, known for innovative work in educational evaluation, in the 1940s posed the idea that objectives ought to drive curriculum and instruction and that new kinds of assessments (beyond paper-and-pencil tests) were needed to transform learning and the nature of educational programs. Other contributions to thinking about evaluation include Benjamin Bloom's 1956 taxonomy of educational objectives, the development of criterion-referenced testing in the 1950s, mastery learning in the 1960s and 1970s, minimum competency testing in the 1970s and 1980s, and performance assessment in the 1990s. All of these, Herman suggested, have been good ideas, but they have not had the effects that might have been hoped for.

Recent Changes in Tests

Most recently, NCLB has had a very clear impact on many aspects of the system, which Scott Marion described. Prior to 2002, for example, states were required to test at one grade each in the elementary, middle, and high school years. NCLB required testing in grades 3 through 8 as well as in high school. Marion argued that this increased testing resulted in some improvements to state standards. The new requirement compelled states to define coherent content standards by grade level, rather than by grade span, and to articulate more precisely what the performance standards should be for each grade. Testing at every grade has also opened up new possibilities for measuring student achievement over time, such as value-added modeling.[2]

The design of states' test forms has also been affected, most notably in the almost complete elimination of matrix sampling designs because they do not provide data on individual students. For a long time test designers have used matrix sampling to produce data about the learning of large groups of students (such as all children in a single grade) across a broad domain. With this sampling approach, tests are designed so that no one student answers every question (which would require too much testing time for complete coverage), but there is sufficient overlap in tested content from student to student to support inferences about how well the group as a whole has learned each aspect

[2] "Value-added modeling" is statistical analysis in which student data collected over time are used to measure the effects of teachers or schools on student learning.

of the domain tested. One advantage of matrix sampling is that each student faces fewer test items—because student-level scores are not produced—and thus there is time to include more complex item types. This approach makes it possible to better assess a broad academic domain because the inclusion of more complex item types is likely to yield more generalizable inferences about students' knowledge and skills.

The types of test questions commonly used have also changed, Marion suggested, with developers relying far less on complex performance assessments and more on multiple-choice items. He cited evidence from the U.S. Government Accountability Office (2009) that the balance between multiple-choice and open-ended items (a category that includes short items) has shifted significantly in favor of the multiple-choice format as states have responded to the NCLB requirements. Many states still use items that could be described as open ended, but use of this term disguises important differences between a short constructed-response item worth a few points and an extended, complex performance task that contributes significantly to the overall score. To illustrate the difference, he showed sample items—a four-page task from a 1996 Connecticut test that called for group collaboration and included 16 pages of source materials, contrasted with mathematics items from a 2009 Massachusetts test that asked students to measure an angle and record their result or to construct a math problem and solve it.

Marion was not suggesting that the shorter items—or others like them—are necessarily of inferior quality. Nevertheless, he views this shift as reflecting an increased focus on breadth at the expense of depth. The nature of post-NCLB state assessments signals that the most important goal for teachers is to ensure that students have an opportunity to learn a broad array of content and skills, even if the coverage is superficial. It is important to ask, if this is the case, whether the types of processes students use to solve multiple-choice items are truly the basis for the 21st century "college- and work-ready" skills that policy makers stress. For example, he pointed out, the 1996 extended item begins by asking the students to break into small groups and discuss their approach to the task—a challenge much closer to what is expected in many kinds of work than those that are posed by most test items.

States have also changed their approaches to high school testing, though the situation is still in flux. There has been a modest increase in the number of states using end-of-course examinations (given whenever students complete a particular course)—as opposed to survey tests that all students take at a certain point (such as at the end of particular grades). A few states have begun using college entrance tests as part of their graduation requirements.

Interim Assessments

Another development has been a marked increase in the use of interim assessments, particularly at the district level, which Margaret Goertz also pointed out. These are assessments that measure students' knowledge of the same broad curricular goals that are measured in annual large-scale assessments, but they are given more frequently and are designed to give teachers more data on student performance to use for instructional planning. Interim assessments are often explicitly designed to mimic the format and coverage of state tests and may be used not only to guide instruction, but also to predict student performance on state assessments, to provide data on a program or approach, or to provide diagnostic information about a particular student. Researchers stress the distinction between interim assessments and formative assessments, however, because the latter are typically embedded in instructional activities and may not even be recognizable as assessments by students (Perie, Marion, and Gong, 2007).[3]

Districts have vastly increased their use of interim assessments in the past 10 years, Goertz noted (see Stecher et al., 2008), and draft guidelines for the Race to the Top Assessment Program encourage school districts to develop formative or interim assessments as part of comprehensive state assessment systems. However, there have been very few studies of how interim assessments are actually used by individual teachers in classrooms, by principals, and by districts or of their impact on student achievement. Moreover, Goertz pointed out, many of the studies that are cited in their favor were actually focused on formative assessments. Studies are also needed to provide technical and other validity evidence to support inferences made from interim assessments.

In surveys, teachers have reported that interim test results helped them monitor student progress and identify skill gaps for their students, and led them to modify curriculum and instruction (Clune and White, 2008; Stecher et al., 2008; Christman et al., 2009). Goertz also noted a study of how teachers used curriculum-based interim assessments in elementary mathematics in two districts that showed that teachers did use the data to identify weak areas or struggling students and to make instructional decisions (Goertz, Olah, and Riggan, 2009). The study also showed that teachers varied in their capacity to interpret the interim assessment data and to use it modify their teaching. However, the researchers found that few of the items in the interim assess-

[3] Formative assessments are those designed primarily to provide information that students can use to understand the progress of their learning and that teachers can use to identify areas in which students need additional work. Formative assessments are commonly contrasted with summative assessments, those designed to measure the knowledge and skills students have attained by a particular time, often after the instruction is complete, for the purpose of reporting on progress. Interim assessments are assessments, which may be formative or summative, that are given at intervals to monitor student progress.

ments provided information that teachers could readily use, and few actually changed their practice even as they retaught material that was flagged by the assessment results. For example, many teachers focused on procedural rather than conceptual sources of error.

Marion noted that the limited research available provides little guidance for developing specifications for interim assessments or for support and training that would help teachers use them to improve student learning. There is tremendous variability in the assessments used in this way, and there is essentially no oversight of their quality, he noted. He suggested that interim assessments provide fast results and seem to offer jurisdictions eager to respond to the accountability imperative an easy way to raise test scores.

Multiple Purposes for Assessment

Another significant change, Goertz pointed out, is that as demands on state-level assessments have increased in a time of tight assessment budgets, tests are increasingly used for a number of different purposes. Table 1-1 organizes some common testing purposes by goal and by the focus of the information collected. In practice these uses may overlap, but the table illustrates the complexity of the roles that assessments play.

TABLE 1-1 Uses of Assessment

Use	Student	Teacher	School
Diagnosis	Instructional decisions, placement, allocation of educational services	Professional development and support	Resource allocation, technical assistance
Inform Teaching and Leaning		Focus, align, redirect content; instructional strategies	Instructional focus, align curriculum to skills or content; school improvement and planning
Evaluation	Certification of individual achievement	Teacher preparation programs, teacher pay	Program evaluation
Public Reporting	Transcripts		Parent or community action
External Accountability	Promotion, high school graduation	Renewal, tenure, pay	Sanctions and rewards

SOURCE: Goertz (2009).

To clarify the implications of putting a single test to multiple uses, Goertz highlighted the design characteristics that are most important for two of these uses, informing instruction and learning and external accountability. For informing instruction and learning, for example, a test should be designed to provide teachers with information about student learning on an ongoing basis, which they can easily interpret and use to improve their instruction. For this purpose, an assessment needs to provide information that is directly relevant to classroom instruction and is available soon after the assessment is given. Ideally, this kind of assessment would provide continuous information—if it is embedded in instruction it can provide continuous feedback. For this purpose, statistical reliability is not as important as relevance and timeliness.

When test data are to be used for external accountability, the assumption is that information about performance will motivate educators to teach well and students to perform to high standards. Therefore, incentives and sanctions based on test results are often used to stimulate action: this raises the stakes for the tests as well as for students and educators. So, when accountability is the goal, several test characteristics are of critical importance: alignment of test questions to standards; standardization of the content, test administration, and scoring to support fair comparisons; and the fairness, validity, and reliability of the test itself.

The tension between these two purposes is at the heart of many of the problems that states have faced with their assessment programs, and it is a key challenge for policy makers to consider as they weigh improvements to accountability systems.

The growing tendency to use assessments for multiple purposes may be explained in part by the loss of assessment staff in a time of tight education budgets. Marion reported that most states have seen an approximately threefold increase in testing requirements without a corresponding increase in personnel (U.S. Government Accounting Office, 2003; Toch, 2006). As a result, many states have moved from internal test design and development to outside vendors, and, he suggested, remaining staff have less time to work with vendors and to think about innovative approaches to testing.

A number of other factors help explain recent changes in the system, Marion suggested. NCLB required rapid results, and the "adequate yearly progress" formula put a premium on a "head-counting" methodology (that is, measuring how many students meet a particular benchmark by a particular time, rather than considering broader questions about how well students are learning). However, the law did not, in his view, provide adequate financial support for ongoing operational costs. He also said that there has been insufficient oversight of technical quality, with the result that the validity of particular assessments for particular purposes has received inadequate attention. The fact that results for multiple-choice and open-ended items are well correlated has been mistaken for evidence that they are interchangeable. Marion said that an era of tight funding has made it easy for policy makers to conclude that open-ended items and other innovative approaches are not worth their higher cost, without necessarily understanding that such items make it possible to assess skills and content that cannot be assessed with multiple-choice items.

THE CURRENT SYSTEM

This outline of some of the important recent changes in assessment systems provided the foundation for a discussion of strengths and weaknesses of the current system and targets for improvement. Goertz and Marion presented very similar messages, which were endorsed by discussant Joan Herman.

Strengths

Attention to Traditionally Underserved Student Populations Including all students in assessments, to ensure that schools, districts, and states are accountable for their results with every group, was a principal goal of NCLB. As a result, although much work still needs to be done, assessment experts have made important progress in addressing the psychometric challenges of testing students with disabilities and English-language learners. Progress has been made in thinking about the validity of assessments for both of these groups—which are themselves very heterogeneous. Test designers have paid attention to the challenges of specifying constructs to be measured more explicitly and removing construct irrelevant variance from test items (for example, by reducing the reading burden in tests of mathematics so that the measure of students' mathematics skills will not be distorted by reading disabilities). As policy makers and psychometricians have worked to strike an appropriate balance between standardization and technical quality, more assessments are available to measure academic skills—not just functional skills—for special populations. Improved understanding of patterns of language acquisition and the relationship between language skills and academic proficiency have supported the development of better tools for assessing English-language learners across multiple domains.

Increased Availability of Assessment Data The premise of NCLB is that if states and districts had more data to document their students' mastery of educational objectives, they would use that information to improve curricula and instructional planning. States and districts have indeed demonstrated a growing sophistication in the analysis and use of data. Improved technology has made it easier for educators and policy makers to have access to data and to use it, and more people are using it. However, the capacity to draw sound inferences from the copious data to which most policy makers and educators now have access depends on their training. As discussed below, this capacity has in many cases lagged behind the technology for collecting data.

Improved Reporting The combination of stricter reporting requirements under NCLB and improved technology has led states and districts to pay more attention to their reporting systems since 2002. Some have made marked improvements in presenting data in ways that are easy for users to understand and use to make effective decisions.[4]

Weaknesses

Greater Reliance on Multiple-Choice Tests In comparison with the assessments of the 1990s, today's state assessments are less likely to measure complex learning. Multiple-choice and short constructed-response items that are machine-scorable predominate. Though valuable, these item types assess only a limited portion of the knowledge and skills in current standards.

More Focus on Tested Content than on Standards Particularly in low-performing schools, test-based accountability has focused attention on standards, especially the subset of academic standards and content domain that is covered by the tests. Although this focus has had some positive effects, it has also had negative ones. States and districts have narrowed their curricula, placing the highest priority on tested subjects and on the content in those subjects that is covered on tests. Research indicates that the result has been emphasis on lower-level knowledge and skills and very thin alignment with the standards. For example, Porter, Polikoff, and Smithson (2009) found very low to moderate alignment between state assessments and standards—meaning that large proportions of content standards are not covered on the assessments (see also Fuller et al., 2006; Ho, 2008).

More Narrow Test Preparation Because of the considerable pressure to make sure students meet minimum requirements on state assessments, many observers have noted an increased focus on so-called "bubble kids," those who are scoring just below cutoff points. A focus on drilling these children to get them above the passing level may often come at the expense of other kinds of instruction that may be more valuable in the long run. Discussants suggested that this test-preparation focus is particularly prevalent in schools serving poor and traditionally low-performing students and that the emerging result is a dual curriculum, in which already underserved children are not benefiting from the rigorous curriculum that is the ostensible goal of accountability (see, e.g.,

[4] Marion cited the Colorado Department of Education's website as a good example of innovative data reporting (http://www.schoolview.org/ [January 2010]).

McMurrer, 2007). It was noted that this approach often results in less attention to the needs of both high- and low-performing students.

Insufficient Rigor Current assessments are regarded as insufficiently rigorous. Analysis of their cognitive demand suggest that they focus on the lower levels of cognitive demand defined in standards and that they are less difficult than, for example NAEP (see, e.g., Ho, 2008; National Research Council, 2008; Cronin et al., 2009).

Challenges

Many of the challenges that the presenters and discussant identified as most pressing mirrored the strengths and weaknesses. They identified opportunities not only to address the weaknesses, but also to build on many of the strengths in the current system.

Purposes of Testing For Goertz, any plans for improving assessment and accountability systems should begin with clear thinking about several questions: "What do we want to test and for what purpose? What kinds of information do we want to generate and for whom? What is the role of a state test in a comprehensive assessment system? What supports will educators need?" Assessments, whatever their nature, communicate goals to students and teachers. They signal what is valued, in terms of the content of curriculum and instruction and in terms of types of learning. All who play a part in the system listen to the signal that testing sends, and they respond accordingly. Goertz suggested that current approaches may be coherent, in a sense, but many are sending the wrong signals to students and teachers.

The System as a Whole A theme of the discussion was that if tests are bearing too much weight in the current system, it is logical to ask whether every element of an accountability system must be represented by a test score. Measures of students' opportunity to learn and student engagement, as well as descriptive measures of the quality of the teaching and learning, may provide a valuable counterbalance to the influence of multiple-choice testing. It is important to balance the need for external accountability against the other important goals for education.

Thus, in different ways, Marion, Goertz, and Herman each suggested that it is important to evaluate the validity of the entire system, seeking evidence that each element of the system serves its intended purpose. The goal for an accountability system should be to provide appropriate evidence for all intended users and to ensure that those users have the capacity and resources to use the information. The key question, as Herman phrased it, is: "Can we engineer the process well enough that we minimize the negative and maximize the positive consequences?" Clearly, she stressed, it does not make sense to rely on an annual assessment to provide all the right data for every user—or to measure the breadth and depth of the standards.

Looking at the system as a whole will entail not only consideration of intended and unintended consequences, but also a clear focus on the capacity of each element of the system to function as intended. But Goertz pointed out that a more innovative assessment system—one that measures the most important instructional goals—cannot by itself bring about the changes that are desired. Support for the types of curriculum and instruction that research is indicating will foster learning, as well as such critical elements as teacher quality, is also needed.

Staff Capacity to Interpret and Act Many at the workshop spoke about the importance of developing a "culture of data use." Even as much more data has become available, insufficient attention has been paid to developing teachers' and administrators' capacity to interpret it accurately and use it to support their decision making. Ideally, a user-friendly information management system will focus teachers' attention on the content of assessment results so they can easily make correct inferences (e.g., diagnose student errors) and connect the evidence to specific instructional approaches and strategies. Teachers would have both the time to reteach content and skills students have not mastered and the knowledge of effective strategies to target the gaps.

System Capacity Looking more broadly at the capacity issue, Marion noted that there has been a "three- or four-fold increase in the number of tests that are given" without any corresponding increase in assessment personnel. Yet performance or other kinds of innovative assessments require more person-hours at most stages of the process than do multiple-choice assessments. These issues were discussed in the next session of the workshop, described in Chapter 2.

Reporting of Results Although there have been improvements in reporting, it has generally received the least attention of any aspect of the assessment system. NCLB has specific reporting requirements, and many jurisdictions have better data systems and better technology as a result. Nevertheless, even the best reports are still constrained by the quality of the data and the capacity of the users to turn these data into information, decisions, and actions.

2

Improving Assessments— Possibilities and Questions

There is no shortage of ideas about how states might improve their assessment and accountability systems. The shared goal for any such change is to improve instruction and student learning, and there are numerous possible points of attack. But, as discussed in Chapter 1, many observers argue that it is important to consider the system as a whole in order to develop a coherent, integrated approach to assessment and accountability. Any one sort of improvement—higher quality assessments, better developed standards, or more focused curricula—might have benefits but would not lead to the desired improvement on its own. Nevertheless, it is worth looking closely at possibilities for each of the two primary elements of an accountability system: standards and assessments.

DEVELOPING STANDARDS THAT LEAD TO BETTER INSTRUCTION AND LEARNING: AN EXAMPLE

Much is expected of education standards. They are widely viewed as a critical tool in public education because they define both broad goals and specific expectations for students. They are intended to guide classroom instruction, the development of curricula and supporting materials, assessments, and professional development. Yet evaluations of many sets of standards have found them wanting, and they have rarely had the effects that were hoped for (National Research Council, 2008).

Shawn Stevens used the example of science to describe the most important attributes of excellent standards. She began by enumerating the most prevalent criticisms of existing national, state, and local science standards—that they include too much material, do not establish clear priorities among the material included, and provide insufficient interpretation of how the ideas included should be applied. Efforts to reform science education have been driven by standards—and have yielded improvements in achievement—but existing standards do not generally provide a guide for the development of coherent curricula. They do not support students in developing an integrated understanding of key scientific ideas, she said, which has been identified as a key reason that U.S. students do not perform as well as many of their international peers (Schmidt, Wang, and McKnight, 2005). Stevens and her colleagues have developed a model for science standards that are based on current understanding of science learning in order to address some of these weaknesses. She described the model as well as a proposed process for developing such standards and a process for using such standards to develop assessments (Krajcik, Stevens, and Shin, 2009).

Stevens and her colleagues began with the recommendations in a National Research Council (2005) report on designing science assessment systems: standards need to describe performance expectations and proficiency levels in the context of a clear conceptual framework, and be built on sound models of student learning. They should be clear, detailed, and complete; reasonable in scope; and both rigorous and scientifically accurate. She briefly summarized the findings from research on learning in science that are particularly relevant to the development of standards.

Integrated understanding of science is built on a relatively small number of foundational ideas that are central across the scientific disciplines—such as the principle that the natural world is composed of a number of interrelated systems—referred to as "big ideas" (see also National Research Council, 1996, 2005; Stevens, Sutherland, and Krajcik, 2009). These are the sorts of ideas that allow both scientists and students to explain many sorts of observations and to identify connections among facts, concepts, models, and principles (National Research Council, 2005; Smith et al., 2006).

Understanding of the big ideas helps learners develop more detailed conceptual frameworks that, in turn, make it possible to undertake scientific tasks, such as solving problems, making predictions, observing patterns, and organizing and structuring new information. Learning complex ideas takes time and often happens as students work on tasks that force them to synthesize new observations with what they already knew. Students draw on a foundation of existing understanding and experiences as they gradually assemble bodies of factual knowledge and organize it according to their growing conceptual understanding.

The most important implication of these findings for standards is that they must be elaborated so that educators can connect them with instruction, instructional materials, and assessments, Stevens explained. That is, not only should a standard describe the subject matter it is critical for students to know, it should also describe how students should be able to use and apply that knowledge. For example, standards not only should describe the big ideas using declarative sentences, but they should also elaborate on the underlying concepts that are critical to developing a solid understanding of each big idea. Moreover, these concepts and ideas should be revisited throughout K-12 schooling so that knowledge and reasoning become progressively more refined and elaborated. Standards need to reflect these stages of learning. And because prior knowledge is so important to developing understanding, it is important that standards are specific about the knowledge students will need at particular stages to support each new level of understanding. Standards should also address common misunderstandings and difficulties students have learning particular content so that instruction can explicitly target them, Stevens noted. For example, standards can identify some of the most common misconceptions students have about particular material—and perhaps point teachers to more detailed documentation of the "non-normative" ideas researchers have identified.

The approach Stevens and her colleagues have developed for designing rigorous science standards reflects this model and is also based on previous work in the design of curricula and assessments, which they call construct-centered design (Wiggins and McTighe, 1998; Mislevy and Riconscente, 2005; Krajcik, McNeill, and Reiser, 2008; Shin, Stevens, and Krajcik, in press). The name reflects the goal of making the ideas and skills (constructs) that students are expected to learn and that teachers and researchers want to measure the focus for aligning curriculum, instruction, and assessment. The construct-centered design process has six elements that function in an interactive, iterative way, so that information gained from any element can be used to clarify or modify the product of another.

The first step is to *identify the construct*. The construct might be a concept (evolution or plate tectonics), theme (e.g., size and scale or consistency and change), or a scientific practice (learning about the natural world in a scientific way). Stevens used the example of forces and interactions on the molecular and nano scales—that is, the idea that all interactions can be described by multiple types of forces, but that the relative impact of each type of force changes with scale—to illustrate the process

The second step is to *articulate the construct, based on expert knowledge of the discipline and related learning research*. This, Stevens explained, means explicitly identifying the concepts that are critical for developing understanding of a particular construct and defining the successive targets students would reach in the course of their schooling, as they progress toward full understanding of the construct. This step is important for guiding the instruction at various levels.

Box 2-1 shows an example of the articulation of one critical concept that is important to understanding the sample construct (regarding forces and interac-

BOX 2-1
Articulation of the Idea That Electrical Forces Govern Interactions Between Atoms and Molecules

- Electrical forces depend on charge. There are two types of charge—positive and negative. Opposite charges attract; like charges repel.
- The outer shell of electrons is important in inter-atomic interactions. The electron configuration in the outermost shell/orbital can be predicted from the Periodic Table.
- Properties such as polarizability, electron affinity, and electronegativity affect how a certain type of atom or molecule will interact with another atom or molecule. These properties can be predicted from the Periodic Table.
- Electrical forces generally dominate interactions on the nano-, molecular, and atomic scales.
- The structure of matter depends on electrical attractions and repulsions between atoms and molecules.
- An ion is created when an atom (or group of atoms) has a net surplus or deficit of electrons.
- Certain atoms (or groups of atoms) have a greater tendency to be ionized than others.
- A continuum of electrical forces governs the interactions between atoms, molecules, and nanoscale objects.
- The attractions and repulsions between atoms and molecules can be due to charges of integer value, or partial charges. The partial charges may be due to permanent or momentary dipoles.
- When a molecule has a permanent electric dipole moment, it is a polar molecule.
- Instantaneous induced-dipole moments occur when the focus of the distribution shifts momentarily, thus creating a partial charge. Induced-dipole interactions result from the attraction between the instantaneous electric dipole moments of neighboring atoms or molecules.
- Induced-dipole interactions occur between *all* types of atoms and molecules, but increase in strength with an increasing number of electrons.
- Polarizability is a measure of the potential distortion of the electron distribution. Polarizable atoms and ions exhibit a propensity toward undergoing distortions in their electron distribution.
- In order to predict and explain the interaction between two entities, the environment must also be considered.

SOURCE: Krajcik, Stevens, and Shin (2009, p. 11).

tions on the molecular and nano scale). This articulation is drawn from research in this particular topic (such research has not been conducted for many areas of science knowledge), and it describes the upper level of K-12 understanding.

The articulation of the standard for this construct would also address the kinds of misconceptions students are likely to bring to this topic, which are also drawn from research on learning of this subject matter. For example, students might believe that hydrogen bonds occur between two hydrogen atoms or not understand the forces responsible for holding particles together in the liquid or solid state (Stevens, Sutherland, and Kajcik, 2009). This sort of information can help teachers make decisions about when and how to introduce and present particular material, and help curriculum designers plan instructional sequences.

The third step is to *specify the way students will be expected to use the understanding that has been identified and articulated*, a step that Stevens and her colleagues call developing "claims" about the construct. Claims identify the reasoning or cognitive actions students would do to demonstrate their understanding of the construct. (Here also, developers would hope to be able to rely on research on learning in the area in question.) Students might be expected to be able to provide examples of particular phenomena, explain patterns in data, or develop and test hypotheses. For example, a claim related to the example in Box 2-1 might be: "Students should be able to explain the attraction between two objects in terms of the generation of opposite charges caused by an imbalance of electrons."

The fourth step is to *specify what sorts of evidence will constitute proof that students have gained the knowledge and skills described*. A claim might be used at more than one level because understanding is expected to develop sequentially across grades, Stevens stressed. Thus, it is the specification of the evidence that makes clear the degree and depth of understanding that are expected at each level. Table 2-1 shows the claim regarding opposite charges in the context of the cognitive activity and critical idea under which it nests, as well the evidence of understanding of this claim that might be expected of senior high school students. The evidence appropriate at a less advanced level, say for middle school students, would be less sophisticated.

The fifth step is to *specify the learning and assessment tasks that students need to demonstrate, based on the elaborated description of the knowledge and skills students need.* The "task" column in Table 2-1 shows examples.

The sixth step of the process is to *review and revise each product to ensure that they are well aligned with one another*. Such a review might include internal quality checks conducted by the developers, as well as feedback from teachers or from content or assessment experts. Pilot tests and field trials provide essential information, and review is critical to success, Stevens explained.

Stevens noted that she and her colleagues were also asked to examine the draft versions of the common core standards for 12th grade English and mathematics developed by the Council of Chief State School Officers and the National Governors Association, to assess how closely they conform to the construct-centered design approach. They noted that both sets of standards do describe how the knowledge they call for would be used by students, but that the English standards do not describe what sorts of evidence would be neces-

TABLE 2-1 A Claim in Context

Critical Idea	Cognitive Activity	Claim	Evidence	Task
Intermolecular Forces	Construct an explanation	Students will be able to explain attraction between two objects in terms of the production of opposite charges caused by an imbalance of electrons.	Student work product will include – Students explain the production of charge by noting that only electrons move from one object to another object. – Students note that neutral matter normally contains the same number of electrons and protons. – Students note that electrons are negative charge carriers and that the destination object of the electrons will become negative, as it will have more electrons than protons. – Students recognize that protons are positive charge carriers and that the removal of electrons causes the remaining material to have an imbalance in positive charge. – Students cite the opposite charges of the two surfaces as producing an attractive force that hold the two objects together.	*Learning Task:* Students are asked to predict how pieces of tape will be attracted or repulsed by each other. *Assessment Task:* Students are asked to explain why the rubbing of fur against a balloon causes the fur to stick to the balloon.

SOURCE: Krajcik, Stevens, and Shin (2009, p. 14).

sary to prove that a student had met the standards. The mathematics standards appeared to provide more elaboration.

DEVELOPING ASSESSMENTS THAT LEAD TO BETTER INSTRUCTION AND LEARNING: AN EXAMPLE

Assessments, as many have observed, are the vehicle for implementing standards, and, as such, have been blamed for virtually every shortcoming in education. There may be no such thing as an assessment task to which no one will object, Mark Wilson said, but it is possible to define what makes an assessment task good—or rather how it can lead to better instruction and learning. He provided an overview of current thinking about innovative assessment and described one example of an effort to apply that thinking, in the BEAR (Berkeley Evaluation and Assessment Research) System (Wilson, 2005).

In Wilson's view, an assessment task may affect instruction and learning in three ways. First, the inclusion of particular content or skills signifies to teachers, parents, and policy makers what should be taught. Second, the content or structure of an item conveys information about the sort of learning that is valued in the system the test represents. "Do we want [kids] to know how to select from four options? Do we want them to know how they can develop a project and work on it over several weeks and come up with an interesting result and present it in a proper format? These are the sorts of things we learn from the contents of the item," Wilson explained. And, third, the results for an item, together with the information from other items in a test, provide information that can spur particular actions.

These three kinds of influences often align with three different perspectives—with policy makers perhaps most interested in signaling what is most important in the curriculum, teachers and content experts most interested in the messages about implications for learning and instruction, and assessment experts most interested in the data generated. All three perspectives need to be considered in a discussion of what makes an assessment "good."

For Wilson, the most important characteristic of good assessment is coherence. A coherent system is one in which each element (including the summative and formative assessments) measures consistent constructs and contributes distinct but related information that educators can use. Annual, systemwide, summative tests receive the most attention, he pointed out, but the great majority of assessments that students deal with are those that teachers use to measure daily, weekly, and monthly progress. Therefore, classroom assessments are the key to effective instruction and learning, he emphasized. Building classroom assessments is more challenging than building large-scale summative ones—yet most of the resources go to the latter.

In some sense, Wilson pointed out, most state systems are coherent. But he describes the current situation in many as "threat coherence," in which "large-scale summative assessment is used as a driving and constraining force, strait-jacketing classroom instruction and curriculum." He maintained that in many cases the quality of the tests and the decisions about what they should cover are not seen as particularly important—what matters is that they provide robust data and clear guidance for teachers. This sort of situation presents teachers with a dilemma—the classroom tests they use may either parallel the large-scale assessment or be irrelevant for accountability purposes. Thus, they can either focus on the tested material despite possible misgivings about what they are neglecting, or they can view preparing for the state test and teaching as two separate endeavors.

More broadly, Wilson said, systems that are driven by their large-scale assessments risk overlooking important aspects of the curricula that cannot be adequately assessed using multiple-choice tests (just as some content cannot be easily assessed using projects or portfolios). Moreover, if the results of one system-wide assessment are used as the sole or principal indicator of performance on a set of standards that may describe a hundred or more constructs, it is very unlikely that student achievement on any one standard can be assessed in a way that is useful for educational planning. Results from such tests would support a very general conclusion about how students are doing in science, for example, but not more specific conclusions about how much they have learned in particular content areas, such as plate tectonics, or how well they have developed particular skills, such as making predictions and testing hypotheses.

Another way in which systems might be coherent is through common items, Wilson noted. For example, items used in a large-scale assessment might be used for practice in the classroom, or slightly altered versions of the test items might be used in interim assessments, to monitor students' likely performance on the annual assessment. The difficulty he sees with this approach to system coherence is that a focus on what it takes to succeed with specific items may distract teachers and students from the actual intent behind content standards.

When the conceptual basis—the model of student learning—underlying all assessments (whether formative or summative) is consistent, then the system is coherent in a more valuable way. It is even possible to go a step beyond this sort of system coherence, to what Wilson called "information coherence." If this is the goal, one would make sure not only that assessments are all developed from the same model of student learning, but also that they are explicitly linked in other ways. For example, a standardized task could be administered to students in many schools and jurisdictions, but delivered in the classroom. The task would be designed to provide both immediate formative information that teachers can use and consistent information about how well students meet a particular standard. Teachers would be trained in a process that ensured a degree of standardization in both administration of the task and evaluation of the results, and statistical controls could be used to monitor and verify the results. Similarly, classroom work samples based on standardized assignments could be centrally scored. The advantage of this approach is that it derives maximum advantage from each activity. The assessment task generates information and is also an important instructional activity; the nature of the assessment also communicates very directly with teachers about what sorts of learning and instruction are valued by the system (see Wilson, 2004).[1]

The BEAR System (Wilson, 2005) is an example of a system designed to have information coherence. It is based on four principles, each of which has a corresponding "building block" (see Figure 2-1). These elements function in a cycle, so that information gained from each phase of the process can be used to improve other elements. Wilson noted that current accountability systems rarely allow for this sort of continuous feedback and refinement, but that it is critical (as in any engineering system) to respond to results and developments that could not be anticipated.

The construct map defines what is to be assessed, and Wilson described it as a visual metaphor for the ways students' understanding develops, and, correspondingly, how their responses to items might change. Table 2-2 is an example of a construct map for an aspect of statistics, the capacity to consider certain statistics (such as a mean or a variance) as a measure of the qualities of a sample distribution.

The item design specifies how students will be stimulated to respond and is the means by which the match between the curriculum and the assessment is established. Wilson described it as a set of principles that allow one to observe students under a set of standard conditions. Most critical is that the design specifications make it possible to observe each of the levels and sublevels described in the construct map. Box 2-2 shows a sample item that assesses one of the statistical concepts in the construct map above.

The "outcome space" is a general guide to the way students' responses to items developed in relation to a particular construct map will be valued. The more specific guidance developed for a particular item is used as the actual scoring guide, Wilson explained, which is designed to ensure that all of the information elicited by the task is easy for teachers to interpret. Figure 2-2 is the scoring guide for the "Kayla" item, with sample student work to illustrate the levels of performance.

The final element of the process is to collect the data and link it back to the goals for the assessment and the construct maps. The system relies on a multidimensional way of organizing statistical evidence of the quality of the assessment, such as its reliability, validity, and fairness. Item response models show students'

[1] Another approach to using assessments conducted throughout the school year to provide accountability data is the Cognitively-Based Assessment of, for, and as Learning (CBAL), a program currently under development at the Educational Testing Service: see http://www.uk.toeic.eu/toeic/uk/news/?news=900&view=detail [April 2010].

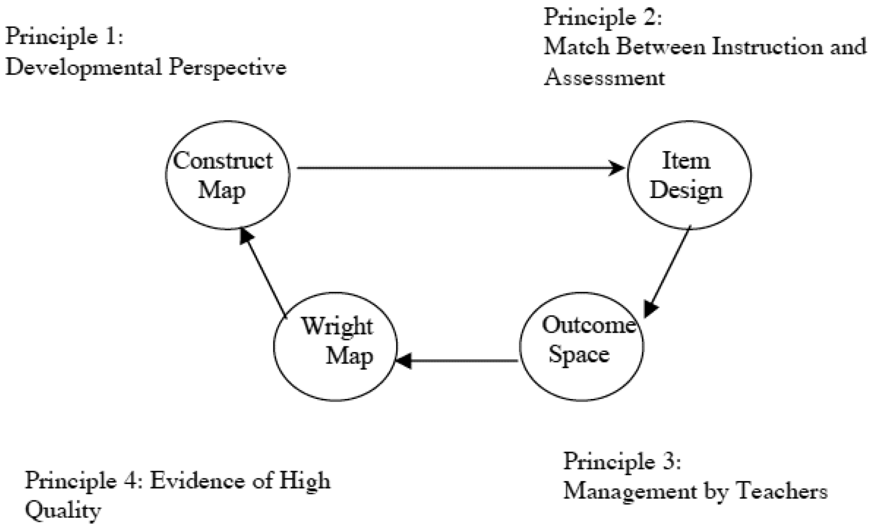


FIGURE 2-1 The BEAR System.
SOURCE: Wilson (2009).

performance on particular elements of the construct map across time and also allow for comparison within a cohort of students or across cohorts.

Wilson closed by noting that a goal for almost any large-scale test is to provide information that teachers can use in the classroom, but that this vision cannot be a reality unless the large-scale and classroom assessments are constructed to provide information in a coherent way. However, he acknowledged that implementing a system such as BEAR is not a small challenge. Doing so requires a deep analysis of the relationship between student learning, the curriculum, and instructional practices—a level of analysis not generally undertaken as part of test development. Doing so also requires a readiness to revise both curricula and standards in response to the empirical evidence that assessments provide.

TECHNICAL CHALLENGES

Stephen Lazer reflected on the technical challenges of pursuing innovative assessments on a large scale from the point of view of test developers. He began with a summary of current goals for improving assessments:

- increase use of performance tasks to measure a growing array of skills and obtain a more nuanced picture of students;
- rely much less on multiple-choice formats because of limits to what they can measure and their perceived impact on instruction;

TABLE 2-2 Sample Construct Map

Conception of Statistics (CoS3): Objective	Student Tasks Specific to CoS3	Student/Teacher Response
CoS3. Consider statistics as measure of qualities of a sample distribution.	CoS3(f) Choose statistics by considering qualities of a particular sample.	– “It is better to calculate median because this data set has an extreme outlier. The outlier increases the mean a lot.”
	CoS3(e) Attribute magnitude or location of a statistic to processes generating the sample.	– A student attributes a reduction in median deviation to a change in the tool used to measure an attribute.
	CoS3(d) Investigate the qualities of a statistic.	– Nick’s spreadness method is good because it increases when a data set is more spread-out.”
	CoS3(c) Generalize the use of a statistic beyond its original context of application or invention.	– Students summarize different data sets by applying invented measures. – Students use average deviation from the median to explore the spreadness of the data.
	CoS3(b) Invent a sharable measurement process to quantify a quality of the sample.	– “In order to find the best guess, I count from the lowest to the highest and from the highest to the lowest at the same time. If I have an odd total number of data, the point where the two counting methods meet will be my best guess. If I have an even total number, the average of the two last numbers of my two counting methods will be the best guess.”
	CoS3(a) Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that the other may not share.	– “In order to find the best guess, I first looked at which number has more than others and I got 152 and 158 both repeated twice. I picked 158 because it looks more reasonable to me.”

SOURCE: Wilson (2009).

- use technology to measure content and skills not easily measured using paper-and-pencil formats and to tailor assessments to individuals; and
- incorporate assessment tasks that are authentic—that is, that ask students to do tasks that might be done outside of testing and are worthwhile learning activities in themselves.

If this is the agenda for improving assessments, he joked “it must be 1990.” He acknowledged that this was a slight overstatement. Progress has been made

BOX 2-2
Sample Item Assessing Conceptions of Statistics

Kayla's Project

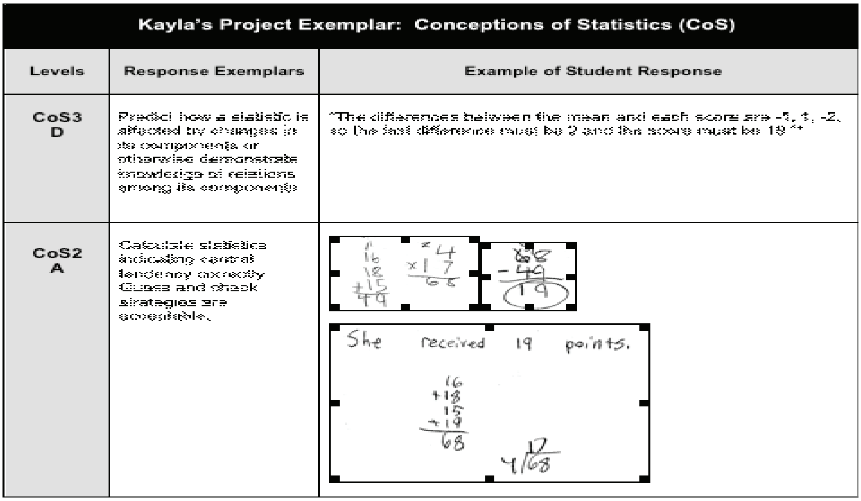
Kayla completes **four projects** for her social studies class. Each is worth **20 points**.

Kayla's Projects	Points Earned
Project 1	16 points
Project 2	18 points
Project 3	15 points
Project 4	???

The mean score Kayla received for ***all four*** projects was **17**.

Use this information to find the **number of points** Kayla received on **Project 4**. Show your work.

SOURCE: Wilson (2009).



Kayla's Project Exemplar: Conceptions of Statistics (CoS)		
Levels	Response Exemplars	Example of Student Response
CoS3 D	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components	"The differences between the mean and each score are -1, 1, -2, so the last difference must be 2 and the score must be 19."
CoS2 A	Calculate statistics indicating central tendency correctly. Guess and check strategies are acceptable.	16 18 +15 49 4 ×17 68 68 −49 19 She received 19 points. 16 +18 15 +19 68 17 4)68

FIGURE 2-2 Scoring guide for sample item.
SOURCE: Wilson (2009).

since 1990, and many of these ideas were not applied to K-12 testing until well after 1990. Nevertheless, many of the same goals were the focus of reform two decades ago, and an honest review of what has and has not worked well during the past 20 years of work on innovative assessments can help increase the likelihood of success in the future.

A review of these lessons should begin with clarity about what, exactly, an innovative assessment is. For some, Lazer suggested, it might be any item format other than multiple choice, yet many constructed-response items are not particularly innovative because they only elicit factual recall. Some assessments incorporate tasks with innovative presentation features, but they may not actually measure new sorts of constructs or produce richer information about what students know and can do. Some computer-based assessments fall into this category. Because they are colorful and interesting, some argue that they are more engaging to students, but they may not differ in more substantive ways from the assessments they are replacing. If students click on a choice, rather than filling in a bubble, "we turn the computer into an expensive page-turner," he pointed out. Moreover, there is no evidence that engaging assessments are more valid or useful than traditional ones, and the flashiness may disguise the wasting of a valuable opportunity.

What makes an assessment innovative, in Lazer's view, is that it expands measurement beyond the constructs that can be measured easily with multiple-choice items. Both open-ended and performance-based assessments offer possibilities for doing this, as does technology. Performance assessments offer a number of possibilities: the opportunity to assess in a way that is more directly relevant to the real-world application of the skills in question, the opportunity to obtain more relevant instructional feedback, a broadening of what can be measured, and the opportunity to better integrate assessment and instruction.

Use of Computers and Technology Computers make it possible to present students with a task that could not otherwise be done—for example, by allowing students to demonstrate geography skills using an online atlas, when distributing printed atlases would have been prohibitively expensive. Equally important, though, is that students will increasingly be expected to master technological skills, particularly in science, and those kinds of skills can only be assessed using such technology. Even basic skills, such as writing, for which most students now use computers almost exclusively whether at home or at school, may need to be assessed by computer to ensure valid results. Computer-based testing and electronic scoring also make it possible to tailor the difficulty of individual items to a test taker's level and skills. Furthermore, electronic scoring can provide results quickly and may make it easier to derive and connect formative and summative information from items.

What are the challenges to using these sorts of innovative approaches? Perhaps most significant is cost. Items of this sort can be time consuming and expensive to develop, particularly when there are few established procedures for this work. Although some items can be scored by machine, many require human scoring, which is significantly more expensive and also adds to the time required to report results. Automated scoring of open-ended items holds promise for reducing the expense and turnaround time, Lazer suggested, but this technology is still being developed. The use of computers may have hidden costs as well. For example, very few state systems have enough computers in classrooms to test large numbers of students simultaneously. When students cannot be tested simultaneously, the testing window must be longer, and security concerns may mean that it is necessary to have wide pools of items and an extended reporting window. Further research is needed to address many of these questions.

Test Development At present, professional test developers know how to produce multiple-choice items with fairly consistent performance characteristics on a large scale, and there is a knowledge base to support some kinds of constructed-response items. But for other item types, Lazer pointed out, "there is really very little in the way of operational knowledge or templates for development." For example, simulations have been cited as a promising example of innovative assessment, and there are many interesting examples, but most have been designed as learning experiments, not assessments. Thus, in Lazer's view, the development of ongoing assessments using simulations is in its infancy. Standard techniques for analyzing the way items perform statistically do not work as well for many constructed-response items as they do for multiple-choice items—and not at all for some kinds of performance tasks. For many emerging item types, there is as yet no clear model for getting the maximum information out of them, so some complex items might yield little data.

The Role of a Theoretical Model The need goes deeper than operational skills and procedures, however. Multiple-choice assessments allow test developers to collect data that supports inferences about specific correlations—for example, between exposure to a particular curriculum and the capacity to answer a certain percentage of a fairly large number of items correctly—without requiring the support of a strong theoretical model. For other kinds of assessments, particularly performance assessments that may include a much smaller number of items, or observations, a much stronger cognitive model of the construct being measured is needed. Without such a model, Lazer noted, or even when a model is well developed, it is also quite possible to write open-ended or computer-based items that are not very high in quality, something he suggested happened too frequently in the early days of performance assessment. The key challenge is not to mistake authenticity for validity; since validity depends on the claim one wants to make, it is very important that the construct be defined accurately and that the item truly measures the skills and knowledge it is intended to measure.

It can also be much more difficult to generalize from assessments that rely on a smaller number of tasks. Each individual task may measure a broader construct than items on conventional tests do, but at the cost of yielding a weaker measure of the total domain, of which the construct is one element. And since the items are likely to be more time consuming for students, they will complete fewer of them. There is likely to be a fairly strong person-task interaction, particularly if the task is heavily contextualized.

It is also important to be clear about precisely what sorts of claims the data can support. For example, it may not be possible to draw broad conclusions about scientific inquiry skills from students' performance in a laboratory simulation related to a pond ecosystem. With complex tasks, such as simulations, there may also be high interdependence among the observations that are collected, which also undermines the reliability of each one. These are not reasons to avoid this kind of item, Lazer said, but he cautioned that it is important to be aware that if generalizability is low enough, the validity of the assessment is jeopardized. Assessing periodically over a time span, or restricting item length, are possible ways of minimizing these disadvantages, but each of these options presents other possible costs or disadvantages.

Scoring Human scoring introduces another source of possible variation and limits the possibility of providing rapid results. In general, the complexity of scoring for some innovative assessments is an important factor to consider in a high-stakes, "adequate yearly progress" environment, in which high confidence in reliability and interrater reliability rates is very important. A certain degree of control over statistical quality is important not only for comparisons among students, but also for monitoring trends. Value-added modeling and other procedures for examining a student's growth over time and the effects of various inputs, such as teacher quality, also depend on a degree of statistical precision that can be difficult to achieve with some emerging item types.

Equating A related problem is equating, which is normally accomplished through the reuse of a subset of a test's items. Many performance items cannot be reused because they are memorable. Even when they can be reused, it can be difficult to ensure that they are scored in exactly the same way across administrations. It could be possible to equate using other items in the test, but if they are of a different type (e.g., multiple choice), the two parts may actually measure quite different constructs, so the equating could actually yield erroneous results. For similar reasons, it can be very difficult to field-test these items, and though this is a mundane aspect of testing, it is very important for maintaining the quality of the information collected.

Challenges for Students Items that make use of complex technology can pose a challenge to students taking the tests as well. Lazer identified two cautions: first, it may take time for students to learn to use the interface and perform the activities the test requires, and, second, the complexity may affect the results in undesired ways. For example, some students may score higher because they have greater experience and facility with the technology, even if their skill with the construct being tested is not better than that of other students.

Conflicting Goals For Lazer, the key concern with innovative assessment is the need to balance possibly conflicting imperatives. He suggested that the list of goals for new approaches is long: assessments should be shorter and cheaper and provide results quickly; they should include performance assessment; they should be adaptive; and they should support teacher and principal evaluation. What this list highlighted for him was that some of the challenges are "know-how" ones—that can presumably be surmounted with additional effort and resources. Others are facts of life. Psychometricians may be working from outdated cognitive models, and this can be corrected. But it is unlikely that further research and development will make it possible to overcome the constraints imposed when reliability and generalizability, for example, are important to the results.

"This doesn't mean we should give up and keep doing what we're doing," he concluded. These are not insurmountable conflicts, but each presents a tradeoff. For him the greatest risk is in failing to consider the choices. To develop an optimal system will require careful thinking about the ramifications of each feature. Above all, Lazer suggested, "we need to be conscious of the limits of what any single test can do." He enthusiastically endorsed the systems approaches described earlier in the day, in which different assessment components are designed to meet different needs, but in a coherent way.

DISCUSSION SUMMARY

A few themes emerged in discussion. First, the time and energy required for the innovative approaches described—specifying the domain, elaborating the standards, validating that model of learning—is formidable. Taking this path would seem to require a combination of time, energy, and expertise that is not typically devoted to test development. However, the BEAR example seemed to marry the expertise of content learning, assessment design, and measurement in a way that offers the potential to be implemented in a relatively efficient way. The discussion of technical challenges illustrated the many good reasons that the current testing enterprise seems to be stuck in what test developers already know how to do well.

This situation raised several questions for presenters and participants. First: Is the $350 million total spending planned for the Race to the Top Initiative enough? Several participants expressed the view that assessment is an integral aspect of education, whether done well or poorly, but that its value could be multiplied exponentially if the resources were committed to develop a coherent system. Second: What personnel should be involved in the development of new assessment systems? The innovation required may be more than existing test publishers could be expected to produce on their own. A participant joked that the way forward might lie in locking cognitive psychologists and psychometricians up together until they resolved their differences—the challenge of balancing the competing imperatives each group raises is not trivial.

A final word from another participant, that "it's the accountability, stupid," reinforced a number of other comments. That is, the need to reduce student achievement to a single number derives from the punitive nature of the current accountability system. It is this pressure that is responsible for many of the constraints on the nature of assessments. Changing that context might encourage states to view assessments in a new light and make use of creative thinking.

3

Innovative Assessment— Lessons from the Past and Present

Chapter 2 described ideas for moving forward and questions about the challenges for innovative assessments. But since the early 1990s many programs pushed past theoretical discussions and small-scale experiments to implement assessments that were in some sense innovative. Some have not been sustainable—for a range of reasons—but others are ongoing. What made these programs innovative? What can be learned from them? Brian Stecher and Laura Hamilton provided an overview of those programs, and a panel of veterans of some of these programs offered their conclusions about them.

The current typical test—multiple choice, paper and pencil—was innovative when it was introduced on a large scale in the early 20th century, Stecher pointed out, but is now precisely the sort that innovators want to replace. So, in that sense, an innovative assessment could be defined simply as one that is not a multiple-choice, paper-and-pencil test. That is, a test might be innovative because it:

- incorporates prompts that are more complex than is typical in a printed test, such as hands-on materials, video, or multiple types of materials;
- offers different kinds of response options, such as written responses, collections of materials (portfolios), or interactions with a computer—and therefore requires more sophisticated scoring procedures; or
- is delivered in an innovative way, usually by computer.

These aspects of the structure of assessments represent a variety of possibilities that are important to evaluate carefully. Several other themes are worth exploring across programs, such as the challenges related to technical quality (e.g., reliability, fairness, and validity), as discussed in Chapter 2. Tests with innovative characteristics (like any tests) send signals to educators, students, and parents, about the learning that is most valued in the system—and in many cases innovative testing has lead to changes in practice. Testing also has costs, including a burden in both time and resources, which are likely to be different for different innovative assessments. Testing also provokes reactions from stakeholders, particularly politicians.

LOOKING BACK

In 1990 performance and other kinds of alternative assessments were popular in the states, with 24 of them using, developing, or exploring possibilities for applying one of these approaches (Stecher and Hamilton, 2009). Today they are much less prevalent. States have moved away from these approaches, primarily for political and budget reasons, but a look at several of the most prominent examples highlights some lessons, as Brain Stecher explained in a synopsis. Individuals who had experience with several of the programs added their perspectives.

Vermont

Vermont was a pioneer in innovative assessment, having implemented a portfolio-based program in writing and mathematics in 1991 (Stecher and Hamilton, 2009). The program was designed both to provide achievement data that would permit comparison of schools and districts and to encourage instructional improvements. Teachers and students in grades 4 and 8 collected work to represent specific accomplishments, and these portfolios were complemented by a paper-and-pencil test.

Early evaluations raised concerns about scoring reliability and the validity of the portfolio as an indicator of school quality (Koretz et al., 1996). After efforts to standardize scoring rubrics and criteria for selecting student work, reliability improved, but evaluators concluded that the scores were not accurate enough to support judgments about school quality.

The researchers confirmed that teachers altered their practice: for example, they focused more on problem solving in mathematics. Many schools began using portfolios in several subjects because they found them useful. However, some critics observed that teachers did not uniformly demonstrate a clear understanding of the intended criteria for selecting student work, and others commented that teachers began overemphasizing the specific strategies that were included on the standardized rubrics. Costs were high—$13 per student just for scoring. The program was discontinued in the late 1990s, primarily because of concerns about the technical quality of the scores.

Kentucky

The Kentucky Instructional Results Information System (KIRIS) was closely watched because it was part of a broad-based response to a state Supreme Court ruling that the education system was unconstitutional (Stecher and Hamilton, 2009). The assessment as a whole covered reading, writing, social science, science, mathematics, arts and humanities, and practical living/vocational studies. The state made significant changes to its schools and accountability system, and it implemented performance assessment in 1992. The program was designed to support school-level accountability; other indicators, such as dropout, attendance, and teacher retention rates, were also part of the accountability system.

Brian Gong described the assessment program, which tested students in grades 4, 8, and 12 using some traditional multiple-choice and short-answer tests, but relied heavily on constructed-response items (none shorter than half a page). KIRIS used matrix sampling to provide school accountability information. Many performance assessments asked students to work both in groups and individually to solve problems and to use manipulatives in hands-on tasks. KIRIS included locally scored portfolios in writing and mathematics.

Evaluations of KIRIS indicated that teachers changed their practice in desirable ways, such as focusing greater attention on problem solving, and they generally attributed the changes they made to the influence of open-ended items and the portfolios (Koretz et al., 1996). Despite the increased burden in time and resources, teachers and principals supported the program.

As with the Vermont program, however, evaluators found problems with both reliability and validity. The portfolios were assigned a single score (in the Vermont program there were scores for individual elements), and teachers tended to assign higher scores than the independent raters did. Teachers reported that they believed score gains were more attributable to familiarity with the program and test preparation than to general improvement in knowledge and skills. This concern was supported by a research finding that teachers tended to emphasize the subjects tested in the grades they taught, at the expense of other subjects. While KIRIS scores trended upward, scores on the National Assessment of Educational Progress (NAEP) and the American College Testing Program for Kentucky students did not show comparable growth (Koretz and Barron, 1998), KIRIS was replaced with a more traditional assessment in 1998 (which also included constructed-response items), although the state continued to use portfolios to assess writing until 2009.

Maryland

The Maryland School Performance Assessment Program (MSPAP), also implemented in 1991, assessed reading, writing, language usage, mathematics, science, and social science at grades 3, 5, and 8 (Stecher and Hamilton, 2009).

The program was designed to measure school performance and to influence instruction; it used matrix sampling to cover a broad domain and so could not provide individual scores. The entire assessment was performance based, scored by teams of Maryland teachers.

There were no discrete items, Steve Ferrara noted. All the items were contained within tasks organized around themes in the standards; many integrated more than one school subject, and many required group collaboration. The tasks included both short-answer items and complex, multipart response formats. MSPAP included hands-on activities, such as science experiments and asked students to use calculators, which was controversial at the time. Technical reviews indicated that the program met reasonable standards for both reliability and validity, although the group projects and a few other elements posed challenges. Evaluations and teacher reports also indicated that MSPAP had a positive influence on instruction. However, some critics questioned the value of the scores for evaluating schools, noting wide score fluctuations. Others objected to the "Maryland learning outcomes" assessed by the MSPAP. The MSPAP was replaced in 2002 by a more traditional assessment that provides individual student scores (which was a requirement of the No Child Left Behind [NCLB] Act).

Washington

The Washington Assessment of Student Learning (WASL), which was implemented beginning in 1996, assessed learning goals defined by the state legislature: reading; writing; communication; mathematics; social, physical, and life sciences; civics and history; geography; arts; and health and fitness (Stecher and Hamilton, 2009). The assessment used multiple-choice, short-answer, essay, and problem-solving tasks and was supplemented by classroom-based assessments in other subjects. WASL produced individual scores and was used to evaluate schools and districts; it was also expected to have a positive influence on instruction.

Evaluations of WASL found that it met accepted standards for technical quality. They also found some indications that teachers adapted their practice in positive ways, but controversy over its effects affected its implementation. For example, the decision to use WASL as a high school exit exam was questioned because of low pass rates, and fluctuating scores raised questions about its quality. The WASL was replaced during the 2009-2010 school year with an assessment that uses multiple-choice and short-answer items. However, the state has retained some of the classroom-based assessments.

A participant with experience in Washington pointed out several factors that affected the program's history. First, the program imposed a large testing burden on teachers and schools. After NCLB was passed, the state was administering eight tests in both elementary and middle schools, with many performance assessment features that were complex and time consuming. Many people were caught off guard by the amount of time the testing consumed. This initial reaction to the program was compounded when early score gains were followed by much slower progress. The result was frustrating for both teachers and administrators.

This frustration, in turn, fueled a growing concern in the business community that state personnel were not managing the program well. The participant said that the initial test development contract was very inexpensive, considering the nature of the task, and when the contract was rebid costs escalated dramatically. And then, as public opinion turned increasingly negative about the program, the policy makers who had initially sponsored it and worked to build consensus in its favor left office, because of the state's term limit law, so there were few powerful supporters to defend the program when it was challenged. This program was also replaced with a more traditional one.

California

The California Learning Assessment System (CLAS), which was implemented beginning in 1993, assessed reading, writing, and mathematics, using performance techniques such as group activities, essays, and portfolios (Stecher and Hamilton, 2009). Some items asked students to reflect on the thinking that led to their answers. Public opposition to the test arose almost immediately, as parents complained that the test was too subjective and even that it invaded students' privacy by asking about their feelings. Differences of opinion about CLAS led to public debate of larger differences regarding the role assessment should play in the state. Questions also arose about sampling procedures and the objectivity of the scoring. The program was discontinued after only 1 year (Kirst and Mazzeo, 1996).

NAEP Higher-Order Thinking Skills Assessment Pilot

An early pioneering effort to explore possibilities for testing different sorts of skills than were generally being targeted in standardized tests was conducted by NAEP in 1985 and 1986 (Stecher and Hamilton, 2009). NAEP staff developed approximately 30 tasks that used a variety of formats (paper and pencil, hands-on, computer administered, etc.) to assess such higher-order mathematics and science skills as classifying, observing and making inferences, formulating hypotheses, interpreting data, designing an experiment, and conducting a complete experiment. NAEP researchers were pleased with the results in many respects, finding that many of the tasks were successful and that conducting hands-on assessments was both feasible and worthwhile. But the pilots were expensive and took a lot of time, and school administrators found them demanding. These item types were not adopted for the NAEP science assessment after the pilot test.

Lessons from the Past

For Stecher, these examples make clear that the boldest innovations did not survive implementation on a large scale, and he suggested that hindsight reveals several clear explanations. First, he suggested that many of the programs were implemented too quickly. Had developers and policy makers moved more slowly and spent longer on pilot testing and refining, it might have been possible to iron out many of problems with scoring, reporting, reliability, and so forth.

Similarly, many of the states pushed forward with bold changes without necessarily having a firm scientific foundation for what they wanted to do. At the same time, the costs and the burdens on students and schools were high, which made it difficult to sustain support and resources when questions arose about technical quality. People questioned whether the innovations were worth the cost and effort.

Another factor, Stecher said, is that many states did not adequately take into account the political and other concerns that would affect public approval of the innovative approaches. In retrospect, it seemed that many had not done enough to educate policy makers and the public about the goals for the program and how it would work. One reason for this lack, however, is that states were not always able to reconcile differences among policy makers and assessment developers regarding the role the assessment was to play. When there was a lack of clarity or agreement about goals, it was difficult to sustain support for the programs when problems arose. And a final consideration for many states was the need to comply with NCLB requirements.

CURRENT INNOVATIONS

Even though many of the early programs did not survive intact, innovative assessment approaches remain in wide use. Laura Hamilton reviewed current examples of three of the most popular innovations: performance assessment, portfolios, and technology-supported assessment.

Performance Assessment

Essays are widely used in K-12 assessments today, particularly in tests of writing and to supplement multiple-choice items in other subjects (Stecher and Hamilton, 2009). Essays have been incorporated into the SAT and other admissions tests and are common in NAEP. They are also common in licensure and certification tests, such as bar examinations.

The K-12 sector is not currently making much use of other kinds of performance assessment, but other sectors in the United States are, as are a number of programs in other countries. One U.S. example is the Collegiate Learning Assessment (CLA), which measures student learning in colleges and universities, is administered on-line, and uses both writing tasks and performance tasks in response to a wide variety of stimuli.

The assessment system in Queensland, Australia, is designed to provide both diagnostic information about individual students and results that can be compared across states and territories. It includes both multiple-choice items and centrally developed performance tasks that can be used at the discretion of local educators and are linked to the curriculum. At the secondary level, the assessment incorporates not only essays, but also oral recitations and other performances. Performance tasks are scored locally, which raises concerns about comparability, but school comparisons are not part of the system, so the pressure is not as heavy on that issue as in the United States. Indeed, Hamilton noted, many aspects of Queensland's system seem to have been developed specifically to avoid the problems associated with the U.S. system, such as score inflation and narrowed curricula.

Other programs use hands-on approaches to assess complex domains. For example, the U.S. Medical Licensing Examination (USMLE) has a clinical skills component in which prospective physicians interact with patients who are trained participants. The trained patient presents a standardized set of symptoms so that candidates' capacity to collect information, perform physical examinations, and communicate their findings to patients and colleagues can be assessed. Hamilton noted that this examination may be the closest of any to offering an assessment that approximates the real-life context for the behavior the assessment is designed to predict—a key goal for performance assessment. Nevertheless, the program has encountered technical challenges, such as limited variability among tasks (the standardized patients constitute the tasks), interrater reliability, and the length of time required (8 hours) to complete the assessment.

These examples, Hamilton noted, illustrate the potential that performance assessment offers, but also demonstrate the challenges in terms of cost, feasibility, and technical quality. For example, sampling remains a difficult problem in performance assessment. Multiple tasks are generally needed to support inferences about a particular construct, but including many tasks can pose a significant burden.

Another difficulty is the tension between the goal of producing scores that support comparisons across schools or jurisdictions and the goal of using the assessment process itself to benefit teachers and students. The Queensland program and the essay portion of the bar exams administered by states both involve local educators or other local officials in task selection and scoring, and this may limit the comparability of scores. When the stakes attached to the results are high, centralized task selection and scoring may be preferred, but at a cost in terms of involving teachers and affecting instruction. Hamilton also noted that none of the examples operated with a constraint such as the NCLB requirement that multiple consecutive grades be tested every year. Indeed, she judged that, “it would be difficult to adopt any of these approaches in today’s K-12 testing system without making significant changes to state policy surrounding accountability.”

Portfolios

Portfolio-based assessments have much less presence in K-12 testing than they once had, but they are used in other sectors in the United States and in a number of other nations. In the United States, the National Board for Professional Teaching Standards (NBPTS), which identifies accomplished teachers (from among candidates who have been teaching for a minimum of several years), asks candidates to assemble a portfolio of information that represents their teaching skills in particular areas, including videotapes of their lessons. This portfolio supplements other information, collected through computer-based assessments, and allows evaluators to assess a variety of teaching skills, including so-called soft skills, practices, and attitudes, such as the capacity to reflect on a lesson and learn from experience. The assessment is time consuming, requiring up to 400 hours of candidates’ time over 12-18 months. Because of the low number of tasks, the program reports relatively low reliability estimates, and it has also raised concerns about rater variability. However, it has received high marks for validity because it is seen as capturing important elements of teaching.

Technology-Supported Assessment

Computers have long been widely used in assessment, although, Hamilton explained, for only a fairly limited range of purposes. For the most part computers have been used to make the administration and scoring of traditional multiple-choice and essay testing easier and less expensive. However, recent technological developments have made more innovative applications more feasible, and they have the potential to alter the nature of assessment.

First, the increasing availability of computers in schools will make it easier to administer computerized-adaptive tests in which items are presented to a candidate on the basis of his or her responses to previous items. Many states had turned their attention away from this technology because NCLB requirements seemed to preclude its use in annual grade-level testing. However, revisions appear likely to permit, and perhaps even encourage, the use of adaptive tests, which is already common in licensure and certification testing.

The use of computerized simulations to allow candidates to interact with people or objects that mirror the real world is another promising innovation. This technology allows students to engage in a much wider range of activities than is traditionally possible in an assessment situation, such as performing an experiment that requires the lapse of time (e.g., plant growth). It can also allow administrators to avoid many of the logistical problems of providing materials or equipment by simulating whatever is needed. Such assessments can provide rapid feedback and make it possible to track students' problem-solving steps and errors. Medical educators have been pioneers in this area, using it as part of the USMLE: the examinee is given a patient scenario and asked to respond by ordering test or treatments and then asked to react to the patient's (immediate) response. Minnesota has also used simulations in its K-12 assessments (discussed in Chapter 4).

Automated essay scoring is also beginning to gain acceptance, despite skepticism from the public. Researchers have found high correlations between human scores and automated scores, and both NAEP and the USMLE are considering using this technology. Moreover, the most common current practice is for computer-based scoring to be combined with human-based scoring. This approach takes advantage of the savings in time and resources while providing a check on the computer-generated scores. However, some problems remain. Automated scoring systems have been developed with various methodologies, but the software generally evaluates human-scored essays and identifies a set of criteria and weights that can predict the human-assigned scores. The resulting criteria are not the same as those a human would use (e.g., essay length, which correlates with other desirable essay traits, would not itself be valued by a human scorer), and may not have the same results when applied across different groups of students. That is, test developers need to ensure that differences between human rater scores and scores assigned by computers do not systematically favor some groups of students over others. Some observers also worry that the constraints of automated scoring might limit the kinds of prompts or tasks that can be used.

In Hamilton's view, technology clearly offers significant potential to improve large-scale assessment. It opens up possibilities for assessing new kinds of constructs and for providing detailed data. It also offers possibilities for more easily assessing students with disabilities and English-language learners, and it can provide an effective means of integrating classroom-based and large-scale assessment.

A few issues are relevant across these technologies, Hamilton noted. If students bring different levels of skill with computers to a testing situation, as is likely, the differences may affect their results: this outcome is supported by some research. Schools are increasingly likely to have the necessary infrastructure to administer such tests, but this capacity is still unequally distributed. Teachers trained to prepare students for these sorts of assessments and accurately interpret the results are also not equally distributed among schools.

Another issue is that the implications of computer-based approaches for validity and reliability have not been thoroughly evaluated.

Looking to the Future

The challenge of developing innovative assessments that are high in quality and cost-effective has not yet been fully resolved, in Hamilton's view. "Recent history suggests that the less you constrain prompts and responses, the more technically and logistically difficult it can be to obtain high-quality results," she observed. Yet past and current work has explored ways to measure important constructs that were not previously accessible, and the sheer number and diversity of innovations has been "impressive and likely to shape future test development in significant ways," she said. Work in science has been at the forefront, in part because problem solving and inquiry are important components of the domain but have been difficult to assess. She suggested that many pioneering efforts in science assessments are likely to find applications in other subjects, over time. This is an important development because the current policy debate has emphasized the role that data should play in decision making at all levels of the education system, from determining teacher and principal pay to informing day-to-day instructional decisions. Assessments that provide deeper and richer information will better meet the needs of students, educators, and policy makers. The possibility of broadening the scope of what can be assessed for accountability purposes is also likely to reinforce other kinds of reform efforts.

Hamilton also pointed out the tradeoffs inherent in any proposed use of large-scale assessment data. Test-score data are playing an increasingly prominent role in policy discussions because of their integral role in statistical analyses that can be used to support inferences about teachers' performance and other accountability questions. Growth, or value-added, models (designed to isolate the effects of teachers from other possible influences on achievement) rely on annual testing of consecutive grades—a need that may mean significant constraints on the sorts of innovative assessments that can be used. Using a combination of traditional and innovative assessments may provide a suitable tradeoff, Hamilton said. She also called for improved integration between classroom and large-scale assessments. "No single assessment is likely to serve the needs of a large-scale program and classroom-level decision making equally well," she argued. A coordinated system that includes a variety of assessment types to address the needs of different user groups might be the wisest solution.

4

Political Experiences and Considerations

As the recurring theme of tradeoffs suggests, political considerations are a very important aspect of almost all decisions about assessment and accountability, and they have played a critical role in the history of states' efforts with innovative assessments. Veterans of three programs—the Maryland School Performance Assessment Program (MSPAP), the Kentucky Instructional Results Information System (KIRIS), and the Minnesota Comprehensive Assessment (MCA) in science—reflected on the political challenges of implementing these programs and the factors that have affected the outcomes for each.

MARYLAND

MSPAP was a logical next step for a state that had been pursuing education reform since the 1970s, Steve Ferrara explained. Maryland was among the first to use school- and system-level data for accountability, and the state also developed the essay-based Maryland Writing Test in the 1980s. That test, one of the first to use human scoring, set the stage for MSPAP. MSPAP was controversial at first, particularly after the first administration yielded pass rates as low as 50 percent for 9th graders. However, the test had a significant influence on writing instruction, and scores quickly rose.

The foundation for MSPAP was a Commission on School Performance established in 1987 by then governor William Schaeffer, which outlined ambitious goals for public education and recommended the establishment of content standards that would focus on higher order thinking skills. The commission also explicitly recommended that the state adopt an assessment that would not rely only on multiple-choice items, and could provide annual report cards for schools.

The governor's leadership was critical in marshalling the support of business and political leaders in the state, in Ferrara's view. Because the Maryland governor appoints members of the state board of education, who, in turn, appoint the superintendent of public instruction, the result was "a team working together on education reform." This team was responding to shifting expectations for education nationally, as well as a sense that state and district policy makers and the public were demanding assurances of the benefits of their investment in education. Ferrara recalled that the initial implementation went fairly smoothly, owing in part to concerted efforts by the state superintendent and others to communicate clearly with the districts about the goals for the program and how it would work and to solicit their input.

Most school districts were enthusiastic supporters, but several challenges complicated the implementation. The standards focused on broad themes and conceptual understanding, and it was not easy for test developers to design tasks that would target those domains in a way that was feasible and reliable. The way the domains were described in the standards led to complaints that the assessment did not do enough to assess students' factual knowledge. The schedule was also exceedingly rapid, with just 11 months between initial planning and the first administration. The assessment burden was great—9 hours over 5 days for 3rd graders, for example. There were also major logistical challenges posed by the manipulatives needed for the large number of hands-on items.

The manipulatives not only presented logistical challenges, they also revealed a bigger challenge. For example, teachers who had not even been teaching science were asked to lead students through an assessment that included experiments. Teachers were also being asked more generally to change their instruction. The program involved teachers in every phase—task development, scoring, etc.—and Ferrara believes that this was one of the most important ingredients in its early success. As discussed above, there is evidence that teachers changed their practice in response to MSPAP (Koretz et al., 1996; Lane, 1999). Nevertheless, many people in the state began to oppose the program. Criticisms of the content and concern about the lack of individual student scores were the most prominent complaints, and the passage of the No Child Left Behind (NCLB) Act in 2002 made the latter concern urgent. The test was discontinued in that year.

For Ferrara, several key lessons can be learned from the history of MSPAP:

- Involving stakeholders in every phase of the process was very valuable, both improving the quality of the program and building political acceptance.

- It paid to be ambitious technically, but it is not necessary to do everything at once. For example, an assessment program could have a significant influence on instruction without being exclusively composed of open-ended items. If one makes a big investment in revolutionary science assessment, there will be fewer resources and less political capital for other content areas.
- It was short-sighted to invest the bulk of funds and energy in test development at the expense of ongoing professional development.

KENTUCKY

The 1989 Kentucky Supreme Court decision that led to the development of KIRIS was intended to address both stark inequities in the state's public education system and the state's chronic performance at or near the bottom among the 50 states, Brian Gong explained. The resulting Kentucky Education Reform Act (KERA), passed in 1990 (which had broad bipartisan and public support), was one of the first state education reform bills and included innovative features, such as a substantial tax increase to fund reform; a restructuring of education governance; the allocation of 10 paid professional development days per year for teachers; and a revamped standards, assessment, and accountability system.

KERA established accountability goals for schools and students (proficiency within 20 years) and called for an assessment system that would be standards based, would rely primarily on performance-based items, and would be contextualized in the same way high-quality classroom instruction is. The result, KIRIS, met these specifications, but the developers faced many challenges. Most critically, the court decision had actually identified the outcomes to be measured by the assessment. These were the basis for the development of academic expectations and the core content for assessment, but the definitions of the constructs remained somewhat elusive. Educators complained that they were not sure what they were supposed to be doing in the classroom, Gong explained, and in the end it was the assessment that defined the content that was valued, the degree of mastery that was expected, and the way students would demonstrate that mastery. But developing tasks to assess the standards in valid ways was difficult, and the assessment alone could not provide sufficient information to influence instruction.

There were other challenges and adaptations. It was difficult to collect data for standard setting using the complex, multifaceted evidence of students' skills and knowledge that KIRIS was designed to elicit. Equating the results from year to year was also difficult: most of the tasks were very memorable and could not be repeated. Alternate strategies—such as equating to administrations in other states, judgmental equating, and equating to multiple-choice items—had psychometric and practical disadvantages. KIRIS also initially struggled to maintain standards of accuracy and reliability in scoring the constructed-response items and portfolios, and adaptations and improvements were made in response to problems. Guidelines for the development of portfolios had to be strengthened in response to concerns about whether they truly reflected students' work. With experience, test developers also gradually moved from holistic scoring of portfolios to analytic scoring in order to gain more usable information from the results. The state also faced logistical challenges, for example, with field testing and providing results in time for accountability requirements.

Nontechnical challenges emerged as well. The state was compelled to significantly reduce its assessment staff and to rely increasingly on consultants. School accountability quickly became unpopular, Gong explained, and many people began to complain that the aspirations were too high and to question the assertion that all students could learn at such high levels. Philosophical objections to the learning outcomes assessed by KIRIS emerged, with some arguing that many of them intruded on parents' prerogatives and invaded students' privacy. The so-called math and reading "wars"—over the relative emphasis that should be given to basic skills and fluency as opposed to broader cognitive objectives—fueled opposition to KIRIS. There were changes in the state's political leadership that decreased support for the program, which did not survive an increasingly contentious political debate and was ended in 1998.

For Gong there are several key lessons:

- Clear definitions of the constructs to be measured and the purposes and uses of the assessment are essential. No assessment can compensate for poorly defined learning targets.
- The design of the assessment should be done in tandem with the design of the intended uses of the data, such as accountability, so that they can be both coherent and efficient.
- The people who are proposing technical evaluations and those who will be the subject of them should work together in advance to consider both intended and unintended consequences, particularly in a politically charged context.
- Anyone now considering innovative assessments for large-scale use should have a much clearer idea of how to meet technical and operational challenges than did the pioneering developers of KIRIS in the 1990s.
- Current psychometric models, which support traditional forms of testing, are inconsistent with new views of both content and cognition and should be applied only sparingly to innovative assessments. The field should invest in the development of improved models and criteria (see, e.g., Shepard, 1993; Mislevy, 1998).

MINNESOTA

Minnesota's Comprehensive Assessment Series II (MCA-II) was developed in response to NCLB, so the state was able to benefit from the experiences of states that had already explored innovative assessment. Assessments already in place in some subjects could be adapted, but the MCA-II in science, implemented in 2008, presented an opportunity to do something new. State assessment staff were also given unusual latitude to experiment, Dirk Mattson explained, because science had not been included in the NCLB accountability requirements.[1]

The result was a scenario-based assessment delivered on computers. Students are presented with realistic representations of classroom experiments, as well as phenomena that can be observed. Items—which may be multiple choice, short or long constructed-response, or figural (e.g., students interact with graphics in some way)—are embedded in the scenario. This structure provides students with an opportunity to engage in science at a higher cognitive level than would be possible with stand-alone items.

Mattson emphasized that the design of the science assessment grew out of the extensive involvement of teachers from the earliest stages. Teachers rejected the idea of an exclusively multiple-choice test, but a statewide performance assessment was also not considered because a previous effort had ended in political controversy. The obvious solution was a computer-delivered assessment, despite concerns about the availability of the necessary technology in schools. The developers had the luxury of a relatively generous schedule. Conceptual design began in 2005, and the first operational assessment was given in 2008. This allowed time for some pilot testing at the district level before field testing began in 2007.

A few complications have arisen. First, Minnesota statute requires regular revision of education standards, so the science standards were actually being revised before the prior ones had been assessed, but assessment revision was built into the process. Second, in 2009 the state legislature, facing severe budget constraints, voted to make the expenditure of state funds on human scoring of assessments illegal.[2] More recently, the state has contemplated signing on to the "common core" standards and is monitoring other changes that may become necessary as a result of the Race to the Top Initiative or reauthorization of the federal Elementary and Secondary Education Act.

Mattson reflected on the process so far, noting that despite the examples

[1] The technical manual and other information about the program are available at http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TechReports/index.html [April 2010].

[2] State money could still be used for machine scoring of assessments. Because Minnesota has no board of education, the legislature is responsible for overseeing the operations of the department of education; federal dollars had been used to support human scoring.

of other innovative assessment programs, many aspects of the MCA were new and had to be developed from scratch. These elements included operational issues, such as a means for conveying what was needed to testing contractors, estimating costs, and supporting research and development efforts. They also included fundamental design issues, and Mattson noted that often the content specialists were far ahead of the software designers in conceptualizing what could be done. Technical challenges—from designing a test security protocol to preparing schools to load approximately 300-475 megabytes of test content onto their servers—required both flexibility and patience. A Statewide Assessment Technology Work Group helped the team identify and address many of the technical challenges, and Mattson pointed to this group as a key support.

For Mattson, it was important that the developers were not deterred by the fact that there were no paths to follow in much of development process. The success of the assessment thus far, in his view, has hinged on the fact that the state was able to think ambitiously. The leaders had enthusiastic support from teachers, as well as grant funding and other community support, which allowed them to sustain their focus on the primary goal of developing an assessment that would target the skills and knowledge educators believed was most important. The flexibility that has also been a feature of the MCA since the beginning—the state assessment staff's commitment to working with and learning from all of the constituencies concerned with the results—should allow them to successfully adapt to future challenges, Mattson said.

POLICY IMPLICATIONS

The three state examples, suggested discussant Lorraine McDonnell, highlight the familiar tension between the "missionaries" who play the important role of seeking ways to improve the status quo and those who raise the sometimes troublesome questions about whether a proposed solution addresses the right problem, whether the expected benefits will outweigh the costs, and whether the innovation can be feasibly implemented. She distilled several policy lessons from the presentations.

First, for an assessment to be successful, it is clear that the testing technology has to be well matched to the policy goals the assessment is intended to serve. Accurate educational measurement may be a technical challenge, but assessment policy cannot be clearly understood independent of its political function. Whether the function is to serve as a student- or school-level accountability device, to support comparisons across schools or jurisdictions, or to influence the content and mode of classroom instruction, what is most important is to ensure that the goals are explicitly articulated and agreed on. McDonnell observed that in many states test developers and politicians had not viewed the function of the state assessment in the same way. As a result, test developers could not meet the policy makers' expectations, and support for the assessment weakened. When policy makers expect the test to serve multiple purposes, the problem is most acute—and policy makers may not agree among themselves about an assessment's function.

It is also very important that the testing system be integrated into the broader instructional system, McDonnell advised. This idea was an element of the argument for systemic reform that was first proposed during the 1990s (Smith and O'Day, 1991) and has been prominent in reform rhetoric, but it has not played a major role in the current common standards movement. She pointed out that although the conditions that truly support effective teaching and learning should be central, they "appear once again to have been relegated to an afterthought." Support is needed to strengthen instructional programs, as well as assessment programs. As numerous participants did throughout the workshop, she highlighted the importance of a comprehensive and coherent system of standards, instruction, and assessment.

States and the federal government have tended, she argued, to select instruments that were easy to deploy—such as tests—and to underinvest in measures such as curricula and professional development that could help to build schools' capability to improve education. Yet unless teachers are provided with substantial opportunities to learn about the deeper curricular implications of innovative assessments and to reshape their instruction in light of that knowledge, the result of any high-stakes assessment is likely to be more superficial test preparation, or what McDonnell called "rubric-driven instruction." This collision between policy pressure for ambitious improvements in achievement and the weak capability of schools and systems to respond was an enduring dilemma during the first wave of innovation, in the 1990s, and McDonnell believes that it has not been resolved.

Another lesson is that policy makers and test designers need to understand the likely tradeoffs associated with different types of assessments and then decide which goals they want to advance and which ones they are willing to forgo or minimize. The most evident tension is between using tests as accountability measures and using them as a way to guide and improve instruction. As earlier workshop discussions showed, McDonnell said, these two goals are not necessarily mutually exclusive, but pursuing both with a single instrument is likely to make it more difficult to obtain high-quality results.

This lesson relates to another, which may be obvious, McDonnell suggested, but appears to be easily overlooked: that if a process is to be successful, every constituency that will be affected by an assessment must have ample opportunity to participate throughout it. The differing perspectives of psychometricians and curriculum developers, for example, need to be reconciled if an assessment system is to be successful, but parents, teachers, and other interests need to be heard as well. If developers fail to fully understand and take into account public attitudes, they may encounter unexpected opposition to their plans. Conveying the rationale for an assessment approach, and the expected benefits and costs, to policy makers and the public may require careful planning and concerted effort. It is often at the district level that the critical communications take place, and too often, McDonnell said, district leaders have been left unprepared for this important aspect of the process. The benefit of clear and thorough communication is that stakeholders are more likely to continue to support a program through technical difficulties, if they have a clear understanding of the overall goals.

Finally, McDonnell stressed, people need to remember that the implementation of innovative assessments takes time. It is very important to build in an adequate development cycle that allows for a gradual, phased roll-out and for adaptation to potential problems. In several of the experiences discussed at the workshop, rushed implementation led to technical problems, undue stress on teachers and students, and a focus on testing formats at the expense of clear connections to curriculum. In several states testing experts acquiesced to political pressure to move quickly in a direction that the testing technology could not sustain. Programs that have implemented innovative features gradually, without dismantling the existing system, have had more flexibility to adapt and learn from experience.

These policy lessons, as well as a growing base of technical advances, can be very valuable for today's "missionaries," McDonnell observed. However, although past experience provides lessons, it may also have left a legacy of skepticism among those who had to deal with what were in some cases very disruptive experiences. Fiscal constraints are also likely to be a problem for some time to come, and it is not clear that states will be able to sustain new forms of assessment that may be more expensive after initial seed funding is exhausted. She also noted that the common standards movement and the Race to the Top Initiative have not yet become the focus of significant public attention, and there is no way to predict whether they will become the objects of ideological controversies, as have past education reforms. None of these are reasons not to experiment with new forms of assessment, McDonnell concluded, but "they are reasons for going about the enterprise in a technically more sophisticated way than was done in the past and to do it with greater political sensitivity and skill."

5

Opportunities for Better Assessment

The present moment—when states are moving toward adopting common standards and a federal initiative is pushing them to focus on their assessment systems—seems to present a rare opportunity for improving assessments. Presenters were asked to consider the most promising ways for states to move toward assessing more challenging content and providing better information to teachers and policy makers and to identify the elements that would need to be in place. Laurie Wise enumerated some of the possible opportunities, Ronald Hambleton described some of the challenges of supporting cross-state comparisons in a new era of innovative testing, and Rebecca Zwick concluded with some historical perspective on cross-state comparisons and suggestions for moving forward.

IMPROVEMENT TARGETS

Laurie Wise began with a reminder of the goals for improvement that had already been raised. Assessments need to support a wide range of policy uses. Current tests have limited diagnostic value and are not well integrated with instruction or with interim assessments. At the same time, however, they are not providing optimal information for accountability purposes because they cover only a limited range of what is and should be taught. Improvements are also needed in validity, reliability, and fairness. For current tests, he observed, there is little evidence that they are good indicators of instructional effectiveness or good predictors of students' readiness for subsequent levels of instruction. Their reliability is limited because they are generally targeted to very broad content specifications, and limited progress had been made in assessing all students accurately. Improvements such as computer-based testing and automated scoring need to become both more feasible in the short run and more sustainable in the long run.

In Wise's view, widespread adoption of common standards might help with these challenges in two ways: (1) by pooling their resources, states could get more for the money they spend on assessment, and (2) interstate collaboration is likely to facilitate deeper cognitive analysis of standards and objectives for student performance than is possible with separate standards.

Cost Savings

The question of how much states could save by collaborating on assessment begins with the question of how much they are currently spending to do it on their own. Savings would be likely to be limited to test development, since many per-student costs for administration, scoring, and reporting, would not be affected, so Wise focused on an informal survey he had done of development costs (Wise, 2009). (Wise's sample included 15 state testing programs and a few test developers, and included only total contract costs, not internal staff costs.) The results are shown in Table 5-1.

Perhaps most notable was the wide range in what states are spending, as shown in the minimum and maximum columns. Wise also noted that on aver-

TABLE 5-1 Average State Development and Administration Costs by Assessment Type

Assessment Type	N	Mean	S.D.	Min.	Max.
Annual Development Costs (in thousands of dollars)					
Alternate	9	363	215	100	686
Regular-ECR	13	1,329	968	127	3,600
Regular-MC Only	5	551	387	220	1130
Administrative Cost per Student (in dollars)					
Alternate	9	376	304	40	851
Regular-ECR	16	26	18	4	65
Regular-MC Only	6	3	3	1	9

NOTES: Extended-constructed-response (ECR) tests include writing assessments and other tests requiring human scoring using a multilevel scoring rubric. Multiple-choice (MC) tests are normally machine scored. Because the results incorporate a number of different contracts, they reflect varying grade levels and subjects, though most included grades 3-8 mathematics and reading.
SOURCE: Wise (2009, p. 4).

age the states surveyed were spending well over a million dollars annually to develop assessments that require human scoring and $26 per student to score them.

A total of $350 million will be awarded to states through the Race to the Top Initiative. That money, plus savings that winning states or consortia could achieve by pooling their resources, together with potential savings from new efficiencies such as computer delivery, would likely yield for a number of states as much as $13 million each to spend on ongoing development without increasing their own current costs, Wise calculated.

Improved Cognitive Analyses

Wise noted that the goal for the common core standards is that they will be better than existing state standards—more crisply defined, clearer, and more rigorous. They are intended to describe the way learning should progress from kindergarten through 12th grade to prepare students for college and work readiness. Assuming that the common standards meet these criteria, states could collaborate to conduct careful cognitive analysis of the skills to be mastered and how they might best be assessed. Working together, states might have the opportunity to explore the learning trajectories in greater detail, for example, in order to pinpoint both milestones and common obstacles to mastery, which could in turn guide decisions about assessment. Clear models for the learning that should take place across years and within grades could support the development of integrated interim assessments, diagnostic assessments, and other tools for meeting assessment goals.

The combination of increased funding for assessments and improved content analyses would, in turn, Wise suggested, support the development of more meaningful reporting. The numerical scales commonly used now offer very little meaningful information beyond identifying students above and below an arbitrary cut-point. A scale that was linked to detailed learning trajectories (which would presumably be supported by the common standards and elaborated through further analysis) might identify milestones that better convey what students can do and how ready they are for what comes next.

Computer-adaptive testing would be particularly useful in this regard since it provides an easy way to pinpoint an individual student's understanding; in contrast, a uniform assessment may provide almost no information about a student who is performing well above or below grade level. Thus, reports of both short- and long-term growth would be easier, and results could become available more quickly. Results that were available more quickly and also provided richer diagnostic information could also improve teacher engagement. Another benefit would be in the increased potential for establishing assessment validity. Test results that closely map onto defined learning trajectories could support much stronger inferences about what students have mastered than are possible with current data, and they could also better support inferences about the relationship between instruction and learning outcomes.

It is clear that common standards can support significant improvements in state assessments, Wise said. The potential cost advantages are apparent. But perhaps more important is that concentrating available brain power and resources on the elaboration of one set of thoughtful standards (as would be possible if a number of states were all assessing the same set of standards) would allow researchers to work together and yield faster and better results.

CROSS-STATE COMPARISONS

One important reason to have common standards would be to establish common learning objectives for students regardless of where they live. But unless all students are also assessed with exactly the same instruments, comparing their learning across states is not so easily accomplished. Nevertheless, the capacity to make cross-state comparisons remains important to policy makers and educators. Ron Hambleton described some of the complex technical issues that surround this challenge, known to psychometricians as "linking": placing test results on the same score scale so that they can be compared.[1]

The basic issue that linking procedures are designed to address is the need to determine, when results from two tests appear to be different, whether that difference means that one of the tests is easier than the other or that one of the groups of students is more able than the other in some way. There are several different statistical procedures for linking the results of different tests (see National Research Council, 1999a, 1999b; see also http://nces.ed.gov/pubs98/linking/c3.asp [June 2010]). If the tests were developed to measure precisely the same constructs, to meet the same specifications, and to produce scores on the same scale, linking procedures are relatively straightforward. They are also very important though, since users count on the results of, say, the SAT, to mean exactly the same thing, year after year. Other—less straightforward—linking approaches are necessary when the tests are developed to different frameworks and specifications or yield scores on different scales. A common analogy is the formula for linking temperatures measured in degrees Fahrenheit or Celsius, but because of the complexity of the cognitive activities measured by educational tests, procedures for linking test scores are significantly more complex.

In general, Hambleton explained, to compare the achievement of different groups of students it is necessary to have some comparability not only in the standards to be measured against, but in the tests and the curricula to which the students have been exposed. While much is in flux at the moment, Hambleton judged that it appears likely that states will continue to use a variety

[1] Hambleton credited two reports from the National Research Council (1999a, 1999b) for much of his discussion.

of approaches to assessment, including varying combinations of paper-and-pencil format, computer-based assessment, and various kinds of innovative assessments. This approach may be an effective way for supporting instruction at the state and district levels, he pointed out, but it will complicate tremendously the task of producing results that can be compared across states. Innovative item types are generally designed to measure new kinds of skills, but the more detailed and numerous the constructs measured become, the greater the challenge of linking the results of one assessment to those of another.

To demonstrate the challenge, Hambleton began with an overview of the complexity of linking traditional assessments. Even in the most strictly standardized testing program, it is nearly impossible to produce tests from year to year that are so clearly equivalent that scores can be compared without using linking procedures to make statistical adjustments. Even for the SAT, possibly the most thoroughly researched and well funded testing program in the world, psychometricians have not been able to get around the need to make statistical adjustments in order to link every different form of the test every year. The basic procedure involved in linking two test forms is, first, to make sure either that some number of people take both test forms or that some number of items appear in both test forms, and, second, to then use statistical procedures to adjust (in one of several ways) for any differences in difficulty between the two test forms that become apparent (see Hambleton, 2009).

The same must also be done if comparisons are to be made across states, Hambleton noted. This is easiest if the states share the same content standards and proficiency standards and also administer the same tests, as is the case with the New England Common Assessment Program (NECAP). However, even when these three elements are the same, it is still possible that items will perform differently across states because of variations in teaching methods, curricula, or other factors, and this possibility must be checked. Any changes to the nature or administration of a test may affect the way items perform. Thus, the addition of new item types, separate sections to measure an individual state's desired content, changes in the positioning of items, or conversion of certain parts of an assessment to computer delivery, may affect equating (see National Research Council, 1999a, 1999b). The timing of the test is also important: even if states have similar curricula, if they administer a common test at different points in the year it will affect the quality of the linking.

Since common items are necessary to perform cross-state linking procedures, some of each type must be given to each group of test takers, and therefore scoring of constructed-response items must be precisely consistent across states. The testing conditions must be as close to identical as possible, including such operational features as test instructions, timing, and test security procedures, as well as larger issues, such as the stakes attached to test results. Even minor differences in answer sheet design or test book layout have the potential to interfere with item functioning (see National Research Council, 1999a, 1999b).

Computer-based testing adds to the complexity by potentially allowing administrations to vary in new ways (in time, in context, etc.) and also by requiring that all items be linked in advance. States and districts are likely to use different operating systems and platforms, which may affect seemingly minor details in the way tests look to students and the way they interact with them (e.g., how students scroll through the test or show their work), and these differences would have to be accounted for in the linking procedures. If there are not enough computers for every student to take an assessment at the same time, the testing window is usually extended, but doing that may affect both the linking and security.

Hambleton stressed that all of these constraints apply to comparisons between just two test forms or states or within a consortium that is using the same assessment. If the comparison is extended across consortia, more significant sources of variation come into play, including multiple different sets of curricula, test design, performance standards, and so forth.

The essential principle of linking, for which Hambleton credited psychometrician Albert Beaton, is “if you want to measure change or growth over time, don’t change the measure.” But this approach is completely impractical in the current context of educational testing, where change—in the content to be tested and in test design—is almost constant.

Moreover, Hambleton explained, most linking procedures rest on the idea that a test is unidimensional (i.e., that it measures skills in a single domain), but new forms of testing are multidimensional (they measure skills in more than one domain). So not only would linking have to account for each of the dimensions assessed, it would have to account for changing combinations of dimensions as tests evolve. Hambleton advised that, at least in assessments used for summative purposes, it will be necessary to place some constraints on these factors to make linking possible.

Hambleton believes that states’ capacity to maintain high-quality linking procedures is already being stretched. With new sorts of assessments it will be critical to take seriously the need to expend resources to sustain the psychometric features that are needed to answer the questions policy makers ask. New research on questions about vertical scaling and linking tests that use multiple assessment modes, for example, will be necessary to support the current goals for assessment reform. A related, but equally important challenge will be that of displaying and reporting new kinds of data in ways that users can easily understand and report. In addition, he said, the number of issues that are involved will complicate the task of supporting valid comparisons of 21st-century assessments, and Hambleton joked that “no psychometrician will be left behind.” He acknowledged, however, that they “can’t just sit in their ivory towers and take five years to solve a problem.”

PERSPECTIVES: THE PAST AND THE FUTURE

The importance of comparisons across states is seldom questioned now, but Rebecca Zwick pointed out that this was not always the case. When the National Assessment of Educational Progress (NAEP) was first developed in the 1960s, she noted, it was not designed to support such comparisons (Zwick, 2009). At that time, the public was very wary of federal involvement in education. Federal enforcement of school desegregation and the passage of the Civil Rights Act of 1964 and the Elementary and Secondary Education Act in 1965 were viewed as the limit of federal involvement in education that the public would accept, and state-by-state comparisons were viewed as too polarizing.

By the 1980s, however, the idea of promoting academic excellence through competition was becoming increasingly popular. President Ronald Reagan's 1984 State of the Union address called for comparisons of academic achievement among states and schools, arguing that "without standards and competition there can be no champions, no records broken, no excellence. . ." (Zwick, 2009). In response, the U.S. Department of Education first developed "wall charts," which visually displayed comparisons of the resources states committed to education and measures of performance, such as average SAT and ACT scores. These comparisons were widely viewed as un-illuminating at best, since the college admissions tests reflected only the performance of college-bound students and were not aligned to curriculum and instruction in the states. By the late 1980s, NAEP had developed its Trial State Assessment, which compared the mathematics performance of 8th graders in 37 states. By the mid 1990s the state assessments were fully operational. Later in the decade, the Clinton administration proposed a "voluntary national test" as a way of collecting comparative information on individual students, but political controversy doomed the project before it was implemented.

The NAEP state assessments received an additional boost when the No Child Left Behind Act (NCLB) required states receiving Title I funding to participate, Zwick noted. (Participation is voluntary, and NAEP has occasionally struggled to secure adequate representation for its matrix-sampling design.) NCLB also required states to meet various requirements for their own assessments and accountability provisions. Comparisons began to be made among states on the basis of the results of their own assessments, even though the assessments used different designs, proficiency standards, and so on. Studies of states' results have shown that states' definitions of proficiency vary significantly and tend to be lower than those used in NAEP (see National Research Council, 2008), and this variation was one of the reasons enthusiasm has grown for the common core standards.

As Laurie Wise discussed, the common core standards are likely to help states achieve economies of scale, and their decisions to form consortia as part of the Race the Top competition will do so as well. With more money to spend on assessment, states should be able to produce tests that are better aligned to standards, contain richer and more innovative items, and are more reliable. Results are likely to be reported more quickly and provide more useful information for diagnosing individual students' needs and guiding instruction. And states could more easily take advantage of a growing body of knowledge and experience as they collaborate to develop new approaches. Funding for research to improve cognitive analyses of test content, validity, and a many other issues could feed the development of more innovative assessments that are also feasible and affordable. And, as the NECAP example has demonstrated, when standards and assessments are shared, states are both somewhat shielded from public reaction to disappointing results and correspondingly less likely to lower their performance standards to achieve nominally better results.

However, as Hambleton made clear, there are significant challenges to making sure that state-to-state comparisons are valid, fair, and useful. One premise of the common core standards is that they will allow states to measure their students against the same expectations, but the rules permit states to supplement the core standards with up to 15 percent of additional content they value. To illustrate, Zwick suggested quantifying the standards and assuming that the 10 states in a consortium share 85 standards, and that each has its own 15 additional unique standards. In that case, out of the total of 85 + 10(15) = 235 standards, only 85/235(100), or 36.2 percent, would be shared among the 10 states. Since no single assessment is likely to cover all of a state's standards at once, the content shared by each of the 10 state assessments could be significantly lower.

As Hambleton explained, every source of variation among state tests can impair the ability to compare results and report them on the same scale. Curriculum and instruction will not be shared, even if states share standards. States are likely to find it very difficult to adhere to common procedures for every aspect of testing (preparation, administration, booklet or format design, accommodations, etc.). States also differ in terms of demographic characteristics, per-pupil expenditures, teacher training, class size, and many other features likely to affect test results.

Zwick had several suggestions for states to support meaningful comparison across the states.

- Provide as much context as possible: Document differences across states in order to provide a context for comparisons.
- Make the test instrument and testing policies as similar as possible: Collaborate on scoring and reporting methods and make sure that format, instructions, protocols for each phase of assessment (including preparation), etc., are uniform. Put unique state items in a separate section to minimize their influence on students' performance of the common items. Develop an adequate pool of items in each area of interest.

- Invest in research and professional development: Provide training to school administrators and teachers in the interpretation and uses of test scores.

Zwick also advocated fostering collaboration among psychometricians, educators, and policy makers, who, despite their differences in perspective, are likely to be pursuing the same ultimate goal—the improvement of teaching and learning in U.S. classrooms.

Some participants suggested that many of the problems with state comparisons could be solved if states opted to use common assessments, but others reiterated that the many sources of variation among states still remain. It was noted that other countries have had success with common curricula and common tests and for some that seems like the best route. Others pointed out that NAEP already provides state comparisons, though without any particular alignment to the standards to which states are working and without scores for individual students.

6

Research Needs

THEORY AND GOALS

There is wide acceptance of the value of a system in which assessments measure student progress in meeting education standards and the test results are used to hold students, schools, educators, and jurisdictions to account for their performance. But, Lorrie Shepard described in her closing remarks two very different theories of action regarding the way such a system will actually bring about improvements have been put forward. Neither the differences between them nor the implications of adopting one or the other have been widely recognized, she added. One, the incentives theory, is that given sufficient motivation, teachers and other school personnel will develop ways to improve instruction. This perspective was the basis for the Elementary and Secondary Education Act of 1994, which required states to establish standards and assessments.

The other approach, which Shepard called the coherent capacity building theory, posited that an additional step, beyond establishing clear expectations and the motivation to meet them, was needed. Educators would also need the capacity, in the form of professional development and other supports, to improve their teaching in order for the accountability measures to have the desired effect (see, e.g., National Research Council, 1995). Shepard believes that the incentives theory is dominant, and that capacity building has consequently been neglected.

Similar imprecision is evident in the possible interpretations of some of the top reform goals of the present moment, Shepard suggested, including

- reforming assessments using conceptually rich tasks,
- integrating 21st-century skills and academic content,
- creating coherence between large-scale and classroom assessments, and
- using data to improve classroom instruction.

For example, treating the first two bullets as distinct enterprises makes little sense, given that a large body of research indicates the importance of teaching content and higher-order thinking skills together. Shepard believes that policy makers do not completely understand that effective teaching relies on a model for how learning proceeds, in which cognitive skills and the knowledge of when and how to use them develop together with content knowledge and understanding of how to generalize from it. She cautioned that, without this theory of learning, policy makers are likely to accept current modes of assessment. They may believe, for example, that narrowing the curriculum is necessary because basic reading and mathematics skills are so important. They may not be aware that excessive drill on worksheets that resemble summative tests does not give students the opportunity to understand the context and purpose for what they are learning—which would enhance their skill development (see Elmore, 2003; Blanc et al., 2010; Bulkley et al., 2010; Olah, Lawrence, and Riggan, 2010). Similarly, although policy makers are in favor of data-driven decision making, Shepard believes, many educators lack the substantive expertise to interpret the available data and use it to make meaningful changes in their practice.

During the workshop discussions, many presenters drew attention to the churning that affects education policy because of shifts in political goals and personnel at the state level. Given that reality, coherence will have to come at a lower level, Shepard argued. The United States does not have a common curriculum, she suggested, because it has no tradition of relying on subject-matter experts in many decisions about education. Psychometricians and policy makers have typically taken the lead in the development of assessments, for example: subject-matter experts have generally been involved in some way, but are not usually asked to oversee the development of frameworks, item development, and the interpretation of results. Now, however, the interests of subject-matter experts and cognitive researchers who have been developing models of student learning within particular disciplines have converged, and this convergence offers the possibility of a coherence that could withstand the inevitable fluctuations in political interests. However, the practical application of this way of thinking about learning is not yet widely understood, Shepard observed. Thus, for Shepard, the opportunity of the present moment is to take the first steps in inventing and implementing the necessary innovations. It is not practical to expect that any one state or consortium could develop an ideal system for all grades and subject areas on the first try, so the focus should be on incremental improvements. She thinks that each consortium grant award should be focused on the development of a system of "next-generation, high-quality" classroom and summative assessments for one manageable area—say, just for mathematics, grades 4 through 8.

She noted that Lauren Resnick had proposed a way of implementing innovative approaches incrementally. Resnick suggested that content-based "modules" that incorporate both a rich curriculum and associated assessments could be adopted one by one and incrementally incorporated into an existing full curriculum. In the near term, this would leave existing assessments unchanged, but, over time, the accumulating body of new modules would eventually lead to a completely transformed system, in which accountability information could be drawn from the assessment components of the innovative curriculum modules. This approach would allow educators to proceed gradually, as the research to support the development of such modules grows, and also to sidestep many of the political and practical challenges that have hampered past programs.

Shepard also emphasized the importance of considering curriculum along with new and improved assessment models. She cautioned that establishing higher standards does not mean only setting cut-points at a higher level, but also incorporating material of a substantively different character into assessments. If this is done without corresponding changes to curriculum and instruction, the result will be predictable—students will not be likely to succeed on the new assessment. In the end, after all, the purpose of the improvements is to "change the character of what we teach and then make those opportunities available to all students and make sure that the assessment can track any changes over time."

Shepard's closed by reminding everyone that "to truly transform learning opportunities in classrooms in ways that research indicates are possible, it will be necessary to remove policy structures—especially low-level tests that misdirect effort; provide coherent curricula consistent with ambitious reforms; and take seriously the need for capacity-building at every level of the education system."

RESEARCH PRIORITIES

Shepard and other discussants were asked to reflect on their highest priorities for research that would support progress in developing and implementing innovative assessments. Many of the ideas overlapped, and they fell into a few categories: measurement, content, teaching and learning, and experimentation.

Measurement Many participants emphasized the need for psychometric models that were developed generations ago to be updated in light of recent research on learning and cognition. New ways of thinking about what should be measured and what sort of information would be useful to educators have been put forward, as described in Chapter 2, and it is clear that current psychometric models do not fit them well. These cognitive models illustrate, for example, the importance of each of the stages that students go through in learning complex material. This idea implies that teachers (and students) need information about students' developing understanding of concepts and facts and how they fit into a larger intellectual structure. Yet educational measurement has tended to focus on one-dimensional rankings according to students' mastery of specific knowledge and skills at a given time. The goals of traditional psychometrics remain important, but perhaps need to be expanded. Means of establishing the validity of new kinds of assessments for new kinds of uses are needed.

Discussants pointed to the need for a strategy for making sure that the information an assessment provides is being used to good effect and a strategy for checking the links in a proposed learning trajectory, to be sure each stage in the progression is reasonable and well supported. The capacity to compare results across assessments is already being stretched, and the introduction of more innovative modes of assessment may present challenges that cannot be solved with current procedures. But the policy demand for comparative information suggests a need for new thinking about the precise questions that are important and the kinds of information that can provide satisfactory answers.

Other fields, one participant noted, have grappled with similar issues. In medicine, for example, simulations are used in credentialing assessments, despite the lack of procedures for equating precisely across assessments. It would be worthwhile to explore the decisions that the medical profession made and their outcomes. It may be, for example, that the technical standards for modes of assessment could vary somewhat, according to the intended purpose to which the results will be put.

A final thought offered on measurement was that the measurement community should be conducting the sort of basic research that addresses not only immediate problems, but also the challenges and technological changes that are likely to emerge a decade from now. Some participants responded that the capacity of the testing profession is already being stretched and that there is little time for this kind of thinking—while others stressed the importance of addressing that problem.

Content The measurement community may need to catch up with advances in cognitive research, but the overall picture of what students should learn is hardly complete. Deeper cognitive analysis of the content to be taught and assessed is still needed. Detailed learning trajectories have been put forward in a few areas of science and mathematics, but much remains to be done. Understanding of the barriers to advancing along a trajectory, and of the efficacy of different approaches to teaching students to overcome those barriers, are in the beginning stages, and other subjects remain far behind science. Without a much broader base of research on these questions, the progress in developing innovative assessments will be hampered. Policy makers are currently working from hypothesized trajectories of how learning in reading, English, language arts, and mathematics progresses from kindergarten through grade 12. These need to be elaborated, and the field needs a plan for gathering data about the validity of the common core standards that are based on them and for improving the descriptions of the trajectories.

Teaching and Learning An important theme of the workshop was the intimate relationship between models of measurement and models of teaching and learning. If assessments are to play the valuable role in education that many envision, they must not only align with what is known about how students think and learn, but also provide meaningful information that educators can use. And if educators are to play their part, their preparation and professional development must encompass this new thinking about assessment and the means to use it. Research is needed to support these changes. Teachers also have much to contribute to evolving thinking about teaching and assessment. Involving them in the research will be critical to ensuring that new kinds of assessment data can really improve instruction.

It is not only data that teachers need, though, some participants pointed out. Their capacity to reflect on and evaluate not only their own practice and capacity to adapt, but also the value of innovations they are asked to try is also important. Working individually, in small groups, as whole departments, or even as schools, teachers can provide a check on such questions as the practical application of theoretical learning trajectories.

Experimentation Several speakers noted that there is no one optimal assessment system waiting to be discovered. A range of international models offer promising possibilities and should be explored in greater detail. The development of state consortia offers the opportunity for the education community to articulate a variety of different models and the theories that underlie them and to work out a variety of ways of addressing key system goals. The idea that educators and policy makers should experiment on students may have negative connotations, but many participants also spoke about the critical importance of taking innovation step by step and learning from each step. In no other field, one participant pointed out, would policy makers overlook the importance of research and development to something as important as redesigning the assessment system. Ideally, the process would begin with a clear picture of the questions that need answers and the development of a strategy for researching those questions and testing hypotheses.

State consortia, individual states, districts, schools, teachers, and students can all contribute to the design of new aspects of assessment systems and the important work of trying them out and collecting information about what works well and what does not. More typical, however, has been a model in which a whole new assessment system is created and presented to the public as ready to be implemented statewide. The big risk in such an approach is that implementation problems could doom an idea with valuable potential before it had a chance to be fully implemented or that individual valuable features of the approach would be thrown out along with features that did not work.

Several participants suggested that retaining some or all of the elements of existing assessment systems, while gradually incorporating new elements, would allow for both the development of political and public acceptance and the flexibility to benefit from experience. An incremental approach may also make it possible to address different aspects of a system in a way that would be too radical to attempt for the whole. Whether the innovations are new instructional units based on core standards, in which assessment is embedded; revised curricula that better map the learning trajectories in new standards; new formats and designs for summative assessments; or some other innovation, it should be possible to gradually construct a coherent system that meets the needs for both accountability and instructional guidance.

References

Blanc, S., Christman, J.B., Hugh, R., Mitchell, C., and Travers, E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 205-225.

Bulkley, K., Christman, J., Goertz, M., and Lawrence, N. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, *85*(2), 186-204.

Christman, J., Neild, R., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., and Travers, E. (2009). *Making the Most of Interim Assessment Data: Lessons from Philadelphia*. Philadelphia, PA: Research for Action.

Chudowsky, N., and Chudowsky, V. (2007*). No Child Left Behind at Five: A Review of Changes to State Accountability Plans.* Washington, DC: Center on Education Policy.

Clune, W.H., and White, P.A. (2008). *Policy Effectiveness of Interim Assessments in Providence Public Schools.* WCER Working Paper No. 2008-10, Wisconsin Center for Education Research. Madison: University of Wisconsin.

Cronin, J., Dahlin, M., Xiang, Y., and McCahon, D. (2009). *The Accountability Illusion.* Washington, DC: Thomas B. Fordham Institute.

Elmore, R.F. (2003). Accountability and capacity. In M. Carnoy, R.F. Elmore, and L.S. Siskin (Eds.), *High Schools and the New Accountability* (pp. 188-209). NewYork: Routledge/Falmer.

Ferrara, S. (2009). *The Maryland School Performance Assessment Performance (MSPAP), 1991-2002: Political Considerations.* Paper prepared for the Workshop of the Committee on Best Practices in State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Steve_Ferrara_Paper.pdf [accessed May 2010].

Fuller, B., Gesicki, K., Kang, E., and Wright, J. (2006). *Is the No Child Left Behind Act Working?: The Reliability of How States Track Achievement.* Working Paper No. 06-1. Berkeley: University of California and Stanford University, Policy Analysis for California Education.

Goertz, M.E. (2009). *Overview of Current Assessment Practices.* Paper prepared for the Workshop of the Committee on Best Practices in State Assessment Systems: Improving Assessment While Revisiting Standards, National Research Council, December 10-11, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Peg_Goertz_Paper.pdf [accessed May 2010].

Goertz, M.E., Olah, L.N., and Riggan, M. (2009). *Can Interim Assessments Be Used for Instructional Change?* CPRE Policy Briefs: Reporting on Issues and Research in Education Policy and Finance. Available: http://www.cpre.org/images/stories/cpre_pdfs/rb_51_role%20policy%20brief_final%20web.pdf [accessed May 2010].

Gong, B. (2010). *Innovative Assessment in Kentucky's KIRIS System: Political Considerations.* Presentation to the Workshop of the Committee on Best Practices in State Assessment Systems: Improving Assessment While Revisiting Standards, National Research Council, December 10-11, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Brian%20Gong.pdf [accessed May 2010].

Hambleton, R.K. (2009). *Using Common Standards to Enable Cross-National Comparisons.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/bota/Ron_Hambleton.pdf [accessed May 2010].

Ho, A.D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Education Researcher, 37*(6), 351-360.

Jennings, J., and Rentner, D.S. (2006). Ten big effects of No Child Left Behind on public schools. *Phi Delta Kappan, 88*(2), 110-113.

Kirst, M., and Mazzeo, J. (1996). The rise, fall, and rise of state assessment in California: 1993-1996. *Phi Delta Kappan, 78*(4), 319-323.

Koretz, D., and Barron, S. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System.* Santa Monica, CA: RAND.

Koretz, D. Mitchell., K., Barron, S., and Keith, S. (1996) *Perceived Effects of the Maryland School Performance Assessment Program*. Final Report, Project 3.2 State Accountability Models in Action. Washington, DC: U.S. Department of Education, National Center for Research on Evaluation.

Krajcik, J., McNeill, K.L., and Reiser, B. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*(1), 1-32.

Krajcik, J., Stevens, S., and Shin, N. (2009). *Developing Standards That Lead to Better Instruction and Learning.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Joe%20Krajcik%20and%20Shawn%20Stevens.pdf [accessed May 2010].

Lai, E.R., and Waltman, K. (2008). *The Impact of NCLB on Instruction: A Comparison of Results for 2004-05 to 2006-07*. IARP Report #7. Iowa City: Center for Evaluation and Assessment, University of Iowa.

Lane, S. (1999). *Impact of the Maryland School Performance Assessment Program (MSPAP): Evidence from the Principal, Teacher, and Student Questionnaires (Reading, Writing, and Science).* Paper presented at the annual meeting of the National Council on Measurement in Education, April 19-23, Montreal, Quebec, Canada.

Lazer, S. (2010). *Technical Challenges with Innovative Item Types.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Steve%20Lazer.pdf [accessed May 2010].

Marion, S. (2010). *Changes in Assessments and Assessment Systems Since 2002.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Scott%20Marion.pdf [accessed May 2010].

Mattson, D. (2010). *Science Assessment in Minnesota.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Dirk_Mattson.pdf [accessed May 2010].

McMurrer, J. (2007). *Choices, Changes, and Challenges; Curriculum and Instruction in the NCLB Era*. Washington, DC: Center on Education Policy.

Mislevy, R. (1998). Foundations of a new test theory. In N. Fredericksen, R.J. Mislevy, and I.I. Bejar (Eds.), *Test Theory for a New Generation of Tests (*pp. 19-38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J., and Riconscente, M. (2005). *Evidence-centered Assessment Design: Layers, Structures, and Terminology.* Menlo Park, CA: SRI International.

National Research Council. (1995). *Anticipating Goals 2000: Standards, Aassessment, and Public Policy: Summary of a Workshop*. Board on Testing and Assessment, Center for Education. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (1996). *National Science Education Standards*. National Committee on Science Education Standards and Assessment. Washington DC: National Academy Press.

National Research Council. (1999a). *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests.* Committee on Embedding Common Test Items in State and District Assessments. D.M. Koretz, M.W. Berthenthal, and B.F. Green (Eds.). Commission on Behavioral and Social Sciences and Education. Washington DC: National Academy Press.

National Research Council. (1999b). *Uncommon Measures: Equivalence and Linkage Among Educational Tests.* Committee on Equivalency and Linkage of Educational Tests. M.J. Feuer, P.W. Holland, B.F. Green, M.W. Berthenthal, and F.C. Hemphill (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2005). *Systems for State Science Assessment.* Committee on Test Design for K-12 Science Achievement. M.R. Wilson and M.W. Berthenthal (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

National Research Council. (2008). *Common Standards for K-12 Education? Considering the Evidence: Summary of a Workshop Series*. A. Beatty, Rapporteur. Committee on State Standards in Education: A Workshop Series. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olah, L., Lawrence, N., and Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*(2), 226-245.

Perie, M., Marion, S., and Gong, B. (2007). *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief*. Aspen, CO: Center for Assessment, The Aspen Institute, and Achieve, Inc. Available: http://www.achieve.org/files/TheRoleofInterimAssessments.pdf [accessed March 2010].

Porter, A.C., Polikoff, M.S., and Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis*, *31*(3), 238-268.

Schmidt, W.H., Wang, H.C., and McKnight, C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies, 37*, 525-559.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405-450.

Shepard, L. (2010). *Research Priorities for Next-Generation Assessment Systems.* Presentation to the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Lorrie_Shepard.pdf [accessed May 2010].

Shin, N., Stevens, S., and Krajcik, J. (in press). *Using Construct-Centered Design as a Systematic Approach for Tracking Student Learning Over Time.* London, England: Routledge, Taylor & Francis Group.

Smith, C.L., Wiser, M., Anderson, C.W. and Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1), 1-98.

Smith, M., and O'Day, J. (1991). Systematic school reform. In S. Fuhrman and B. Malen (Eds.), *The Politics of Curriculum and Testing* (pp. 233-267). Philadelphia: Falmer Press.

Stecher, B., and Hamilton, L. (2009). *What Have We Learned from Pioneers in Innovative Assessment?* Paper prepared for the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, National Research Council, December 10-11, Washington, DC. Available: http://www7.nationalacademies.org/bota/Brian_Stecher_and_Laura_Hamilton.pdf [accessed May 2010].

Stecher, B.M., Epstein, S., Hamilton, L.S., Marsh, J.A., Robyn, A., McCombs, J.S., Russell, J.L., and Naftel, S. (2008). *Pain and Gain: Implementing No Child Left Behind in California, Georgia, and Pennsylvania, 2004 to 2006*. Santa Monica, CA: RAND.

Stevens, S., Sutherland, L., and Krajcik, J.S., (2009). *The Big Ideas of Nanoscale Science and Engineering: A Guidebook for Secondary Teachers*. Arlington, VA: National Science Teachers Association.

Sunderman, G. (Ed.). (2008). *Holding NCLB Accountable: Achieving Accountability, Equity, and School Reform*. Thousand Oaks, CA: Corwin Press.

Toch, T. (2006). *Margins of Error: The Testing Industry in the No Child Left Behind Era*. Washington, DC: Education Sector.

U.S. Department of Education. (2009). *Race to the Top Program Executive Summary.* Available: http://www.ed.gov/programs/racetothetop/resources.html [accessed January 2010].

U.S. Government Accountability Office. (2009). *No Child Left Behind Act: Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments.* GAO Report-09-911. Available: http://www.gao.gov/cgi-bin/getrpt?GAO-09-911 [accessed November 2009].

U.S. Government Accounting Office. (2003). *Characteristics of Tests Will Influence Expenses: Information Sharing May Help States Realize Efficiencies.* GAO Report-03-389. Available: http://www.gao.gov/cgi-bin/getrpt?GAO-03-389 [accessed November 2009].

Wiggins, G., and McTighe, J. (1998). *Understanding by Design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wilson, M. (Ed.). (2004). *Towards Coherence between Classroom Assessment and Accountability, 103rd Yearbook of the National Society for the Study of Education, Part II.* Chicago, IL: The University of Chicago Press.

Wilson, M. (2005). *Constructing Measures: An Item-Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). *Developing Assessment Tasks That Lead to Better Instruction and Learning.* Presentation to the Workshop for the Committee on Best Practices in State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/bota/Mark_Wilson.pdf [accessed May 2010].

Wise, L. (2009). *How Common Standards Might Support Improved State Assessments.* Paper prepared for the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Laurie_Wise_Paper.pdf [accessed May 2010].

Zwick, R. (2009). *State Achievement Comparisons: Is the Time Right?* Paper prepared for the Workshop of the Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards, December 10-11, National Research Council. Washington, DC. Available: http://www7.nationalacademies.org/bota/Rebecca_Zwick_Paper.pdf [accessed May 2010].

Appendix A

Workshop Agenda

Best Practices for State Assessment Systems
Workshop 1
December 10-11, 2009

National Academy of Sciences
2100 C Street, NW
Washington, DC
Auditorium

Thursday, December 10

OPEN

9:15-9:45 **Welcome**
Stuart Elliott, Director, Board on Testing and Assessment
Judith Rizzo, Executive Director and CEO, James B. Hunt, Jr. Institute for Educational Leadership and Policy

Overview of Workshop Goals
Diana Pullin, *Chair,* Committee on Best Practices for State Assessment Systems

- **Précis of previous workshop series and report**
- **Overview of the goals and plans for the current workshop series**
- **Introduction of the idea of innovative assessment**
- **Discussion of the current status of the common standards movement**

9:45-11:30 **Session I. Examining the Status Quo: What Are the Benefits and Limitations of the Current Approaches to Assessment in This Country?**
Moderators: Diana Pullin, Dirk Mattson

(9:45-10:15) Overview of Current Assessment Practices
This session will provide a review of the current test-based accountability system, the goals and purposes it has developed to serve, and its strengths and limitations.
Presenter: Peg Goertz, University of Pennsylvania

(10:15-10:45) Changes in Assessments and Assessment Systems Over the Past Decade
This session will review the ways assessments and approaches to assessment have changed over the past decade, including changes in item types, uses of local and interim assessments, and advancements in assessment of special populations.
Presenter: Scott Marion, National Center for Improvement of Educational Assessment (NCIEA)

(10:45-11:15) Synthesis of Key Ideas
Discussant: Joan Herman, National Center for Research on Evaluation, Standards, & Student Testing (CRESST)

(11:15-11:30) Focused Discussion
Moderators lead focused discussion with presenters and audience members.

11:30-12:15 **Working Lunch**

12:15-3:45 **Session II. Changing the Status Quo**
Moderators: Joan Herman, Rebecca Maynard

(12:15-1:00) Developing Standards That Lead to Better Instruction and Learning
This session will discuss ways to specify standards so that they (1) more accurately delineate the skills and knowledge to be learned and (2) can be more accurately and readily translated into instruction and assessment. Examples will be drawn from the draft common core standards.
Presenters: Joe Krajcik and Shawn Stevens, University of Michigan

(1:00-1:45) Developing Assessment Tasks That Lead to Better Instruction and Learning
This session will explore ways to use more elaborated standards to develop assessment tasks that accurately measure the intended skill and knowledge, with a particular focus on

ways to ensure that assessments measure higher-order, critical thinking skills using a variety of item types.
Presenter: Mark Wilson, University of California, Berkeley

(1:45-2:30) Technical Challenges of Implementing Innovative Assessments
This session will explore the technical challenges associated with developing more innovative assessment tasks that measure challenging content and skills, tradeoffs associated with these kinds of assessments/tasks, and ways that the information gathered from innovative assessments might be used to support better decision making about students and instruction.
Presenter: Stephen Lazer, Educational Testing Service

2:30-2:45 **Break**

2:45-3:15 **Synthesis of Key Ideas**
Discussant: Scott Marion, NCIEA

3:15-3:45 **Focused Discussion**
Moderators lead focused discussion with presenters and audience members.

3:45-5:00 **Session III.A What Is the Status of Innovative Assessment?**
Moderators: Diana Pullin, Mark Wilson

(3:45-4:30) Lessons from the Past and Current Efforts
This session will provide an overview of the experiences of pioneers in the area of innovative assessment, such as programs developed for Kentucky (KIRIS), Maryland (MSPAP), Vermont (Portfolio Assessment Program), and California (CLAS performance assessment) which are no longer in operation. Examples from currently operational assessment programs, international assessments, and in fields other than K-12 education will also be discussed.
Presenters: Brian Stecher and Laura Hamilton, RAND

(4:30-5:00) Focused Discussion
Moderators lead focused discussion with presenters and audience members.

5:00 **Presentation by Lauren Resnick, Learning Research and Development Center, University of Pittsburgh**

5:30 **Adjourn Workshop**
Reception

6:00 **Working Group Dinner (in Lecture Room)**

Friday, December 11

OPEN

8:30-10:30 **Session III.B What Is the Status of Innovative Assessment? (cont.)**
Moderators: Diana Pullin, Mark Wilson

(8:30-9:30) Panel Discussion: Political Considerations
This session will explore the political/practitioner perspective on the pioneer program discussed in Part A. Panelists representing several of the programs will address the following questions:

- What was the motivation for the assessment? Why was it considered? Who wanted it? Who wasn't in favor of it?
- What was involved in getting the assessment adopted? What, if any, obstacles were encountered? How were they overcome?
- What was involved in developing the assessment? What, if any, obstacles were encountered? How were they overcome?
- What issues were encountered with implementation of the assessment? What, if any, obstacles were encountered? How were they overcome?
- What were primary reasons for the demise of the program?

Panelists: Steve Ferrara (MSPAP), Brian Gong (KIRIS), and Dirk Mattson (Minnesota)

(9:30-10:00) Synthesis of Key Ideas
Discussant: Lorraine McDonnell, University of California, Santa Barbara

(10:00-10:30) Focused Discussion
Moderators lead focused discussion with presenters and audience members.

10:30-10:45 Break

10:45-2:15 Session IV. Exploring the Opportunities
Moderators: Rebecca Maynard, Dirk Mattson

(10:45-11:30) What Opportunities Does the Common Standards Movement Offer for Improving Assessment?
This session will explore the opportunities the common standards movement might offer for moving to more innovative assessments that assess challenging content and also give more information to teachers and local decision makers. The presentation will address technical issues and potential benefits of collaboration across states, drawing on examples from the experiences of Vermont, New Hampshire, Rhode Island, and Maine (New England Common Assessment Program)
Presenter: Laurie Wise, HumRRO

11:30-12:30 Working Lunch

Session IV. Exploring the Opportunities (cont.)

(12:30-1:15) Using Common Standards to Enable Cross-State Comparisons
This session will focus on the elements that would need to be in place in order for test results to be compared across states, including issues associated with adding state-specific items. The presentation will address the inferences that policy makers and test users might want to make and what is required to support each kind of inference.
Presenter: Ron Hambleton, University of Massachusetts, Amherst

(1:15-1:45) Synthesis of Key Ideas
Discussant: Rebecca Zwick, ETS and University of California, Santa Barbara

(1:45-2:15) Focused Discussion
Moderators lead focused discussion with presenters and audience members.

2:15-2:30 **Break**

2:30-4:00 **Session V. Setting Research Priorities**
Moderators: Diana Pullin, Scott Marion

(2:30-3:00) Research Priorities
The U.S. Department of Education has set aside $350 million for developing tests to measure common standards. This panel will listen to the workshop discussions and consider the implications for research. The presenter and discussants will address the following questions:

- Given the issues raised during the workshop, what are realistic priorities for research?
- What projects/efforts are most in need of research?
- How would you proportionally allocate the funding?

Presenter: Lorrie Shepard, University of Colorado

(3:00-3:45) Responses
Discussants: Laurie Wise, HumRRO; Joan Herman, CRESST; Rebecca Maynard, University of Pennsylvania

(3:45-4:15) Focused Discussion
Moderators lead focused discussion with presenters and audience members.

4:15 **Closing Remarks, Adjourn**
Diana Pullin, *Chair*

Appendix B

Workshop Participants

Joan Abdallah, American Association for the Advancement of Science
David Abrams, New York State
Frank Adamson, Stanford University
Martha Aliaga, American Statistical Association
Jaime Allentuck, Government Accountability Office
Guy-Alain Amoussou, National Science Foundation
Allison Armour-Garb, Rockefeller Institute of Government
Sally Atkins-Burnett, Mathematica
Margaret Bartz, Chicago Public Schools
Alix Beatty, National Research Council
Rolf K. Blank, Council of Chief State School Officers
Bonnie Bracey Sutton, Emaginos
Molly Broad, American Council on Education
Gina Broxterman, National Center for Education Statistics
J.B. Buxton, Metametrics
Peg Cagle, California Teachers Advisory Council
Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Doug Christensen, Nebraska Department of Education
Julia Clark, National Science Foundation
Sherri Coles, Family Support Center on Disabilities
Bruce W. Colletti, Northern Virginia Community College
Jere Confrey, North Carolina State University
Tim Crockett, Measured Progress
Christopher Cross, Cross and Joftus, LLC

Jerome Dancis, University of Maryland
Linda Darling-Hammond, Stanford University
Stephanie Dean, James B. Hunt, Jr. Institute
George E. DeBoer, American Association for the Advancement of Science
Gabriel Della-Piana, University of Utah
Betty Demarest, National Education Association
Craig Deville, Measurement Incorporated
Pasquale DeVito, Measured Progress
Shelby Dietz, Center on Education Policy
Chris Domalski, National Center for the Improvement of Educational Assessment
Nancy Doorey, Educational Testing Service
Kelly Duncan, National Research Council
Janice Earle, National Science Foundation
John Easton, U.S. Department of Education
Tracey Edou, Department of Energy
David Egnor, U.S. Department of Education
Lisa Ehrlich, Measured Progress
Stuart Elliott, National Research Council
John Ewing, Math for America
Florence D. Fasanelli, American Association for the Advancement of Science
Steve Ferrara, CTB/McGraw-Hill
Michael Feuer, National Research Council
Rebecca Fitch, U.S. Office for Civil Rights
Beth Foley, National Education Association
Pat Forgione, Educational Testing Service
Denise Forte, U.S. House of Representatives, Committee on Education and Labor
Gavin Fulmer, National Science Foundation
Randall Garton, Shanker Institute
Michael Gilligan, James B. Hunt, Jr. Institute
Alan Ginsburg, U.S. Department of Education
Margaret Goertz, University of Pennsylvania
Brian Gong, National Center for Improvement of Educational Assessment
Mandi Gordon, George Mason University
Mark D. Greenman, National Science Foundation
Eunice Greer, National Center for Education Statistics
Laura Hamilton, RAND
Pierce Hammond, Office of Vocational and Adult Education
Mariana Haynes, National Association of State Boards of Education
Andres Henriquez, Carnegie Corporation
Joan Herman, CRESST
Richard Hill, National Science Foundation

Margaret Hilton, National Research Council
Gene Hoffman, Human Resources Research Organization
Yung-chen Hsu, American Council on Education
Bruce Hunter, American Association of School Administrators
Kirk Janowiak, U.S. Department of Education
Arundhati Jayarao, Einstein Fellow, Office of Senator Kirsten Gillibrand (D-NY)
Michael Jennings, University of Alabama
Wyn Jennings, National Science Foundation
Allan Jones, Emaginos
Barb Kapinus, National Education Association
Martin Kehe, GED Testing Service
Tom Keller, National Research Council
William Kelly, American Society for Engineering Education
Eugenia Kemble, Shanker Institute
Julie Kochanek, Learning Point Associates
Judith Koenig, National Research Council
Ken Krehbiel, National Council of Teachers of Mathematics
Melissa Lazarin, Center for American Progress
Steve Lazer, Educational Testing Service
Anne Lewis, K-12 Assessment and Performance Management Center, Educational Testing Service
Dane Linn, National Governors Association
Alan Maloney, North Carolina State University
David Mandel, National Center on Education and the Economy
Scott Marion, National Center for the Improvement of Educational Assessment
Dirk Mattson, Minnesota Department of Education
Rebecca Maynard, University of Pennsylvania
Lorraine McDonnell, University of California, Santa Barbara
Jamie McKee, Bill & Melinda Gates Foundation
Tiah McKinney, George Mason University
Raegen Miller, Center for American Progress
Sherri Miller, ACT
Zipporah Miller, National Science Teachers Association
Chris Minnich, Council of Chief State School Officers
Hassan Minor, Howard University Middle School of Mathematics and Science
William Montague, Association of Independent Schools of Greater Washington
Scott Montgomery, Council of Chief State School Officers
Jean Moon, National Research Council
Patricia Morison, National Research Council
Lesley Muldoon, Achieve
F. Howard Nelson, AFL-CIO

Rose Neugroschel, National Research Council
Alexander Nicholas, National Science Foundation
Steven Obenhaus, Office of Senator Joseph I. Lieberman (ID-CT)
Carol O'Donnell, U.S. Department of Education
Cornelia Orr, National Assessment Governing Board
Ray Pecheone, Stanford University
Anthonette Pena, National Science Foundation
Marianne Perie, National Center for the Improvement of Educational Assessment
Ashley Clark Perry, James B. Hunt, Jr. Institute
Kristina Peterson, House Committee on Education and Labor
Valena Plisko, National Center for Education Statistics
Gerrita Postlewait, Stupski Foundation
Diana Pullin, Boston College
Sam Rankin, American Mathematical Society
Joseph Reed, National Science Foundation
Sue Rigney, U.S. Department of Education
Judith Rizzo, James B. Hunt, Jr. Institute
Roy Romer, The College Board
Robert Rothman, Alliance for Excellent Education
Gerhard Salinger, National Science Foundation
Eugene Schaffer, University of Maryland, Baltimore
Elizabeth Schneider, Alliance for Excellent Education
Susan Sclafani, National Center for Education and the Economy
Kelly Scott, The Aspen Institute-Commission on NCLB
Robert Scott, Department of Education, Texas
Barbara Shannon, California Teachers Advisory Council
Lorrie Shepard, University of Colorado
Elena Silva, Education Sector
Malbert Smith, MetaMetrics
Patty Sobecky, University of Alabama
Nancy Spillane, National Science Foundation
Brian Stecher, RAND
Jack Stenner, MetaMetrics
Marc Sternberg, U.S. Department of Education
Shawn Stevens, University of Michigan
Justin Stone, Associate for the American Federation of Teachers
Martin Storksdieck, National Research Council
Vic Sutton, Emaginos
Kevin Sweeney, The College Board
Monica Thammarath, Southeast Asia Resource Action Center
Doua Thor, Southeast Asia Resource Action Center
Tom Toch, Association of Independent Schools of Greater Washington

LeRoy Tompkins, Office of the State Superintendent of Education, DC
Elizabeth VanderPutten, National Science Foundation
Dave Vannier, National Institutes of Health
Joyce VanTassel-Baska, Center for Gifted Education
David Wakelyn, National Governors Association
Michael Wallace, Howard University
Wanda Ward, Education and Human Resources-National Science Foundation
Denny Way, Pearson
Susan Weigert, U.S. Department of Education
Joanne Weiss, U.S. Department of Education, Race to the Top Initiative
Antoinette Wells, NASA
Ann Whalen, U.S. Department of Education
April White, James B. Hunt, Jr. Institute
Amber Wilke, Northwest Evaluation Association
Joe Willhoft, Office of Superintendent of Public Instruction, Washington
Mark Wilson, University of California, Berkeley
Bob Wise, Alliance for Excellent Education
Lauress Wise, Human Resources Research Organization
Steve Wise, Northwest Evaluation Association
Zhijian Wu, University of Alabama
Judy Wurtzel, U.S. Department of Education, Office of Planning, Evaluation and Policy Development
Raymond Yeagley, Northwest Evaluation Association
Rebecca Zwick, University of California, Santa Barbara, and Educational Testing Service