



**Assessment in Support of Instruction and Learning:
Bridging the Gap Between Large-Scale and
Classroom Assessment - Workshop Report**
Committee on Assessment in Support of Instruction and
Learning, Committee on Science Education K-12,
National Research Council

ISBN: 0-309-52616-7, 76 pages, 6 x 9, (2003)

**This free PDF was downloaded from:
<http://www.nap.edu/catalog/10802.html>**

Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online for free
- Purchase printed books and PDF files
- Explore our innovative research tools – try the [Research Dashboard](#) now
- [Sign up](#) to be notified when new books are published

Thank you for downloading this free PDF. If you have comments, questions or want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to comments@nap.edu.

This book plus thousands more are available at www.nap.edu.

Copyright © National Academy of Sciences. All rights reserved.

Unless otherwise indicated, all materials in this PDF file are copyrighted by the National Academy of Sciences. Distribution or copying is strictly prohibited without permission of the National Academies Press [<http://www.nap.edu/permissions/>](http://www.nap.edu/permissions/). Permission is granted for this material to be posted on a secure password-protected Web site. The content may not be posted on a public Web site.

ASSESSMENT IN SUPPORT OF INSTRUCTION AND LEARNING

Bridging the Gap Between **Large-Scale** and **Classroom Assessment**

Workshop Report

Committee on Assessment in Support of Instruction and Learning

Board on Testing and Assessment
Committee on Science Education K-12
Mathematical Sciences Education Board
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 study was supported by Contract/Grant No. ESI-0102582 between the National Academy of Sciences and the National Science 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 0-309-08978-6 (Book)

International Standard Book Number 0-309-52616-7 (PDF)

Additional copies of this report are available from the 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>

Copyright 2003 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America.

Suggested citation: National Research Council (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment. Workshop report*. Committee on Assessment in Support of Instruction and Learning. Board on Testing and Assessment, Committee on Science Education K-12, Mathematical Sciences Education Board. 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. Bruce M. Alberts 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. Wm. A. Wulf 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. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

**COMMITTEE ON ASSESSMENT IN SUPPORT OF
INSTRUCTION AND LEARNING**

J. MYRON ATKIN (*Chair*), Center for Educational Research, Stanford
University

EVA L. BAKER, School of Education, University of California, Los Angeles

JAN DE LANGE, Freudenthal Institute, Utrecht University, The Netherlands

TOM KELLER, Maine Department of Education, Augusta

JAMES MINSTRELL, Talaria, Inc., Seattle

MARGE M. PETIT, National Center for the Improvement of Educational
Assessment, Portsmouth, New Hampshire

ANTHONY SCOTT, Chicago Public Schools

LORRIE A. SHEPARD, School of Education, University of Colorado, Boulder

GUADALUPE VALDES, Department of Spanish and Portuguese, Stanford
University

MERYL W. BERTENTHAL, *Study Director*

ANDREW E. TOMPKINS, *Research Assistant*

MICHAEL DECARMINE, *Project Assistant*

BOARD ON TESTING AND ASSESSMENT

EVA L. BAKER (*Chair*), The Center for the Study of Evaluation, University of California, Los Angeles

LORRAINE MCDONNELL (*Vice Chair*), Departments of Political Science and Education, University of California, Santa Barbara

LAURESS L. WISE (*Vice Chair*), Human Resources Research Organization, Alexandria, Virginia

CHRISTOPHER F. EDLEY, JR., Harvard University Law School

EMERSON J. ELLIOTT, Independent Consultant, Arlington, Virginia

MILTON D. HAKEL, Department of Psychology, Bowling Green State University

ROBERT M. HAUSER, Institute for Research on Poverty, Center for Demography, University of Wisconsin, Madison

PAUL W. HOLLAND, Educational Testing Service, Princeton, New Jersey

DANIEL M. KORETZ, Graduate School of Education, Harvard University

EDWARD P. LAZEAR, Graduate School of Business, Stanford University

RICHARD J. LIGHT, Graduate School of Education and John F. Kennedy School of Government, Harvard University

ROBERT J. MISLEVY, Department of Measurement and Statistics, University of Maryland

JAMES W. PELLEGRINO, University of Illinois, Chicago

LORRIE A. SHEPARD, School of Education, University of Colorado, Boulder

KENNETH I. WOLPIN, Department of Economics, University of Pennsylvania

PATRICIA MORISON, *Acting Director*

LISA ALSTON, *Administrative Associate*

COMMITTEE ON SCIENCE EDUCATION K-12

CARY SNEIDER (*Chair*), Boston Museum of Science

CARLO PARRAVANO (*Vice Chair*), Merck Institute for Science Education,
Rahway, New Jersey

TANYA ATWATER, Department of Geological Sciences, University of
California, Santa Barbara

FRANCISCO AYALA, Department of Ecology and Evolutionary Biology,
University of California, Irvine

CAROL BREWER, Division of Biological Sciences, University of Montana,
Missoula

JUANITA CLAY-CHAMBERS, Detroit Public Schools

KATHLEEN COMFORT, WestEd, San Francisco

DAVID CONLEY, Center for Educational Policy Research, University of
Oregon, Eugene

ALAN FRIEDMAN, New York Hall of Science, Corona

JEFFREY FRIEDMAN, Friedman Lab, Rockefeller University, New York

BARBARA GONZALEZ, Department of Chemistry and Biochemistry,
California State University, Fullerton

LINDA GREGG, TERC, Cambridge, Massachusetts

PATRICIA HARMON, San Francisco Unified School District, California

JENIFER HELMS, Educational Consultant/Evaluator, Denver

ANNE JOLLY, The Regional Laboratory at SERVE, Mobile, Alabama

JUDITH JONES, East Chapel Hill High School, North Carolina

TOM KELLER, Maine Department of Education, Augusta

OKHEE LEE, School of Education, University of Miami, Florida

JAMES MINSTRELL, Talaria, Inc., Seattle

MARY JANE SCHOTT, Charles A. Dana Center, Austin, Texas

JERRY VALADEZ, Fresno Unified School District, California

JEAN MOON, *Study Director*

JULIE SCHUCK, *Research Associate*

LASHAWN SIDBURY, *Senior Project Assistant*

MATHEMATICAL SCIENCES EDUCATION BOARD

JOAN LEITZEL (*Chair*), President Emerita, University of New Hampshire
JERE CONFREY (*Vice Chair*), Department of Curriculum and Instruction,
University of Texas, Austin

JUDY ACKERMAN, Montgomery College, Rockville, Maryland
DEBORAH LOEWENBERG BALL, School of Education, University of
Michigan

THOMAS BANCHOFF, Department of Mathematics, Brown University
JAN DE LANGE, Freudenthal Institute, Utrecht University, The Netherlands
LOUIS GOMEZ, School of Education and Social Policy, Northwestern
University

DOUGLAS A. GROUWS, Curriculum and Instruction, University of Iowa
ARTHUR JAFFE, Department of Mathematics, Harvard University
ERIC JOLLY, Education Development Center, Newton, Massachusetts
DANIEL KENNEDY, The Baylor School, Chattanooga, Tennessee
JIM LEWIS, Department of Mathematics and Statistics, University of
Nebraska, Lincoln

KAREN LONGHART, Flathead High School, Kalispell, Montana
GEORGE MCSHAN, National School Boards Association, Harlingen, Texas
KAREN MICHALOWICZ, The Langley School, McLean, Virginia
JUDITH MUMME, WestEd, Camarillo, California
CASILDA PARDO, Valle Vista Elementary School, Albuquerque
SUE PARSONS, Department of Mathematics, Cerritos College, Norwalk,
California

MARGE PETIT, The National Center for the Improvement of Educational
Assessment, North Fayston, Vermont

DONALD SAARI, Distinguished Professor of Economics and Professor of
Mathematics, University of California, Irvine

RICHARD SCHEAFFER, Professor Emeritus, University of Florida
WILLIAM STEENKEN, Hamilton, Ohio

FRANCIS SULLIVAN, Center for Computing Sciences, Bowie, Maryland
HUNG HIS WU, Department of Mathematics, University of California,
Berkeley

CAROLE LACAMPAGNE, *Study Director*

VICKI STOHL, *Research Associate*

DIONNA WILLIAMS, *Senior Project Assistant*

Preface

The National Research Council (NRC) Workshop on Bridging the Gap Between Large-Scale and Classroom Assessment was convened during a period of rising attention in education policy circles to matters of testing and assessment. At this juncture in American education history, the emphasis is increasingly on large-scale examinations developed outside the classroom to gauge what students know. Their aim is primarily to strengthen public accountability. This kind of assessment, which now is projected at orders of magnitude much greater than anything yet seen in this country, is already having profound effects. There are serious consequences—financial and otherwise—for students, parents, teachers, schools, and districts associated with the test results. Tests have also been shown to have powerful influences on curriculum and teaching methods.

One problem with relying exclusively on tests designed to examine millions of students is that they do not easily conform to curricula devised to match state and national standards for mathematics or science. Nor do they do much to promote the kind of student learning that is reflected in those standards. Additionally, these external assessments may have little relation to what students are learning and teachers are teaching in their classrooms. Most important, at present the system does not usually incorporate forms of assessment that have been shown, when done well, to have a direct and positive influence on how much students learn: specifically, the assessments that are part of a teacher's everyday classroom practice and that are integrated into instruction.

To quote from a recent publication, *Knowing What Students Know* (NRC, 2001c) from the Board on Testing and Assessment, one of three NRC standing boards and committees that joined to organize the present workshop, “The cur-

rent imbalance of mandates and resources should be redressed by shifting from an emphasis on external forms of assessment to an increased emphasis on classroom formative assessment to assist learning” (p. 310).

The NRC workshop reported here addressed that gap between external and classroom assessment. During the workshop we heard about issues associated with designing an assessment system that meets the demands of public accountability and, at the same time, improves the quality of the education that students receive day by day. The workshop focused on assessment that addresses both accountability and learning.

What guidelines or criteria might be developed to take advantage of the strengths and potential inherent in large-scale examinations, on the one hand, and everyday assessment in the classroom on the other? How might steps be taken to minimize the sometimes counterproductive nature of some assessment practices—indeed, to maximize the potential of each practice? What are the challenges? What is gained and what is lost as the states and the nation try to create a coherent and integrated assessment system? These are some of the many questions raised.

The heart of the workshop was an opportunity to learn about approximately a dozen programs in which attempts are being made to bridge the gap. It should be recognized that the workshop was exploratory. It was not a showcase. None of the programs that were described and discussed is perfect. Few are exemplary, except in the goals they are trying to accomplish. Most face serious challenges. The members of the committee that planned the workshop are deeply indebted to those who agreed to talk about the current state of their work in a setting that encouraged probing questions. All the participants recognized that it would take hard and steady effort to construct a high-quality system.*

A further goal of the workshop was to establish clearer directions for specific NRC initiatives in the months and years ahead to inform the larger education community about issues associated with assessment, learning, and accountability. Therefore the genesis of the workshop is relevant. Three of the constituent bodies of the NRC’s Center for Education joined to plan the two-day meeting. They will be involved in whatever initiatives grow out of the workshop deliberations.

For ten years, the Board on Testing and Assessment has been producing insightful publications on improving large-scale examinations. It has helped the education community and the public to recognize the strengths and limitations of such examinations. It has led the way in synthesizing research on the topic, making recommendations, and pointing out areas that need additional serious study. The Mathematical Sciences Education Board and the Committee on Science Education K-12, while not inattentive to assessment issues, have focused primarily on matters of curriculum and teacher education. Bringing these three

*See Appendix C for sources of further information about the programs discussed.

groups together to lead this workshop, and inform the NRC's future work in this arena, helps to ensure the kind of scope and comprehensiveness needed around the topic of assessment for both learning and accountability.

In planning this workshop the Committee on Assessment in Support of Instruction and Learning benefited tremendously from the contributions and goodwill of many people, and the committee is grateful for their support. First, we wish to acknowledge the National Science Foundation (NSF), which sponsored this workshop through a grant to the Center for Education. We particularly thank Janice Earle, who served as the link between the NSF and the committee. The Board on Testing and Assessment, the Committee on Science Education K-12, and the Mathematical Sciences Education Board—the units within the National Research Council that launched this workshop—were instrumental in shaping the project and in providing general guidance and support along the way.

Within the NRC, a number of individuals supported the project. Michael Feuer, executive director of the Division of Behavioral and Social Sciences and Education; Patricia Morison, associate director of the Center for Education; and Jay Labov, deputy director of the Center for Education, provided support and encouragement along the way. The committee expresses particular gratitude to the members of the NRC project staff for contributing their intellectual and organizational skills throughout the life of the project. Meryl Bertenthal, the project's study director, helped to conceptualize the workshop and provided guidance and support to the committee. Judy Koenig was responsible for planning the committee's first meeting and, at the workshop, proved to be a skilled note taker and exacting timekeeper. Andrew Tompkins provided excellent research support and adeptly handled all of the logistics related to the workshop. We were particularly impressed by his knowledge and use of technology, which allowed us to feature more than twenty speakers and their slides without a single glitch. Michael DeCarmine ably assisted Andrew in ensuring that the committee's work proceeded smoothly. The committee is extremely grateful to Alix Beatty for her skillful writing of this workshop summary.

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 NRC's Report Review Committee. 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 study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report: Paul J. Black, Department of Education and Professional Studies, King's College, London; Peggy Carlisle, Teacher, Pecan Park Elementary School, Jackson, Mississippi; Sharon Sikora, Center for Learning and Teaching of the West, Colorado State University; and

Gary Sykes, Education Administration and Teacher Education, Michigan State University

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the final draft of the report before its release. The review of this report was overseen by Marshall S. Smith, Education Program, The William and Flora Hewlett Foundation, Menlo Park, California. Appointed by the National Research Council, 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 authoring committee and the institution.

Finally, I would like to thank all of the committee members, who generously contributed their time and intellectual efforts to this project. The organization of such a large workshop and the conceptualization of the criteria for selecting programs to feature was an extraordinary challenge that they met extremely well.

J. Myron Atkin, *Chair*

Contents

Introduction	1
The Criteria in Context	5
The Ideal, 10	
Large-Scale Assessments, 11	
Classroom Assessments, 13	
The Nature of the Gap	16
Some International Examples	19
Australia, 20	
Queensland, 21	
Great Britain: Enhanced Formative Assessment, 22	
The International Baccalaureate (IB) Diploma Programme, 23	
Programme for International Student Assessment (PISA), 23	
Assessment to Improve Learning	26
Nebraska: School-Based Teacher-Led Assessment Recording System, 26	
Delaware: Comprehensive Science Assessment, 27	
Vermont: The Vermont Assessment System and the Partnership for the Assessment of Standards-Based Science, 29	
Wyoming: Body of Evidence System, 30	

Maine: Comprehensive Assessment System, 32	
Washington: Adapting a Traditional Assessment, 34	
Berkeley Evaluation and Assessment Research System, 35	
Northern California Mathematics Assessment Collaborative, 36	
Facet-Based Assessment, 38	
Model-Based Assessment, 39	
Concluding Thoughts and Possible Next Steps	42
References	45
Appendices	
A Workshop Agenda	47
B Workshop Participants	51
C Resources for Further Information	55

1

Introduction

Educational assessments are a major feature of the educational landscape in the United States. They serve many purposes—policy makers and administrators use them to monitor both the progress of schools and systems and the relative success of educational policies, for example, and also to answer questions about individual students for placement and other purposes. These purposes, for which large-scale, standardized, assessments are usually used, generate the most public discussion, but assessments are also used by teachers, in both formal and informal ways on a daily basis, to monitor students' learning and to identify specific areas in which further work is needed. Classroom assessments are an important tool for providing feedback to students so they can adjust their learning; they also help teachers to identify student misconceptions and to modify their instruction accordingly.¹ Whatever form it takes, classroom assessment is a critical component of effective instruction.

Although both kinds of assessments have a very important role to play, they are not often accorded equal weight by policy makers or in public discussion. Large-scale assessments have become increasingly politicized, at both the local and national levels. Their results have been used in political campaigns and other venues to make points they were not designed to support. Large-scale test results are also widely used to make both formal and informal evaluations of local

¹In discussing classroom assessment the committee is thinking of the assessments that are part of ongoing classroom life, such as written or oral weekly quizzes, end-of-semester examinations, portfolios, and comments and grades on homework assignments (NRC, 2001b).

schools (and thus can influence property values). As states work to comply with the testing provisions of the No Child Left Behind Act, the nation is likely to see both a greater quantity of large-scale tests, and heightened attention to their results.

For all these reasons, and, perhaps, simply because they are so much more visible, large-scale tests are far more frequently on the public agenda than their classroom counterparts. Moreover, the two kinds of tests are seldom aligned in such a way that they can support one another. Indeed, classroom teachers do not always recognize the potential of large-scale assessments because the assessments their students are given are not directly relevant to their instructional goals, and also in many cases because teachers have not had sufficient training in assessment issues to understand fully how best to use such tests and the data they generate. The feedback from large-scale assessments is often too general for teachers to use in making future curricular and instructional decisions and often arrives so long after the assessment that it cannot be applied to current students.

At the same time, large-scale assessment programs rarely seem to tap into the insights about students' learning that classroom teachers are in a unique position to offer through their own assessments. Though classroom assessments are often focused on what are known as "formative" purposes—to provide immediate feedback that can shed light on student learning—they can also provide "summative" evidence about students that can be used to classify or place them, for example. In a number of contexts, as will be discussed below, educators have found that classroom assessments, if properly designed, can be used for the broader accountability purposes that are more typical of large-scale assessments. At present, however, there is an apparently large gulf between the two types of assessments as they are used in the United States; close inspection of this gap reveals an array of interrelated issues.

The gap between classroom and large-scale assessments has caught the attention of several National Research Council (NRC) committees, and one result has been a clear consensus that instruction and learning are best supported in educational systems when large-scale and classroom assessments are aligned with each other and with standards, curriculum, instruction, and professional development.² A three-year study of the implications of new information about learning and cognition for educational assessments resulted in a report, *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001c), which lays out several features that would characterize an educational system that achieves this seamless integration.

²By an aligned system, the committee means one in which each of the key elements has been designed both with reference to one another and with reference to overarching system goals. In such a system, the elements work together rather than, as can easily happen in a large, complex enterprise such as a public school system, at cross purposes.

As the committee that wrote that report recognized, the ideal of seamless alignment has proved difficult to achieve in practice. To better understand how the ideal of alignment is conceptualized in practice, three NRC boards—the Mathematical Sciences Education Board, the Committee on Science Education K-12, and the Board on Testing and Assessment—formed a joint steering committee, the Committee on Assessment in Support of Instruction and Learning, to plan a workshop that would bring together leading experts in measurement and assessment with international, state, and local program directors to illustrate some ways in which classroom and large-scale assessments can work together conceptually and operationally to better support student learning.

The goal of the workshop was to highlight current efforts to align classroom and large-scale assessments with each other and with instruction, standards, curriculum, and professional development. To accomplish this, the workshop featured discussions of the relative successes and challenges of science and mathematics assessment systems that are attempting to bridge the gap between classroom and large-scale assessments; it also included discussions of research-based visions of effective assessment programs that have not yet been put into practice on a large scale. Featured programs would be selected based on their potential to provide insight into the ways in which more coherent assessments could be designed and implemented. Selected workshop speakers would also explore practices in other countries, alternatives to standardized tests as sources of data for accountability purposes, and opportunities and advances in our understanding of cognition and learning.

The intent of the workshop would not be to evaluate the programs presented, but rather to gain a better understanding of the ways in which the ideals of a coherent assessment system, as described in the research literature and synthesized in a number of NRC reports (1993, 1998, 2000, 2001a, 2001b, 2001c, and 2002), might be implemented in practice.

Planning for the workshop was shaped by a set of specific criteria, discussed below, that might characterize an ideal system. These criteria were distilled from the reports listed above as well as from other relevant research, for example, *National Science Education Standards* (NRC, 1996); *Assessment Standards for School Mathematics* (National Council of Teachers of Mathematics, 1995); *Configuring Curricula for Instructionally Supportive Assessment* (Popham, in press); and *Building Tests to Support Instruction and Accountability* (Commission on Instructionally Supportive Assessment, 2001).

At the workshop, held January 23-24, 2003, presentations on programs developed in seven states as well as other examples, including some from abroad, stimulated lively discussion. Questions were raised not only about how ideal goals translate into practice, but also about the different kinds of obstacles to success in these efforts. (See Appendix A for the workshop agenda, Appendix B for a list of the workshop participants, and Appendix C for contact information.)

While the committee made no effort to systematically evaluate the success of the programs presented, it did learn much of interest about how those involved see the challenges before them, and about some of the strategies they have devised for overcoming them. The purpose of this report is to provide an account of the discussions, and to use some of the examples presented as a way of putting flesh on the bones of the concepts that were introduced in *Knowing What Students Know*. The committee recognizes that no existing program has yet been able to meet all of the ambitious goals it identified. In this report the committee does not intend to signal endorsement of the examples discussed. Rather, the intention is to illustrate what different ways of attempting to meet the goals suggested by the criteria can look like.

2

The Criteria in Context

The steering committee began the process of planning the workshop by considering the characteristics of an assessment system in which classroom and large-scale assessments work together to support learning. It agreed with earlier committees that, to be effective, assessment systems must do more than provide valid data. They must also be designed so that the information produced can be used to improve both the educational system and the teaching and learning process. In such a system a single assessment does not function in isolation but rather within a coordinated system in which the state, the district, the school, and the classroom each play a role.

The specific criteria the committee identified are listed and briefly described here. They are elaborated further in the discussion of workshop presentations later in this report. The steering committee made no attempt to evaluate the relative importance of each of the criteria, nor did it use the criteria to evaluate programs. Rather, the intent was to use the experiences of workshop presenters as a vehicle for thinking about the ways in which each of the criteria can contribute to the establishment of a coherent system.

The following are the ideal characteristics of assessment systems that the committee identified:¹

- **Comprehensive:** A comprehensive system is one in which a range of measurement approaches are used to provide a variety of evidence to

¹The first three criteria are adapted from *Knowing What Students Know* (NRC, 2001c); the last two were distilled from other reports listed in the Introduction.

support educational decision-making. A well-designed system includes both formative (to support students' ongoing learning and help teachers make instructional decisions) and summative (to evaluate students' level of achievement at the completion of a phase of learning) assessments that move students toward a manageable and clearly articulated set of outcomes. Measures might also include those that assess the quality of instruction, and provide evidence that improvements in tested achievement represent actual gains in learning as opposed to improved test-taking skills, for example.

- **Coherent:** A coherent system is one in which the conceptual base or models of learning underlying the assessments used at all levels (large-scale or classroom) are compatible. Furthermore, the content, processes, and skills measured by different assessments across the system are compatible. For a system to be coherent, alignment is needed among standards, curriculum, instruction, and professional development so that each element contributes to a common set of learning goals.
- **Continuous:** In a coordinated system, assessments measure student progress over time—for example, over a school year, over several grades, or over a student's school career. Assessments are ongoing and seamlessly integrated into instruction.
- **Integrated:** An assessment system is integrated if it is carefully designed to fit into a larger, coherent educational system that provides resources and professional development to ensure that teachers have the capacity to do what is expected of them based on the standards in place.
- **Includes High-Quality Assessments:** All of the assessments included in the system should be of high quality, by which is meant, first, that they must adhere to relevant professional standards. To further illustrate what high quality means, the committee has identified a set of specific characteristics that large-scale and classroom assessments can exhibit, which are summarized in Boxes 2-1 and 2-2.

These criteria address the educational assessment environment as a whole, and certainly it is not possible to talk about the relative effectiveness of large-scale or classroom systems without considering the contexts in which they are designed to operate. Nevertheless, there are many choices of approach for assessing students, and the workshop began with an overview of current thinking about both large-scale and classroom assessments. The discussion was grounded in professional thinking on the purposes that each kind of assessment serves best, and offered an overview of their potential, as well as their limitations.

BOX 2-1 Shared Characteristics of Large-Scale and Classroom Assessments

Large-Scale Assessments

Shared model of student learning:

- Models of learning should include developmental progressions over time.

Shared conception of disciplinary knowledge and competence:

- Focus on assessing what is most highly valued rather than what is easy to measure.
- Focus on evaluating understanding and reasoning, rather than on rote recall.
- Assess enabling skills and procedural knowledge in contexts of application.
- Signal to teachers and students what is important for them to teach and learn.
- Base assessments on standards that are clearly written so that teachers, students, parents, and the public understand what it is that is being assessed and what constitutes mastery.
- Measure a manageable body of knowledge and limited number of the most important skills so each can be assessed fully and thoroughly.
- Target both general forms of cognition, such as problem solving and inductive reasoning, and forms that are more domain-specific, such as deduction and proof in mathematics or the systematic manipulation of variables in science.
- Move away from a preponderance of assessment items that are short, skill-focused, single-answer, and decontextualized towards greater

Classroom Assessments

Shared model of student learning:

- Models of learning should include developmental progressions over time.

Shared conception of disciplinary knowledge and competence:

- Focus on assessing what is most highly valued rather than what is easy to measure.
- Focus on evaluating understanding and reasoning, rather than on rote recall.
- Assess enabling skills and procedural knowledge in contexts of application.
- Signal to students what is important for them to learn.
- Base assessments on standards that are clearly written so that students, parents, and the public understand what it is that is being assessed and what constitutes mastery.
- Measure a manageable body of knowledge and limited number of the most important skills so each can be assessed fully and thoroughly.
- Target both general forms of cognition, such as problem solving and inductive reasoning, and forms that are more domain-specific, such as deduction and proof in mathematics or the systematic manipulation of variables in science.
- Move away from a preponderance of assessment items that are short, skill-focused, single-answer, and decontextualized towards greater

continued

BOX 2-1 Continued

use of tasks that are context based, measure rich and well-structured knowledge, are open to multiple approaches (and, in some cases, to multiple solutions), are complex in the responses they demand, and are drawn from a wide spectrum of concepts and processes.

Designed to be valid and useful to support large-scale educational decisions:

- Are technically sound and timely.*
- Are designed in accordance with the purpose for which the results will be used.
- Measure the skills and knowledge they purport to measure.
- Are designed in accordance with accepted practices that include a detailed consideration of the reliability, validity, and fairness of the inferences that will be drawn from the test results.
- Report results in enough detail to reveal needed instructional changes and to highlight deficiencies in system resources that can lead to improved instruction.
- Focus on knowledge that students gain through instruction rather than on learning that takes place outside of school, or is a function of individual talents, socioeconomic status, or test preparation activities.
- Provide opportunities for students with different background experiences to connect their knowledge resources to relevant school expectations.

use of tasks that are context based, measure rich and well-structured knowledge, are open to multiple approaches (and, in some cases, to multiple solutions), are complex in the responses they demand, and are drawn from a wide spectrum of concepts and processes.

Designed to be valid and useful to support classroom decisions:

- Are technically sound and timely.*
- Are designed in accordance with the purpose for which the results will be used.
- Measure the skills and knowledge they purport to measure.
- Are designed in accordance with accepted practices that include consideration of the reliability, validity, and fairness of the inferences that will be drawn from the test results and use of follow-on evidence to redress inaccuracies.
- Report results in enough detail to reveal needed instructional changes and enable students to improve performance.
- Focus on knowledge that students gain through instruction rather than on learning that takes place outside of school, or is a function of individual talents, socioeconomic status, or test preparation activities.
- Provide opportunities for students with different background experiences to connect their knowledge resources to relevant school expectations.

*The standards for technical accuracy and immediacy of feedback are quite different for large-scale and classroom assessments but both levels of assessment must meet their respective standards in these areas.

continued

BOX 2-1 Continued

A range of measurement approaches used to provide a variety of evidence to support educational decision making:

- Provide opportunities for students to demonstrate competence in a variety of ways.

A range of measurement approaches used to provide a variety of evidence to support educational decision making:

- Provide opportunities for students to demonstrate competence in a variety of ways.

SOURCE: Adapted from NRC (1993, 1996, 1998, 2000, 2001a, 2001b, 2001c, and 2002), National Council of Teachers of Mathematics (1995), Commission on Instructionally Supportive Assessment (2001), and Popham (in press).

BOX 2-2 Unique Characteristics of Large-Scale and Classroom Assessments

Large-Scale Assessments

Provide comparative data, both normative and standards based, that allow policy makers, teachers, parents, and students to make judgments about the adequacy of performance and the specific curricular and instructional areas where improvement is needed.

Provide quality feedback to teachers about patterns of errors that could be the target for instructional interventions in the future.

Must be cost-effective and feasible; in particular, the benefit to students from information gain must be worth the instructional time lost to testing and test preparation.

Classroom Assessments

Must be ongoing and integrated seamlessly into instruction so that teachers and students are receiving frequent but unobtrusive feedback about their progress.

Assess some desired proficiencies in each knowledge domain that cannot be effectively assessed on a large-scale assessment, such as a student-designed experiment or a piece of creative writing revised over time.

Provide quality ongoing feedback to teachers about patterns of errors that could indicate the need for modification of instructional strategies.

Help teachers to identify and reconstruct students' misconceptions.

Provide quality feedback to students about their performance and specific guidance about how to improve (most useful when students are given descriptive, criterion-

continued

BOX 2-2 Continued

Results must be reported to stakeholders so as to enable meaningful use of assessment data and forestall misinterpretations.

based feedback rather than merely providing number or letter grades):

- Help students to identify and reconstruct their misconceptions.

Help students to assess their current levels of understanding in relation to well-articulated learning goals and what they, as students, clearly understand to constitute quality work:

- Involve peer- and self-assessments as well as teacher judgments.
- Place more emphasis on allowing students to participate in developing and analyzing the results of the assessments rather than viewing assessments as something that is done to them by teachers.

SOURCE: Adapted from NRC (1993, 1996, 1998, 2000, 2001a, 2001b, 2001c, and 2002), National Council of Teachers of Mathematics (1995), Commission on Instructionally Supportive Assessment (2001), and Popham (in press).

THE IDEAL

While no current assessment programs have been identified that satisfy all of the attributes described above, some can be seen as making significant progress in implementing specific features of a high-quality program. To explore what it might be like to teach and learn in a coherent and balanced assessment environment, where assessments, curriculum, instruction, and professional development are fully aligned with standards, the committee invited Gail Burrill, a teacher and teacher educator at Michigan State University and former president of the National Council of Teachers of Mathematics, to inaugurate the workshop by simulating such a situation for the workshop audience.

Describing an array of embedded, formative assessment techniques, Burrill illustrated for the workshop participants how assessment can help to shape learning and direct instruction. Examples from Japan, the Netherlands, and China helped to illustrate the ways in which assessments can circumscribe both what is taught and how it is learned. Burrill used these international examples to make

the point that educators in the United States are often leery of expecting students to transfer their knowledge to new contexts. In the examples she discussed, assessments were more challenging in that they called on students to use cognitive processes on unfamiliar material, but she argued that U.S. students could handle this kind of challenge.

To be sure that there is correspondence between what is taught and what is valued, Burrill suggests, input from many sources is necessary. Subject area experts, curriculum developers, researchers, teachers, cognitive scientists, and assessment developers need to work together to develop the standards and the assessments that will be used to measure student mastery of the specified competencies. Key for Burrill is that teachers be able to make choices as they implement a curriculum, and that assessments serve as an appropriate guide to what is taught. Coherent assessments will foster coherent curriculum and effective instruction; lack of coherence leads to unfocused learning and shallow understanding.

LARGE-SCALE ASSESSMENTS

While large-scale assessments can be controversial, and are easily misused, they are an important way of obtaining certain kinds of extremely valuable information about students. Large-scale assessments, those that are designed to provide evidence about large numbers of students, are the primary means by which accountability evidence is obtained in the United States. Indeed, there is little dispute that accountability—the provisions made for those who use, fund, and oversee public education to review and evaluate its effectiveness—is a crucial element in the continued success of public education.

As Lorrie Shepard of the School of Education, University of Colorado, Boulder, outlined at the workshop, there are three particular uses for which large-scale tests are essential. The first is *program diagnosis*. Assessments that make it possible to compare the performance of a large number of students can be used to identify patterns of strengths and weaknesses that are in turn critical for identifying any needed improvements in curriculum or instruction. Assessments developed for large-scale use, to provide evidence about district- or statewide performance, can also *exemplify*, as Shepard termed it, the educational goals described in standards and curriculum documents. In other words, assessment tasks and examples of student work make concrete just what students will actually know or be able to do if they meet defined standards. Large-scale assessments are also useful for one-time *certification* or screening; for example, to identify students who are not ready for grade-level work in reading and who need follow-up targeted assessment to determine their specific needs for remediation.

Shepard also noted that large-scale assessments often provide teachers an opportunity for effective professional development. Development of tests, scoring, curriculum development, and standards-based professional development are

all occasions when efforts to improve classroom assessment strategies can be woven into the program. Shepard argues that more could be gained through these opportunities if teachers had improved access to materials that model teaching for understanding, such as extended instructional activities, formative assessment tasks, and scoring rubrics with summative assessments built in to them.

While the value of large-scale assessments for these purposes is clear, it is equally clear that they are not useful for many other important educational purposes, particularly that of providing detailed understanding of individual students' performance. Professional standards are firm on the point that it is not a test itself that can be established as valid, but particular inferences that may be made from the test data (see *National Science Education Standards* (NSES) Standard 13.2, NRC, 1996).

Nevertheless, administrators who are pressed for both time and resources are often tempted to find tests that can serve more than one purpose. While this can be done, it necessarily entails compromises. Noting, "Ironically, the questions that are of most use to the state officer are of the least use to the teacher" (NRC, 2001c, p. 224), the Committee on the Foundations of Assessment framed the problem as a trade-off in assessment design between supporting accountability for schools and systems and supporting the need for specific guidance about individual students.

As Shepard stated, "The best way to help policy makers understand the limitations of an external, once-per-year test for instruction is to recognize that good teachers should already know so much about their students that they could fill out the test booklet for them." Shepard listed some of the contrasts, shown in Box 2-3, between large-scale and classroom assessments that make clear why different instruments are usually needed for different purposes.

BOX 2-3 Contrasts Between Large-Scale and Classroom Assessments

Large-Scale Assessments

Need to be standardized
Given on a uniform date
Must show independent performance
Delayed feedback
Stringent requirements for technical accuracy

Classroom Assessments

Need to be dynamic
Given as needed
Can show assisted performance
Immediate feedback
Less stringent requirements

SOURCE: Shepard (2003, January).

Many large-scale assessments are what psychometricians call “norm-referenced,” which means that one of their functions is to provide evidence of how students compare to one another. The resulting scores can be used to spread students’ performance out along a scale. The SAT is a good example of such a test: it is designed not to assess particular knowledge or content, but to provide college and university admissions officials with a means of ranking students based on their potential to succeed at college-level work. The questions are carefully selected, based on pretesting results, to present a range of difficulty, so that very few students are likely to succeed at either all or none of them, and so that the students will be spread out along the scale. Performance on such tests is often expressed in terms of percentiles, with a particular score reflecting performance that is better than that of a certain percentage of other test takers.

Other assessments are called “criterion-referenced” because their scoring “refers” not to the past performance of other students but to a fixed body of knowledge. Good examples of this kind of testing include professional licensure tests, which often identify minimum acceptable levels of mastery. With such tests, it does not matter how well other students have done; it matters only that a prospective airline pilot or surgeon has mastered a particular body of knowledge deemed essential. Assessments used with K-12 students can be of either type, and in some cases may blend the two. For example, states that use tests developed by national companies, which are often norm-referenced and offer the state the opportunity to determine how its students compare to those of the same age across the country, may also wish to assess their students’ knowledge of particular aspects of their standards. A state may add sections to the norm-referenced portion or make other modifications to adapt the test to the multiple purposes it has identified, though, as noted above, such an approach entails compromise.

CLASSROOM ASSESSMENTS

Discussion of classroom assessments has been somewhat less tidy, in part because the definition of such assessments is less precise, and the range that the term covers was evident at the workshop. Teachers make assessments of their students’ learning every day, by noting the misconceptions or insights that underlie a question, for example, or observing the way a student makes use of materials provided for a task. They also assess them more formally, with particular questions in mind, and it is through the teacher’s aim in assessing that presenter Dylan Wiliam, professor of education at King’s College in London, defines classroom assessment, or, in his phrase, assessment for learning. That is, if the aim of the assessment is to improve the student’s learning in some direct way, rather than to rank, evaluate, or certify some aspect of performance, then it is properly in the realm of classroom assessment.

For Wiliam, it is the feedback provided to the student that is critical to the success of this enterprise, and he describes it as a three-part process. First the

teacher must find out where the student is in relation to the goals for the class; next, he or she must clearly convey to the student what those goals are. Perhaps most important, the teacher must then help the student in concrete ways to move toward those goals. Assessments that are intended primarily to provide feedback to students and to shape their learning are often called formative assessments, and distinguished from summative ones, which are intended primarily to evaluate students. This mode of categorizing assessments shares some aspects with the dichotomy between classroom and large-scale assessments that is the subject of this report, but it is important to remember that a large-scale assessment could serve formative purposes, just as a classroom assessment can serve summative purposes.

Presenter Jan de Lange, professor and director of the Freudenthal Institute at the University of Utrecht, The Netherlands, addressed the issue of classroom assessments used in teaching mathematics, using a description of a project carried out in Philadelphia and Milwaukee by the Freudenthal Institute to highlight several points. The project's goal was to influence the quality of learning and instruction by changing classroom assessment methods, and it used an Assessment Pyramid to depict the different levels of mathematical competencies that students display. In the pyramid, level 1 covers reproduction and facts, level 2 is making connections and simple problem solving, and level 3 is complex problem solving and mathematical reasoning.

Teachers involved in the project were given a variety of supports, including both assessment materials and training, through which they could help their students think more deeply about mathematics. At the same time, teachers' thinking about what constitutes effective classroom assessment, scoring, and other issues was expanded. The pyramid was the basis for defining expectations for student performance, for structuring instruction, and for giving students useful feedback in relation to learning goals and competency levels.

The pyramid was derived from the framework used in the Organisation for Economic Co-operation and Development's Programme for International Student Development (PISA) (PISA's assessment program is described in Chapter 4). De Lange argued that the alignment between the pyramid used in the classroom and the large-scale PISA demonstrated for teachers that a comprehensive, coherent, and continuous assessment is possible. At the same time, by working with the pyramid the teachers became skilled at recognizing and analyzing quality assessment. Through the two-year study, de Lange explained, teachers changed their approaches to both classroom assessment and the teaching of mathematics in significant ways.

For committee chair J. Myron Atkin, professor at the Center for Educational Research, Stanford University, the key is the teacher's unique capacity to monitor students' progress over time. In his presentation, which focused on the way classroom assessment functions in science education, Atkin asked workshop participants to consider the many different opportunities a teacher has to assess what

students know and can do in the course of a project that takes place over several weeks or months.

As an example, Atkin cited a project in which a group of students monitored the state of a pond near their school and investigated the nature and possible causes for an algal bloom that occurred in the course of their study. Not only were they conducting original research, in the sense that no scientists had previously studied that particular pond, the students were also able to respond to unpredictable events. The project afforded them many opportunities to demonstrate their capacity to bring prior knowledge and experience to bear on a problem, their proficiency with available methods and tools, and their resourcefulness in drawing on available sources of data and interpretation. Their teacher was able to monitor their progress through formal output, such as field notes and reports, as well as in countless informal interchanges that revealed the students' thinking and their development over time.

This project exemplified for Atkin how a teacher can develop an "assessment culture," in which the focus is on inquiry—a key element of both the content and skill standards included in the NSES (NRC, 1996). The teacher was able to assess students on skills and knowledge that are deemed essential by NSES, and yet are impossible to measure using a one-time performance assessment. The challenge Atkin identified is to find more ways to make systematic use, for purposes of accountability beyond the classroom, of the information about students that teachers are in a unique position to obtain.

3

The Nature of the Gap

Some workshop participants quibbled with describing the problem at hand as a gap between two kinds of assessment that needs to be bridged, favoring instead the notion that systems need to be better balanced. Workshop discussion made clear that more than one kind of gap can be identified, and that achieving balance among different elements of an educational system is indeed an important and challenging goal. The gaps considered at the workshop include those between:

- large-scale and classroom assessments;
- formative assessments, designed to enhance learning, and summative assessments, designed to evaluate student performance;
- the goals of assessment for accountability and assessment for learning;
- the complexity of the science of large-scale assessment and the professional development provided for teachers on the topic;
- the rich potential of classroom assessment strategies and the professional development and time available for teachers to take advantage of it;
- the curriculum dictated by state and district standards and the classroom preparation made necessary by external assessments;
- the data provided by many large-scale assessments and teachers' day-to-day needs for information about their students;
- the ambitious goals identified in most standards documents and the time available to address them in the classroom;
- the knowledge and skills identified in standards documents as important to master and the content that can be assessed using currently available large-scale instruments;

- the demands placed on teachers and teachers' available time and resources; and
- the numbers and kinds of resources used to support external forms of assessment and those allocated for classroom formative assessment to assist learning.

To begin with just one of these gaps, the pressure on teachers to prepare students for large-scale tests developed for accountability purposes can clearly be very great, yet such tests may not bear a close relationship to what is happening in any given classroom. When this happens there is often a large gap between the objectives teachers and administrators would naturally develop, and those dictated by the inherently circumscribed nature of the external test. Not only the objectives are at odds in this situation; there also can be, more broadly, a gap between the vast domain of skills, knowledge, and cognitive processes that have been identified as important for students to master and described in standards documents, and the far narrower sets of skills and knowledge that can be assessed using the instruments currently available.

Focusing on this way of framing the problem, James Popham, a psychometrician at the University of California, Los Angeles, and chair of the Commission on Instructionally Supportive Assessment empaneled by five major educational organizations,¹ described what he sees as the urgent need for state departments of education to limit and prioritize the goals they include in the standards and curricula. In his view, most such documents identify so many goals that meeting the desired standards would be literally impossible. Popham suggested that when a standards document fails to provide clear guidance as to what knowledge and skills are essential, the results are quite the reverse of what policy makers hope for.

First, there are mismatches between what is taught and what is tested, Popham argued. Second, the material teachers have covered well tends to get eliminated from future tests because, since these tests are designed to spread students out across a range, items on which most kids succeed tend to get dropped from the pool. Thus the very material which teachers have presumably been most successful at teaching often gets eliminated from future tests. As teachers detect the absence of particular content from the test, they are likely to lessen their emphasis on it and turn to other material that is tested. Finally, Popham argued, because traditionally constructed achievement tests strive to create sufficient score-spread to permit accurate comparative interpretations to be made, many of the items that are included for the purpose of spreading students out are linked to students'

¹The American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, and National Middle School Association.

socioeconomic status or to students' inherited academic aptitudes. As a consequence, in such situations, it is impossible to tell whether students' test performance is the result of what they were taught at school or the result of characteristics they brought with them to school.

One key to bridging the gap, suggested Popham, is to identify "a modest number of truly significant outcomes." These must be conceptualized in terms of how they might be taught, and identified as constructs that can be assessed in easily reportable ways. In addition, these objectives must be explained in terms that are very clear to teachers. At the same time, however, he argued that standards should not be defined—as they often are by default—as those elements of the content that are easily assessed using existing instruments. The gap, said Popham, is perhaps best described as that between the goals of accountability and instruction, and it may be a symptom of the undisciplined way in which standards have often been developed at both the state and district levels.

While Popham saw the gap in terms of the way standards are defined, it was clear that participants found numerous ways of defining it, and, as noted, few were comfortable with a single formulation. Indeed, elaboration of the possible sources and characteristics of the gap was a recurring theme, particularly in the question and answer sessions following many of the presentations. It was from some of these exchanges that the importance of paying attention to the particular circumstances in which each of the programs presented was developed became so evident. This point arose in the concluding session and will be discussed further in the last chapter.

4

Some International Examples

The primary focus of the steering committee's efforts was to find examples of the many forms that an assessment program built around improving learning can take. The committee looked at programs in seven states, several international examples, and three programs developed by researchers: the Berkeley Evaluation and Assessment Research assessment model, Facet-Based Assessment, and Model-Based Assessment. Presenters for each of these programs were asked to discuss not only the goals and characteristics of their programs, but also the ways in which the programs exemplify the criteria the committee had identified. They were also asked to talk about problems and obstacles they had encountered, as well as successes they believed they had achieved and methods of securing evidence of their results. In this chapter, the examples from abroad are discussed.

The notion of gaps between different elements and goals of the educational system may not have been as much on the minds of education officials in other countries, but the assessment systems in several countries nevertheless seem to have much to offer the discussion in the United States. Two different Australian systems, for example, offer interesting ways of thinking about alignment and coherence. Studies from Great Britain demonstrate a way teachers can use assessments to help students make progress in their learning, while the International Baccalaureate program shows the role that teachers can play in a widely dispersed system. The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) demonstrates one way in which diverse constituents can focus on the material that is most important to assess.

AUSTRALIA

At the national level, Australia has built a large-scale assessment on the basis of a preexisting framework, or “map of progress,” that outlined the knowledge and skills students should develop. Geoff Masters, chief executive officer at the independent, nonprofit Australian Council for Educational Research (ACER), explained to workshop participants that the resulting system developed almost by happenstance, yet has many interconnected and mutually supporting parts.

The original framework took the form of a detailed matrix showing levels of competence in different aspects of each subject area. In English, for example, the first subject for which a framework was developed, descriptions of competence in reading, writing, listening, speaking, and viewing were developed. The framework describes eight different levels of competency for each skill and is designed to cover the years of compulsory schooling.

ACER recognized that teachers needed some guidance in monitoring student progress along the framework. On its own initiative, ACER developed an assessment resource for teachers that they could use in making their own assessments of how children were progressing in terms of the framework. The resource kits, which were sold to schools around the country, included activities and materials and a range of assessment methods to be used individually and with groups of students.

When the national government later decided to conduct a national survey of primary children’s literacy skills, to obtain data similar to that provided in the United States by the National Assessment of Educational Progress, ACER submitted a proposal to develop an assessment based on the model they had already devised for the teachers’ assessment kits. Government officials agreed to adopt the ACER model, thus establishing a national assessment system that relied on teachers to conduct and score the assessments.

A number of means of ensuring consistency and fairness were built into the system. First, as with the original resource kits, the assessment supplied guidelines and scoring rubrics. A group of experienced external assessors trained and monitored teachers in the use of the assessment methods. These assessors also visited schools and monitored a subset of the assessments as they were conducted. Second, all the student work generated for assessment purposes was collected for further monitoring at a central office in Melbourne. The work was sampled and, where discrepancies were found, rescored.

The initial assessment was successful, yielding results for nearly 9,000 students in the third and fifth grades. Student performance was shown in terms of their progress along the matrix; the relative performance of socioeconomic subgroups was also shown.

Unfortunately, as Masters explained, the national government was surprised in the end to find no indication of how many students had “passed.” Since the assessment was designed only to show how far groups of students had progressed

through the stages identified in the matrix, no cutpoints had been identified for either grade. However, ACER was able to go back and conduct a standard-setting exercise to determine what minimum level of competency in reading and writing should be expected at each grade. Pass rates could then be determined retroactively, and although the results turned out to be controversial, the exercise demonstrated the adaptability of the assessment system for the accountability purposes that are particularly important to policy makers and politicians.

QUEENSLAND

Richard Shavelson, professor of education and psychology at Stanford University, described for the audience the somewhat different situation in the Australian state of Queensland, whose system he has studied. There the state had for many years relied on a set of “A-level” examinations prepared by the University of Queensland, similar to those used in Great Britain, both to determine how well students were prepared for college study in different subjects and as an element in the college selection process. In 1970-1971, concern began to mount that the exams were too difficult and were the cause of an undesirable narrowing of the curriculum. Queensland decided to replace the A levels with formative assessments that would more directly address students’ needs, and then to build on those to obtain summative information about student performance that would be of value beyond the classroom.

In essence, as Shavelson explained, Queensland officials decided to develop “a system for auditing the local implementation of curriculum and assessment and accountability.” Teachers and local schools are responsible for both curriculum and assessment and their work is monitored to ensure that it is consistent across the state and meets standards for quality. An infrastructure was set up to accomplish the monitoring, which includes a Board of Senior Secondary Studies, which set the syllabi—the essential goals for content, cognitive skills, and domain-specific skills—for each subject and the general methods for conducting assessments. The board is also responsible for moderation of scores, a process by which teachers’ scores are calibrated with one another to achieve consistency across classes and schools. Below this board, a series of district-level content panels in each of the A-level subjects provides more direct support to schools and teachers. Each school is then free to develop its own two-year, A-level curriculum in each subject, as well as a culminating exam. The exams are scored according to a Queensland-wide, five-point, domain-referenced scale, and moderated.

Thus, schools and teachers are given a considerable amount of both direction and latitude. They use formative and summative assessments throughout the two A-level years, based on guidelines provided by Queensland, using both kinds (and students are always aware of the purpose of a particular assessment) to help students understand in detail the expectations they are striving to meet. To

Shavelson, the key to the system's apparent success over thirty years is the very close link made between the curriculum and the content of the assessments.

To American eyes, one striking aspect of both Australia's national assessment system and the Queensland model is the degree to which each, in its way, accords significant value to the judgments of teachers about their students. In these systems, teachers have many different opportunities for training and development to improve the knowledge and skills they need to play a key role in the assessment program. They can become involved in development and scoring of assessments (as are many of their counterparts in the United States), and receive the trust necessary to develop evaluative assessments of students on their own.

GREAT BRITAIN: ENHANCED FORMATIVE ASSESSMENT

Dylan Wiliam of King's College, London, described efforts in Great Britain to focus closely on the ways teachers can use assessments to help students make progress in their learning. He began by describing an overview of approximately 250 studies that explored the effectiveness of a formative classroom assessment (also sometimes called assessment for learning) in which clear evidence of a positive effect on learning was found. Specifically, Wiliam explained, when teachers provide students with clear feedback that gives them guidance on the steps they need to take to improve, students progress at a greater rate than they do in response to other kinds of feedback.

Wiliam also described a study in which a group of twenty-four mathematics and science teachers were asked to develop their use of formative assessment with one class in several specific ways: by making greater use of higher-order questioning, providing task-involving rather than ego-involving feedback, developing the use of peer- and self-assessment strategies, and exploring the use of summative tests for formative purposes (Black, Harrison, Lee, Marshall, and Wiliam, 2002). For each class, the local class that could best be used as a control was identified so that any improvements in learning could potentially be measured, and in this study as well evidence of a positive effect was found.

While the methods sound simple—allowing a longer wait time while students consider how to answer a question, for example—Wiliam stressed the importance not of the methods themselves, but of the insights into how students learn that led to them. The idea, he explained, is to initiate students into a culture of learning in which they not only take responsibility for their learning but are supported in the steps they need to take to progress. At the same time, teachers' capacity to make useful inferences about their students are enhanced, just as their opportunities to use these inferences are increased (Black et al., 2002).

THE INTERNATIONAL BACCALAUREATE (IB) DIPLOMA PROGRAMME

The International Baccalaureate (IB) Diploma Programme offered workshop participants an additional way to think about the role of teachers in assessment. George Pook, head of assessment for the International Baccalaureate Organisation, explained that the IB was developed to provide a common curriculum for students around the world, as well as a grading system that would be recognized and understood by colleges and universities around the world. Thus, consistency is very important to the success of the program, but at the same time there is a need to entrust considerable responsibility to widely dispersed schools and teachers.

The IB uses a variety of assessment strategies for summative purposes. For example, students must complete an extended essay on a topic of their own choosing at the end of the program, which is scored centrally. Examinations may include tasks ranging from multiple-choice questions to full-length essays, as appropriate for each subject. Oral presentations are also required in language subjects, and these are scored by teachers using criteria supplied by the IB program. All of the results are reported in terms of a seven-point scale that is linked to defined levels of performance that program administrators try to keep consistent from year to year as well as across participating schools around the world, who of course work in different languages. The points on the scales describe content and skills, and the scoring is intended only to indicate how well students have mastered them, not to spread students out for comparative purposes.

Internal, teacher-generated assessments play a significant role in the program for both formative and summative purposes. Teacher-generated assessments address a different range of subject matter and skills than the IB-generated assessments do. The two types are intended to complement one another in creating an overall measure of a student's achievement. Teachers' ongoing formative assessments are viewed as opportunities for students to see how they are progressing along the criteria defined in the seven-point scale. Released test questions, rubrics, and student work are all used to provide this feedback. Many IB teachers serve as external assessors for other schools, and also have opportunities to review and revise the curricula in their disciplines. All IB teachers receive support in the form of resource materials, workshops, and an online curriculum center. Moderators are available to give teachers feedback on their internal assessment methods, as well as their assignments and their grading.

PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PISA)

The OECD, which was formed as part of the Marshall Plan after World War II, is composed of thirty nations, all of which are democratic market economies. As Barry McGaw, director for education at OECD, explained at the workshop, a

primary function of the OECD is to collect data in a number of policy areas, and in the late 1980s the organization began a process of upgrading its statistical work in education, with the particular goal of ensuring that the data used to represent national systems become more comparable. While the OECD had been using data regarding educational outcomes supplied by the International Association for the Evaluation of Educational Achievement for a number of years, it began to gather data of its own in the mid-1990s through PISA. The focus is on summative data that can be used to make useful comparisons among the member nations.

The primary initial goal for PISA was, as McGaw explained, “to estimate the yield of national education systems,” and he acknowledged that this is a grand ambition. Yield is an economic concept not generally used in the study of education, but it led the developers of PISA to focus on what students can do with what they have learned, and thus avoid the difficulty of identifying the material that had been covered in common across many countries. Thus PISA assesses the “literacy” of fifteen-year-olds in reading, mathematics, and science. They use a variety of measurement approaches—multiple-choice questions as well as open-ended short questions and written pieces, but the assessments are not intended to be used for formative, classroom purposes.

McGaw provided some examples of the kinds of questions that can be considered using PISA data, using tables and graphs, for example, to show how the member countries vary in terms of the balance they achieve between equity and quality. He also showed graphically that countries vary considerably in terms both of how much spread they have between their lowest and highest performing students, and also in terms of how much of that spread occurs within schools and how much occurs across schools. Probing that question even deeper, he presented a table that broke down the variation that occurs across schools according to whether it was intended—that is, the result of deliberate tracking of students into academic or vocational programs, for instance—or unintended. Data such as these, McGaw explained, are very useful for helping countries see that there are alternatives to the way they are structuring their education systems. South Korea, for example, has been remarkably successful at achieving both high quality and high equity; it has the lowest degree of spread among high- and low-performing students, while overall performance is high.

Although PISA does not fit particularly well with the criteria laid out by the committee, McGaw noted that it does offer formative possibilities in a system context. Denmark, he noted, has found that though it spends among the largest amounts per students, its average student performance figures are quite low. As a consequence, the ministry of education is working to make the system operate more efficiently and improve student performance. Doing so, of course, implies that it is confident that the constructs measured by PISA are genuinely important, even though they are not directly linked to the curriculum taught in Denmark or any other country.

It is in this sense that PISA's experience might be most useful to educators looking for ways to bridge the gaps. The process of developing PISA was an extensive effort to build a framework that defined a reasonable set of expectations for fifteen-year-olds in each of the domains. International groups drew on assessments from around the world and worked through cultural and language differences to come up with two versions of the test, one in English and one in French, that represented their best effort to assess what is really important for fifteen-year-olds to be able to do. McGaw suggested that any concepts that got past the double translations and other reviews, field tests, differential item functioning (DIF) analyses,¹ and other screens were likely to be truly key concepts. He does not believe that PISA focuses mostly on what is easy to assess, rather than what is important, and does believe that it assesses understanding and reasoning, not factual recall.

¹DIF analyses flag test questions that perform differently for a particular subgroup of test takers than for the group as a whole. Thus, for example, if students in one country, or those who are native speakers of a particular language, have difficulty with a question for cultural reasons rather than because of their skill with its content, it can be identified so that it need not count against them.

5

Assessment to Improve Learning

The U.S. programs presented at the workshop were selected based on an informal review of efforts in states and districts to put into practice the goal of aligning their assessment systems with standards and curricula. The committee acknowledges that there are many more states and districts that are working to bridge the gap between classroom and large-scale assessments, but exploring more than a few at the present workshop was beyond its charge. The selected programs are, however, exemplary in that they are making progress towards goals of the kind identified by the committee. Not all of the programs have articulated their goals in the same terms the committee had identified, but all share a commitment to using assessments to improve learning, and were seen as evidently meeting at least one of the criteria (see summary in Chapter 2, and Boxes 2-1 and 2-2).

NEBRASKA: SCHOOL-BASED TEACHER-LED ASSESSMENT RECORDING SYSTEM

Nebraska is an interesting state to consider first because it had no statewide assessment program at all until 2000, and thus had the benefit of many years to observe the efforts of other states before initiating its own program. As Patricia Roschewski, director of assessment for the Nebraska Department of Education, explained, the state had first developed academic standards in 1998, and had decided that it needed an assessment program for two reasons. First, the state wanted to collect information about student performance that could be used to improve instruction. Second, it wanted accountability data that could be shared

with the public. Nebraska was clear in wanting the primary stakeholders in the system to be students and teachers, rather than policy makers, and this decision led it to give teachers a key role in the assessment program.

The Nebraska program's title, the School-Based Teacher-Led Assessment Recording System (STARS), is a very brief summary of the goals the state had developed, and, as Patricia Roschewski explained, the focus on teachers led it to devote a considerable proportion of the available resources to professional development and support. Many Nebraska districts had developed their own assessment systems, mostly criterion-referenced and classroom-based, but the state perceived that teachers and administrators generally had had very little training in assessment issues. STARS is essentially a way of building on existing local assessments to meet the new statewide goals.

Under STARS, goals based on the state standards are set for each district and school, and clearly articulated so that students, parents, and everyone else concerned understands the expectations for learning. For accreditation purposes, each district gives a norm-referenced test, such as the Terra Nova or the SAT 9, which typically covers some 35-40 percent of the standards. The remaining 60-65 percent is measured using classroom-based assessments developed by teachers; local teachers and administrators can blend these with activities and assessments dictated by district curricula in whatever ways they choose. The state monitors these assessments using a national advisory panel made up of assessment experts and Nebraska educators. This panel reviews and rates assessment portfolios prepared by the districts over a period of several months each summer. Districts whose methods are not successful receive further support and training; exemplary methods are shared around the state.

The system, Roschewski explained, works in part because the local curricula and state standards are closely aligned and clearly understood, and in part because intensive training has built the "assessment literacy" of the educators who are responsible for the bulk of the assessment. Nebraska teachers who once had little reason to think about issues such as validity and reliability are now responsible for ensuring that they assess their students in ways that stand up to professional scrutiny. The state had not thought in terms of bridging a particular gap, said Roschewski, but rather had sought to focus on balancing and integrating a new element—the desire for more feedback that teachers and students could use to improve learning and for information that could be used for accountability—into a system without disrupting the balance it had already achieved.

DELAWARE: COMPREHENSIVE SCIENCE ASSESSMENT

Delaware provides another example of a program that involved a significant amount of teacher training to increase assessment literacy. Its comprehensive science assessment grew out of the state's commitment to improve science learning. Rachel Wood, education associate at the state's Science Resource Center,

described a process that began in 1992 with the development of state standards in science. A needs assessment revealed that science was indeed getting short shrift; in the elementary grades it was often taught for as little as forty-five minutes a week. Curriculum materials were developed, and the state focused on identifying explicit learning goals for the topics outlined in the standards. For example, a requirement that fourth graders study electricity was broken down into precise descriptions of the key concepts related to electricity that were to be mastered. Attention was paid at the same time to both cognitive and practical factors that would affect articulation—so that the prerequisites for meeting the curricular objectives were accomplished grade by grade.

Once the state was pleased with its curricular units, it took a look at the accompanying end-of-unit assessments, and was not satisfied. In particular, it found that the scoring rubrics were generic and provided little useful diagnostic information. The state wanted to obtain summative information that could be used for accountability purposes from assessments that were closely linked to the curriculum, but also wanted the assessments to give teachers clear feedback they could use to improve their instruction. Delaware wanted specific data about how students were faring with particular elements of the curriculum, and it wanted assessments that would be part of a continuous loop of feedback and improvement, thus fostering a community in which teachers and students shared a sense of the purpose of and expectations for science learning. The state made the decision that teachers should be heavily involved in the assessment process, and that a significant investment in professional development was needed.

One of the innovations Delaware instituted was in direct response to the need for diagnostic assessment data. Using a system of double-digit scoring rubrics, modeled after a strategy used in the performance component of the Third International Mathematics and Science Study, educators could collect not only data showing how well students did with particular items, but also data on the kinds of misconceptions that kept them from complete understanding. In this system, the first digit works in the same way many rubrics do, indicating that a response is completely or partially correct. The second digit indicates the nature of misconceptions expressed in the answer (and raters are trained to recognize and code these) so that teachers can see what is missing in their students' understanding. Moreover, widespread misconceptions can often be traced to areas of the curriculum that are not adequately addressed, or to ambiguities in texts or materials.

The state recognized that few teachers had sufficient background in assessment issues to meet the emerging needs. With some outside resources, the state provided intensive professional development for a cadre of teachers, who could then branch out and work with other teachers. Not only did teachers undergo training to improve their understanding of assessment issues as well as their capacities for making use of both formative and summative assessments, they were also increasingly linked together in the kind of community of learners referred to earlier by Dylan Wiliam. Using shared materials available online,

including assessments, rubrics, and student work, as well as professional development activities, teachers were encouraged to share ideas about specific goals for student learning and ways to help their students meet them.

VERMONT: THE VERMONT ASSESSMENT SYSTEM AND THE PARTNERSHIP FOR THE ASSESSMENT OF STANDARDS-BASED SCIENCE

Vermont was the subject of national attention in the spring of 2002, when it announced that it was considering foregoing public education funds so that it would not have to comply with all of the assessment requirements of the No Child Left Behind Act. The state subsequently decided to accept the funds and is now trying to work out a way to satisfy the new federal requirements using locally designed assessments as well as statewide, large-scale assessments, as it has been doing for a number of years. Vermont's existing assessment program was designed to rely in part on a formal set of assessment tools developed or selected by districts, or, in some cases, by schools, to meet their specific needs.

As described by Bud Myers and David White, assessment coordinators in the Vermont Department of Education, the state's goals for its local assessments are very clear. Assessments are to

- be linked to state and local content standards,
- provide information that is valued at the local level,
- support teaching and learning,
- meet tough standards of reliability and validity, and
- be part of a continuum of assessment strategies that serve a range of purposes at the national, state, district, school, and program level, including both evaluation and feedback to students.

Vermont has developed an infrastructure both to support teachers and administrators in carrying out assessments and to ensure that the local assessments meet quality standards. Technical advisory panels oversee the quality of local assessments. Materials are provided to guide the development of local assessments, and exemplary assessment tools, item banks, and other resources are posted on a website accessible throughout the state. Review panels continuously evaluate assessment tools, and summer institutes help teachers keep up to date on assessment strategies. The state has built professional development for both teachers and administrators into the system, and has developed master's degree programs for teachers with an incomplete command of the mathematics and science knowledge needed to teach the content outlined in the standards.

A part of Vermont's assessment system is the Partnership for the Assessment of Standards-Based Science (PASS) program, which is a commercially available standards-based science assessment developed by WestEd. Kathy Comfort, prin-

cial investigator and director of PASS, described how the program fits Vermont's goals and dovetails with the larger question of integrating large-scale and classroom assessments. PASS was originally developed as a large-scale assessment that states and districts could use to measure their students' performance and growth in science against national standards and learning goals. PASS also meets the science assessment requirements of the No Child Left Behind act. The PASS assessment is aligned with the content recommendations of the *National Science Education Standards* (NRC, 1996) and the American Association for the Advancement of Science's *Benchmarks for Science Literacy* (1993). It incorporates multiple measures—enhanced multiple-choice questions, hands-on performance tasks, constructed-response investigations, and open-ended questions—to get at different kinds of knowledge and skills. WestEd staff worked closely with Vermont officials to customize the assessment to Vermont's standards and learning goals.

In response to feedback from PASS users, WestEd is developing ways that the program could also be used to help inform instruction and guide professional development. WestEd is using PASS to conduct research on the relationship among different assessment components, instructional practices, and student achievement, and on teachers' understanding of large-scale assessment results and the uses they make of the results in their classroom practice. Vermont teachers develop school and classroom science assessments using the methodology and learning goals of the PASS assessment. Teachers are also involved in developing items and in scoring, which provides an opportunity for large numbers of them to focus on specific performance expectations, and to share information and ideas.

While Vermont is proud of what it has done to make local assessments an integral part of its system, Bud Myers discussed some of the issues that are still of concern. Questions have arisen about how to keep the local assessments secure, and also about ways to make sure all the stakeholders find them credible. Perhaps foremost, however, is the question of resources. A significant degree of professional development, in both content and assessment issues, has been required to achieve current levels of competence. Myers raised concerns about both the funding and time that will be required to keep the program moving forward. He also cited the requirements of the No Child Left Behind Act, noting that they are not readily compatible with a system that relies as heavily as Vermont does on local assessments. Adding additional assessments to meet the requirement would substantially increase the assessment costs the state will have to bear.

WYOMING: BODY OF EVIDENCE SYSTEM

Wyoming's newly approved system grew out of the desire to make sure that graduating students had mastered the content specified in the state standards. Scott Marion, former director of assessment for the Wyoming Department of

Education, described how, in lieu of an end-of-school exit exam, the state decided to develop the Body of Evidence System (BOE). Under the system, students will, over time, establish that they have mastered the material required for graduation—performance standards in nine content areas. They will be able to meet these requirements as early as eighth grade, and typically will complete most by the end of tenth grade. Multiple sources of evidence will be acceptable.

An important goal for the BOE system was to improve teaching, learning, and classroom assessment; at the same time, Wyoming hoped to avoid some of the negative consequences other states had encountered using single high-stakes exams to make sure students had mastered graduation requirements. The state has asked local districts to design the measures by which students would demonstrate their mastery, based on a set of five assessment design principles arrived at through a deliberative process. Each district's program will be evaluated in terms of:

- alignment with the state's content and performance standards;
- consistent and reliable application;
- fairness, in that it is not biased against any subgroups and uses accommodations and alternate assessments appropriately; and in that it provides students with multiple opportunities, using different formats, to demonstrate their knowledge and skills;
- standard-setting, as revealed in the strength of its rationale for its method of choosing cut scores,¹ and how closely they are linked to performance standards; and
- comparability, through evidence that requirements are applied in comparable ways across classrooms, programs, schools, and the district. (Wyoming decided not to evaluate comparability from district to district, since each would be meeting minimum requirements.)

Recognizing that in most cases local educators lack the expertise to design the innovative measures Wyoming wanted to see in use, the state has begun providing considerable professional development and technical support for this endeavor. Moreover, it decided to use peer review to evaluate local systems, in part because of the many opportunities this would provide for professional development; reviewers are drawn from every one of Wyoming's districts and some serve as team leaders throughout the state. The reviewers work with national experts, Marion explained, and the review process has already helped those involved grapple with the real meaning of alignment, coherence, and other assessment design principles. In addition, to address the sometimes poor quality of locally developed assessments, the state formed the Body of Evidence Consor-

¹A cut score is a score point below which performance is deemed unacceptable for a particular purpose.

tium, a partnership of almost all of the districts, Wyoming's Department of Education, and national assessment experts, which disseminates assessment knowledge and skills through workshops and other activities.

Marion discussed what he perceives as the most difficult challenges the state has faced in implementing the BOE system. As noted, the state was initially disappointed with the quality of many local assessments, and efforts to address that problem have led in many cases to a deeper conversation about theories of learning and modes of teaching. While this is an ongoing challenge, Marion was pleased to find veteran teachers seeking guidance on how to modify their teaching in light of what they had learned through the BOE process. On a more practical note, the state has found that aggregating the various kinds of evidence to make fair decisions about students across districts has been a challenge, as has setting standards.

Reflecting on how Wyoming's system looks in light of the criteria presented by the committee, Marion concluded that the BOE system has focused on finding a variety of workable summative assessments. Consequently, it places relatively little emphasis on classroom assessment—the state hopes that the BOE system will foster classroom discourse and the kinds of ongoing feedback that teachers and students need, but it has not made that a requirement. He suggested that while a system can try to address all of the criteria the committee identified, and perhaps come close on many of them, there is a fundamental choice that needs to be made in the end between the unique characteristics and demands of large-scale assessment and those of classroom assessment. Marion expressed concern that there is a contradiction between the goal of assessing the few, carefully chosen, big ideas and the goal of assessing in a way that provides frequent and unobtrusive feedback. As he affirmed, “You can't assess big ideas very frequently unless you are assessing parts of the big ideas, and then are they still big ideas?”

MAINE: COMPREHENSIVE ASSESSMENT SYSTEM

Like many states, Maine developed a new assessment system after new standards were put into place. Jill Rosenblum and Pam Rolfe, assessment coordinators at the Maine Mathematics and Science Alliance and the Maine Department of Education, respectively, described the state's efforts. Maine had three principal goals for its assessment program, as outlined in 1997 legislation, but it highlighted as the first producing “high quality information about student performance that will inform teaching and learning.” The other two goals are monitoring schools and administrative units and holding them accountable for their success at making sure students meet the state standards, and certifying that students have met the content standards.

Maine was determined to meet those goals with a system that delegated a considerable amount of the assessment work to schools and districts. The state administers a large-scale assessment in six subjects at grades four, eight, and

eleven, and participates in the National Assessment of Educational Progress. While the state expects that it will need to further modify its system to meet the requirements of the No Child Left Behind Act, it currently relies on local educators to devise their own strategies for all the remaining assessments required to meet Maine’s three goals. Table 5-1, provided by Rolfe and Rosenblum, summarizes the basic structure of the system.

To unify its system, Maine developed a very specific “alignment protocol,” which spells out in detail the relationship between the assessments at all levels and the state standards. All assessments are to be linked to learning targets described in the standards documents, and they are conducted at the classroom, school, district, and state levels, as well as at all grades. It is left to the discretion of local educators to determine when they think their students have mastered a particular body of material and are ready to be assessed on it. Students are assessed using a wide variety of methods, and are given multiple opportunities to demonstrate their knowledge, understanding, and developing skills. The assessments are in many cases common instruments but are tailored to fit local curriculum and instruction, and provide immediate feedback to teachers and students.

The state is now completing the pilot testing of its assessment plan, which uses a combination of anchor tasks, common tasks, and assessments developed and selected at the local level. Thus, in Rosenblum’s view, Maine avoided the need to make the basic choice between large-scale and classroom objectives that Scott Marion identified in Wyoming. Maine, she argued, has taken a middle

TABLE 5-1 Characteristics of Maine’s Assessment System

	Primary Purpose	Selected or Developed by	Scored by
Classroom assessment	Informing teaching and learning	Individual teacher	Individual teacher
School or district assessment	Informing and monitoring	Groups of teachers and administrators	Groups of teachers (and others)
State assessment	Monitoring and evaluating programs to ensure accountability	Groups of administrators, and/or policy makers	Scorers outside the district
Assessment system	Informing teaching, monitoring and evaluating, certification	District assessment leadership	Both internal and external

SOURCE: Maine Department of Education (2003).

path: the school and district assessments have shared features but are firmly grounded in the curriculum.

Professional development has been a key to making the system work, according to Rosenblum and Rolfe. For teachers to succeed with this new kind of responsibility, Rosenblum explained, they need to make assessment concepts such as validity and reliability a part of their day-to-day thinking. They need to internalize the links between the content in the standards, the local curriculum, their own instructional models, and the purposes and nature of the assessments they are carrying out.

Maine bolstered teachers' capacity to do this through a series of regional seminars that tackled assessment issues and presented the details of the way the system was to operate. At summer institutes for assessment development, educators had many opportunities to build their base of knowledge, share ideas, and participate in scoring sessions that helped them focus on performance expectations. Maine considers the work it has done in professional development to be one of the key successes of the program, and cites not only improved assessment literacy, but also improved instruction and a broad-based sense of shared responsibility for the program's success.

WASHINGTON: ADAPTING A TRADITIONAL ASSESSMENT

Greg Hall, assistant superintendent of assessment and research in the Office of the Superintendent of Public Instruction, Washington State Department of Education, explained that the principal purpose of Washington's assessment system is to provide the state, districts, schools, parents, and other stakeholders with evidence of how well students are meeting state standards. The state made the decision to use an assessment program—it is using a criterion-referenced test developed jointly with a commercial testing company—to lead an effort to reform and improve its system. Articulated as an effort to make Washington competitive internationally, the reform goal was not initially popular in a state that had previously been characterized by strong local control of education. Many initially saw the assessment program that was to drive the reform as secretive and out of touch with classroom needs.

The state identified professional development as the potential bridge that could link teachers and classrooms into the potential benefits of the new assessment system, and has found a number of ways to involve teachers in the process. First, they are participants in all stages of test development. The test contractor was asked to conduct all item-writing workshops in the state and to involve only Washington teachers. Teachers also pilot the assessments and are involved in review of the pilot data; they have also conducted the scoring, which has provided ongoing opportunities for them to focus on performance benchmarks. Through regional learning and assessment centers, national assessment experts provide training in assessment issues and methods of interpreting data. Teacher

assessment leadership teams help disseminate the knowledge they gain at the centers, and provide support to other teachers in their home districts and schools.

Washington also strives to help its teachers make use of the data they can obtain from the large-scale assessment. Reports that are provided to every school and district include data linked to each learning target and strand in the state standards, as well as item analyses by school, district, and state. A companion document contains the language of the learning target, so that educators can track patterns in performance on different elements of the standards. The supporting document also provides guidance on how to analyze the data and how to use the released items that are included.

Hall told the workshop that Washington expects that now that teachers are developing competence with large-scale assessment issues, and becoming more comfortable with the data that they can provide, the state will be able to further develop teachers' assessment literacy and, in turn, improve their classroom assessment skills.

BERKELEY EVALUATION AND ASSESSMENT RESEARCH SYSTEM

The Berkeley Evaluation and Assessment Research (BEAR) Center has developed a science assessment system, BEAR, that is based on close links between assessment and curriculum. Indeed, explained Mark Wilson of the University of California at Berkeley, and one of the system's contributing researchers, the idea guiding BEAR is that a large-scale assessment that is not coherent with classroom assessment cannot effectively improve instruction because any gains students make on it will be superficial. At the same time, he added, if classroom assessments are not linked to large-scale assessments, teachers will be faced with the need to teach two curricula, another recipe for failure.

Developed in tandem with a middle school science curriculum, the Issues, Evidence, and You (IEY) program, BEAR is based on a developmental perspective on students' science learning. It is structured around what Wilson calls "progress variables," definitions of the steps students take as they develop higher levels of competence and deeper understanding of the material they are studying. The teacher uses the progress variables to guide instruction and to provide direct feedback to students. The assessment component consists of opportunities to observe student performance, through tasks that are embedded in the instructional program and linked to particular progress variables, and through "link tests," which assess similar skills in different contexts. Thus link tests provide a kind of check on the information gained through the embedded assessments; teachers evaluate both using common, generic scoring guides and examples of student work.

These different sorts of items are then scaled so that student progress on the multiple progress variables that define the curriculum can be monitored. These results are used to establish that the assessments achieve high standards of

reliability and validity (for example, that the classroom-based IEY assessments have reliabilities similar to those archived on standardized tests). The results can be displayed in a variety of ways that can help teachers with planning and instructional activities—for example, by showing an individual’s progress over a year, the state of a class at a particular time, or detailed results on each item for a particular student.

Scoring sessions, in which teachers collaborate to calibrate their expectations, have been a crucial part of the program. The teachers not only learn from one another about performance standards and ways of working with students, they also use the opportunity to have deeper conversations about the educational implications of the assessments and other issues related to teaching. At the same time, Wilson explained, these sessions have been the principal way teachers have made the system their own and internalized its goals and overall approach. Teachers have also conducted similar moderating sessions in their classrooms to help students understand the performance expectations and enter into the goals of the program.

In describing the genesis of the BEAR program—it was developed primarily by graduate students in measurement working with curriculum developers—Wilson noted the ways in which that process encapsulated the gaps the present workshop attempted to address. He observed that the curriculum developers functioned in a sense as artists do, working to assemble a set of experiences that would provoke thinking and have effects on the participants. They had little instinct for the prime concern of the measurement specialists, who focused on finding valid and reliable evidence of particular outcomes. Yet these two groups were able to find common ground using concrete notions of what students would be doing in the form of the progress variables. Using that common framework, they were able to combine their disparate goals into a coherent system.

NORTHERN CALIFORNIA MATHEMATICS ASSESSMENT COLLABORATIVE

The Mathematics Assessment Collaborative (MAC), an initiative of the California-based Noyce Foundation, is made up of thirty school districts in the San Francisco Bay area that share the goal of using high-quality mathematics performance assessments to improve both instruction and student learning.² Participating districts assess 65,000 students every year in grades three through ten. Linda Fisher, who directs MAC, and David Foster, mathematics program director of the Noyce Foundation, described the way the collaborative’s assess-

²The MAC is one of several related projects designed to support mathematics instruction that have been sponsored by the Noyce Foundation. It is considered a component of the Silicon Valley Mathematics Initiative, which addresses all aspects of mathematics instruction and learning.

ment program works and provided a detailed look at the kinds of feedback teachers get about their students from the assessments.

The assessments used by the collaborative are produced by a commercial test publishing company (CTB/McGraw-Hill), together with the Mathematics Assessment Resource Service (MARS), which is a joint endeavor of a number of universities to write performance exams, scoring guides, and score reports that are aligned with the national standards produced by the National Council of Teachers of Mathematics. The collaborative has been administering a performance-based assessment system since 1998; it provides both formative and summative data.

Foster began by setting the collaborative's use of MARS in the context of California's assessment program. He noted that the performance of California students on the SAT 9, a commercially available, norm-referenced test, had increased steadily from 1998 to 2002, but that there were significant discrepancies between student performance on that test and on the MARS. A comparison of the results showed that although both assessments were based on the same standards, students who performed well on the SAT 9 did not necessarily perform well on MARS, the performance-based assessments. The findings for seventh graders, for example, showed that half of the students who performed well on the norm-referenced test did not meet national standards for seventh graders according to the MARS results. These results, Foster explained, demonstrate the critical importance of using multiple measures to assess student performance—without them, educators and administrators can be seriously misled about their students' learning.

The MARS assessment program was designed not only to provide multiple measures of achievement, but also to provide tools teachers can use to target their instruction. The focus on teachers meant both that significant opportunities for professional development were incorporated into the program, and also that the assessment results were produced in a way that was meaningful for teachers in the classroom as well as for more summative purposes. Fisher presented a number of assessment tasks, and some of the data produced from them, to illustrate the "Tools for Teachers" that the MARS program includes.

Box 5-1 is a sample of the results teachers get for each task; it shows results for point four on a ten-point scale. The goal in providing this kind of detail is to encourage teachers to be "reflective about their practice" Fisher explained. What the organizers of the collaborative have found is that as teachers work with such feedback, and consider ways to use it with their students, they become curious about research that might help them understand the misconceptions their students showed and suggest techniques to help them in addressing these problems.

Sessions with teachers to go over the assessment data also yielded broader insights about the kinds of professional development that might best help teachers improve instruction. Fisher explained, for example, that in sessions focused on the textbooks students were using, teachers quickly identified links between the way many of them oriented the information they presented and some of the

**BOX 5-1 Student Performance on Geometry Task Shapes,
MARS 5th Grade Exam, 2002**

Points	Understandings	Misconceptions
4	About half the students correctly drew the rectangle and the square. The other half could draw the square and met partial success with the other shapes.	Many students put together shapes that made triangles, but not right triangles. Students also put together two shapes that did not make triangles at all for the last part.

SOURCE: Fisher and Foster (2003, January).

student misconceptions they had discovered through the assessments. They brainstormed ways to use the textbooks differently so they could anticipate and forestall the misconceptions.

Teachers involved in the collaborative have a variety of other sources of support and development. Summer workshops as well as training sessions during the school year, supporting materials (the “Tools for Teachers,” which include targeted questions for them to use in evaluating their test results and lesson plans), opportunities to participate in scoring the assessments, opportunities for one-on-one coaching and classroom observations, and schoolwide debriefing sessions, are all part of the program. Both Fisher and Foster stressed that the various ways in which teachers are involved and encouraged to learn and change are key elements of the program.

FACET-BASED ASSESSMENT

Jim Minstrell, a former high school physics teacher in Washington state, described a system he has created for teaching physics according to a model of students’ developing understanding. The facet-based system is based on the cognitive principle that students come to physics with ideas and preconceptions that teachers need to identify and build on. To describe the basic units of thought, Minstrell chose the word “facets”—meaning pieces of knowledge, reasoning, or beliefs that students have—because he wanted to include both correct ideas and the incorrect, naïve, or incomplete ideas that students typically have along the way to complete understanding. He chose not to use the word “misconceptions” for the incorrect or incomplete ideas because these ideas often reflect important steps along the way to full understanding that teachers can use to advantage.

Facet clusters, then, are sets of facets related to a particular topic that include both the learning target and a complete and accurate understanding of a complex

principle or other topic, as well as students' evolving notions, arranged in the approximate order that developing understanding usually follows. The facets and clusters have been identified through research, teacher observations, and analysis of student work. Using this means of organizing the content, Minstrell and his colleagues developed a set of tools with which teachers can structure instruction and assessment.

The system provides teachers with tasks, activities, preassessments, and scoring procedures that help them discover which facets their students are using, and then guide students toward complete understanding. All of the activities and assessment tools are linked to some part of the facet cluster for a particular topic and are also coded so that they can be easily analyzed. The codes work with multiple-choice as well as short-answer questions: distractors (incorrect choices) and other student-generated responses are linked to the naïve or incomplete facets identified for the topic. Thus, when a teacher sees that a group of students misunderstand, for example, the effect of ambient air on weight, he or she is prepared: the facet-based system will likely supply a "prescriptive activity" the teacher can use to address this shared misunderstanding in the classroom.

To make the system accessible to more than just a handful of teachers, Minstrell and his colleagues developed a website for Washington teachers and their students. Teachers can find elements such as preinstruction activities for eliciting naïve understandings, "checkout" questions to monitor students' development, tools for interpreting and using assessment results, and other resources and support. Students can also log on to do activities and get feedback about their progress.

Teachers who have used the system have shown measurable improvements in results for individual units, but Minstrell has found it difficult to involve teachers as extensively as he had hoped. Web access in schools has presented a practical obstacle: many schools have outdated systems that are slow or cannot navigate the site, and in many schools students have only limited web access.

A perhaps larger problem has been that many teachers who were intrigued by facet-based assessment were not sure they could manage to incorporate it and still cover all the material their students would need to meet state requirements. While the facet clusters are linked to Washington performance benchmarks for physics, Minstrell recognizes that teachers will need more support if they are to make full use of the program. He and his colleagues are currently conducting research to better understand what kinds of professional development and teacher and district support will be needed to make the program more readily accessible.

MODEL-BASED ASSESSMENT

In her presentation on the Los Angeles Unified School District's application of a National Center for Research on Evaluation, Standards, and Student Testing (CRESST) program, Eva Baker discussed some ideas she believes are critical to

the goal of using assessments to support learning. For Baker, professor in the School of Education, University of California, Los Angeles, the goal of assessment is to produce both usable and useful knowledge, and she explained what she meant by the distinction. Usable knowledge is in a form that can be understood and applied, it is timed appropriately, and it may cause rethinking of the problem. Useful knowledge yields a new solution, based on rethinking of the problem. It is adapted to the situation, it is sufficient to provide a solution, and it can yield an improved outcome.

Some schools are much more successful than others at using assessment knowledge for several reasons. They focus on the learning of both students and adults. They make constant use of appropriate information, drawn from both formal and informal assessments, and they focus on feedback and change. Learning and change are publicized and the entire learning community takes pride in its achievements. The CRESST program, called Model-Based Assessment (MBA), is rooted in this understanding of the ways in which assessments can benefit a learning community.

MBA takes research-based understanding of thinking skills and applies it to different content areas. MBA's key elements of learning are

- content understanding,
- problem solving,
- metacognition (consciousness about one's thought processes),
- communication, and
- teamwork and collaboration.

With MBA, these basic principles were intended to guide both the design of assessments and instruction. Models were developed that could be used as templates and transferred to many subject areas, and were designed so that new teachers can easily be trained to use and score them; they are also reusable and thus relatively inexpensive and easy to adapt. The models, or templates, include tasks, formats, prompts, scoring guides, directions, and samples.

The scoring and performance expectations are based on a research-based model of the way experts in particular domains think and work in their area of expertise. Experts make use of principles or themes in organizing their existing knowledge as well as new information. They draw on prior knowledge, identify explicit relationships among ideas or pieces of information, and avoid misconceptions.³ Baker illustrated the application of this understanding of expertise with several sample templates, showing how the prompts were derived from an understanding of expertise in particular domains, such as using primary documents to organize an essay.

³The expert model is discussed more fully in *Knowing What Students Know* (NRC, 2001c).

Despite the challenges it presented, the opportunity to try out MBA in Los Angeles was welcome, as the assessment's creators were very eager to find out how well the program could operate on a large scale. Initially the plan was to use MBA in four subjects at three grade levels and in two languages. The program is currently being administered in grades two through nine. CRESST staff have trained a large cadre of teachers to score the assessments and to train other teachers. Despite pressures to provide more concrete accountability and to address mandated curriculum packages, Baker has hopes that the program will continue.

CRESST has been conducting validation studies and pursuing a number of research efforts to help it refine the program. Baker cited several key elements to their success in running MBA on such a large scale. Because of the vital importance of cost and time factors, CRESST worked from the start of the program to maintain a low cost per student, and thus benefited from the crucial support of both the school board and teachers' union. Finally, because MBA was designed to be easily transferable, responsibility for the program could be shifted relatively easily to the school district staff, which had many important benefits. Los Angeles educators were much better able to implement the knowledge gained from the assessments because they felt responsible for the program. Moreover, teachers learned and benefited from their participation, and the MBA was more easily meshed with other educational mandates by those within the system than it could have been by CRESST staff.

6

Concluding Thoughts and Possible Next Steps

At the end of the two-day program, a panel of participants was convened to synthesize what they had heard and try to identify some of the key messages. Referring back to the criteria the committee had asked the presenters to consider with regard to their programs, the discussants noted that it was clear that selecting any one of them as most important could not be the key to bridging the kinds of gaps that were discussed. Rather, the criteria emerged as important ways of considering the strengths and limitations of different approaches.

Each of the programs presented was tailored to suit a particular set of circumstances, and to address particular challenges, and some of the differences among them were striking. Some served students in relatively disadvantaged circumstances; others served greater numbers of advantaged students. The programs ranged in scale, in their methods, and in the goals they were trying to meet. Thus no one starting point would make sense for all of them.

The discussants also noted that few of the presenters provided much evidence of the effectiveness of their programs.¹ Moreover, with a few exceptions, little effort has yet been made to transfer these programs to other settings with different characteristics. The discussants noted that such follow-up work is badly needed. The programs discussed for the most part struck them as very promising, but many are still in early stages of their development. It will be very important,

¹This may have been partly because the agenda was very full. Such evidence may be available for many of the programs, and interested readers are encouraged to seek it using the contact information in Appendix C.

the discussants agreed, to see which elements of them are of use beyond their own contexts.

At the same time, however, several important common threads were apparent in the presentations and in the discussion. These are presented not as recommendations to those responsible for assessment programs, but rather as a distillation of the experiences described at the workshop which may be helpful to others:

- **Responsibility lodged with teachers.** In virtually every one of the programs, the responsibilities that devolve to teachers seem to be critical to the success of the enterprise. Many speakers were struck by the extent to which these programs were dependent on teachers who were prepared to change their thinking and their practice. Teachers were asked to master new concepts and techniques for assessment, and also, in many cases, to change other elements of their work as they adapted to the needs of the assessment program. Perhaps most important, teachers' judgments about student performance, how and when to assess, and many other issues are being sought and used in these programs to an extent not often seen.
- **Commitment to professional development.** Presenter after presenter spoke about how important an investment in professional development was to their programs. The developers of many of the programs that placed new responsibilities on teachers realized from the start that teachers had not had sufficient training in measurement to succeed with the new requirements without targeted training up front. Ongoing support of many kinds—through summer workshops, shared websites, mentoring networks, and the like—is another key element in many. Experienced teachers were enlisted in many cases to spread their knowledge to colleagues, and teachers were offered opportunities to participate in test development and scoring sessions. Several presenters expressed concern that resources to maintain this level of commitment may be at risk but all seemed convinced that it was crucial.
- **Clear descriptions of expectations for students.** Many of the programs had in common descriptions of the expectations for students that are unusually concrete and detailed. Using frameworks, matrices, or some other structure, many of the examples that were discussed provide teachers with clear definitions of the stages students are likely to move through as they progress to mastery of chosen academic objectives. Breaking the learning process down in this way seemed to be a particularly useful way of meshing the goals of instructions and accountability.
- **Plentiful feedback to teachers and students.** In many of the programs, provision of usable feedback is built into the system, and often careful thought has gone to the form the feedback will take. Reports of assessment results often include analysis that breaks down the student work to reveal specific misconceptions and gaps in knowledge. The feedback is,

in many cases, designed to be folded back into both the teachers' decisions and the students' thinking about where they stand and what they need to do.

- **Summative assessments do not stand alone.** Though none of the examples discussed was perhaps initiated with the explicit goal of bridging a gap—or, certainly, of meeting the criteria the committee has described—they do mostly share the notion that summative assessments ought not to be stand-alone exercises, but elements of an integrated system. While few would likely disagree with that notion, the programs here have taken a variety of specific steps to try to make it a reality. As each of the programs proceeds, evidence of their success may influence other states and districts that are recognizing the consequences of having a system that is not as coherent and integrated as it could be.
- **Adherence to professional standards.** In many of the programs discussed, the explicit assistance of measurement experts was sought either to review new assessment plans or to work with and train the teachers and officials who would be developing and carrying out the program. The developers of these programs recognized that they were attempting something ambitious and that taking particular care that the technical innovations passed professional muster would be important. At the same time, content specialists were often involved in developing the detailed expectations for students discussed above. The programs that were developed by researchers were of course also grounded in high professional standards. Although high professional standards are important to any assessment program, the participation of experts is perhaps especially important where educators and administrators are trying to meet expectations for accountability in new ways.

This workshop is just a step toward the National Research Council's goal of fostering the understanding of and commitment to assessment for learning. The information presented here will be used in studies just being initiated by all three of the boards that sponsored the workshop. The Board on Testing and Assessment is overseeing a project that will help states design the science assessments that will be required under the No Child Left Behind Act. The Mathematical Sciences Education Board has a study of mathematics assessments underway, and the Committee on Science Education K-12 is conducting a study on science learning. The committee hopes that the examples presented here will stimulate the thinking of each of these committees as they consider the tensions presented by assessment systems with multiple goals.

References

- American Association for the Advancement of Science, Project 2061. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom*. London: Department of Education and Professional Studies, King's College.
- Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability: A guide for policymakers*. Available: <http://www.nea.org/accountability/buildingtests.html> [June 24, 2003].
- Fisher, L., and Foster, D. (2003). *Using a large-scale assessment system to improve teaching and learning in mathematics*. Presentation for National Research Council's Committee on Assessment in Support of Instruction and Learning workshop Bridging the Gap Between Large-Scale and Classroom Assessment, January, Washington, DC.
- Maine Department of Education. (2003). *Characteristics of Maine's assessment system*. Available: <http://www.state.me.us/education/lsalt/compassess.htm> [June 24, 2003].
- National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.
- National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment*. Mathematical Sciences Education Board. Washington, DC: National Academy Press.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- National Research Council. (1998). *Learning about assessment, learning through assessment*. Mathematical Sciences Education Board. Mark Driscoll and Deborah Bryant (Eds.). Center for Science, Mathematics, and Engineering Education. Washington, DC: National Academy Press.
- National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Committee on Developments in the Science of Learning. John Bransford, Ann Brown, and Rodney Cocking (Eds.). Committee on Learning Research and Educational Practice. Suzanne Donovan, John Bransford, and James Pellegrino (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

- National Research Council. (2001a). *Adding it up: Helping children learn mathematics*. Mathematics Learning Study Committee. Jeremy Kilpatrick, Jane Swafford, and Bradford Findell (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2001b). *Classroom assessment and the national science education standards*. Committee on Classroom Assessment and the National Science Education Standards. J. Myron Atkin, Paul Black, and Janet Coffey (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2001c). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. James Pelligrino, Naomi Chudowsky, and Robert Glaser (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in American high schools*. Committee on Programs for Advanced Study of Mathematics and Science in American High Schools. Jerry Gollub, Meryl Bertenthal, Jay Labov, and Philip Curtis (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Popham, W. James. (in press). *Crafting curricula aims for instructionally supportive assessment*. Available: <http://education.umn.edu/nceo/Presentations/CraftingCurricula.pdf> [August 2003].
- Shepard, L. (2003). *Large-scale assessment: Key policy and psychometric issues*. Presentation for National Research Council's Committee on Assessment in Support of Instruction and Learning workshop Bridging the Gap Between Large-Scale and Classroom Assessment, January, Washington, DC.

APPENDIX A

Workshop Agenda

Bridging the Gap Between Large-Scale and Classroom Assessment
January 23–24, 2003
The National Academies
500 Fifth Street, NW, Room 100
Washington, DC

Thursday, January 23

Time	Topic	Speaker
7:30–8:00	Continental Breakfast	
8:00–8:45	Welcome and Workshop Overview	Michael Feuer, National Research Council J. Myron Atkin, Stanford University Marge Petit, National Center for Improving Educational Assessment
8:45–9:30	Teaching and Learning in a Coherent Educational Environment	Gail Burrill, Michigan State University

9:30–10:00	Large-Scale Assessment: Laying Out the Territory	Lorrie Shepard, University of Colorado, Boulder
10:00–10:45	Classroom Assessment in Support of Learning and Instruction	Dylan Wiliam, King’s College, London
11:00–12:00	A Report from the Commission on Designing Instructionally Supportive Assessment	James Popham, University of California, Los Angeles
12:00–12:45	Lunch	
12:45–1:30	Bridging the Gap: Aligning Formative and Summative Assessments	Richard Shavelson, Stanford University
1:30–2:15	Special Issues in Mathematics and Science Assessment	Jan de Lange, Freudenthal Institute, Utrecht University, The Netherlands J. Myron Atkin, Stanford University
2:15–3:00	Northern California Mathematics Assessment Collaborative	Linda Fisher and David Foster, The Noyce Foundation
3:00–3:45	Delaware Comprehensive Science Assessment	Rachel Wood, Delaware Department of Education
3:45–4:45	Building Bridges Between Large-Scale and Classroom Assessment with PASS; The Vermont Assessment Program: An Exemplar	Kathleen Comfort, WestEd Bud Meyers and David White, Vermont Department of Education

4:45–5:45	International Comparative Perspective	Barry McGaw, Organisation for Economic Co-operation and Development
5:45	Adjourn	

Friday, January 24

Time	Topic	Speaker
7:30–8:00	Continental Breakfast	
8:00–8:45	Washington State Assessment Program	Greg Hall and Dawn Billings, Washington State Department of Education
8:45–9:30	International Baccalaureate Diploma Programme	George Pook, International Baccalaureate Organisation
9:30–10:30	A Classroom/National Literacy Assessment System	Geoff Masters, Australian Council for Educational Research
10:30–11:30	BEAR Assessment System	Mark Wilson, University of California, Berkeley
11:30–12:15	Facet-Based Assessment	Jim Minstrell, Talaria, Inc.
12:15–1:00	Lunch	
1:00–1:45	Model-Based Assessment: Why, What, How, How Good, and What Next?	Eva Baker, University of California, Los Angeles

1:45–3:45	Nebraska: STARS Program	Patricia Roschewski, Nebraska Department of Education
	A Proposed High School Graduation Assessment System to Blend Local Initiatives and State Requirements: Experience from Wyoming	Scott Marion, Wyoming Department of Education (<i>formerly</i>)
	Maine’s Comprehensive Assessment System	Jill Rosenblum, Maine Mathematics and Science Alliance Pam Rolfe, Maine State Department of Education
3:45–5:15	Panel Discussion	James Popham, Mark Wilson, Barry McGaw, Lorrie Shepard, Jill Rosenblum, Dylan Wiliam, Eva Baker, and J. Myron Atkin
5:15–5:45	Next Steps	Continued Panel Discussion
5:45	Adjourn	

APPENDIX B

Workshop Participants

Alegria, Adelina
University of California, Los
Angeles

Asp, Elliott
Douglas County School District
Colorado

Banilower, Eric
Horizon Research, Inc.

Barchfeld-Venet, Penny
TERC

Beane, DeAnna
Association of Science-Technology
Centers, Inc.

Bertani, Al
Chicago Public Schools

Binder, Wendy
The LASER Center

Bricker, Leah
Project 2061
American Association for the
Advancement of Science

Campbell, Jay
Educational Testing Service

Carlisle, Peggy
Pecan Park Elementary School
Jackson, Mississippi

Carlson, Jim
U.S. Department of Education

Carmody, David
District of Columbia Public Schools

Chatman, Liesl
Science and Health Education
Partnership
University of California, San
Francisco

Clemens, Bev Douglas County School District Colorado	Garfield, Joan University of Minnesota
Coffey, Janet Stanford University	Gartzman, Martin Chicago Public Schools
Countryman, Lyn University of Northern Iowa	Gerretson, Helen University of Northern Colorado
DeBoer, George Project 2061 American Association for the Advancement of Science	Glidden, Heidi American Federation of Teachers
Della-Piana, Gabriel National Science Foundation	Good, Dan Ohio State Department of Education
Driesler, Stephen Association of American Publishers, School Division	Hammond, Peirce U.S. Department of Education
Earle, Janice National Science Foundation	Haney, Michael National Science Foundation
Espinoza, Anna United Independent School District Laredo, Texas	Hansen, Phillip Chicago Public Schools
Fields, Ray National Assessment Governing Board	Harmon, Patricia San Francisco Unified School District
Franklin, Christine University of Georgia	Hollinger Martinez, Debra National Center for Education Statistics
Freitag, Patricia National Science Foundation	Jackson, Tamara Office of Science and Technology Policy
Fry Bohlin, Carol California State University, Fresno	Jolly, Anne SERVE
	Kirby, Sheila RAND Corp.

Krebs, Susan Academy School District 20 El Paso County, Colorado	Medhurst, Kristin Department of Defense Education Activity
Kubo Della-Piana, Connie National Science Foundation	Miller-Jones, Dalton Portland State University
LaPointe, Archie Educational Testing Service	Minner, Daphne Center for Science Education Education Development Center, Inc.
Leitzel, Joan University of New Hampshire	Mitchell, Monica National Science Foundation
Lemke, Mariann National Center for Education Statistics	Naftel, Scott RAND Corp.
Lopez-Ferrao, Julio National Science Foundation	O'Connell Ross, Patricia U.S. Department of Education
MacGregor, Ian National Science Resources Center Smithsonian Institution	Palaez, Nancy California State University, Fullerton
Malwitz, Jaime National Science Foundation	Perez-Pelaez, Anna Association of Science-Technology Centers Inc.
McGiffert, Laura Achieve	Pugsley, Ronald U.S. Department of Education
McKinney, Wilhelmina Department of Defense Education Activity	Pyke, Curtis George Washington University
McMunn, Nancy SERVE Regional Educational Laboratory	Raizen, Senta National Center for Improving Science Education
Mecca, Peter Department of Defense Education Activity	Salley, Columbus CTB/McGraw-Hill

Sikora, Sharon
Center for Learning and Teaching of
the West

Simutis, Len
Ohio State University

Sipes, Sunny
Human Resources Research
Organization

Snowwhite, Larry
Houghton Mifflin Co.

Sprigg, Nancee
Douglas County School District
Colorado

Sroufe, Jerry
American Educational Research
Association

Stein, Sondra
National Institute for Literacy

Stites, Regie
SRI International

Suiter, Marilyn
National Science Foundation

Swanson, Elisabeth
Montana State University, Bozeman

Sweet, David
U.S. Department of Education

Tappan, Richard
Center for Assessment

Thiemann, Alan
Thiemann Aitken & Vohra, LLC

Tuomi, Jan
Mid-continent Research for
Education and Learning

Turner, Ken
Academy School District 20
Colorado Springs, Colorado

Weedon, Jason
Achieve

Winokur, Marc
Center for Learning and Teaching of
the West

Wood, Kim
Horizon Research, Inc.

Worth, Karen
Education Development Center, Inc.

APPENDIX C

Resources for Further Information

The Ideal

Presenter: Gail Burrill
Academic Specialist/Mathematics
Division of Science and Mathematics Education
Michigan State University
116 North Kedzie Lab
East Lansing, MI 48823
(517) 432-2152 ext. 133
burrill@msu.edu
<http://www.dsme.msu.edu/>

Large-Scale Assessments

Presenter: Lorrie Shepard
Dean, School of Education
Professor of Education
School of Education, Room 124
University of Colorado at Boulder
249 UCB
Boulder, CO 80309-0249
(303) 492-6937
lorrie.shepard@colorado.edu
<http://www.colorado.edu/education/faculty/lorrieshepard/>

Classroom Assessments

Presenter: Jan de Lange
Freudenthal Institute
Utrecht University
Aidadreef 12
3561 GE Utrecht
Netherlands
(31) 30 263 55 55
jan@fi.ruu.nl
<http://www.fi.uu.nl/en/welcome.html>

Presenter: J. Myron Atkin
Professor
Center for Educational Research at Stanford
Stanford University
520 Galvez Mall
Stanford, CA 94305-3084
(650) 723-4385
atkin@stanford.edu
<http://ed.stanford.edu/suse/>

The Nature of the Gap

Presenter: W. James Popham
Professor Emeritus, UCLA
1706 Keoniloa Place
Koloa, HI 96756
wpopham@ucla.edu
<http://www.nea.org/accountability/buildingtests.html>

Australia

Presenter: Geoff Masters
Chief Executive Officer
Australian Council for Educational Research Ltd.
Private Bag 55
Camberwell VIC 3124
Australia
(61) 3 9277 5511
masters@acer.edu.au
<http://www.acer.edu.au/>

Queensland

Presenter: Richard J. Shavelson
Professor of Education and Psychology
308 School of Education Bldg.
485 Lasuen Mall
Stanford University
Stanford, CA 94305-3096
(650) 723-4040
richs@stanford.edu
<http://www.stanford.edu/dept/SUSE/SEAL/>

Great Britain: Enhanced Formative Assessment

Presenter: Dylan Wiliam
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
dwilliam@ets.org
<http://www.kcl.ac.uk/depsta/education/hpages/dwilliam.html>

The International Baccalaureate Programme

Presenter: George Pook
Head of Assessment, International Baccalaureate Organisation
Fortran Road, St. Mellons
Cardiff, Wales CF30WB
(44) 29 20 54 7777
georgep@ibo.org
<http://www.ibo.org/>

Programme for International Student Assessment

Presenter: Barry McGaw
Director for Education
Organisation for Economic Co-operation and Development
2 rue André-Pascal
75775 Paris Cedex 16
France
(33) 1 45 24 92 10
barry.mcgaw@oecd.org
<http://www.pisa.oecd.org/>

Nebraska: STARS Program

Presenter: Patricia Roschewski
Director of Statewide Assessment
Nebraska Department of Education
301 Centennial Mall S.
Lincoln, NE 68519
(402) 471-2495
proschew@nde.state.ne.us
<http://www.nde.state.ne.us>

Delaware: Comprehensive Science Assessment

Presenter: Rachel Wood
Education Associate, Science/Environment Education
Science Resource Center
Curriculum & Instructional Improvement Branch
655 Glenwood Avenue
Smyrna, DE 19977
(302) 653-3448
rwood@doe.k12.de.us
<http://www.doe.state.de.us/>

Vermont: Partnership for the Assessment of Standards-Based Science

Presenter: Kathy Comfort
Principal Investigator/Project Director PASS and RISSA
WestEd
730 Harrison Street
San Francisco, CA 94107-1242
(415) 615-3161
kcomfort@wested.org
<http://www.wested.org/cs/wew/view/pj/278>

Vermont Assessment Program

Presenter: Herman “Bud” Meyers
Standards and Assessment Coordinator
Vermont Department of Education
120 State Street
Montpelier, VT 05620-2501
(802) 828-5101
bmeyers@doe.state.vt.us
<http://www.state.vt.us/educ/>

Presenter: David White
Science Assessment Coordinator
Vermont Department of Education
120 State Street
Montpelier, VT 05620-2501
(802) 828-0154
dwhite@doe.state.vt.us

Wyoming: Body of Evidence System

Presenter: Scott Marion
Former Director of Assessment
Wyoming Department of Education
2300 Capitol Avenue
Cheyenne, WY 82002-0050
<http://www.k12.wy.us/>

Maine: Comprehensive Assessment System

Presenter: Pam Rolfe
Local Assessment Coordinator
Maine Department of Education
23 State House Station
Augusta, ME 04333
(207) 624-6785
pam.rolfe@state.me.us
<http://www.state.me.us/education/homepage.htm>

Presenter: Jill Rosenblum
Assessment and Evaluation (K-12)
Maine Mathematics and Science Alliance
PO Box 5359
Augusta, ME 04332
(207) 287-6644
jrosenblum@mmsa.org
<http://www.mmsa.org>

Washington State Assessment Program

Presenter: Greg Hall
Assistant Superintendent
Assessment and Research
Office of Superintendent of Public Instruction
PO Box 47200
Olympia, WA 98504-7200
(360) 725-6336
ghall@ospi.wednet.edu
<http://www.k12.wa.us>

Presenter: Debra Brown
Executive Assistant
Assessment and Research
Office of Superintendent of Public Instruction
Old Capitol Building
PO Box 47200
Olympia, WA 98504-7200
(360) 725-6334
dabrown@ospi.wednet.edu

Berkeley Evaluation and Assessment Research System

Presenter: Mark Wilson
Professor
Graduate School of Education
University of California
Berkeley, CA 94720
(510) 642-7966
mrwilson@socrates.berkeley.edu
<http://www-gse.berkeley.edu/research/BEAR/>

Northern California Mathematics Assessment Collaborative

Presenter: Linda Fisher
Director of the Mathematics Assessment Collaborative
237 Navigator Drive
Scotts Valley, CA 95066
(831) 430-0506
lfisher@noycefdn.org
<http://www.noycefdn.org/math/mac.htm>

Presenter: David Foster
Program Director, Math
The Noyce Foundation
17485 S. Monterey Boulevard, Suite 301
Morgan Hill, CA 95037
(408) 776-1645
dfoster@noycefdn.org

Facet-Based Assessment

Presenter: Jim Minstrell
Talaria, Inc.
821 2nd Avenue, Suite 1150
Seattle, WA 98104
(206) 748-0443
jimminstrell@talariainc.com
<http://www.talariainc.com/k12.html>

Model-Based Assessment

Presenter: Eva Baker
Director/Co-director
University of California, Los Angeles
CSE/CRESST
GSE&IS Building, Mailbox 951522
Los Angeles, CA 90095-1522
(310) 206-1530
eva@ucla.edu
<http://www.cse.ucla.edu/index1.htm>

