

## Keeping Score

### DETAILS

---

96 pages | 7 x 10 | PAPERBACK  
ISBN 978-0-309-06535-1 | DOI 10.17226/9635

### AUTHORS

---

by Ann Shannon; Mathematical Sciences Education Board, National Research Council

BUY THIS BOOK

FIND RELATED TITLES

### Visit the National Academies Press at [NAP.edu](http://NAP.edu) and login or register to get:

---

- Access to free PDF downloads of thousands of scientific reports
- 10% off the price of print titles
- Email or social media notifications of new titles related to your interests
- Special offers and discounts



Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. (Request Permission) Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences.

# **KEEPING SCORE**

**by Ann Shannon**



MATHEMATICAL SCIENCES EDUCATION BOARD  
CENTER FOR SCIENCE, MATHEMATICS, AND ENGINEERING EDUCATION  
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS  
Washington, D.C.

**NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, NW • Washington, DC 20418**

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.

The Center for Science, Mathematics, and Engineering Education (CSMEE) was established in 1995 to provide coordination of all the National Research Council's education activities and reform efforts for students at all levels, specifically those in kindergarten through twelfth grade, undergraduate institutions, school-to-work programs, and continuing education. The Center reports directly to the Governing Board of the National Research Council.

The Mathematical Sciences Education Board was established in 1985 to provide a continuing national capability to assess the status and quality of education in the mathematical sciences and is concerned with excellence in education for all students at all levels. The Board reports directly to the Governing Board of the National Research Council.

Development, publication, and dissemination of this report were supported by a grant from The Carnegie Corporation of New York. Any opinions, findings, or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of The Carnegie Corporation of New York.

International Standard Book Number 0-309-06535-6

Permission for limited reproduction of portions of this book for education purposes but not for sale may be granted on receipt of a written request to the National Academy Press, 2101 Constitution Avenue, NW, Washington, DC 20418.

Additional copies of this report may be purchased from the National Academy Press, 2101 Constitution Avenue, NW, Lock Box 285, Washington, DC 20055. (800) 624-6242 or (202) 334-3313 (in the Washington Metropolitan Area). This report is also available online at <http://www.nap.edu>.

Printed in the United States of America

Copyright 1999 by the National Academy of Sciences. All rights reserved.

# THE NATIONAL ACADEMIES

National Academy of Sciences  
National Academy of Engineering  
Institute of Medicine  
National Research Council

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. William 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. Kenneth I. Shine 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. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.



**NATIONAL RESEARCH COUNCIL  
CENTER FOR SCIENCE, MATHEMATICS, AND ENGINEERING EDUCATION  
MATHEMATICAL SCIENCES EDUCATION BOARD  
JULY 1, 1998 - JUNE 30, 1999**

**HYMAN BASS** (MSEB Chair)  
Columbia University

**JERE CONFREY** (MSEB Vice Chair)  
University of Texas at Austin

**RICHARD A. ASKEY**  
University of Wisconsin–Madison

**SHERRY BACA**  
Prescott Unified School District

**DEBORAH BALL**  
University of Michigan

**BENJAMIN BLACKHAWK**  
St. Paul Academy and Summit School

**RICHELLE BLAIR**  
Lakeland Community College

**PATRICIA CAMPBELL**  
University of Maryland

**INGRID DAUBECHIES**  
Princeton University

**KAREN ECONOMOPOULOS**  
TERC

**SUSAN EYESTONE**  
National Parent Teachers Association

**LEE JENKINS**  
Antioch Unified School District

**GLENDA T. LAPPAN**  
Michigan State University

**MIRIAM MASULO**  
IBM Corporation

**DAVID MOORE**  
Purdue University

**MARI MURI**  
Connecticut Department of Education

**RICHARD NORMINGTON**  
TQM Services Group

**MARK SAUL**  
Bronxville Public Schools

**RICHARD SCHOEN**  
Stanford University

**EDWARD A. SILVER**  
University of Pittsburgh

**WILLIAM TATE**  
University of Wisconsin–Madison

**JERRY UHL**  
University of Illinois at Urbana-Champaign

**SUSAN S. WOOD**  
J. Sargeant Reynolds Community College

**Project Staff**

**RODGER BYBEE**  
Executive Director, CSMEE, through  
June, 1999

**SUZANNE WOOLSEY**  
Acting Executive Director, CSMEE

**JOAN FERRINI-MUNDY**  
Director, MSEB, through June, 1999

**GAIL BURRILL**  
Director, MSEB

**BRADFORD FINDELL**  
Program Officer/Editor

**GALE MOORE**  
Financial & Admin. Associate

## Reviewers

---

This report has been reviewed 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 authors and the NRC in making the 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 content of 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 participation in the review of this report:

**HUGH BURKHARDT**  
Shell Centre  
University of Nottingham, England

**MARIETA HARRIS**  
Memphis City Schools  
Memphis, TN

**EDWARD T. ESTY**  
Consultant  
Chevy Chase, MD

**DEBORAH SPENCER**  
Education Development Center, Inc.  
Newton, MA

While the individuals listed above have provided many constructive comments and suggestions, responsibility for the final content of this report rests solely with the authoring committee and the NRC.

## Table of Contents

---

<b>Acknowledgments</b>	<b>viii</b>
<b>Preface</b>	<b>ix</b>
<b>Chapter 1: Introduction</b>	<b>1</b>
<b>Chapter 2: A model for assessment development: Achieving balance</b>	<b>10</b>
<b>Chapter 3: Assessment and opportunity to perform</b>	<b>31</b>
<b>Chapter 4: Assessment and opportunity to learn</b>	<b>55</b>
<b>Chapter 5: Alignment and standards-based assessments</b>	<b>71</b>
<b>References</b>	<b>81</b>



## Acknowledgments

---

The development of this manuscript owes much to many. At New Standards, I am especially indebted to Pam Beck, Phil Daro, Elizabeth Stage, Dick Stanley, and Bokhee Yoon who provided encouraging comments on various drafts. In addition, the support of Bob Agee, Harold Asturias, and Mishaa Degraw is warmly acknowledged. I would also like to thank Claudia Alfaro for her help in producing this manuscript.

Members of the Balanced Assessment Project also played a role in the development of the ideas addressed here. I would like to thank Alan Schoenfeld, Sandy Wilcox, Judi Zawojewski, Alan Bell, Hugh Burkhardt, Rita Crust, John Gillespie, Daniel Pead, Richard Phillips, and Malcolm Swan.

It would not be practicable to name each of the teachers and mathematics educators who have contributed to this work. A special word of acknowledgment is due, however, to Diane Briars, Steve Leinwand, Joanne Mosier, Marge Pettit, as well as Dot Dow and her staff in the mathematics department at Dublin High School.

At the NRC, I would like to thank Joan Ferrini-Mundy for giving me the opportunity to collect together and write down these ideas that have been percolating over the last 10 years. Special thanks are also due to Gail Burrill and Brad Findell for their help editing each draft of this booklet and in coordinating its review and production. This manuscript benefited immensely from the comments of the reviewers, who, of course, remained anonymous to me throughout the process.

Finally, I would like to acknowledge the loving help, support, and encouragement that I received throughout from Andrew Bell.

---

Ann Shannon  
New Standards  
University of California, Office of the President

## Preface

---

Curriculum reform, performance assessment, standards, portfolios, and high stakes testing—what’s next? What does this all mean for me in my classroom? Many teachers have asked such questions since mathematics led the way in setting standards with the publication of the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989). This seminal document and others that followed served as catalysts for mathematics education reform, giving rise to new initiatives related to curriculum, instruction, and assessment over the past decade. In particular, approaches to classroom, school, and district-wide assessment have undergone a variety of changes as educators have sought to link classroom teaching to appropriate assessment opportunities.

Since the publication of *Everybody Counts* (National Research Council [NRC], 1989), the Mathematical Sciences Education Board (MSEB) has dedicated its efforts to the improvement of mathematics education. A national summit on assessment led to the publication of *For Good Measure* (NRC, 1991). This statement of goals and objectives for assessment in mathematics was followed by *Measuring Up* (NRC, 1993a), which provided prototypical fourth-grade performance assessment tasks linked to the goals of the NCTM’s *Curriculum and Evaluation Standards*. *Measuring What Counts* (NRC, 1993b) demonstrated the importance of mathematics content, learning, and equity as they relate to assessment. The MSEB is now prepared to present perspectives on issues in mathematics education assessment for those most directly engaged in implementing the reform initiatives on a daily

basis—classroom teachers, school principals, supervisors, and others in school-based settings.

The MSEB, with generous support and encouragement from the Carnegie Corporation of New York, seeks to bring discussion of assessment to school- and district-based practitioners through an initiative called Assessment in Practice (AIP). Originally conceived as a series of “next steps” to follow the publication of *Measuring Up* and *For Good Measure*, the project, with assistance from an advisory board, developed a publication agenda to provide support to teachers and others directly involved with the teaching and assessment of children in mathematics classrooms at the elementary, middle, and high school levels.

Through a pair of resource booklets, AIP presents an exploration of issues in assessment. These booklets are specifically designed to be used at the school and school district level by teachers, principals, supervisors, and measurement specialists. Because these booklets are commissioned works, the opinions and recommendations they contain are those of the authors and not necessarily of the MSEB or the NRC. The first booklet, *Learning About Assessment, Learning Through Assessment*, written by Mark Driscoll and Deborah Bryant, discusses ways to assist teachers in learning about assessment and how student work can be a rich resource in professional development. This booklet, *Keeping Score*, written by Ann Shannon, discusses issues to be considered while developing high-quality mathematics assessments. Much of the raw material and analysis in this booklet grew out of the author’s work with two projects, New Standards and Balanced Assessment, that are intended to produce mathematics assessments supporting the NCTM *Standards*. The publication of this booklet is not an NRC or MSEB endorsement of these projects, but rather a suggestion that many of the ideas may be useful to people who are rethinking the roles that assessment has played in mathematics education.

As we continue in our efforts to understand the implications of standards-based curriculum, instruction, and assessment, it is critical that teachers and others involved with the practice of instruction have the opportunity to reflect on how to best achieve the ultimate goal of improving student learning in mathematics. The MSEB welcomes this opportunity to provide resources in the area of assessment.

---

Hyman Bass, Chair  
Mathematical Sciences Education Board  
April, 1999

## Chapter 1: Introduction

---

Assessment has long been viewed as a tool for effecting change in classrooms, and consequently as a factor that can influence the learning of mathematics (Resnick & Nolan, 1995; Romberg, Zarinnia, & Williams, 1990). Since the release of the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989), many individuals and organizations have been working to develop assessments that represent the kind of mathematics and support the kind of mathematical teaching and learning envisioned by those standards. Now, after 10 years, some of that work has come to fruition with the availability of some examples of large-scale standards-based assessments. In both purpose and design, these assessments are very different from traditional norm-referenced tests. Much has been learned about standards-based assessment during the development process.

This booklet is intended to highlight a few lessons that have relevance and implications for the classroom. It is directed toward an audience of supervisors of mathematics teachers, mathematics teachers, designers of mathematics tasks and assessments, and administrators. Particular issues arise with assessments designed to support “reformist” approaches to the teaching and learning of mathematics. The implications of these issues for task development and assessment design, classroom practice, and assessment policy will be discussed here. Task designers should be able to draw upon the ideas presented here in creating balanced and equitable assessments for all students. Mathematics teachers and their supervisors will be able to use these lessons in designing and

administering classroom assessments. The materials will help administrators provide teachers with the support necessary for enhancing the teaching and learning of mathematics. This chapter begins by defining standards-based assessment and other key terms, and then compares standards-based assessment with traditional norm-referenced tests. The remainder of the chapter describes the organization of the booklet.

In this booklet, experience and research from two assessment development projects, Balanced Assessment and New Standards, will be drawn upon to offer guidance about the development and implementation of assessments. Throughout, four particularly important ideas will be addressed: balance, opportunity to perform, opportunity to learn, and alignment. The intent is not to describe all of the details associated with large-scale assessment development but rather to discuss how these four basic ideas can be used to inform the development and implementation of innovative assessment instruments.

### **Standards-based assessments**

The assessments we are concerned with here are those that are necessary in a standards-based system of education. In such a system, assessments need to be designed to assess whether students meet publicly negotiated and agreed-upon standards. The standards to which an assessment is referenced might be those of a state, a district, or the National Council of Teachers of Mathematics (i.e., NCTM, 1989). It does not necessarily matter what or whose standards are chosen, but in our view what is important is that the selected standards promote a broad and balanced approach to learning mathematics, where conceptual understanding and mathematical skills are both emphasized.

Standards will be more likely to have a positive effect on mathematics learning if assessment, curriculum, and instruction are aligned with them (Webb, 1997). Assessments that are aligned to standards are referred to as *standards-based assessments*. When curriculum, instruction, and assessment are aligned to the same set of standards, those *standards* lay out what a student should know and be able to do. *Curriculum and instruction* provide opportunities for students to learn the mathematics that the standards ask them to learn and to acquire the know-how for them to show what they have learned. *Assessments* provide opportunities for students to perform and allow inferences to be made about what students know and can do in mathematics.

When instruction and assessment are aligned to standards, the role of the teacher is central. What goes on in the classroom is crucial to the enhancement of learning mathematics (Black &

William, 1998). The teacher is a critical interpreter in this effort. She must interpret what the standards are asking students to do and translate these expectations into learning opportunities for students. Through classroom assessment, she must interpret how students are learning what they are supposed to learn. Then, in the light of this assessment, she must give students feedback and create subsequent learning opportunities for the whole range of students under her care. An important aspect of the teacher's role is to enable all students, not just the most academically talented, to learn the mathematics that is specified.

Unlike traditional tests, standards-based assessments will be tightly rather than loosely linked to the curriculum. Some examples of assessments are the tightly linked to curriculum are the system of Advanced Placement (AP) Examinations that are designed by the College Board. For these exams, a syllabus or course guide specifies what students are to learn, and then the end-of-course examinations assess how well they have learned that material. Most other countries have exams that are based directly on syllabi that often comprise two years of study. These high stakes tests include very specific and often demanding questions that require a variety of types of responses. The U.S., in contrast, depends predominantly on multiple-choice tests.

In a standards-based system, the student is actively involved in the learning process. The student can strengthen opportunities to learn by working to understand mathematics rather than being concerned only with completing the work and getting a grade. Students need to learn that standards-based assessments are designed so students may succeed through hard work. It is then the student's job to learn *how* to do that hard work.

### **Differences between norm-referenced and standards-based assessments**

**Purpose.** Traditional norm-referenced tests are intended to sort students by rank or to compare the performance of one cohort of students against that of another cohort. Thus, on a norm-referenced test, half the students are expected to have scores that are below average. Standards-based assessments, on the other hand, are intended to assess whether students can do what a specified set of standards asks them to do. On a standards-based assessment, then, it is possible for all students to meet the standard.

Norm-referenced tests report how a student or group of students compare to the "norm." In contrast, standards-based assessments are *criterion-referenced*, in the sense that test designers set an absolute level of performance and report whether a student has met that level. The standards, then, are the criteria against which student performance is measured.

**Preparation.** One of the crucial differences between traditional norm-referenced tests and standards-based assessments is that students *are* expected to study for standards-based assessments, and students who work hard to meet clearly defined learning expectations should be able to do well. Traditional norm-referenced tests are descended from the earliest intelligence tests, for which subjects were never expected to study. The expectation that students should succeed on standards-based assessments through hard work has important implications for their design, and the *Assessment Standards for School Mathematics* (NCTM, 1995) goes a long way toward defining explicit standards for such assessments.

**Range of skills.** Standards-based assessments need to be designed to assess a much broader range of mathematics than has traditionally been assessed. If assessment is to drive a renewed approach to teaching and learning, it needs to incorporate a broad and significant range of mathematics to provide all students with the opportunities to solve a variety of worthwhile problems, reason mathematically, understand concepts, develop technical skills, make connections among mathematical ideas, and communicate about mathematics (NCTM, 1995, p. 11).

**Variety of tasks.** To assess a considerably broader range of mathematics in depth, a standards-based assessment needs to include short or elaborate *constructed-response* items, in which students must formulate and perhaps explain their answers. For this reason, standards-based assessments call for a shift away from reliance on multiple-choice and other *selected-response* items, for which the student chooses the correct answer from a list. Unfortunately, the term *open-ended* is sometimes used to describe constructed-response tasks. This can be confusing, because *open-ended* better describes one specific type of constructed-response task. Three types of constructed-response tasks are described as follows:

- Closed—these are usually short with one obvious solution path that leads to a single correct answer.
- Open-middle—these have more than one solution path, although they all lead to a single correct answer.
- Open-ended—these have multiple solution paths that lead to many different answers.

Closed and open-middle constructed-response tasks are useful in creating an examination that must be standardized across an entire state or district. Open-ended items are more useful for assessment that does not need to be entirely standardized across a whole cohort of students.

**Variety of student products.** This greater variety of task types brings in its wake a greater variety of student products. In traditional tests, student products usually conform to a single type: the selection of a correct response from among a number of choices. In contrast, in much broader and balanced standards-based assessments, students are asked to construct the correct responses to some tasks and to select the correct response to others.

Standards-based assessment calls for a greater variety of student products in order to give them the opportunity to show the full range of what they know and can do. In making this change, an assessor is embracing the challenge not only of assessing those aspects of mathematics that are *easy* to assess, but also of finding ways to assess those aspects of mathematics that are *essential* to assess.

**Balancing the instrument.** Besides these essential differences between task types and student products, there are important differences in the design and construction of standards-based assessments and traditional tests. Briefly, to construct a traditional norm-referenced test, a set of items is drawn from a large pool of multiple-choice items. Earlier field trials of this large pool allow a  $p$ -value to be assigned to each item. (An item's  $p$ -value is the proportion of students that select the correct response.) Items are then selected for a test such that the distribution of  $p$ -values approximates a normal distribution and so that the domain to be assessed is covered. In constructing standards-based assessments, such statistical analysis plays a lesser role. Instead, what is key is the selection of tasks that reflect the depth, range, and structure of mathematics represented in the student expectations (or standards) on which the assessment is based. Each task is developed to ensure that it has a reasonable score distribution across populations and that it does not unfairly disadvantage or advantage any one particular student group. Each task is then scored using a rubric—a scoring guide—referenced to an explicitly stated set of standards.

**Performance sampling.** These differences in design of standards-based assessments and traditional norm-referenced tests have implications for performance sampling—choosing a reasonable set of items from among the many that assess part of the domain to be tested. Virtually all assessments entail a sampling of student performance, unless the goals are few and are narrowly specified. Norm-referenced tests purport to cover the essential ingredients by using as many selected-response items as possible. A 50-minute test can include as many as 40 items because students are given only a short time to select a response to each multiple-choice item. Standards-based assessments, in contrast, seek to sample the essential mathematics by carefully selecting a balanced set of tasks. Students need considerably longer than a few minutes to construct a response to a worthwhile complex task,



and so a 50-minute assessment might be made up of only a few items. But a few complex tasks cannot sample the domain in the *same way* as a large number of multiple-choice items. A judicious way of handling the dilemma of performance sampling is to create an assessment that contains a combination of elaborate constructed-response tasks, selected-response items, and short constructed-response exercises. The specific combination of tasks that would then be placed on an examination would be selected to reflect the depth, range, structure, and balance of the standards to which the examination was referenced. The inclusion of a variety of task types can also allow assessment of aspects of mathematics, such as mathematical communication, that are difficult to assess with an exam that consists only of multiple-choice items, for example.

**Scoring as value judgments.** Another important difference between norm-referenced and standards-based assessments is the location and visibility of the value judgments necessarily associated with assessment. All assessment involves value judgments about what is important to assess and how important is each aspect. When a student's performance on an assessment task is evaluated, feedback can be given or a score can be assigned. Both of these involve value judgments, and in this sense, evaluation of student performances defines what counts as important. In standards-based assessments, these value judgments can be explicit and can be shared among teachers and students. In fact, making the necessary value judgments can become the responsibility of the professional peer group. In contrast, with norm-referenced tests, the necessary value judgments remain the responsibility of test designers because they are embedded in the design and choice of items and their distracters (the available incorrect responses), and are less visible to teachers and students.

**Score reporting.** Norm-referenced and standards-based assessments also differ in the kinds of data produced. Norm-referenced data report, often through a percentile score, how a student or group of students compare to a reference population. Standards-based assessments, on the other hand, report data that are referenced to some agreed-upon standard. The sets of scores that are generated by the task rubrics are aggregated to provide a total mathematics score and perhaps some sub-scores (subsets of the total score). When the assessment is standards-based, an individual's aggregate scores can be given meaning in terms of the standards. For example, non-overlapping ranges of aggregated scores can be designated as *exceeds the standard*, *meets the standard*, *nearly meets the standard*, or *far below the standard*. Such score classification is based on professional judgments, tasks, rubrics, and score distributions, and so is arrived at by a process that includes judgmental, standards-referenced, and normative elements.

## A hidden danger

If the function of standards-based assessments is to raise standards, care must be taken to ensure that they do not have a narrowing effect on curriculum and instruction (Schoenfeld, 1988). Indeed, the optimism surrounding the alignment of instruction and assessment is tainted by concern that when student learning expectations are defined by overly narrow assessments with consequences for teachers, students, or both, the impact on instruction and learning will not be uniformly positive (Romberg, Zarinnia, & Williams, 1990).

There is reason to worry that the pressure of consequences attached to high-stakes assessments will lead teachers and students to seek the most efficient way they can conceive of to reach their specified targets. Unfortunately, this often results in an enormous investment of classroom time in preparation for the test. For example, the *New York Times* recently carried this report:

The Bronx Division of High Schools asked Kaplan to train its teachers to help students tackle the state's new English Regents exam, which is being introduced in June and will become a condition for graduation next year. (Hartocollis, 1999, p. B1)

If the test is too narrow or omits important aspects of mathematical learning, there is little doubt that so much focus on test-taking strategies will create a learning environment antithetical to the wider educational goals that are envisaged in a standards-based system. If, however, teachers were asked about the most efficient strategies for preparing students for worthwhile and challenging assessments, it is quite likely they would stress that there is no alternative to *learning the mathematics that is specified*. It is this quality that standards-based assessments are intended to cultivate.

## Organization of this booklet

Chapter 2 outlines a model for standards-based assessment. The model incorporates elements that we feel are essential if the assessment is to enhance instruction and learning. The main thrust of this chapter is that assessments must be balanced. To be balanced, assessments must assess those aspects of mathematics that are important, rather than confining themselves to those aspects that might be easy to assess. The sense of balance that informs our model follows from standards and principles advanced by the Mathematical Sciences Education Board (MSEB) and NCTM (NCTM, 1989, 1995, 1998; NRC, 1989, 1993b).

The model presented in Chapter 2 was developed with the benefit of many years of assessment development experience of a

large number of professional designers and teachers. Other models have been developed and used, although many of the differences are only in the details. The process, the experience, and the research from which this model is derived are presented in Chapter 3.

Chapter 3 draws upon work with hundreds of teachers and thousands of students to construct a practical guide to assessment development. Here, the reader can come to understand the experience and research that has guided our process. This chapter analyzes and illustrates those aspects of task design that have been found to detract from providing students with real opportunities to perform. It also describes how the theory and practice of assessment development have reached new understandings of the interactions between students and assessment tasks, so that when seemingly good tasks fail to produce good results in field trials and the source of the failure is extraneous to the mathematics, the task may often be revised in ways that maintain the important mathematical ideas that the task was intended to assess. When students' performances on assessment tasks are scored, the results lead to inferences about what they do or do not know. Essentially, this chapter deals with issues that are relevant to the validity of these inferences.

In Chapter 4, the focus shifts away from the task on its drawing board and toward the task as seen in the social milieu of the classroom. Here, the main focus is on issues that influence opportunity to learn. The chapter illustrates just how difficult students find non-routine assessment tasks when they do not work on them regularly in class. Students often make tasks more difficult when they try to force the use of the mathematics that they have learned most recently. And students' own evaluations of assessment tasks often indicate that they have very clearly defined notions about what counts as appropriate behavior for the mathematics classroom.

Chapter 4 also covers two central problems teachers confront:

- how little mathematics many of their students seem to know, and,
- how much content they, as teachers, are expected to cover.

These two pressures have led many teachers to adopt various coping strategies, even though such strategies are considered to be far from perfect. For example, their students' lack of mathematical knowledge has led some teachers to concentrate on basic skills rather than broaden the aspects of mathematics that they teach in their classrooms. The pressure to cover an exceedingly large amount of material also has led many teachers to make less than satisfactory choices about their curriculum. Paul Black and Dylan Wiliam's recent article *Inside the Black Box—Raising Stan-*

*Standards Through Classroom Assessment* (1998) is used to identify the additional support that teachers must be given if they are to create opportunities for students to learn mathematics in a balanced way. Black and Wiliam urge those interested in raising standards to open up the black box of the classroom. They argue persuasively for the efficacy of formative assessment in raising standards. If formative assessment is to be effective, Black and Wiliam note, it must be integrated into rather than bolted on top of current instructional practices.

Finally, Chapter 5 addresses the issue of aligning instruction and assessment with standards. The presentation here is substantially informed by the recent and comprehensive monograph *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). Webb leaves little doubt that, for assessment to be effective, instruction must be aligned with standards.

## Chapter 2: A model for assessment development: Achieving balance

---

The *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), indicated a clear need for new mathematics assessments that would embrace mathematical goals that had not been considered in traditional norm-referenced tests. Because it is tempting, in any process of change, to concentrate solely on aspects that were previously ignored, the main theme of this chapter is achieving *balance* in assessing various aspects of mathematical knowledge, including skills that were adequately assessed by traditional tests as well as newer goals such as problem solving and mathematical communication. Chapter 1 describes many of the differences between traditional norm-referenced tests and standards-based assessments, but such descriptions are not sufficient for building a balanced assessment, and they provide too little guidance for evaluating the extent to which an assessment fits with a set of standards. There are a variety of ways to achieve balance. This chapter presents one way that is based upon a distinction among mathematical skills, conceptual understanding, and problem solving, but that remains quite flexible in its interpretation and use. The chapter begins by describing several different formulations of assessment principles that collectively frame the construction of a “model for balance” to guide assessment development work. The chapter continues with descriptions of several types of assessment, according to their origin and use, followed by a detailed description of the model. The chapter closes with recommendations for using the model to evaluate the balance of an assessment system.

## Conceptual frameworks for mathematics assessment

Encouraged by the publication of the NCTM *Standards*, several groups began defining principles that could guide the development of balanced pictures of mathematical accomplishment.

In *Measuring What Counts*, the MSEB put forward three fundamental assessment principles that would support effective learning of mathematics (NRC, 1993b):

- **The Content Principle**—assessment should reflect the mathematics that is most important for students to learn.
- **The Learning Principle**—assessment should enhance mathematics learning and support good instructional practice.
- **The Equity Principle**—assessment should support every student's opportunity to learn important mathematics.

At the same time, the New Standards and Balanced Assessment organizations produced a framework for balance, based on NCTM's *Standards*. The seven principal dimensions of the resulting framework are outlined below (Schoenfeld, Burkhardt, Daro, & Stanley, 1993):

- **Content**—assessment should reflect content in a broad sense and include concepts, senses, procedures and techniques, representations, and connections.
- **Thinking processes**—assessment should engage students in a wide range of thinking processes that include conjecturing, organizing, explaining, investigating, formulating, and planning.
- **Student products**—assessment should require a variety of student products that include models, plans, and reports.
- **Mathematical point of view**—assessment should present mathematics as an interconnected body of knowledge, by engaging students in mathematics that is connected to realistic, illustrative, and pure contexts.
- **Diversity**—assessment should be sensitive to issues of access.
- **Circumstances of performance**—assessment should vary according to time allocated, whether it is performed individually, in pairs, or in groups, and whether there is opportunity for feedback and revision.
- **Pedagogy and aesthetics**—assessment tasks should be engaging, believable, and understandable, and should not disenfranchise the common sense of the student.

Two years later, the NCTM *Assessment Standards for School Mathematics* (1995) produced a set of standards for assessment

that incorporated MSEB’s assessment principles. NCTM’s six standards for assessment are listed below:

- **Mathematics**—assessment should reflect the mathematics that all students need to know and be able to do. The content of assessment must be shaped by important mathematics that is broad and balanced.
- **Learning**—assessment should enhance mathematics learning. Assessment tasks should offer students an opportunity to learn important mathematics.
- **Equity**—assessment should promote equity. All students should have the opportunity to learn the mathematics that is to be assessed. And all students should have an equal opportunity to show what they know and can do on assessments.
- **Openness**—assessment should be an open process. Students should either be able to do what the standards are asking them to do or know how far they are from meeting the standards and understand what they need to do to close the gap.
- **Inferences**—assessment should promote valid inferences. Assessment allows us to make more or less valid inferences about what students know and can do. It is more difficult to make valid inferences about those aspects that students do not appear to know.
- **Coherence**—assessment should be a coherent process. The cohesion of assessments should reflect the cohesion of instruction and teaching.

The frameworks produced by these organizations are complementary rather than competing, and so there is a great deal of overlap among the three. The linkages between the details of the three frameworks are illustrated by the diagram in Table 1.

Together, these three conceptual frameworks suggest the development of a broad and balanced approach to assessment. The challenge for task designers, therefore, becomes that of translating what is being advanced by these conceptual frameworks into worthwhile assessment opportunities for all students.

### Types of assessment

An assessment that carries out these visions for balance must take a broad view of assessment, encompassing various types of student products and circumstances of performance. For the purposes of this document, school mathematics assessment will be described by two types: on-demand assessment and classroom-embedded assessment.

<b>Table 1. Comparison of Principles</b>		
<b>MSEB Principles</b>	<b>BA/NS Dimensions</b>	<b>NCTM Standards</b>
Content	Content	Mathematics
Learning	Thinking processes Student products Mathematical point of view Pedagogics and aesthetics	Learning
Equity	Diversity Circumstances of performance	Equity Openness Inferences Coherence

*On-demand assessment* refers to testing that is completed in time-limited conditions. Some on-demand assessments are standardized in order to see how different students perform under the same conditions. A truly standardized test, however, is impossible to create—there will always be aspects of a test that will advantage or disadvantage some but not all students. On-demand norm-referenced tests are used for the purpose of sorting or ranking students, but on-demand assessments can be standardized and used to serve other purposes. For example, standards-based on-demand examinations can be standardized (as far as possible) and used to determine whether students meet a set of standards.

On-demand assessments also are constructed by teachers to assess the taught curriculum. A teacher-constructed examination can be standards-based. In this case, the standards to which the examination is referenced are the standards of the individual teacher (or group of teachers teaching the same course). These individual standards may or may not coincide with publicly negotiated standards.

*Classroom-embedded assessment* refers to assessment that teachers use in their daily work to assess what students know and can do. This type of assessment is useful in enabling teachers to understand how students are progressing and to design or adjust on-going instructional strategies. Embedded assessment also can be used as a component of portfolio assessment. In Vermont, for example, embedded assessments are used to compile student portfolios, and these contribute to the statewide assessments in mathematics. One of the strengths of embedded assessment is that it can bridge the gap between instruction and assessment.

In this document the assessment issues that are discussed are applicable to each of these assessment arenas. Chapter 3 concentrates on task design issues that have been identified as part of the



task development efforts conducted by Balanced Assessment and New Standards. The rest of this chapter examines the conditions that need to be present for an assessment system, test, or examination to be balanced. What follows here is an overall model for mathematics assessment.

### **A model for a balanced assessment in mathematics**

The first question that must be asked of an assessment system is, “What mathematics do we want to assess?” The answer will depend on the mathematics that is specified in a state’s or district’s standards. But this question also needs to be answered in a way that is specific enough to enable teachers to provide opportunities for students to learn the mathematics that is to be assessed. The document *Every Child Mathematically Proficient: An Action Plan* provides an outline of what students should be able to do by the end of the ninth grade (Learning First Alliance, 1998, p. 11-12). Many other outlines also are feasible, but this one provides a useful starting point for developing the content specifications of an examination.

The next question that must be asked of an assessment system is, “What do we want students to be able to do with the mathematics described in the content specifications?” The *New Standards Performance Standards* ask students to show that they have a repertoire of mathematical skills, that they understand mathematical concepts, that they can use mathematics to solve problems, and that they can put mathematics to work by completing extended projects. In the *New Standards Reference Examinations*, tasks are designed with the following categories in mind:

- Mathematical Skills
- Conceptual Understanding
- Mathematical Problem Solving

These three categories are comparable to the mathematical abilities categories used by the National Assessment of Educational Progress [NAEP] in their assessment of mathematics: procedural knowledge, conceptual understanding, and problem solving (National Assessment Governing Board, 1995).

Because many tasks require a combination of skills, concepts, and problem-solving strategies, classification under these headings often requires identification of the *task hurdle*: the primary mathematical skill, concept, or strategy that the student must employ in order to demonstrate some success on the task. Still, the boundaries that separate these categories are not well defined (nor do they need to be). The utility of these categories lies not in their precision, but in their role in developing a balanced view of

mathematics assessment. Important here is the realization that to de-emphasize one of these aspects of assessment would be to promote a distorted and impoverished view of school mathematics. The domain need not be organized in this precise way, but these three categories do provide a useful model.

### Mathematical skills as an assessment target

When the assessment target is mathematical skills, tasks that assess students' knowledge of important facts, routines, or algorithms are needed. Usually, mathematical skills tasks can be solved by recalling an important idea or well-practiced routine.

The task *Right Triangle* is designed to assess whether students can find the length of the hypotenuse of a right triangle. The hurdle in *Right Triangle* is procedural. Students will typically use the Pythagorean Theorem to find the length of the hypotenuse in a right triangle, a procedure that all students who have been taught the Pythagorean Theorem will have learned. The choice of numbers and the absence of an elaborate context helps ensure that the conceptual demand of the task is kept to a minimum.

The task *Find the Volume* is also a straightforward task that measures skills. To find the volume, the student must know (or know how to find) and use the formula for the volume of a rectangular prism. The task is not designed to reveal what the student *understands* about volume. A student solving this task might know little about volume, beyond the formula.

Mathematical skill tasks make little or no conceptual or strategic demands on students. This approach allows assessment of students' factual and procedural capabilities without confounding them with other capabilities.

### Conceptual understanding as an assessment target

When tasks focus on conceptual understanding, they require students to use an idea, reformulate it and express it in their own terms. Tasks that call only for students to apply well-practiced routines or algorithms are not sufficient for the purpose of assessing conceptual understanding. Conceptual understanding tasks require that students represent, use, or explain a concept.

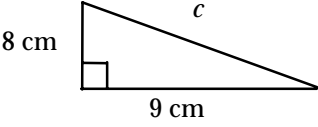
Figure 1. Right Triangle
 <p>The diagram is not drawn to scale. What is the length of <math>c</math>? Give your answer to the nearest mm.</p>

Figure 2. Find the Volume
<p>Measurements taken from a box give the length, width, and height to be 36, 14 and 16 inches.</p> <p>Find the volume of the box.</p>

*Figures 1 and 2 reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).*

Good conceptual understanding tasks should not be solvable from an understanding of mathematics that is inherently fragile and composed of decontextualized or fragmented slivers of mathematical knowledge—no matter how well learned and remembered.

### Figure 3. Almost Right

The sides of a triangle have lengths 8, 9, and 12.  
One of the angles in this triangle is either a right angle or close to it.  
Decide if this angle is exactly a right angle, a little larger, or a little smaller.  
Give reasons to support your decision.

*Reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).*

The task *Almost Right* is presented as an assessment of conceptual understanding. In responding to *Almost Right*, students are to decide whether an angle that is very close to 90 degrees is right, obtuse, or acute. In our trials, the majority of students used the converse of the Pythagorean Theorem to reason that the triangle is not a right triangle because  $64 + 81 \neq 144$ . Then, in order to decide whether the angle is obtuse or acute, students use the fact that  $64 + 81 > 144$ , and some make the correct inference that the angle is acute. Clearly, no procedural

use of the Pythagorean Theorem will facilitate this level of fluency with the concepts; as a consequence, *Almost Right* is classified as an assessment of conceptual understanding.

The task *Gutter* (Figure 4) also assesses conceptual understanding. The conceptual hurdle to be navigated here is that of expressing the radius of the semi-circular base in terms of  $W$ . This can be accomplished by setting up the equation  $\pi r = W$  and then solving for  $r$ . This task is not wholly free of procedure. To arrive at the solution, a student must manipulate this equation correctly and then carry out the required substitutions. Nevertheless, the symbolic representation required far outweighs the symbolic manipulation that is required. And it is negotiating the hurdles posed by representation and re-expression in this task that indicates fluency in understanding of the relationship between radius and circumference. Clearly, this is a task that cannot be solved through algorithmic manipulations, but it may be solved by students who have had the opportunity to work with and reflect on the relationships between diameter and circumference and between the area of the base and the volume of a solid that has regular cross-sections.

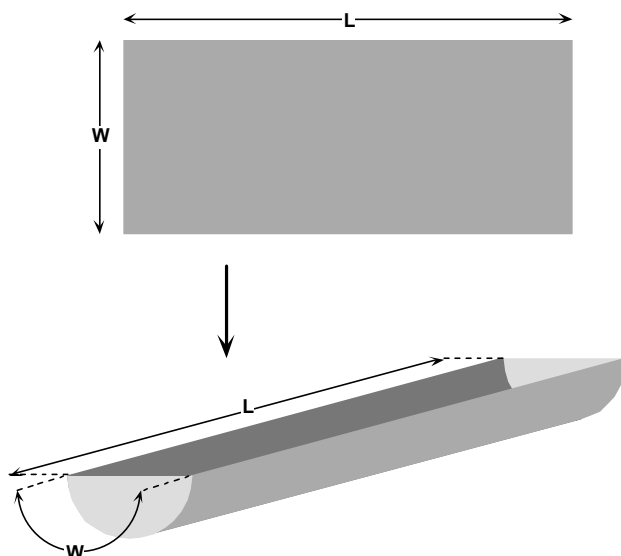
Conceptual understanding tasks make little or no procedural or strategic demands on students. This approach allows assessment of students' conceptual understandings without confounding them with other capabilities.

### Problem solving as an assessment target

When mathematical problem solving is to be assessed, tasks must make these requirements of students:

**Figure 4. Gutter**

A rectangular sheet of metal with length  $L$  and width  $W$  is rolled up into a gutter with semi-circular cross section. End pieces are added so that it will hold water.



Show step by step how to develop a formula for the volume of water in the gutter when it is full.

Use the symbols  $L$  and  $W$  in your formula.

*Reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).*

- formulate an approach to a problem;
- select the mathematical procedures, concepts, and strategies necessary, and then deploy these when implementing a solution; and
- draw a conclusion.

If students are to formulate an approach, the task should not provide too much structure or directive instruction (either explicitly or implicitly). This is because if a task contains a large amount of structure, this structure will provide an approach for the students. Chapter 3 provides examples of tasks whose structure dictates a particular approach to the task, even though the problem may be solved in a variety of ways.

If students are to be required to select procedures, facts, and concepts, and then use these to solve a problem, care must be taken to ensure that the skills and concepts that students are asked to use are ones that they have already fully absorbed. When a task

requires students to try to make high-level use of unassimilated skills and concepts, they are often unable to make any headway. This is usually because the strategic use of unabsorbed skills and concepts causes the total cognitive demand of the task to be unreasonably high. Such assessment situations can be demoralizing for students, can lead to false-negative inferences, and are usually a waste of assessment time.

Tasks that work well to assess problem solving in an on-demand situation are ones that ask students to bring together skills and concepts that they understand well. For example, when the examination is to be designed for the tenth grade, the best tasks to use to contribute to the problem-solving score are those that draw largely on the skills and concepts that students will have learned in ninth and even eighth grade. Of course, when the assessment of problem solving is embedded in classroom practice and not strictly time-limited, much more complex tasks can be administered. In classroom situations (and when the stakes are not high), these more complex problem-solving tasks can be used to help students learn and assimilate skills and concepts.

Choosing problem-solving tasks that require students to make high-level use of skills and concepts that they have learned one or even two years previously does not mean that standards are necessarily going to be lowered. On the contrary, grade-level appropriate assessment of skills and concepts can be assessed by tasks that are specifically designed to do just that, and so the standards of technical skill and conceptual understanding can be protected.

The next task, *Snark Soda* (Figure 5), is an example of a longer task that assesses problem solving. This problem-solving task contains few or no directive steps. Also in *Snark Soda*, the skills and concepts that the student is to draw on are likely to be well understood by high school students. *Snark Soda* exemplifies well what we mean by high level use of well-assimilated concepts, facts, and skills.

To be successful on *Snark Soda*, the student must be able to identify and manage these three main components:

- *fitting shapes* of solid geometry such as cylinders and hemispheres to the different parts of the given bottle;
- *measuring the diagram* (described as an accurate and full-size drawing) to find values for parameters such as the radius and height for these shapes;
- *computing the volume* of these shapes, using the measured values.

**Figure 5. Snark Soda**

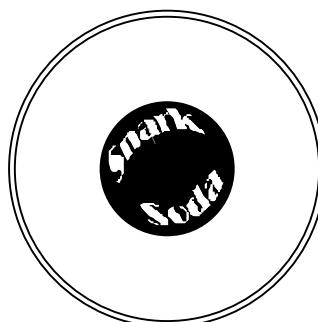
*In this task, use geometric shapes to model as closely as you can the volume of the liquid in the bottle shown below.*

The pictures show top and side views of the bottle. They are accurate and drawn full size.

FRONT VIEW



TOP VIEW



1. Figure out as accurately as you can a good approximation for the volume of the liquid in the bottle.

*Use a ruler to measure the bottle.*

2. Discuss the accuracy of your model by talking about where it gives over-estimates or underestimates.

You will be assessed on:

- how clearly you showed what you did using diagrams, formulas, and words;
- how easy it is for someone who was not in your classroom today to repeat what you did and check your approximation for the volume.

*Reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).*

Problem-solving tasks should make medium-level procedural and conceptual demands on students. Worthwhile problem-solving tasks can assess the way in which students use skills and concepts that they have fully absorbed.

### **Why develop different types of task?**

One type of task is not enough to give a broad and balanced vision of mathematics. Until recently, the only widely available tests were various norm-referenced tests that were characterized by short, closed, multiple-choice items that assessed procedural skills, at the expense of *any* assessment of conceptual understanding or problem solving. Early attempts to address this

imbalance were characterized by the development of assessment tasks that emphasized contextual and mathematical connections or that reflected the richness, elegance, and beauty of mathematics. Once again, a single task type emerged. These assessment tasks were often complex constructed-response tasks that focused simultaneously on aspects of conceptual, procedural, and strategic knowledge. As will be explained further in Chapter 3, however, when skills and concepts are assessed through problem-solving tasks, there is the risk of making false-negative errors about students' procedural and conceptual knowledge. Frequently, students' responses to problem-solving tasks can lead to the inference that they lack proficiency in basic skills and concepts, even though further analysis will reveal that students understand the concepts. Apparently, students are often unable to use procedures and concepts in situations where the strategic demands of the task are high.

Separate procedural, conceptual, and problem-solving tasks provide a balanced yet unconfounded assessment of these important capabilities. The balance of an assessment is therefore found in the assessment as a whole and not in the individual tasks that make up the assessment.

### **One content area, three different tasks**

The tasks *Find the Volume*, *Gutter*, and *Snark Soda* are presented here to illustrate assessment tasks that measure mathematical skill, conceptual understanding, and mathematical problem solving, respectively. They also illustrate how differently structured tasks that all focus on the same content area can be used to measure procedural, conceptual, and strategic aspects of mathematical learning.

If a test or classroom practice confines itself to just one or two of these aspects of mathematical learning, then the test or classroom practice will not be sufficiently balanced because it will not give students the opportunity to show what they can do across all three; and all three are important. Phil Daro, Executive Director of New Standards, makes this point by using a three-legged stool as a metaphor for the curriculum (personal communication, March, 1999). One leg represents skills, a second conceptual understanding, and a third problem solving. If one of the legs were missing, the stool would clearly fall in the direction of the missing leg. Analogously, if one of these three critical curricular dimensions is missing, then the curriculum will fall in the direction of the missing dimension.

If, however, the test or classroom practice ensures that all three aspects are addressed in a way that is accessible to students

and enables students to show what they can do across all three, then that test or classroom practice has gone a substantial way toward balance.

### **The same task might measure different aspects of mathematics**

Thus far the discussion has been about balance as a property of a test and about task categorization—as problem solving, conceptual understanding, or skills—based on properties of the task. But as pointed out earlier, the boundaries between these categories are not precise. Furthermore, the classification depends on the mathematics level of the students. In other words, what a task is assessing must be evaluated in relation to the group for whom the task is intended (Wilson, Fernandez, & Hadaway, 1993; Schoenfeld, 1985). Problem-solving tasks, for example, are often identified as tasks that are non-routine. But what is non-routine to one student might be routine for another. The same holds for the classification of tasks that are designed to assess conceptual understanding. For example, the task *Almost Right* was designed to assess conceptual understanding. If this task were administered to a group of students who were accustomed to using tasks of this type in class to practice applying the Cosine Law, then it would assess skills rather than conceptual understanding.

### **Maintaining skills while achieving depth and balance**

Many traditional mathematical skills remain essential and need to be maintained, rather than ignored, in any effort to promote conceptual understanding and problem solving. Using different tasks for each of these assessment objectives makes it possible to report separate scores for mathematical skills, conceptual understanding, and mathematical problem solving. Such a score report makes visible the importance of each of these capabilities in a broad and balanced assessment. In addition, this visible balance reassures the public that skills remain critical and, as Massell, Kirst, and Hoppe (1997) argue, aids in gaining public support for the current attempt to encourage a broader view of mathematics education than that which is associated with the traditional curriculum.

Curriculum development benefits are also afforded by reporting separate scores for separate learning objectives. When an examination is designed to report scores for important aspects of mathematical learning according to well-chosen categories, the information that teachers, supervisors, and administrators receive is far more specific than the information that might be gained from a single mathematics score. For example, when states or districts use the *New Standards Reference Examination*, the information that is provided in the individual, class, school, district, and state



score reports is organized according to three reporting categories: Mathematical Skills, Conceptual Understanding, and Mathematical Problem Solving. This means that a teacher or other interested party can use these sub-scores to determine how extra support might best be allocated. For example, if students' mathematical skills scores indicate that a large proportion of students has met the standards, while mathematical problem-solving scores indicate that a large proportion of students has not met the problem-solving standards, this would indicate that additional classroom instructional efforts may need to be directed at creating opportunities for the students to develop greater problem-solving capabilities.

### **The whole is greater than the sum of its parts**

It is important to stress that procedural, conceptual, and problem-solving knowledge are not mutually exclusive constructs. For example, mathematical problem-solving tasks should not be constructed so that they are devoid of mathematical skills and concepts. Problem-solving tasks meriting placement in a balanced mathematics assessment *will* incorporate skills and concepts, but these skills and concepts *should not* constitute a major hurdle of the task. By the same token, neither skills nor conceptual understanding tasks will totally lack a formulation component: Every task, even those that are simply cast, will require some amount of figuring out what to do. Nonetheless, when the desired measurement target is either skills or conceptual understanding, strategy formulation should not constitute a significant hurdle.

On the basis of the model that is described so far, we would not wish to advance a model of assessment that is in the form of a two-dimensional array—with mathematical skills, conceptual understanding, and mathematical problem solving along one axis and topics such as number, measurement, geometry, algebra, statistics, and probability along the other. To do so would be to neglect at least two other dimensions: mathematical connections and mathematical communication, both of which are of critical importance because they are both strong indicators of mathematical power.

### **Assessing mathematical connections**

A balanced assessment must provide evidence about whether students are able to use mathematical skills and concepts as they are connected within mathematics or to some real-world context. There are three important kinds of mathematical connections:

- Concept-concept connections
- Concept-context connections
- Concept-representation connections

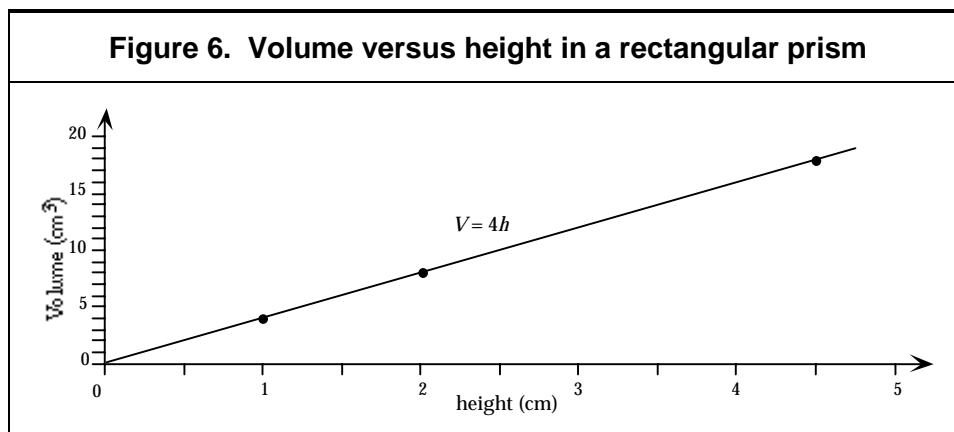
**Concept-concept connections** are often referred to as connections *within* mathematics. For a student to make concept-concept connections, the student will need to experience and to understand the rich interplay among mathematical ideas, and to begin to see mathematics as an integrated whole. An emphasis on concept-concept connections through instruction is important in enabling students to learn concepts and to learn about the interconnectedness of mathematics as a discipline. Certainly, an instructional emphasis on concept-concept connections between and within grades is a prerequisite for success on the kinds of problem-solving tasks that require students to make high-level use of concepts and skills that they have learned in previous grades. Achieving success on such problem-solving tasks will help students to absorb prior knowledge more fully, strengthen their understandings, and help them to accommodate new ideas.

Examples of types of tasks that capture the spirit of the concept-concept connections are provided in the Core Assignments developed recently by the National Center on Education and the Economy (NCEE). In the task *Volume of Sand in a Rectangular Prism* (adapted from NCEE's *Core Assignment: Volume*, 1998) students are given three congruent rectangular prisms that are filled to various heights with sand and are asked to plot the volume of sand in the container as a function of the height of the sand. Next, students are asked to address the following questions or discussion points:

- What kind of a function is it? Say why!
- What is the slope of the line?
- How can the slope be interpreted in terms of the sand in the containers?
- What is the equation of the line?
- How does this equation relate to the volume formula of rectangular prisms?
- Suppose the container were turned so that one of its sides became the base, how would the graph of the function change?

In this task, conceptual understanding about rectangular prisms and their volume is connected to concepts of linear function and slope. Students must graph volume as a function of the changing height of sand in a rectangular prism. Then they must find and interpret the slope of the graph in terms of the area of the base of a rectangular prism.

Students and teachers alike have found this task quite difficult. Although both students and teachers had previously demonstrated that they could find the slope of a straight line, many were initially unsuccessful doing so in the context of this task. Once they had



calculated the slope of the line, some had difficulty attaching meaning to it. According to some teachers, their difficulties with this task reflected their lack of experience working with the concepts of linear function and volume at the same time.

As a worthwhile learning opportunity for students, teachers could present a graph representing the volume of sand in a container as a function of height. Figure 6 serves as an example for such an approach.

Students can see that the graph is a straight line with equation  $V = 4h$ , and they are asked to use the graph to address the following questions:

- What can be said about the shape of the container that holds the sand?
- What can be said about the shape of the base?
- What can be said about the size of the base?
- What can be said about the shape of the container?

Instead of using mathematical connections in working from the physical structure to its mathematical representation, students are asked in this exercise to use the mathematical representation to make inferences about the physical structure. Both teachers and students have reported that this approach has helped them to better understand how the slope of the graph is related to the area of the base of the container.

**Concept-context connections** are made when mathematics is extracted from a context outside of mathematics. Such contexts can deepen students' understanding of important mathematics (NCTM, 1989; NRC, 1993b, 1998). When students are given the opportunity to use conceptual and procedural knowledge in contexts outside of mathematics, they also are given the opportunity to strengthen their existing understandings and hone their mathematical power.

*Measuring Up* (NRC, 1993a) and *High School Mathematics at Work* (NRC, 1998) are good sources for high-quality, complex tasks that are mathematically rich and contextually relevant and that can be used as either instructional or assessment tasks. (If used for an assessment however, most of these extended tasks are more appropriate for assessment that is spread out over several days rather than examinations that are time-limited.) *Snark Soda* is one example of a task that makes use of contexts outside mathematics and that could be used in an on-demand assessment. Chapter 3 contains several tasks in which the essential mathematics is contextualized: Students must model shopping carts or paper cups, extract mathematics from a forester's diameter tape, deal with physical constraints, or grapple with issues of percent increase and decrease. Each of these tasks presents enormous challenges for large numbers of students in middle and high school. Many teachers are surprised by the lack of mathematical power that is evident when students attempt to use basic mathematics to solve problems set in a context. These inadequacies cannot be addressed if students are not encouraged to connect mathematics to the world around them.

**Concept-representation connections** are those that give students the opportunity to translate among different representations, such as between a formula and a graph. Research suggests that forms of representations need not be taught as ends in themselves but can be useful both for achieving understanding and for communicating that understanding (Greeno & Hall, 1997).

The crucial role that mathematical connections play in instruction and learning does not, however, mean that every assessment task should make connections either within or outside mathematics. What is necessary is for the entire set of tasks in an assessment package or test to be balanced with respect to mathematical connections. Certainly, if a test were devoid of mathematical connections, there would be less incentive for teachers to make mathematical connections a part of their regular classroom practice. Traditional norm-referenced tests generally do not make use of mathematical connections in the way that we are advocating here. As a consequence, traditional mathematics curricula also have failed to place much emphasis on mathematical connections. If assessment practice continues to ignore mathematical connections, important opportunities to enhance mathematics learning and to support good instructional practice will be lost (NRC, 1993b).

### **Assessing mathematical communication**

In contrast to mathematical connections, all assessment tasks require at least some communication. In the case of selected-

response or short, closed, procedural tasks, communication is frequently trivial. As tasks become more complex, the communication requirements become more significant.

In the world beyond school, it is often important to communicate to an outside audience, such as a boss, a client, a politician, or a friend. Thus, students should be able to communicate about mathematical ideas by describing mathematical concepts and explaining reasoning and results. When developing mathematical communication, students should use the language of mathematics, its symbols, notation, graphs, and expressions, to communicate through reading, writing, speaking, and listening. The power of mathematical communication lies in its capacity to foster deeper understanding of mathematics (Cobb & Lampert, 1998; NCTM, 1998; Zucker & Esty, 1993). It is incumbent upon assessments, therefore, to incorporate a broad and balanced concept of mathematical communication. Similarly, it is important that classroom instruction provide the opportunity for students to communicate mathematical ideas. If students do not communicate mathematical ideas regularly, it is unlikely that they will know what to do when this is required in an assessment.

### **Circumstances of performance**

The dimensions of balance that are described here cannot be addressed adequately in an examination or test—where students are expected to work individually, with few if any resources, and in a strictly time-limited situation. On-demand examinations and tests simply cannot assess problem solving, mathematical connections, or mathematical communication adequately (Arcavi, Kessel, Meira, & Smith, 1998). Also, on-demand examinations cannot assess students' capabilities in problem-posing, careful revision of argument, extended work, presentation, or organization of material from outside sources. For these reasons, it will almost always be necessary to vary the circumstances under which assessments are performed if the goal is to assess everything that is specified in the standards (Webb, 1997). These variations in circumstances can include:

- On-demand tests or examinations created by the teacher or by an external body.
- Classroom-embedded assessment that is completed by the student as part of the day-to-day requirements of a mathematics course. Some components of this can be completed in collaboration with peers, and other portions should be completed individually.
- Long-term projects and investigations, with opportunity for feedback and revision. Such projects provide opportunity for

students to put mathematics to work in an extended way, where they have opportunity to use resources around them, and where problem solving is more like problem solving in the real world.

These variations in circumstances of performance will yield a variety of student products, including open-ended constructed-response items. The variety can be greater with the realization that constructed responses do not need to be written, especially for long-term projects. For example, a constructed response might be spoken or built. The product could even be a video or part of a group assessment.

### **Recommendations for evaluating the balance of an assessment system**

Many different schemes can be used to organize and evaluate an assessment program. One model is to construct charts as templates that might be used by an assessment specialist when evaluating the balance of a school, district, or state assessment system. Although in some schemes, such charts might focus on aspects of learning such as thinking processes, the charts that follow build from the ideas presented earlier in the chapter on circumstances of performance, on-demand assessment, and mathematical connections.

Chart 1 helps develop an overview of the balance of the key aspects of the standards or learning expectations that are represented in the assessment system. It is important to work from expectations to assessments, rather than from assessments to expectations, because this will help to ensure that assessment does not rely solely on those aspects that are easy to assess and neglect those desirable aspects that are intrinsically more difficult to assess.

The NCTM's 9-12 standard on functions, for example, lists expectations that students can

- ❑ model real-world phenomena with a variety of functions;
- ❑ represent and analyze relationships using tables, verbal rules, equations, and graphs;
- ❑ translate among tabular, symbolic, and graphical representations of functions;
- ❑ recognize that a variety of problem situations can be modeled by the same type of function;
- ❑ analyze the effects of parameter changes on the graphs of functions. (NCTM, 1989, p. 154)

<b>Chart 1. Circumstances of Performance</b>			
	On-demand Assessment	Embedded Assessment	Long-term projects and investigations
Skills	Translate ... Represent ...	Translate ...	
Conceptual understanding	Analyze ...	Recognize ...	
Problem solving and communication	Model ...		Model ...
Mathematical connections		Translate ...	

The teacher or assessment coordinator might decide to distribute these as shown in Chart 1. Note that some expectations appear more than once in the chart. It is critical, of course, that all expectations appear at least once.

This chart gives the assessment coordinator the opportunity to flag those critical dimensions that cannot be assessed in an examination. For example, Chart 1 provides the opportunity to map content expectations onto the assessment system, allowing some of those expectations to be developed through extended work with feedback and revision, oral presentations, and the construction of responses that are not necessarily written.

Chart 2 is to be used for focusing on the on-demand components of an assessment system. Illustrative entries are chosen from among the tasks discussed in this chapter. The column headings in Chart 2 might be general content strands, such as geometry or algebra or specific topics within a standard, depending on the scope of the assessment. It is very difficult to do justice to the assessment of problem solving in a test or examination. However, problem solving should not be dropped altogether from tests or examinations. Although examinations may be a far from perfect place to measure problem solving, putting problem solving in the assessment increases the likelihood that it will be addressed in the classroom (Webb, 1997).

The amount of time needed for a specific on-demand assessment is largely controlled by the amount of time needed to assess each of the different learning expectations. Groups of tasks that take about two to five minutes each to complete can provide a good assessment of procedural or conceptual knowledge. Assessment of problem solving requires the use of tasks that take considerably longer to complete. The minimum amount of time that can reasonably be allocated to a problem-solving task is about ten

<b>Chart 2. On-Demand Assessment Components</b>				
	<b>Geometry</b>	<i>Other topics or content strands ...</i>		
Skills	Find the Volume			
Conceptual understanding	Almost Right			
Problem solving and communication	Gutter			

minutes. But fifteen-minute tasks are long enough to give students the opportunity to show how they can carry their work through to a solution, and yet short enough to ensure that not too much assessment time is wasted if the student is unable to formulate any approach to a particular task. The smallest number of tasks that might contribute to a reliable problem-solving score is about six. Therefore, it is reasonable to allot 90 minutes to an on-demand assessment of problem solving. The minimum amount of time that might be spent assessing procedural and conceptual knowledge in an on-demand context is approximately 75 minutes. The *New Standards Reference Examination* accomplishes this in three sittings of about 55 minutes each.

On-demand examinations scores do not need to be based upon data obtained from traditional closed-book responses, prepared by students without access to any external resources that might be useful for solving mathematical problems. Examination scores could, as an alternative, be made up of two scores—one that is derived from the on-demand component, and one that is derived from coursework; that is, work that is generated as a course requirement, and could include a range of responses that are written, oral, video, or built.

Chart 3 can be used to trace the path of mathematical connections throughout the entire assessment system. The emphasis given to each cell would depend on the mathematical connections that are emphasized in the learning expectations. The development of mathematical connections is vital to an enhanced understanding of mathematics. Absent the development of such connections, mathematics remains fragmented, cluttered, and decontextualized.

The recommendations in this chapter can be used to address the development of a single test or a complete assessment program. Chapter 3, addresses specific task development issues. The research and experience presented in Chapter 3 will provide additional support for the assessment model and recommendations presented in this chapter.



Chart 3. Mathematical Connections				
		Concept- Concept Connections	Concept- Context Connections	Concept- Representation Connections
<b>On-demand</b>	Skills			
	Conceptual understanding	Volume of Sand (linearity)		
	Problem solving and communication		Snark Soda (measurement)	
Embedded assessment				
Long-term projects and investigations				

### Chapter 3: Assessment and opportunity to perform

---

This chapter focuses on issues that arose during the New Standards' task development process, undertaken to build balanced assessments. The research and task development experience discussed here also offers additional insight about the foundations of the model for assessment that is presented in Chapter 2. The target audience for the assessments included both students who have encountered a standards-based curriculum as well as students in classrooms with a traditional curriculum. As part of the development process, information about instructional experiences was gathered so that results could be interpreted meaningfully and defensibly for both groups of students. In presenting several task development case studies, including analysis of some notable *failures* as well as successes, more information is provided to help the reader see why the model for a balanced assessment is defined as it is. Although the examples in this chapter are drawn from work with high school students, the ideas apply across grade levels.

One of the most striking aspects of task development is just how hard students find many tasks that are designed to assess conceptual understanding or problem solving. Time and again when tasks were piloted in classrooms—tasks that appeared to provide students the opportunity to show what they know—the tasks were for some reason inaccessible for most students. One explanation for this result is that many of the tasks may not closely resemble those that students are accustomed to completing in class.

In light of the evidence generated in the classroom, we had to make choices about how to proceed. One option, for example, was to declare that such tasks are too ambitious and to abandon them in favor of assessment tasks similar to those that students are accustomed to completing in class and for homework. Because the goal of the project was to produce tasks and assessments that would enhance instruction and student learning, we decided instead to advance the craft of task development sufficiently to provide students access to what had been previously inaccessible tasks.

Meeting that challenge required looking closely at the students' performances and attempting to determine what was making the tasks so difficult. Two broad themes emerged. First, students are sometimes not given sufficient *opportunity to perform*, by which we mean some aspects of a task prevent students from showing what they have learned. Opportunity to perform is a primary focus of this chapter. Second, students are sometimes not given sufficient *opportunity to learn*, by which we mean the students' classroom experiences have not left them well equipped to succeed on certain kinds of tasks. Opportunity-to-learn issues are addressed in Chapter 4.

Of course, opportunity-to-perform and opportunity-to-learn issues are inextricably linked. If students have not had the opportunity to learn, then it will be difficult to identify task characteristics that could prevent students from showing what they know and can do. Nonetheless, it is important to try to separate these issues and to recognize where the responsibility for each lies. Responsibility for opportunity to perform lies with the task and the task developer, and opportunity to learn is primarily the responsibility of administrators, teachers, parents, and students.

In our work, opportunity-to-perform issues emerged as a recasting of the time-honored concept of task validity—whether a task measures what it is intended to measure—because unless students are given sufficient opportunity to perform it is not possible to make valid inferences about what the students know and can do. Thus, when draft versions of a task failed to produce expected results in field trials, we questioned the task's face validity and asked what the task *did* measure. The challenge was to determine the source of the difficulty and then to revise the task in ways that maintain the important mathematical ideas the task was intended to assess.

This chapter briefly describes the task-development process and then illustrates several key concepts that emerged while attempting to construct tasks that provide students with opportunities to perform. In particular, some tasks create *cognitive overload* by attempting to assess skills, conceptual understanding,

and problem solving simultaneously. When tasks seem inaccessible, there are ways to *create access* while maintaining task integrity. One such way is to provide carefully constructed *scaffolding*. When placing tasks in contexts, the context sometimes obscures the mathematics. Another issue is *over-zealous assessment*—the temptation to assess everything that is possible from a given context. When many students gave incorrect responses, sometimes the source was a *task miscue*—an element of the task presentation that leads students to give an incorrect response. Some tasks are stated in such a way that an incorrect solution is obvious and enticing. Without thinking the problem through, most students will respond with that solution. Such task presentations contained what can be called *elephant traps*. The chapter closes with a list of recommendations for avoiding these obstacles. This list might serve as a starting point for those readers who wish to develop assessment tasks that maximize students’ opportunities to perform.

### The development process

The New Standards task development process is designed to produce candidate assessment tasks for a series of standards-based examination that are referenced to the *New Standards Performance Standards*. The following outline briefly describes the high school task development process that evolved in the course of this work:

Task kernels are solicited from teachers and professional assessment developers in the U.S., Europe, and Australia, and also from U.S. curriculum development projects; for example, The Interactive Mathematics Project, Core Plus, Modeling Our World, College Preparatory Mathematics, Connected Mathematics, and Mathematics in Context.

Task kernels are tried out in a small number of classrooms under the observation of a task designer who makes rudimentary judgments about the tasks’ measurement targets.

The preliminary tasks are sent to an expert review panel, together with the initial judgments about their measurement targets. The review panel is composed of a curriculum developer, a mathematician, a grade level appropriate mathematics teacher, and a mathematics educator who has a special interest and expertise in identifying and addressing equity issues.

Following this initial review, the tasks are organized into balanced packages comprising roughly 45 minutes worth of assessment and are sent to three teachers who live and work in different educational environments in the U.S. These “co-developers” then administer the candidate task packages to their own students or observe their colleagues administering the tasks to appropriate groups of students.

At a task development meeting the co-developers work with New Standards staff and members of the expert review panel to revise the tasks in the light of the classroom trials, to identify or verify the tasks' measurement targets, and to create rudimentary scoring rubrics for each task.

The tasks that survive the task development meeting are sent to a second set of co-developers for further classroom trials. At a second task development meeting, the tasks, their measurement targets, and their rubrics are again revised as necessary.

Finally, a balanced and robust set of tasks is selected and used to create a version of the *New Standards Reference Examination*. These examinations are field tested in a stratified sample of schools.

Once the data from the field-test is available, the examination tasks are returned to the expert panel for final review and any necessary final revisions.

The final examinations are compiled and put before an equity review panel prior to being published.

As can be seen from this description, the task-development process provides many opportunities for learning.

### **When cognitive overload stymies opportunity to perform**

The task *Hang Glider* (Figure 7) simultaneously requires mathematical skill, conceptual understanding, and mathematical problem solving. As such, it is a good example of *cognitive overload*.

Question 1 is relatively straightforward, requiring only some fairly primitive modeling. Students must realize that to estimate the area of sail needed, they must multiply their body weight by 1 square foot per pound of body weight and add the weight of the frame. (A more complex task emerges if the weight of the sail itself is considered, but no student in our sample took this direction.)

In Question 2, the complexity of the task soars. The diagram in Figure 8 illustrates one possible approach to solving for the total sail area. The left-hand side of the hang glider is decomposed into two triangles that are rotated and reflected in order to reconstitute the right hand side into a rectangle of length  $l$  and width  $w/2$ , where  $l$  and  $w$  are the length and width of the hang glider. In this question, both the conceptual and the strategic hurdles are quite high.

Question 3 adds another dimension. One route towards success is to recognize and then solve equations relating  $l$  and  $w$ . For example, if the answer to Question 1 was 130 square feet, the formula from Question 2 gives  $130 = \frac{1}{2}wl$ .

**Figure 7. Hang Glider**

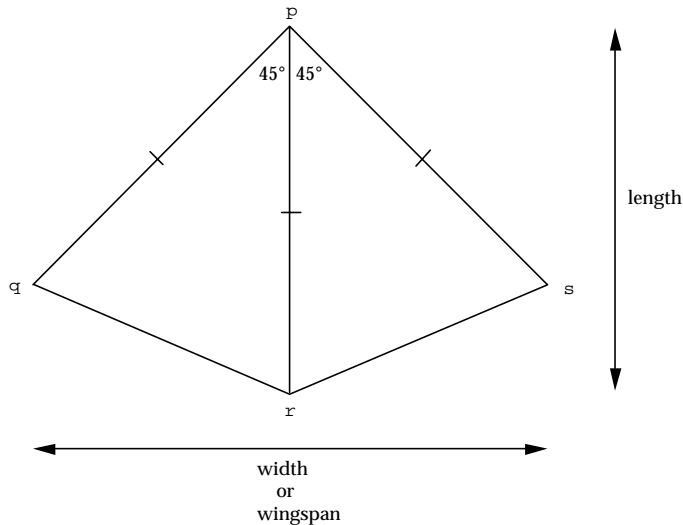
*In this task you will be asked to figure out the area and dimensions of a sail for a hang glider.*

The sail is shaped so that:

$$pq = pr = ps$$

$$\angle rpq = \angle rps = 45^\circ.$$

pr is a line of symmetry



The frame for the sail will weigh about 35 pounds. One square foot of sail is needed for each pound of weight that is to be lifted.

1. About how many square feet of sail would you need to make a hang glider for yourself? *Be sure to show all your calculations and reasoning.*
2. Explain fully why the area of the sail is given by the formula:

$$\text{Area} = \frac{\text{length} \times \text{width}}{2}$$

Use a diagram to help you, if you wish, but be sure to show all your work.

3. Calculate suitable dimensions for a sail that will lift you.  
You should find suitable values for pq and qr.

Again, be sure to show all your calculations and reasoning.

*Reprinted with permission from the Balanced Assessment project, University of California, Berkeley.*

Finding a second equation relating  $l$  and  $w$  is much more difficult. One way to do this is to use the diagram to find a relationship between  $w$  and  $l$ :

$$pq^2 + ps^2 = qs^2, pq = ps = l, \text{ and } qs = w.$$

$$\text{So, } l^2 + l^2 = w^2.$$

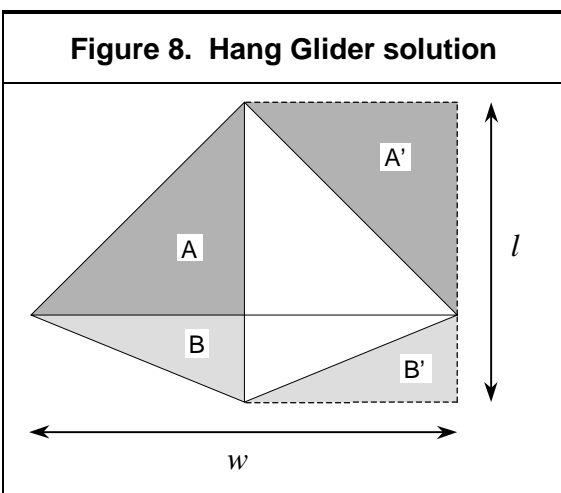
$$\text{Then } w = \sqrt{2}l.$$

Finally,  $w = \sqrt{2}l$  can be substituted into the equation  $wl/2 = 130$ , giving  $260 = \sqrt{2}l^2$ . Solving this gives  $l \approx 13.5$ . Clearly, the solution to this portion of the task involves highly non-trivial conceptual and manipulative demands.

The length of  $qr$  still needs to be determined. This can be done using the law of sines:

$$\frac{\sin 45^\circ}{qr} = \frac{\sin 67.5^\circ}{13.5}.$$

But the student has no chance of reaching this point without sufficient success on Question 1 and Question 2, to be able to draw upon those solutions to set up the equation that acts as the springboard to Question 3.



This task was piloted with 184 high school students. Using a four-point scoring rubric that defines a score of '1' as *little or no success*, just 17 students in the entire pilot group were able to achieve a score of '2' or higher. Not a single student was able to fully accomplish the task, and just one was able to provide a response that could be marked as "ready for revision."

It is difficult to make reasonable inferences about the specific nature of the obstacles that stand between the students and success on this difficult task. Is the obstacle that the students

were unable to formulate successful approaches to the problem? Or was it that the students were unable to handle the total skill and concept demands? As it was given, *Hang Glider* indicated neither what students know and can do nor what students do not know and cannot do.

*Hang Glider* demands that students make very high-level use of mathematical ideas. The data suggest that only the most talented of students will have enough experience to access these concepts and to use them in the sophisticated way that *Hang Glider* demands. In other words, *Hang Glider* is a task that asks students to make strategic use of concepts that are, for the majority of tenth grade students, not fully integrated into the students' existing conceptual frameworks (Hiebert & Carpenter, 1992).

Our developmental experience shows that when students work simultaneously at the cutting edge of both their strategic domain

and their conceptual domain, the result is cognitive overload, and only the most talented can demonstrate success. Because all students can and should benefit from studying to prepare for standards-based assessments, the assessments should be designed so that students may be successful not only through special insight but also through hard work. This is not to say that the concepts entailed in *Hang Glider* will never be fair game in an assessment. Concepts such as these *are* fair game, but care must be taken to ensure that they are assessed in an arena that does not have confounding strategic hurdles. Tasks entailing a high strategic hurdle often provide a false negative assessment of what students understand about underlying concepts.

### Creating access while preserving task integrity

One of the design challenges associated with developing almost any non-routine task is that of creating access without radically altering the intended measurement target. Responses to the challenge can be caricatured as follows:

A task that seems appropriate for a specified cohort of students turns out to be almost totally inaccessible. Initial classroom trials reveal that almost no students can make any sensible headway on the task. As a result, the task is subjected to a series of creative revisions. In subsequent classroom trials, the task produces a distribution of responses that is considerably more palatable. All involved are happy.

That is, all involved are happy until someone asks, What is still being assessed by the revised task? Does it still exemplify the kind of challenging and non-routine task that students should be able to do? Or, have the task revisions taken away the most interesting and mathematically challenging parts of the task? Often, creative task revisions introduced to promote access actually produce a less challenging and significantly more routine task that measures something quite different from the originally intended assessment target.

One example of a seemingly inaccessible task emerged in early trials of the now successful and relatively accessible<sup>1</sup> task *Snark Soda* (Figure 5, p. 19). Initial pilots of this task produced virtually no success among large numbers of high school students. The following complaint typified the response of almost all students who attempted the original version of this task:

---

<sup>1</sup> In Chapter 4, data obtained from grade 11 students show that the version of *Snark Soda* (Figure 5, page 19) still provides an enormous challenge to many high school students. The implications of this are discussed in terms of the current neglect of solid geometry in the high school curriculum.



*“There are no numbers, and without numbers you cannot find the volume of anything.”*

Apparently, students did not think to measure the drawing of the soda bottle, even though the drawing was described as being *full size* and *accurate*. Clearly, if this were the only thing that *Snark Soda* was going to tell us about students’ thinking, then it was not going to emerge as an informative assessment task.

The design challenge was to create a version of the task that would lead students to recognize the measurements they needed to make without destroying the core ideas behind the task. Initial suggestions identified ways that the diagram of the bottle could be labeled with judiciously selected measurements. One argument supporting this particular revision was that using rulers to measure diagrams can be quite alien to the culture of many American high school mathematics classes. Some teachers reported that they caution their students not to use rulers to measure diagrams in traditional geometry classes. The problem with this particular direction for task revision, however, was that it would completely carry out the primary modeling component of the task. In other words, to provide measurements for the bottle (including deciding *which measurements* would need to be made) would have been tantamount to doing the most interesting and challenging part of the task for the students. This change would have radically altered the assessment target of *Snark Soda*, shifting it from problem solving to skills.

To preserve the integrity of the task as a problem-solving one—where students would decide where to take measurements on the bottle and how to decompose it into familiar geometric shapes—it was decided that measurements should not be supplied. Instead, the task was more subtly modified, by adding the words *use a ruler to measure the bottle*. With this revision, students could be directed to find the necessary measurements, but the heart of the task was not altered.

One might ask why this simple solution was not suggested as the initial “fix” for *Snark Soda*. Perhaps the reluctance results from the long tradition of creating assessments composed entirely of bite-sized tasks and parceling out bite-sized assignments for students to do in their mathematics classes. Classrooms need to become places where students are given the opportunity to learn and then practice how to formulate and implement their own approaches to challenging, non-routine tasks. Assessments need to provide opportunities for students to showcase their mathematical understanding in ways that are challenging and non-routine.

## Scaffolding—guidelines and some case studies

Scaffolding is a technique that is used frequently in task development to regulate the accessibility of tasks. *Snark Soda* (as presented on page 19) is an example of a relatively unscaffolded task. It could be turned into a highly scaffolded task by offering, for example, the following instructions to the student:

1. Divide the drawing of the bottle into good approximations of regular geometric shapes. Sketch the geometric shapes you have chosen.
2. Measure the drawing of the bottle and mark the dimensions on your sketches.
3. Use your sketches, measurements, and formula sheet to find a good approximation of the volume of liquid in the bottle.

If this more-scaffolded version of the task were administered to students, the challenge for the student would probably be radically different from the challenge offered by the less-scaffolded version of *Snark Soda*. The scaffolding suggested here would shift the assessment target of the task away from problem solving and toward mathematical skills.

Several small-scale research studies have been conducted to investigate systematically the influence of scaffolding on students' performance on problem-solving tasks. In these studies, two different versions of the same task were administered to several different classes of students. The tasks were identical in all aspects except the amount of scaffolding.

In the first study (Shannon & Zawojewski, 1995), students were presented with two versions of a task involving shopping carts. The relatively unscaffolded version was called *Supermarket Carts* (Figure 9). The scaffolded version was called *Shopping Carts*, and it was identical to *Supermarket Carts* except that Questions 1 and 2 were replaced by the following questions:

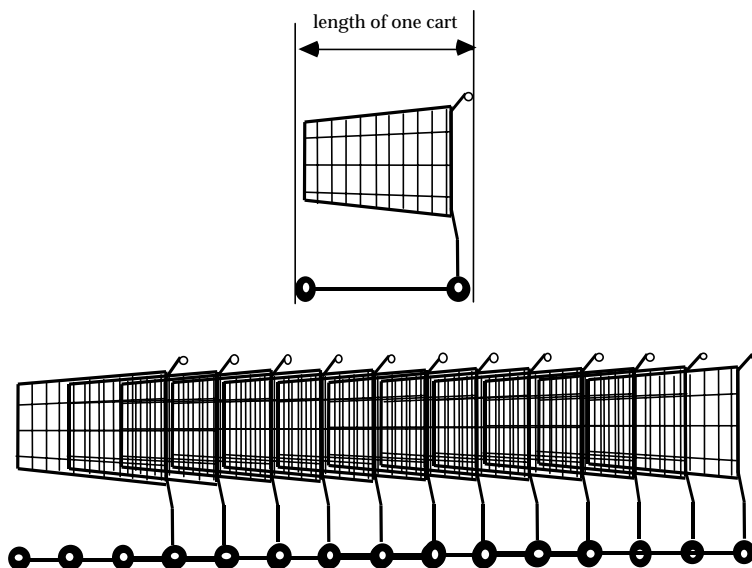
1. What is the length in centimeters of one full size shopping cart?
2. When they are “stacked,” by how much distance does each shopping cart stick out beyond the next one in the line? Show in a rough sketch of nested carts what this distance refers to.
3. What would be the total length of a row of 20 nested carts?
4. How many nested carts could fit in a space 10 meters long?
5. Create a formula that will tell you the length  $S$  of storage space needed for carts when you know the number  $N$  of shopping carts to be “stacked.” You will need to show HOW you built your rule; that is, we will need to know what information you drew upon and how you used it.

**Figure 9. Supermarket Carts**

*In this task you are asked to think mathematically about supermarket carts. You are asked to create a rule that can be used to predict the length of storage space needed given the number of carts.*

The diagram below shows a drawing of a single cart. It also shows a drawing of 12 shopping carts that have been “stacked” together.

The drawings are  $\frac{1}{24}$  th real size.



1. Create a formula that will tell you the length  $S$  of storage space needed for carts when you know the number  $N$  of shopping carts to be “stacked”. You will need to show HOW you built your rule; that is, we will need to know what information you drew upon and how you used it.
2. Now create a rule that will tell you the number  $N$  of shopping carts that will fit in a space  $S$  meters long.

*Later versions of these tasks appear in Balanced Assessment for the Mathematics Curriculum: Middle Grades Assessment Package 2 and High School Package 1, © 1999, The Regents of the University of California. All rights reserved. Published by Dale Seymour Publications, an imprint of Pearson Learning. Used by permission. Further information on these packages can be obtained from the publisher or the project Web site, [www.educ.msu.edu/MARS](http://www.educ.msu.edu/MARS).*

6. Now create a formula that will tell you the number  $N$  of shopping carts that will fit in a space  $S$  meters long.

Teachers divided their classes into two comparable groups. One version of the task was administered to each group within the same classroom. Students worked on the task individually under the impression that only one task was being administered.

In response to the scaffolded *Shopping Carts*, almost all students managed to develop an appropriate linear function to model the nested carts, but in response to the unscaffolded *Supermarket Carts*,

few students were able to do so. It seems reasonable to speculate that had the students who attempted *Supermarket Carts* been given the opportunity to attempt *Shopping Carts*, they would have shown a similar level of competence in developing an appropriate linear function.

The results of this study illustrate the role scaffolding plays in altering both the assessment target and the challenge of tasks. *Shopping Carts* is a scaffolded task. Implicitly, an approach to the task is outlined by the directive questions that target specific skills and concepts. *Supermarket Carts* is a less-scaffolded task. No auxiliary questions suggest or direct an approach. Students are told what to produce, but they are not told *how* to produce it. *Supermarket Carts* does not ensure that specific skills and concepts will be targeted. The less-scaffolded nature of *Supermarket Carts* provides the opportunity for exploration of the general strategies that students deploy in developing their approach to a non-routine task that calls for a mathematical model of a physical structure. In recognition of its substantial strategic hurdle, *Supermarket Carts* would be primarily a problem-solving task. The carefully constructed questions that direct an approach in *Shopping Carts*, on the other hand, reduce the strategic hurdle considerably, so that it would be categorized as an assessment of conceptual understanding.

Investigations of these and other tasks suggest that using student responses to less-scaffolded tasks to make judgments about students' basic competencies is to run the risk of making *false negative* judgments. Tasks such as the unscaffolded *Supermarket Carts* that seem to be good means of assessing general problem-solving strategies will probably underestimate students' proficiency in dealing with underlying skills and concepts. For example, when teachers were asked to administer only *Supermarket Carts* to their students, they expressed little doubt about its appropriateness in assessing what their students had learned about linear functions. In fact, those who had recently completed work in linear functions with their students fully expected that most of their students would be able to rise to the demands of this task. When it emerged instead that few students were able to model successfully the length of the stack as a linear function of the number of carts in the stack, the teachers expressed disappointment and feared that perhaps their students had learned little if anything about linear functions. However, the research suggests that these students probably *did* learn about linear functions but simply were not yet able to select and deploy this knowledge in a non-routine task with a high strategic hurdle.

Investigations of the role of scaffolding also suggest that giving students the opportunity to practice solving scaffolded tasks such

as *Shopping Carts* does not breed success on unscaffolded tasks such as *Supermarket Carts*. Furthermore, if tasks are often scaffolded to make them more accessible for students, students also must be given the opportunity to practice solving other tasks that are unscaffolded, non-routine, and challenging. Scaffolding, if too widely used, will thwart efforts to implement a broad and balanced system of mathematics instruction and assessment.

### Contextual challenge and barriers to performance

The small-scale study described in the previous section concentrated on the role of scaffolding in altering the structure of a task and thus regulating task challenge. Another series of small-scale studies was designed to investigate the role of context in altering the challenge of a task. The surprising results of these studies indicate the importance of determining empirically (rather than through professional judgment) how the context can inhibit opportunity to perform. Scaffolded and unscaffolded versions of the task *Storage Containers* were produced by replacing the carts in each of *Shopping Carts* and *Supermarket Carts* with stackable storage containers. This pair of tasks (scaffolded and unscaffolded) involving *Storage Containers* was of the same form as *Shopping Carts* and *Supermarket Carts*, differing only in minor contextual features. The effect of replacing carts with containers was explored by comparing student performance on unscaffolded containers with unscaffolded carts and scaffolded containers with scaffolded carts. In each of these paired sets of tasks, *Storage Containers* emerged as significantly less challenging than its counterpart involving carts. When student performance on the unscaffolded containers tasks was compared with performance on the scaffolded version of carts, the containers task again emerged as less challenging. From a performance standpoint, these two pairs of tasks are clearly *not* of the same form, and so there must be additional differences that could explain the relative performance hurdles.

The following is a list of some of the subtle differences in the tasks:

- The drawings of the storage container that are provided are much less complicated geometrically than the drawings of the shopping carts.
- In carts, the stack is horizontal and as each new cart is added the length of the stack increases horizontally. In storage containers, the stack is vertical, and as each new container is added the height of the stack increases vertically.
- In carts the scale factor is  $1/24$ ; in containers the scale factor is  $1/10$ .

Careful attention to students completing the cart- and container-based tasks revealed that it was easier for the students to take measurements from the stack of containers than from the stack of carts. Physical characteristics of the carts, such as the wheels and handles, seem to add unnecessary complication and confusion to the task of measuring the carts.

Students would physically mime the growth of the containers as they grappled with its representation but did not use any similar action with respect to carts. The vertical increases of the stack of containers seemed to be easier for students to visualize than the horizontal increases of the stack of carts. One speculation, therefore, is that the greater ease in visualizing the vertical growth of the containers helps students construct the corresponding symbolic representation. In some part, this may account for the superior performance of students on the tasks involving containers.

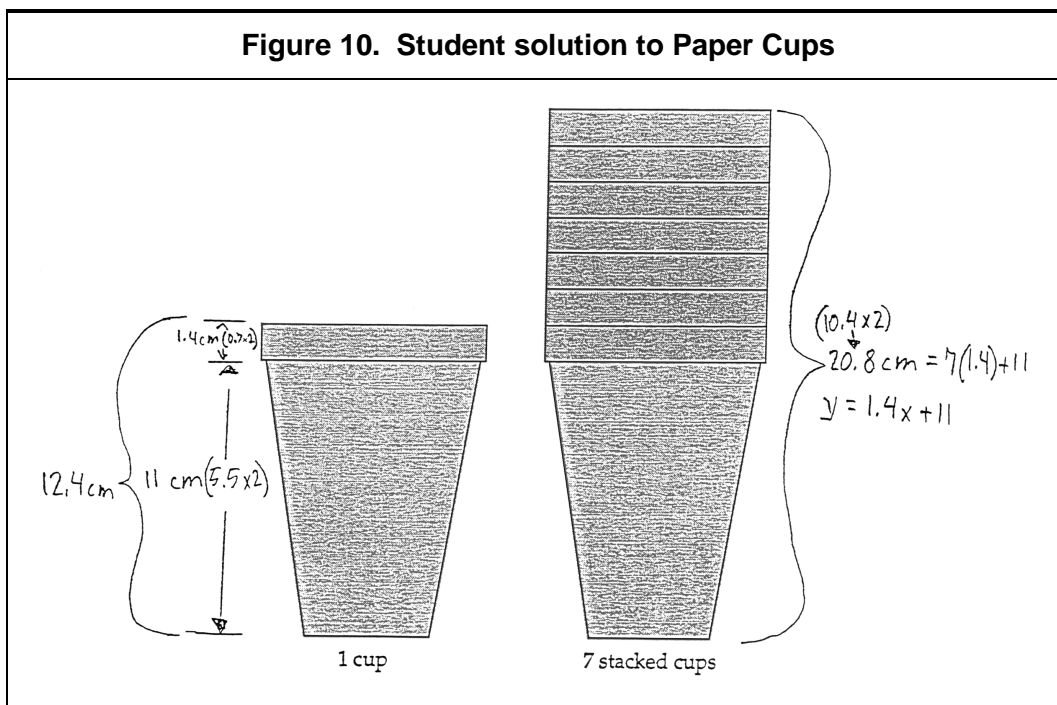
In addition, the choice of scale factor,  $1/24$  in the drawings of the carts and  $1/10$  in the drawings of the containers, emerged as a strong influencing characteristic (Shannon & Zawojewski, 1995). The scale factor of  $1/24$  provided a greater hurdle in the unscaffolded *Supermarket Carts* than in the scaffolded *Shopping Carts*. However, the scale factor of  $1/10$  did not emerge as a significant hurdle in either version of *Storage Containers*.

This series of comparisons of tasks involving shopping containers and carts demonstrates how contextual issues can be used (advertently or inadvertently) to alter the challenge of a task without altering its general structure or lowering its strategic hurdle. These contextual issues relate to opportunity to perform.

To continue investigation of these issues, the parallel task *Paper Cups* was produced by replacing the storage containers with paper cups. The paper cups were drawn to  $1/2$  actual size. The challenge of *Paper Cups* did not seem as great as the challenge of *Storage Containers*. So, as before, scaffolded *Paper Cups* was compared with scaffolded *Storage Containers*, and unscaffolded *Paper Cups* with unscaffolded *Storage Containers*.

Both versions of *Paper Cups* emerged as more accessible than the corresponding *Storage Containers* tasks, and this relative ease also can be explained in terms of contextual characteristics.

When many students attempted *Storage Containers*, they measured the size of one container (2 centimeters scaled up to give the actual size of a container as 20) and then measured the amount that each additional container *stuck out* above the one below (0.5 centimeters scaled up to give the actual size of a *stick out* as 5). Then they tried to represent the height of the stack in terms of the height of one container plus the height of  $x-1$  *stick outs*; algebraically,



Reprinted with permission from *New Standards™*. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).

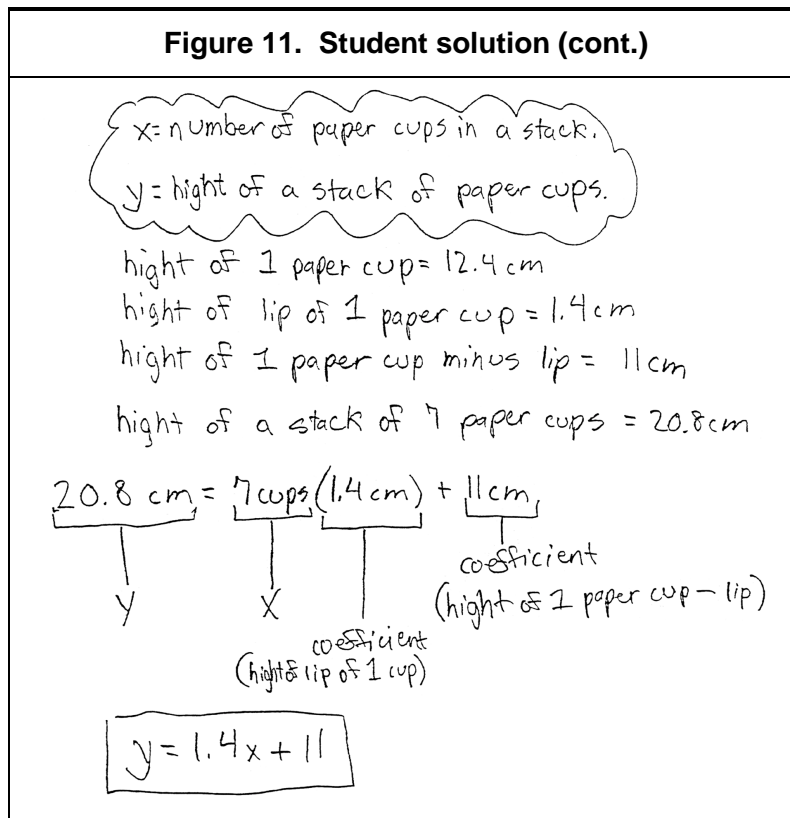
as  $H = 20 + 5(x-1)$ , where  $H$  represented the total height of the task, and  $x$  represented the number of containers in the stack.

When students produce a formula of this type, it is clear that they have successfully navigated what we refer to as the  $x-1$  aspect of this family of tasks. This is an aspect of *Shopping Carts* that very few students manage to process correctly. The mistake that occurs when students do not successfully navigate the  $x-1$  aspect of *Storage Containers* is usually expressed as follows:

$H = 20 + 5x$ , where  $H$  represents the total height of the task, and  $x$  represents the number of containers in the stack.

In contrast to both *Shopping Carts* and *Storage Containers*, however, many students working on *Paper Cups* will immediately decompose the cup into the following two parts, which they sometimes label as the *body* and the *brim*, as illustrated in the student solution in Figure 10. This decomposition enables many students to create the required formula directly in terms of the height of body of one cup plus the height of  $x$  brims, as illustrated in the remainder of this student's response (Figure 11).

Clearly, the structure of the cup lends itself to this decomposition, which enables students to finesse the  $x-1$  aspect of the task. The specific features of the cup reduce the conceptual demands of the problem. We say this because the specific features of the cup enable students to deal with  $x$  lips rather than  $x-1$  cups, and dealing with



Reprinted with permission from *New Standards™*. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).

$x$  is less sophisticated than dealing with  $x-1$ . Put in another way, the contextual factors associated with the cup provide greater opportunity for students to perform.

*Paper Cups* emerges, therefore, as a task that has a relatively high strategic hurdle, is appropriately challenging, and yet can be presented without relying on any directive questioning. It is a type of task that can be quite beneficial to use with students who are not accustomed to solving context-based problems. It also is a good introduction to non-routine tasks because it may be solved with lower levels of tenacity, it encourages perseverance, and it enables students to show what they can do rather than what they cannot do.

These findings have important implications for the model of assessment that is advanced in Chapter 2, which recommends separating assessment of mathematical skills, conceptual understanding, and problem solving. In this family of problem-solving tasks that involved stacks, students showed the most success when the conceptual demand of the task was reduced. Therefore, task developers should take care that the conceptual demands of a problem-solving task do not prevent students from showcasing their problem-solving capabilities.



These findings also demonstrate one way to increase access to a task *without* using scaffolding to structure or dictate an approach to the task, thereby reducing its strategic demands. Access may be improved by reducing the conceptual demand of the task while keeping the strategic demand of the task intact. Of course this does not mean that assessment of conceptual understanding is to be sacrificed in the interest of assessing problem solving. Remember, the model advanced in Chapter 2 recommends creating specific tasks to assess conceptual understanding. In these specially designed tasks, the conceptual demands will need to be as deep and as far ranging as the conceptual demands of the standards on which the assessment is based.

### **The implications of contextual challenge on opportunity to learn**

Considerable attention has been invested in examining both the obvious and the more subtle differences among the cart, container, and cup tasks. Given current recommendations to situate some mathematical learning and assessment activities in realistic contexts (e.g., NRC 1993b; NCTM, 1995), it is worthwhile to explore in detail the ways in which specific contexts outside of mathematics can facilitate or challenge mathematical thinking. In assessment, particularly when the stakes are high, it is imperative to discern the ways in which the contexts affect opportunity to perform and consequently issues such as equity and fairness. Because any context will be more familiar to some students than others, some bias is inevitable, but bias can be reduced through continual review and input from equity experts who can detect biases not apparent to the task designer.

Some connections with the world outside of mathematics are recommended for learning as well as for assessment (e.g., NCTM, 1989, 1995; NRC, 1993b). In view of this recommendation, the research into the relative effects of replacing carts with containers and then containers with cups leads to questions about the relative effects of the specifics of linear function tasks that rely on contexts such as car rental charges, phone call charges, and electricity charges. When each of these involves an initial value in the form of a fixed charge and a constant increase in the form of a fixed charge per day, or fixed charge per minute of call, or fixed cost per kilowatt-hour of electricity, each of these situations can be modeled by  $y = kx + b$ . These types of problems are now commonly used in schools to teach linear functions. The issue is how the specifics of these situations might count for or against student learning. A comparison of student performance on this type of task relative to student performance on *Paper Cups* or other stacking applications would probably lead to interesting insights. At this stage, it seems that the context of the tasks involving stacking would make the underlying concepts *more* accessible to students. This is because

the quantities that are to be related to each other in *Paper Cups* (number and height) seem to be much more tangible for students than the quantities to be related in tasks situated in contexts involving rental car, phone call, or electricity charges.

In addition, examples of stacking enable the students to trace the structure of the stack in different representations (i.e., verbal, table, and diagram) and this makes it possible to use the structure to demystify the translation to more abstract representations (i.e., graph, formula). And it is this attribute of these structures that suggests their use in the initial teaching of concepts such as linear functions. The variables that need to be represented in physical structures comprised of cups or books are more concrete and more *visible* for students than are the quantities such as cost and time in tasks involving rental cars and telephones or quantities such as cost and kilowatt-hours in tasks involving utility bills. If students are taught about linear functions using contexts they can visualize in a concrete tangible way, it is hoped that they will be able to apply the ideas they have learned to less obvious situations. A related point is that the call for connections with the world outside mathematics has led to frequent use of contexts such as rental cars, phone calls, and utility bills, and our assessment development experience suggests that these contexts probably differ greatly in their abstractness and in their ability to serve as learning tools. Explorations should be carried out into the effectiveness of frequently used contexts, to determine whether these contexts are truly suitable for initial learning purposes.

### **Over-zealous assessment**

Sometimes task designers, in their eagerness to create worthwhile tasks, try to assess everything that it is possible to assess in a given situation. This phenomenon can be called *over-zealous assessment*. The problem of this affliction is most apparent in assessment opportunities where *less* might actually mean more.

Through scaffolding, for example, task designers sometimes try to wrench every possible detail out of a given context or scenario. Overuse of scaffolding generally dictates a solution path for the student, and serves to control what the student uses in the assessment. One advantage of tightly controlled assessments is that they tend to have better measurement characteristics. For example, if the intention is to build a large-scale assessment that can be standardized, then the scores will be more reliable and generalizable when the test comprises many tightly controlled items rather than smaller numbers of less well-specified problems. The disadvantage of tightly controlled assessments is that the task designer effectively specifies the mathematics, and all that remains for the student is to be led through a series of steps dictating the solution

to a task that might once have been interesting and challenging. If tests comprise only tightly controlled tasks, then the assessment will not include the full range of tasks that is necessary for a balanced assessment. Highly scaffolded tests greatly restrict the opportunity to assess strategy formulation, tenacity, high-level use of skills and concepts, communication and mathematical connections. Overuse of scaffolding, therefore, decreases the capability of assessments to improve the ways in which the teaching, learning, and assessment of mathematics is enacted (as envisioned by NCTM, 1989, 1991, 1995; NRC, 1989, 1990, 1993b).

Scaffolding is not the only means of assessing everything in a given situation. It is sometimes tempting to pose a problem-solving task with a substantial strategic hurdle, then go on to load the task down with additional mathematically important ideas. Question 2 in *Supermarket Carts* (p. 40), for example, asks students to manipulate the function they were asked to create in Question 1. Undoubtedly this sort of skill is important and does not deserve to be embedded in a larger problem. The equity issues are obvious—it is unlikely that students stymied by Question 1 will be able to even attempt Question 2. In such cases, it would be unreasonable to make inferences about the students' ability to manipulate symbolic expressions. This is not to say that short closed questions such as these have no place in an assessment. On the contrary, important skills such as these should have their *own place* in an assessment—but not tucked away at the bottom of a larger assessment of strategic use of mathematics. Indeed, their place is in assessment tasks designed specifically to assess mathematical skills and concepts, and these assessment tasks might well be those that use scaffolding intentionally to target specified aspects of mathematics.

### **Turning task miscues into opportunity to perform**

The effort that is put into developing assessment tasks and identifying their assessment targets will be wasted if similar effort is not paid to the interpretation of student work. Hiebert and Carpenter have noted that the assessment of understanding relies heavily on indirect inference from student responses to a task (Hiebert & Carpenter, 1992). Our own work in the development of assessments has shown that great care must be taken before inferring causal linkages between student understanding and students' responses to a given task; for example, it might not be reasonable to infer from a completely incorrect solution that the student does not understand the underlying concept.

There is a large amount of evidence illustrating how task characteristics can *miscue* students, leading them to provide an incorrect response. Miscues manifest themselves in a whole range of aspects, including graphics that mislead, sentence structure

that miscommunicates, and assumptions that are not shared between task-doers and task-makers. Miscues also can be a source of bias when different groups of students have differential familiarity with some aspect of the task presentation. Although some bias is inevitable, task designers should make every effort to detect and reduce it whenever possible.

One example of a miscue is provided by a recent attempt to assess probability. Students were asked to analyze a game that was presented as having been devised to raise money for the school library. The designer's intent was to ask students to estimate how much money the game would raise and to say how the game should be changed to raise more money for the library. An unfortunate choice of question, *How could they raise more money for the library?*, stimulated a whole host of creative money raising suggestions, but few of these dealt with the intended mathematical activity of changing the odds of the game. The problem here was the task's miscue rather than students' conceptions or misconceptions about probability.

Examples of miscue founded on unshared assumptions were provided by attempts to use the context of a forester's *Diameter at Breast Height* (DBH) tape to explore student understanding of the relationship between diameter and circumference. A DBH tape is used to provide a direct reading of the measure of the diameter of a tree. The tape is wrapped around the circumference, and the measure of the tree's diameter is read directly from the tape (based on appropriately scaled markings).

An initial version of the assessment task asked students to explain how they would create such a tape. This prompted students to provide a plethora of explanations including: the tape would need to be long because trees can be incredibly large, the material would need to be flexible so that it could be wrapped around a tree, and marks would need to be put on at least one end so that the measurements could be read. Here was a classic case of task miscue, which had more to say about task presentation than about students' understanding of the relationship between the diameter and circumference of a circle. Ultimately, a useful version of this task removed the miscues by providing students with a diagram showing part of a tape that was calibrated in centimeters and part of a special tape that was blank. Students calibrated the special tape so that it could be used to measure the diameter of trees directly.

Examples of graphics that miscue abound in task development work. Many of the students who respond to the tasks speak English only as a second language, and so graphics can be a useful way of reducing the reading challenge of a task. Such graphics are of two main types:

- those that are essential for communicating the mathematics intrinsic to the task; cups, carts, and containers, for example, are intrinsic graphics because these represent the physical structures to be modeled; and
- those that are used for cosmetic purposes or with the intent of reducing the reading challenge of the task.

In a task that asked students to design a circular ice-skating rink, according to a set of given constraints, a graphic depicting a skater on a circular ice-rink was inserted to reduce the reading challenge of the task. Yet early trials of the task produced large numbers of student responses based on a rink that was square rather than circular. These responses led us to notice that the graphic showing the circular ice-rink was framed by a dark square border. This border may have been the most perceptually salient feature of the graphic, and as a consequence it had unintentionally miscued the students.

Another interesting example of miscue by graphic occurred with the task *Shoelaces*. A large one-half scale drawing of a shoe was provided to serve an equity purpose; in early trials of the task, some students were able to use the lace holes on their own shoes as props when they were working on the task. For equity purposes, therefore, it seemed important that all students have access to a realistic drawing of a shoe with lace-holes and laces. This graphic caused no difficulty and is intrinsic to the task. The difficulties centered on a smaller cosmetic graphic. The most perceptually salient characteristic of this graphic, for many students, turned out to be its right-angular heel. This aspect served as an invitation for an unexpectedly large number of students to try applying the Pythagorean Theorem to this task. When the square root of the square of the height of the shoe plus the square of the length of the shoe did not seem to produce a reasonable final result, many students then multiplied this by the number of lace holes. When the graphic was adjusted so that it no longer had the appearance of a right triangle, no further applications of the Pythagorean Theorem to this linear function task emerged. More important, once students were freed from the unintended task miscue, they were able to show what they did know or could figure out about modeling the length of the lace needed as a function of the number of lace holes.

Once miscues have been identified, they serve to remind task developers that task development is a humbling experience. These episodes stress the importance of trying out different versions of a new task with small groups of students and peers, and of taking seriously those responses that appear odd or inexplicable, regardless of how few of them occur.

**Figure 12. Broken Plate (version 1)**

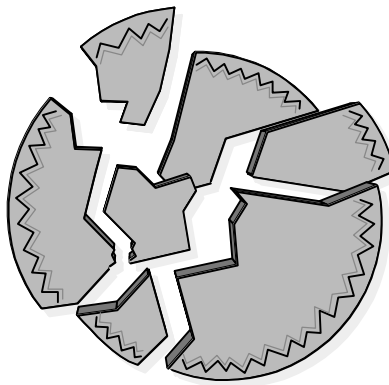
Imagine that you are visiting your friend Robin and you break one of a pair of flat circular plates. You want to replace it.

The two plates were identical.

The remaining plate has a diameter of 18 cm.

You want a potter to make an exact replacement of the one that you broke.

You know that the diameter of the tile will shrink by 16% when it is “fired” in the kiln.



Write a note telling the potter what size to make the diameter of the plate before firing.

Remember that the plate will shrink by 16% when it is fired.

Include a diagram of the plate before and after firing.

Mark the size of the diameter on each diagram.

*Reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or [www.ncee.org](http://www.ncee.org).*

### Turning elephant traps into learning opportunities

With regard to assessment, the term *elephant trap* refers to an unintended task hurdle or a task hurdle that provides no information other than the observation that large numbers of students consistently arrive at a common incorrect response. The task *Broken Plate* (Figure 12) provides an example of this phenomenon.

When this task was administered to a stratified sample of high school students, there was a remarkable convergence among the student responses. Students invariably decided that the diameter of the plate before firing should be 20.88 centimeters, because 16% of 18 is 2.88, and  $18 + 2.88 = 20.88$ . Students had obviously fallen into the trap of thinking that an  $x\%$  increase followed by an  $x\%$  decrease will get you back where you started. This task does not encourage students to demonstrate what they know, but instead traps them into showing what they do not know.

Perhaps the most useful characteristic of the task *Broken Plate* is that it highlights aspects of percentage increase and decrease as problematic and identifies an area of instructional need. A second version of *Broken Plate* (Figure 13) was piloted with another sample of students. This version incorporates an incorrect response that

**Figure 13. Broken Plate (version 2)**

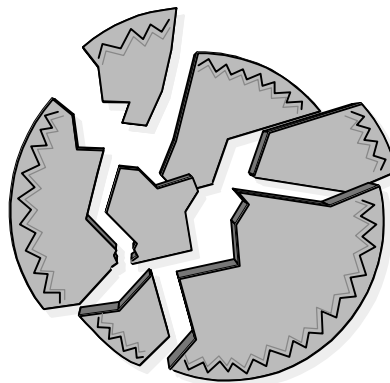
Imagine that you are visiting your friend Robin and you break one of a pair of flat circular plates. You want to replace it.

The two plates were identical.

The remaining plate has a diameter of 18 cm.

You want a potter to make an exact replacement of the one that you broke.

You know that the diameter of the tile will shrink by 16% when it is “fired” in the kiln.



**Here is Robin’s plan.**

Let us ask a potter to make the diameter of the new plate 20.88 cm because that is 16% more than 18 cm. Then when the 20.88 cm diameter shrinks by 16%, we will have a plate exactly like the one we broke. Look here’s how I figured it out.

$$18 \times 0.16 = 2.88$$

and

$$18 + 2.88 = 20.88$$

**Robin’s solution is wrong.**

1. Show that when the new plate with diameter 20.88 cm shrinks by 16% its diameter will NOT be 18 cm.
2. State a method that will give a plate that is the correct size. Include a diagram of the plate before and after firing. Mark the size of the diameter on each diagram.

*Reprinted with permission from New Standards™. For more information contact National Center on Education and the Economy, 202-783-3668 or www.ncee.org.*

typified student performance on the initial version. The response to this later version was remarkable:

- students’ responses spanned a range of answers rather than conforming to a single type,
- the majority of students’ responses to Question 2 were correct, and
- students were able to use *the typical incorrect* response to develop a correct one.

We would argue that this technique of incorporating an incorrect response and identifying it as such can frequently be

used to enable a task to function as a learning opportunity. The juxtaposition of the incorrect response and the student's misconception will create cognitive conflict for the student. The student is given the opportunity to reflect on an incorrect response, resolve the conflict, and produce the correct response.

The technique of giving students a wrong answer and asking them to supply a correct one is recommended in a recent publication commissioned by the NAEP Validity Studies (Jakwerth, Stancavage, & Reed, 1999). This technique was used with considerable success in the development of the Key Stage 3 mathematics tests that were developed to assess the National Curriculum for Mathematics in England and Wales (Close, 1996). The technique of situating common misconceptions in assessments is one that provides a direct opportunity for assessment to enhance learning and so heed the call that is expressed in The Learning Principle (NRC, 1993b) and the Learning Standard (NCTM, 1995).

Not every student will successfully resolve the conflict posed by such an approach. Indeed, some refuse the opportunity by asserting that the student response labeled as incorrect is in fact not wrong! What reason do students give for this? Often, they simply assert that the incorrect response is correct because it coincides with what they would have done or that it coincides with what they believe to be true. Nonetheless, what we have identified here is how cognitive conflict can be used to convert an elephant trap into an opportunity to learn. The constructive use of mathematical errors and misconception corroborates previous research in using student misconception as a learning tool (Bell, 1993; Bell, Swan, Onslow, Pratt, & Purdy, 1985; Borasi, 1996; Graeber & Campbell, 1993).

### **Recommendations for task development**

What follows is a list of recommendations for those who are interested in creating assessments or evaluating the quality of assessments. This list includes those recommendations from the NAEP Validity Studies (Jakwerth, Stancavage, & Reed, 1999) that appear appropriate to mathematics assessment. The NAEP Validity Studies investigation was conducted by interviewing students immediately after they had completed the eighth-grade 1998 national NAEP assessments in reading and civics about their test-taking behaviors and their reasons for omitting questions. Many students find constructed-response tasks difficult in general, and particularly difficult when they are asked to complete such tasks under time-limited conditions. The NAEP Validity Studies report that in the 1996 NAEP mathematics assessment, omission rates at grade eight were as high as 25 percent on some questions, with the highest omission rates on the extended-response questions items. There is a need to create extended-response mathematics tasks



that are as accessible as possible. Recommendations for task development are as follows:

- Select contexts that create rather than restrict access. Do not assume that a realistic context will facilitate access. It is possible to explore the accessibility of a particular context by trying out the same mathematical idea in a range of different contexts.
- Keep the reading challenge of the task low. Use diagrams to communicate the demands of the task. Test out the graphics: they should not include irrelevant variables that might mislead the student.
- Use clear and unambiguous vocabulary.
- Avoid esoteric abbreviations or idioms that might not be familiar to all students.
- Use scaffolding to create access but evaluate the effect on the assessment target.
- Beware of over-zealous assessment where there is the temptation to load a task down with too many parts. If students have been unsuccessful on the first or second part of a task, they are unlikely to attempt parts that come later.
- Beware of cognitive overload. Try tasks out with students to make sure that the cognitive demands of the tasks are aligned with the expectations laid out in the standards and that the demand is appropriate with the circumstances of performance that are required. More can be expected in a situation where the circumstances of performance are characterized as research-feedback-and-revision than on a timed situation.
- Locate talented task designers. In addition to developing its own tasks, New Standards sought kernel tasks from many sources, including The Balanced Assessment Project, mathematics teachers from across the U.S., curriculum developers, and task developers in Australia and England.

Perhaps the most sound practical advice is that all revisions to high-stakes assessments should be tried out with students to explore the effect of these revisions on opportunity to perform. *No one can guess reliably how students will respond to a particular version of a task.*

## Chapter 4: Assessment and opportunity to learn

---

In this chapter, the focus shifts away from task development and issues that impinge on students' opportunity to perform and toward assessment tasks as seen in the social milieu of the classroom because that is where students are (or are not) provided opportunity to learn. Providing such opportunity means creating access for students to the procedural, conceptual, and strategic knowledge to support a deep and robust understanding of mathematics and the *know how* necessary to demonstrate this multifaceted knowledge. A number of key research reports have focused attention on the importance of collecting opportunity-to-learn data to inform the interpretation of assessment data (NCTM & NRC, 1997; NRC, 1989, 1997, 1998; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993; Schmidt, McKnight, Valverde, Houang, & Wiley, 1997; Stigler & Hiebert, 1997). In particular, opportunity-to-learn issues emerge as an essential strategy in any effort to improve education and address equity issues (Massell, Kirst, & Hoppe, 1997; Black & Wiliam, 1998).

The concept of opportunity to learn is linked to the concept of opportunity to perform as described in Chapter 3. If students lack opportunities to perform, they will not be able to show what they know and can do. If students lack opportunities to learn, they will not be able to avail themselves of opportunities to perform. In Chapter 3, issues concerning opportunity to perform led to a focus on the development of assessment tasks and the dimensions of a balanced assessment. Here, issues concerning opportunity to learn lead to a focus on teaching and learning.

The task development experience that underpins both the model for balanced assessment in Chapter 2 and the discussion of opportunity-to-perform issues in Chapter 3 was not conducted behind the closed doors of assessment designers' offices. Instead, much of this experience has been obtained in mathematics classrooms across the country, because each assessment task is put through several rounds of systematic classroom trials, including initial task trials implemented in two or three mathematics classrooms, and also large-scale field tests where the tasks are put through trials with a large stratified sample of students.

Each classroom trial is observed by at least one of the following:

- a full-time assessment developer,
- a full-time mathematics teacher who is participating as a co-developer in the assessment development process,
- a classroom teacher who is responsible for providing written evaluations of the task in action, or
- a teacher who is participating in a professional development program focusing on assessment.

This chapter draws upon this extensive body of classroom-generated experience to present a series of recommendations and conclusions based on observations of assessment tasks in the social context of the classroom. Many of the sections of this chapter highlight barriers to opportunity to learn, such as tight sequencing of teaching and testing, inappropriate emphasis on skills acquisition activities, inappropriate task modification, the need to cover the curriculum, preconceptions of teachers, and gaps in the curriculum. Each of these discussions is framed by locating it within the context of relevant and recent research. Where appropriate, larger implications for the teaching and learning of mathematics are identified. Other sections discuss how complex tasks may be used to enhance learning opportunities in the classroom. While working through such a task, for example, misconceptions and mistakes may be viewed as opportunities to learn rather than as complications to be avoided. Furthermore, class work on complex tasks is necessary to develop problem-solving tenacity and the ability to communicate about mathematics. The purpose of this chapter is to identify how efforts to improve assessment might be used to improve mathematics instruction and learning. Therefore, the primary concern in this chapter is not just finding better ways to assess students but finding ways to enable students to perform better on worthwhile assessments. The chapter closes with a list of recommendations based upon these issues and informed by Black and Wiliam's (1998) contention that learning is driven by what teachers do in classrooms.

## Tight sequencing of teaching and testing

A common obstacle to success on non-routine tasks is the tendency of students to attempt to apply the specific mathematics that they are currently studying to the task at hand, whatever that task might be. For example, when students were presented with either *Shopping Carts* or *Paper Cups* immediately after they had studied area or volume, many began their work by trying to find the area or the volume of the cart or cup. Similarly, when a large number of students tried to set up and solve a system of linear equations in response to one of these two tasks, it turned out that the class had just been studying systems of linear equations. When a disproportionate number of students in a class provided solutions involving  $y = mx + b$  to *Broken Plate*, a task investigating the relationship between percent decrease and percent increase, their teacher confirmed that his students were working on the slope-intercept form of the equation of a line. When considering what might have inspired such seemingly incongruous responses, it seemed that students were using whatever tool was most readily at hand rather than grappling with and making sense of the task. In fact, the premise on which most classroom assessment rests is to assess what has just been taught.

The problem with this *what we just studied* phenomenon is that it usually does not work well in creating access to the task, and often makes the task less accessible for the student. In the words of one teacher commenting on her students' work: "Most students made the task harder than it was, when they tried to make it fit with what we were currently studying."

When teachers discussed this at professional development workshops on assessment or at assessment co-developers meetings, they realized that they shared a common problem. Teachers are frequently quite taken aback by this realization and hypothesize that it is their common practice of teaching a topic then testing it, teaching another topic and then testing, that might cultivate this behavior in their students. These teachers acknowledge that they rarely test the mathematics that their students have learned twelve, six, or even just two months previously. Therefore, it is also rare that their students are required to attempt challenging non-routine tasks.

This *what we just studied* phenomenon is not reserved for those students who are under-prepared in mathematics or for those who find learning mathematics difficult, but can be observed even in honors classes where student participation is usually marked by a high level of success. Many highly successful students, well on their way to successful completion of Algebra II and Trigonometry courses, failed to solve a problem such as that posed by the unscaffolded version of *Shopping Carts* because, rather than trying to make

sense of the task, they attempted to bring only their most recently learned but inappropriate mathematics to bear. It was as if these students had in their heads a directory of template problems that they had learned how to solve. Instead of thinking about the task at hand and making decisions about the mathematics that might be needed to solve it, these students simply forced aspects of one template problem after another onto *Shopping Carts*. When students of this caliber work on non-routine assessment tasks, it is evident that they know a lot of mathematics—far more than is actually needed to solve the task. But it also is evident that their mathematical understanding is fragile and inflexible (Lesh, Lamon, Lester, & Behr, 1992). Further, it is evident that these students have had little practice either making sense of mathematics or using mathematics in a practical fashion. These observations also reinforce an earlier conclusion stated by Schoenfeld: “Knowing’ a lot of mathematics may not do some students much good if their beliefs keep them from using it” (1987, p. 198).

The most serious aspect of this *what we just studied* phenomenon is that when students attempt to make only their most recently learned mathematics relevant to the task at hand, they are providing evidence of a routine that may characterize all of their learning of mathematics. The students’ habit is to do mathematics without having to *think* about the task. Unfortunately, these students are not only doing what they usually do, but also are doing what usually works for them.

The coupling of teaching and testing in this way has consequences both for the mathematics that is learned and for students’ perception of the learning of mathematics. On the one hand, it leaves students ill-equipped to tackle non-routine tasks where an important hurdle is the selection of relevant mathematics. On the other hand, it instills in students the notion that mathematics can be learned by applying recently learned mathematics without a great deal of thought. It also runs the risk of teaching students that making mathematical sense or using common sense are not appropriate behaviors for the mathematics classroom.

Coupling teaching and testing in this way also has consequences for what a teacher can say about her students’ learning of mathematics. How does a teacher know whether her students have really learned the mathematics? How does the teacher know that her students will retain what they have been taught over the longer term? How does the teacher know what her students can do with the mathematics they have learned?

The view of teaching and learning that is evidenced by such tight coupling of teaching and testing also has been criticized by Schoenfeld:

All too often we focus on a narrow collection of well-defined tasks and train students to execute those tasks in a routine, if not algorithmic fashion. Then we test the students on tasks that are very close to the ones that they have been taught. If they succeed on those problems, we and they congratulate each other on the fact that they have learned some powerful mathematical techniques. In fact, they may be able to use such techniques mechanically while lacking some rudimentary thinking skills. To allow them, and ourselves, to believe that they understand the mathematics is deceptive and fraudulent. (Schoenfeld, 1988, p. 30)

One of the more far-reaching effects of this practice, as shown by our own experience and addressed in the discussion by Schoenfeld, emerges most acutely when narrowly defined tests are used for state- or district-wide accountability purposes. In such cases, teachers report that they find themselves under increasing administrative pressure to spend greater and greater amounts of time preparing for the test (Romberg, Zarinnia, & Williams, 1990). This can breed an ever-expanding culture of test preparation and, at its most extreme, runs the risk that test preparation could completely replace instruction. Mathematics classes might then become characterized by students working repetitively on set after set of questions that mimic those that are on the test.

When narrowly defined tests are used in this way to address accountability needs, the consequences for learning are inevitable. Students' opportunity to learn is replaced by the opportunity only to practice a narrow range of test questions. There is a well-grounded fear that this approach will fail to prepare students for higher level mathematics courses. Such an approach will do little to inculcate a mathematical disposition or to encourage students to invest in further study of mathematics. Finally, the *costs* of this kind of testing can become hidden—large amounts of teacher time, and classroom resources are diverted away from teaching and learning and are used instead to prepare students for narrow tests that are at best loosely connected to a balanced curriculum.

### **Inappropriate emphasis on skills acquisition activities**

Teachers often say that although some non-routine tasks are interesting, rich, and target worthwhile mathematics, they are not appropriate for *their* students.

When we explore this perception further, we find that many of their students are considered by these teachers to be under-prepared in mathematics. In the view of their teachers, these students lack basic skills. Teachers described how, in an effort to rectify this situation, they restricted their students to sets of short, closed, procedural exercises. They have the perception that their students must acquire some basic level of achievement in rudimentary

mathematics before they can be permitted to attempt challenging non-routine tasks. In these teachers' views, the full range of tasks illustrated here would be far too challenging for their students, and so they believe it necessary to restrict students to skills-based tasks.

One serious problem with this common approach in teaching mathematics to students who are under-prepared is that there is very little evidence that it works to do anything more than teach simple calculation procedures, terms, and definitions (Hiebert, 1999). Hiebert draws on the most recent National Assessment of Educational Progress (NAEP) to answer the question, "*What are students learning from traditional instruction?*" He reports:

In most classrooms, students have more opportunities to learn simple calculation procedures, terms, and definitions than to learn more complex procedures and why they work or to engage in mathematical processes other than calculation and memorization. (Hiebert, 1999, p. 12)

Another serious problem is that it is simply inequitable for large numbers of students to emerge from high school without ever having had the opportunity to engage in mathematics work that has been designed to develop conceptual and strategic capabilities. Clearly, the intention is not to deny students this opportunity. Teachers usually intend to shift to a more interesting gear after their students provide evidence that they have acquired the basic skills. Unfortunately, this frequently does not happen, and many students leave school without ever having been given the opportunity to learn mathematics in a broad and balanced way.

The problem can be approached somewhat differently. There is increasing evidence that the memorization of decontextualized fragments of mathematics does not work well in helping students learn mathematics (Hiebert, 1999). But there also is increasing evidence that students can learn when instruction regularly emphasizes engagement with challenging tasks (Stein & Lane, 1996; Schoen & Ziebarth, 1998), or when teachers regularly use technological tools to develop mathematical ideas (Heid, 1988; Hiebert & Wearne, 1996). The Carnegie Learning Program makes extensive use of technology and is currently demonstrating great success in motivating reluctant learners in large urban districts (Hadley, personal communication, February, 1999).

### **Inappropriate task modification**

As noted above, teachers will often argue that tasks of the type presented in Chapters 2 and 3 are more appropriate for students who are better prepared mathematically. Perhaps as a consequence, when teachers administer these tasks to their students, many of them massage the challenge of each task, in the hope that

their students will become neither too frustrated nor too confused by its demands. This practice of massaging the challenge of particular tasks to close the gap between the teachers' perception of what their students know and the perceived demands of the task has also been reported by others (Doyle, 1988; Henningsen & Stein, 1996).

Many teachers are particularly adept at deploying gap-closing strategies. They will often provide directive hints, construct pertinent demonstrations, introduce task scaffolding, and when all else fails they will sometimes try to *walk* their students through the task. The evidence presented in Chapter 3 shows how scaffolding and other well-intended challenge reduction techniques can radically alter the assessment target of the task. This evidence suggests that teachers' gap-closing processes can have far-reaching implications for students' opportunity to learn through challenging non-routine tasks, and that these strategies can restrict the actual range of tasks that their students will truly have the opportunity to tackle.

### **Covering the curriculum**

Another factor that can inhibit the use of worthwhile assessment tasks in classrooms is teachers' perception of the length of time that it will take their students to do the tasks. Teachers sometimes fear that, if they were to invest the time necessary for administering rich assessment tasks, they might be unable to cover large portions of the material they are expected to cover. Choices about the allocation of precious classroom time are difficult. However, many involved in the reform of mathematics teaching and learning urge teachers to cover less but spend more time going deeper, thus creating a broader and more balanced system of instruction (NCTM, 1989, 1995; NRC 1993b; Schmidt, McKnight, & Raizen, 1997; Schmidt, McKnight, Valverde, Houang, & Wiley, 1997; Stigler & Hiebert, 1997).

The implementation of worthwhile assessment tasks in classrooms is not the only innovation that is labeled as time-consuming. Indeed, most effective teaching strategies are time-consuming and therefore regarded as untenable by teachers who are faced with a large amount of material to cover. There is little doubt that if teachers are to be freed to provide opportunity to learn for all, they must be freed from the burden of covering large amounts of material.

### **Preconceptions of teachers**

It is interesting to observe teachers as they consider assessment tasks with a view toward possibly embedding them in their instruction. Frequently, teachers will work through a task and



then draw extensively on this experience in their appraisal of the task's appropriateness. As a consequence, this process leads some teachers to reject certain tasks outright. One teacher said,

This task would not be appropriate for my students. If it took *me* this long to complete the task, my students would never be able to stay at it.

Another stated,

This task is too abstract, I had to really think about this task. My students would never be able to start it.

Clearly, teachers are very concerned about overwhelming their students and about selecting appropriately demanding tasks for them. This is not surprising given the large number of students who give up all too quickly when they are presented with an assignment that does not immediately resemble one that they have been taught how to do. These findings about teachers' perceptions of the appropriateness of such assessment tasks in their classrooms corroborates research that addresses teachers' perceptions of the appropriateness of instructional materials. For example, teachers' perceptions have been found to be affected by both their perceptions about their students' backgrounds and abilities and the mathematical knowledge of the teachers themselves (Floden, 1996).

Some teachers do recognize that even tasks that challenge the teachers themselves sometimes can be appropriate for their students. It is difficult, however, to persuade other teachers that almost all of their students *can* learn to do challenging mathematics tasks and that students can learn mathematical skills at the same time that they are working on challenging tasks. This is in contrast to what seems to be a deep-seated belief that students' ability (or inability) to do mathematics is immutable and not something that can be improved upon by creating new or enhanced opportunities to learn.

Through classroom observations and interviews about assessment tasks, some revealing aspects of student beliefs about learning mathematics have also been identified. Many students have clearly defined and somewhat narrow views of what counts as appropriate behavior for the mathematics classroom. For example, many students have great difficulty in formulating a workable approach to a non-routine task. When students evaluate such tasks and describe how the tasks might be improved, they almost invariably judge the tasks as not giving them a clear enough indication of what they are supposed to do. They gave responses such as,

"Be more specific about what you want us to do on paper."

"Tell us more information on what we are actually supposed to figure out."

“You do not make it clear what you want us to do. It is better if you say—do this, then do this.”

In these responses, the students have revealed that they do not expect to have to formulate an approach to challenging tasks. They expect that their assignments will make clear not only what they are supposed to do but also the steps that they should take to do it. Many students simply do not perceive *doing* challenging mathematics as appropriate work for mathematics classrooms. Many students just want to be told what to do by their teachers. By the same token many teachers believe that, with so much content to cover, there is little time to do anything but *tell* their students as much as possible. Many students also will express a lack of confidence in teachers who wish to delve deeply into mathematics rather than rush through larger amounts of material at great speed (Borasi, 1996). Far from relishing the opportunity to dwell on fundamental mathematics with new eyes, students are often concerned that they will never be able to cover the given curriculum, or that focusing on specific aspects of mathematics in greater depth will adversely affect their final grade.

### Gaps in the curriculum

Developing assessment tasks sometimes highlights limitations in the ways in which curriculum content is determined. For example, the study of solids and their volume is a content area that is often de-emphasized in the current high school curriculum, and students invariably find our tasks involving solids and their volume difficult. As an illustration, Table 2 shows the distribution of scores for responses to *Snark Soda* (Figure 5, p. 19).

To earn a score of 4, a student must *fully accomplish* the task. To do so, the student must model the entire bottle using two or more solids, consider the curvature of the top and bottom of the bottle, address accuracy, and communicate each step of the work. This level of success requires significant integration of mathematical skill, conceptual understanding, and problem solving, but it is reasonable to expect that students in the eleventh grade will have fully absorbed these specific skills and concepts. Therefore it is disappointing that so few students are able to make use of these skills and concepts to fully accomplish the task.

Score	Off task	Score 1	Score 2	Score 3	Score 4
N = 877	2	431	291	129	24
%	0.3	49.1	33.2	14.7	2.7

To earn a score of 3, a student must prepare a response that, while not fully complete, can be characterized as *ready for revision*. It should be reasonable to infer that the student has the mathematical knowledge and ability to solve the task. The student can show this by modeling the entire bottle using two or more solids and addressing the curvature of either the top or the bottom of the bottle. The student might or might not address the accuracy of the volume and might not fully communicate each step of the work. Even so, just one student in six was able to reach or exceed this level of achievement on the task.

To earn a score of 2, a student must show *partial success* by modeling the bottle using more than one geometric solid (e.g., two cylinders). The student might not address either curvature and may use a combination of area and volume formulas. For most of the students who were able to achieve any significant success with this task, this level of achievement was as far as they got.

To earn a score of 1, a student must *engage with the task* but will have done so with little or no success. For example, the response might use only one cylinder to model the entire bottle. When the response contains only words or drawings that are unrelated to the task, the response is scored as “off task.” Notice that these two categories account for almost one-half of all of the eleventh-grade student responses.

What can be said about such disappointing performance? In this version of the task, students were advised to use a ruler, so the problem was not that students did not think to use a ruler to measure the bottle. One hypothesis is that the issue has less to do with specific task characteristics and more to do with students’ experience with solids in their classroom.

Many teachers readily confide that they often have only a few days left at the end of the tenth grade to devote to volume. Others indicate that they feel that the large body of knowledge they are obliged to cover during the tenth grade sometimes makes it impossible to cover volume at all. Why would a teacher choose to leave out volume rather than any other topic? It appears that this decision often reflects teachers’ perceptions of what is or is not necessary for the next mathematics course their students will take. For many students studying geometry, Algebra II is the next course in the sequence, and there seems to be a belief that a study of solids and their volume is not critical for success in Algebra II. As a consequence, the topic is often neglected. Unfortunately, even though the study of solids and their volume may not be a prerequisite for Algebra II as it is traditionally defined, a sound conceptual understanding of this subject area truly is a prerequisite for calculus. Clearly, if high school curriculum is determined solely by a perception

of what is required for the next course, then the longitudinal coherence of school mathematics is jeopardized.

Another place where a study of solids and their volume is de-emphasized is in large-scale assessments. In New York, for example, the Spring 1997 Pilot Questions that are used by many teachers to prepare their students for the Mathematics A Examination (which will soon replace Course I in the New York Regents sequence) suggest that a study of solids and their volume will be confined to finding the volume of a rectangular prism. Undoubtedly, this de-emphasis at the assessment level will bring about a de-emphasis on all other solids in the curriculum.

Study of solids and their volume should not be added to the curriculum in a cursory way because the problems described here cannot be addressed without placing such study at the firmly in the mathematics curriculum. The marginalization of solids within the curriculum has unfortunate consequences that extend beyond student preparation for the study of calculus. A study of solids and their volume provides an abundance of useful material for those seeking to enhance the learning of mathematics through an emphasis on connections, both within mathematics and with worthwhile and relevant contexts outside of mathematics. The *Principles and Standards for School Mathematics: Discussion Draft* (NCTM, 1998) goes a long way toward placing the study of solids and their volume firmly in the curriculum, and does so in a way that provides a coherent sequence across the Pre-K–12 curriculum.

Unfortunately, the problems identified by this assessment work are not confined to the study of solids and the tenth-grade curriculum. Schmidt and Cogan write:

Review of the TIMSS U.S. mathematics achievement and curriculum analysis results forms a rather compelling notion that the fundamental problem with our mathematics education system lies not with students or teachers but primarily with the way in which we think about and develop our mathematics curriculum. (Schmidt & Cogan, 1999, p. 7)

The TIMSS study, which characterized the U.S. curriculum as repetitive and lacking depth, indicates that the learning of mathematics can be tackled effectively only after something is done to re-conceptualize the mathematics curriculum. Assessment development experience also has demonstrated that a fragmented and cluttered curriculum puts teachers under enormous pressure and restricts their opportunities to deepen the focus of their instruction.

## Misconceptions and mistakes as opportunities to learn

Attempts to develop tasks designed to assess the robustness of students' understanding of mathematics have been met by an interesting mix of teacher reaction. For some teachers, this approach has validated their own classroom practice, characterized by a constructive focus on the robustness of conceptual development. In part, such an approach requires putting a constructive focus on misconceptions that are brought into the open by presenting students with thought-provoking and sensitive tasks. According to these teachers, assessments that made misconceptions visible were an invaluable aid to long-term learning and retention.

For many other teachers, however, conceptually oriented tasks that have the power to realize student misconceptions are to be avoided, lest these confuse students who already find learning mathematics difficult. In recent development work, New Standards staff prepared a formative assessment package to be used in a conceptual approach to the teaching and learning of slope (NCEE, 1998). Students use this package to investigate slopes of ramps, slopes of stairs, and slopes of lines, and to work through a range of challenging and conceptually oriented assignments on slope. A final assignment invites students to imagine a world where slope is defined not as rise over run but as run over rise. Students are asked to discuss the implications of this redefinition of slope. This final task is designed to assess the robustness of students' conceptual understanding of slope. Many teachers have reacted vehemently, arguing that this will run the risk of confusing students, or even leave them with the erroneous view that slope *is* defined as run over rise.

To steer away from conceptually oriented tasks is to adopt a view of student learning that is characterized by memorization of isolated fragments of knowledge, inherently fragile and unlikely to be retained. Indeed, teachers should use assessments that *do* operationalize student misconceptions, not to confuse students, but as part of the process of developing mathematical understandings that are robust and can withstand both the test of time and of counter-argument. Tasks that ferret out student misconceptions will provide insight into how the student has internalized the body of knowledge that the teacher is attempting to teach. It is only from a clear sense of what the student understands that subsequent instruction can be tailored to benefit the student. Seen in this way, tasks that illuminate student misconceptions will be crucial to the process of benchmarking growth in student understanding (Borasi, 1996).

### **Assessment practice makes perfect—developing tenacity**

When teachers regularly administer quality non-routine tasks to their students and provide feedback to students about their progress, it becomes possible to distinguish those aspects of student performance that are more resistant to change from those that are less so.

One aspect of student performance where there are real opportunities to foster improved learning behaviors is that of tenacity—student readiness to stay with non-routine problems. If students can be led to recognize and accept that it often takes time and effort to know what to do when they look at a task, they are less likely to give up prematurely. Teachers can foster this recognition and acceptance by giving the following directions each time their students are asked to work on a non-routine assessment task:

This task is designed to assess how well you can solve non-routine problems.

You will not have learned how to solve this problem in class. But you will have learned the mathematics that you will need to solve this problem.

Remember that when you look at this task for the first time you will probably not know what to do. This task is designed to see what you do when you don't know immediately what to do. So don't give up immediately—read the question again and again. Try to say in your own words what you are being asked to do.

Teachers report that some variation on this theme was important in focusing student attention on the task, reducing student frustration, and increasing student tenacity.

### **Assessment practice makes perfect—developing communication**

Communication is one aspect of student performance that is frequently difficult to develop. It is common to see a group of students making substantial inroads in solving a challenging task, where classroom discourse is characterized by focused mathematical discussion and quality thought. It is disappointing to read student responses later, only to find that little of their engaging work has been committed to paper. Even when the work is recorded, it is often difficult to see complete chains of thought.

To encourage students to communicate more effectively and so earn the credit their work suggests they deserve, we recommend providing students might be provided with the following opportunities:

- to score other student responses to tasks, trying to follow the line of reasoning, and to provide feedback to the student;

- to be given true mathematical statements and asked to explain why they are true;
- to represent ideas using more than one form of mathematical representation;
- to represent mathematical ideas using their own words and to practice writing down the main tenets of these ideas.

When these strategies are used with students, the students gain a better understanding of how to communicate their efforts. For example, when students scored other students' work, they were provided with a model of different levels of communication and were able to use this model to evaluate the effectiveness of responses. When students were asked to say why a given statement was true, they were relieved of the manipulative challenge of the task and could concentrate entirely on communicating their understanding.

### What can be done—some recommendations

This section presents recommendations for those who are interested in using assessment to enhance instruction, including several recommendations advanced by Black and Wiliam (1998), who make the centrally important point that *learning is driven by what teachers and students do in classrooms*.

Black and Wiliam draw upon a great number of research studies to argue persuasively that to enhance learning, specific attention must be paid to formative assessment. This is an aspect of teaching that Black and Wiliam posit as indivisible from effective teaching, and indeed as *the heart of effective teaching*:

We use the general term *assessment* to refer to all those activities undertaken by teachers—and by their students in assessing themselves—that provide information to be used as feedback to modify the teaching and learning activities. Such assessment becomes *formative assessment* when the evidence is actually used to adapt the teaching to meet student needs [italics in original]. (Black & Wiliam, p. 140)

Nonetheless, for formative assessment to be effective, it cannot be simply bolted on to existing practice. Instead, there is a need for a radical re-conceptualization of what teachers and students do in the classroom. Teachers will need a great deal of support as they attempt to rethink and restructure their classroom practice. Here are some key recommendations for enhancing instruction and learning:

- Provide teachers with a rich and varied supply of worthwhile assessment tasks to be used as classroom-embedded instruction and that will provide students with opportunities to perform.

- Encourage schools to move toward the use of high-quality, end-of-course assessments that are standardized across an entire school, district, or state. This will help teachers appreciate the importance of teaching to a set of publicly agreed upon and challenging standards.
- Encourage teachers, parents, and students to de-emphasize grades and emphasize feedback to students.
- Provide professional development for teachers that will help them provide their students with useful and constructive feedback that can improve learning rather than compare or rank students.
- Structure professional development to enable teachers to recognize and appreciate student growth. All students can learn to complete challenging mathematics tasks. Student work that demonstrates growth in tenacity, communication, and procedural, conceptual, and strategic knowledge should be generated and shared with teachers in the school, district, or state.
- Work with teachers to develop a view of the student as an active rather than as a passive learner.
- Provide professional development that will enable teachers to incorporate student self-assessment as a useful tool for learning.
- Provide teachers with tasks for class and homework that are aligned with standards or learning expectations.
- Demonstrate that all students can learn mathematics (either by using videos or student work that shows growth). This is important in encouraging teachers to regard students as having potential to be tapped rather than having innate inability.
- Encourage students and teachers to become willing participants in a diagnostic approach to learning, where errors and misconceptions are exposed and resolved, rather than left as unacknowledged and invisible obstacles to learning.
- Create approaches to learning where students are given time to communicate, to explore, to receive feedback on, and to re-orient their evolving understanding. In such classrooms, students can work for mathematical understanding rather than only for coverage.
- Develop approaches to curriculum adoption that are coherent within and across grades.
- Enable teachers to use a curriculum that encourages a more integrated and connected approach to learning mathematics. Many of the curricula developed with funding from the National Science Foundation are excellent resources.



- Encourage teachers to avoid textbooks that take a superficial approach to mathematical connections.
- Consider organizing teaching in a way that enables teachers to teach across an entire grade span. For example, a teacher might continue to teach the same cohort of sixth-grade students through grades seven and eight. This would provide continuity for students and enable teachers to develop a longitudinal view of the larger curriculum.

## Chapter 5: Alignment and standards-based assessments

---

Textbooks are a critical component of the curriculum. Supervisors and others involved in choosing textbooks need to ensure that the textbooks they adopt are aligned to the curricular guidelines or frameworks of their system. Today, however, those who choose textbooks also need to consider the type and quality of assessments that are either created or adopted by their districts or states. More and more districts and states are attempting to implement assessment systems that are aligned to standards (e.g., Chicago, Maryland, Texas). Unfortunately, many who are involved in this process have no clear models of what an aligned system might look like.

When it comes to evaluating whether a particular assessment is aligned to a given set of standards, it is really not all that useful to rely solely on the test vendor's evaluation. It is far too tempting for the vendor to think creatively and envisage the enormous number of ways in which the test might meet the needs of the client. Just as it is possible to link a given textbook to various state standards in a cursory way (or to link a whole range of standards to a single textbook) and call it standards-based, it is also possible to map any given test or assessment instrument to a wide range of standards in the same fashion. A far more sound way to proceed is for a test-selection committee to apply a set of alignment criteria and to make its own professional judgment about the degree of alignment between the standards and an assessment. Districts or states should want standards and assessment to be

aligned in a way that encourages students to learn the mathematics specified in the standards.

Standards-based assessment should improve instruction and learning, rather than sustain current inadequacies or present new ones. Indeed, if a standards-based assessment program were simply to reinforce the status quo, it would do little or nothing to enhance the learning of mathematics (NCTM, 1989, 1991, 1995; NRC, 1993b).

This chapter draws extensively on Norman Webb's *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997) to illustrate the most salient issues in aligning standards and assessments. Webb provides a coherent way of thinking about and evaluating the alignment of standards and assessments. He provides useful criteria that can be used by district or state assessment adoption committees. This chapter highlights some of Webb's criteria, applies them to the *New Standards Performance Standards* and corresponding *Reference Examinations*, and discusses more general issues that might arise in their application.

## The alignment of standards and assessment

There are two overarching components that need to be addressed by any well-formulated alignment analysis. First and foremost, alignment criteria are needed to evaluate *Balance* and *Equity and Fairness*.

### Balance in learning and assessment

There is usually little argument that balance is a critical consideration for learning and assessment. Balance can help ensure that assessments no longer confine themselves to those aspects of the curriculum that are easy to assess.

Webb's criteria for alignment of expectations and assessments in mathematics and science education (Webb, 1997) provide a comprehensive and accessible guide to those seeking systemic reform through standards and standards-based assessments. Six of Webb's categories that can contribute to an evaluation of balance are concerned with the following aspects: categorical concurrence, depth of knowledge, range of knowledge, structure of knowledge, balance of representation, and dispositional consonance. Taken as a group, these categories provide a powerful tool kit for qualitative alignment analysis. Each category has well-specified agreement criteria that can be used by professionals in their alignment evaluations.

In Webb's system, *categorical concurrence* describes the degree of agreement between the content categories of the standards and

the content categories of the assessment. According to Webb, alignment of categories would be judged as *stringent* when the organizing categories of the standards are the same as the reporting categories of the assessment.

By way of example, the alignment between the organizing categories of the *New Standards Performance Standards* and the reporting categories of the *New Standards Reference Examination* is illustrated by the following diagram:

<b>New Standards Performance Standards</b>	<b>Reference Examination Reporting Category</b>
1. Number and Operation	Conceptual Understanding
2. Geometry and Measurement	
3. Functions and Algebra	
4. Statistics and Probability	
6. Mathematical Skills and Tools	Mathematical Skill
5. Mathematical Problem Solving and Mathematical Reasoning	Mathematical Problem Solving
7. Mathematical Communication	
8. Putting Mathematics to Work	

There is clearly no stringent alignment between these performance standards and the reporting categories. For example, *Putting Mathematics to Work* cannot be assessed using the *Reference Examination* because this standard asks students to use mathematics to complete extended tasks and projects. This standard can instead be assessed using the New Standards' portfolio system, or any other portfolio system that assesses the ways in which students can put mathematics to work. Furthermore, the *Reference Examination* does not report separate scores for each of the content standards, nor does it report a score for mathematical communication. Instead, the content standards contribute to a combined conceptual understanding score, while the communication standard is assessed as part of the mathematical problem-solving score and also by a portfolio system. If a reliable score were to be reported for each of the standards individually, the examination would need to incorporate many more items than it currently contains, and it would need to be considerably longer than its current approximate three-hour length (Shavelson, Gao, & Baxter, 1993; Linn, 1994). As an alternative, each *New Standards Reference Examination* reports three scores that are statistically reliable, and the alignment between these and the categories of the *New Standards Performance Standards* would rate as *acceptable* in Webb's system.

Webb's focus on the alignment of categories establishes a conceptual basis for rudimentary evaluation of the link between assessment and standards. Such an evaluation can be particularly useful to a curriculum or assessment specialist who is in the process of selecting an examination or assessment system that is aligned to the school's, district's, or state's standards. Suppose, for example, that an examination offers stringent alignment between its reporting categories and the standards. If there are more than three or four reporting categories, it would be important to ask, "what are the *costs* of this stringent alignment?" Or, "what trade-offs are being made here?" It could be that stringent alignment of categories is achieved by using a large number of only very short exercises on the examination. Thus, the assessment might contain no complex tasks similar to those described in Chapters 2 and 3. The tradeoff would be between the variety and complexity of tasks that students will be expected to complete and stringent alignment of categories. For many, this particular tradeoff would be unacceptable because it very often sends the message to those involved in teaching and learning that it is *not* important for students to solve complex tasks. Alignment criteria should consider the necessary tradeoffs implied by choices and evaluate their consequences for teaching and learning.

Webb's *Depth of Knowledge* category judges how well the cognitive demands of the assessments match the expectations outlined in the standards on which they are based. This category calls for an expert judgment on whether the assessments contain tasks comparable to what the standards suggest that they should be able to do. If an assessment consisting entirely of multiple-choice items were constructed to assess the expectations of either the NCTM *Standards* or the *New Standards Performance Standards*, it is quite unlikely that the cognitive demands of the communication component of either of these standards could be realized.

Beck (1998) defines criteria to be used in evaluating the alignment between tests and standards. One of these criteria also judges the match between the cognitive demand of the task and the cognitive demand of the standards. Both Beck and Webb take care to ensure that standards-based assessments call for a level of mathematical functioning that is just as cognitively demanding as are the standards on which the assessment is based. This category is important to implement for any system seeking to improve upon the *mile wide, inch deep* nature of many mathematics programs and to instill a system of instruction and learning that is more reflective of current reform principles (e.g., NCTM, 1989; NRC, 1993b).

If both the adopted standards and the corresponding assessments call upon students to go beyond short procedural exercises and template problems, the hope is that classrooms will then

provide the opportunity for students to learn and to practice doing the kinds of things that both the standards and the assessments are asking them to do. High expectations are for *all* students, and conditions need to be created to enable all students to tackle mathematical problems that are challenging and reflect the depth of the standards.

When a mathematics curriculum or assessment specialist is evaluating the depth category in selecting an examination, to understand the demand of the tasks, it is important to work through each task, paying attention to the time taken and the mathematics assessed by each one. This also will provide some idea about the opportunity to perform that is afforded by the examination. Frequently, tasks look easier than they really are.

The *Range of Knowledge* category evaluates bias in the way the components of an assessment sample the expectations expressed in the various components of the corresponding standards. To be aligned, the range of expectations should be reflected in the range of assessments.

The range of knowledge category serves as a reminder that the main function of standards-based assessment is not to discriminate between or sort students but to determine the gap between what students know and can do and what the standards are asking them to do. In some cases there may be no gap, which does not mean that there is something wrong with the assessment or the standards. This category should remind us that standards and standards-based assessments offer a radically different way of thinking about and conceptualizing our notions of assessment.

The range of knowledge category also ensures that assessments do not ignore those aspects of standards that are important and yet are straightforward and can be easily demonstrated by all students. Standards-based assessment developers and administrators need not worry if there is a ceiling effect (that is, all students or almost all students demonstrate success) associated with a particular item or set of items. This is to say, even if all students demonstrate success on a particular item, it does not mean that the item has been placed on the assessment in error. As long as the item can be mapped onto the standards and evaluated as assessing a depth of knowledge that is consistent with that required by the standards, then ceiling effects can be celebrated.

In the instrument that Beck devises to evaluate the alignment of examinations to standards, she also considers range of knowledge. In her instrument, sets of items that are mapped to each standard are judged by how well they reflect the challenge of that standard. In this way, Beck's range of challenge category is comparable to Webb's range of knowledge category.

The *Structure of Knowledge* category goes to the heart of the changes in mathematics education reflected in the NCTM *Standards*. If a standards-based assessment is designed to align with the model for assessment presented in Chapter 2, the structure of knowledge category ensures that the assessment will incorporate problem solving, connections (both from within and from outside of mathematics), mathematical communication, and representation. If the complete set of assessments designed to assess problem solving were of the heavily scaffolded type described in Chapter 3, the structure of the assessments would not be comparable with the structure of the problem-solving standard (NCTM, 1989, 1998).

The structure of knowledge category shifts emphasis from the pieces to the overarching structure. It ensures that when the structure of knowledge in standards and assessments are aligned, the assessments will not permit success for students whose understanding of mathematics is fragile or based on disconnected fragments. If this shift is to lead to a corresponding shift in classroom practice, there will need to be a shift away from the tightly sequenced teaching and testing practices described in Chapter 4. This “teach it then test it” coupling of instruction and learning may help students do template problems or exercises, but it also inhibits them when they are called upon to apply their learning to challenging or non-routine tasks.

The *Balance of Representation* category requires that expectations be given comparable emphasis in both the standards and the assessment. The balance of representation category is important because it ensures that attributes of the standards are not simply paid lip service in the assessment but instead are assigned weights that reflect their actual importance. Beck’s alignment analysis includes a similar category, which she calls *content representativeness*. It is usually not feasible for a timed test to assess everything. Instead, the assessment designer must sample the entire domain. Beck’s category specifies:

A set of items mapped to a particular standard is judged as content representative to the degree that the elements of the standard represented in the set of items are strongly connected to the elements that are not sampled directly in the examination. (Beck, 1998, p. 16)

It would be very difficult for any examination, taken all by itself, to satisfy completely the balance of representation category. Instead, it is important to include other forms of assessment that can probe aspects of problem solving, communication, and mathematical connections as part of an overall assessment plan. Extended tasks of the type found in *Measuring Up* (NRC, 1993a) and *High School Mathematics at Work* (NRC, 1998) would be useful in portfolio assessment or for the purpose of developing classroom-embedded

assessments. Such tasks can help avoid the narrowing of the curriculum that can occur when classroom practice is overly concerned with routinized preparation for the test.

The final category addressed under Webb's Content Focus concerns *Dispositional Consonance*. This category is designed to ensure that assessments do not simply ignore those aspects of disposition that the standards attempt to cultivate, including, for example, belief that mathematics is valuable and confidence in one's own ability. If aspects of disposition are important enough to be included in the standards, they deserve to be observed or monitored. And if they are never observed, monitored, or reported, it is very likely that they will fall through the cracks. Black and Wiliam (1998) found that student self-assessment was a useful component of formative assessment. Therefore, if standards expect students to become independent learners, then self-assessment will be essential in cultivating such a disposition.

The sense of balance derived from Webb's and Beck's work on alignment is essential if standards and assessment are to have a positive influence instruction and learning. Certainly, this sense of balance is necessary in providing opportunity to learn for all students. It also is essential to define a balanced set of curricular activities for all students and especially so for those who may have been restricted to a diet of short exercises. Bond makes this case strongly when he writes:

The concentration on teaching basic skills to disadvantaged students has blinded us as educators to the capabilities of such students for sophisticated thought and complex problem solving. The vast majority of students can only learn what they are taught, and can master only what they practice. They do not learn what they are not taught, and they do not master what they do not practice. (Bond, 1995, p. 23)

According to Bond, therefore, a balance in teaching is an imperative for the opportunity to learn.

### **Equity and fairness in learning and assessment**

Webb (1997) defines *Equity and Fairness* as being one of five general categories for judging alignment. This category, which reinforces the need for multiple measures of assessment, draws upon a series of research studies that show that the format of an assessment can have an adverse impact on students' opportunity to perform (Shavelson & Baxter, 1992; Shavelson, Gao, & Baxter, 1993; Shavelson, Webb, & Rowley, 1989). In addition to promoting multiple measures, Webb's *Equity and Fairness* alignment category focuses attention on the role of culture, ethnicity, and gender in restricting students' opportunity to perform.



Webb's equity and fairness category also embraces the concept of opportunity to perform, as defined and discussed in Chapter 3. There, issues of opportunity to perform were separated from those of opportunity to learn in an attempt to identify and take responsibility for those aspects of assessment tasks that inhibit opportunity to perform. Broadly, opportunity to perform was defined as the opportunity created by the task for the students to show what they had learned.

The assessment research and experience described in Chapter 3 identified a multitude of factors that must be recognized and controlled to develop quality assessments. In particular, total cognitive load, task miscue, overuse of scaffolding, contextual challenge, and over-zealous assessment can count against students' opportunity to perform. This research and experience reinforces Shavelson and Baxter's conclusion:

Performance assessments are very delicate instruments. They need to be carefully crafted.... Shortcuts taken in development of these assessments will produce poor measuring devices. (Shavelson & Baxter, 1992, p. 24)

In Beck's alignment instrument, source of challenge is defined as an important aspect that directs attention toward issues of opportunity to perform. She argues:

In a set of items with appropriate sources of challenge, the greatest challenges in the items lie in the mathematics targeted by the standard (as opposed, for example, to challenges of reading comprehension or interpreting item context). (Beck, 1998, p. 13)

Because all students can and should benefit from studying to prepare for standards-based assessments, it is important that assessments do not (advertently or inadvertently) create tasks that trick or trap. Esoteric abbreviations, contexts, reading comprehension, or any of the other barriers to performance identified in Chapter 3 should not prevent students from being able to showcase their understanding. Similarly, it is important that assessments be designed so that success depends upon mathematical understanding and not *only* upon unusual talent or special insight.

Bond reminds us that it is not enough only to eliminate bias that counts against some portions of the population:

It is important to note that specific issues of bias and fairness, and the more general issues of unintended negative consequences, involve not only the elimination of elements in assessment that unduly *disadvantage minority persons* but also the elimination of construct irrelevant elements that may subtly *advantage majority persons over others* [italics in original]. (Bond, 1995, p. 23)

## Using alignment criteria—some recommendations

In a state or district where there is an aligned system of instruction and learning, the challenge is to activate those opportunities for learning that are afforded by the alignment of standards and assessment. Webb writes:

Through understanding the link between expectations and assessment, teachers are more likely to find ways to translate what is being advanced ... into their daily work with students. (Webb, 1997, p. 2)

The important connection described here is between standards and assessment and *not* between instruction and assessment. When standards and assessment are aligned, appropriate instruction can be delineated. When instruction and assessment are aligned *without* regard to learning expectations, there is a danger that instruction will become too narrowly focused on the assessment rather than on the depth, range, structure, of the learning expectations themselves. Here are some recommendations that can be used in aligning standards and assessments:

- Recognize that tradeoffs will always need to be made to achieve alignment.
- Evaluate how any one tradeoff might affect students' opportunity to learn.
- Evaluate how any one tradeoff might affect equity and fairness.
- When adopting an assessment that is aligned with standards, make sure that the adoption committee works from the standards to the assessment, rather than from the assessment to the standards. This will help ensure that important aspects of the standards are not inadvertently overlooked.
- Use multiple assessment measures whenever possible. A single examination is seldom sufficiently fair and equitable.
- Introduce teachers to alignment criteria through professional development. When teachers understand the link between standards and assessment, they are more likely to align their teaching to both standards and assessment. A useful exercise is to compare different assessment systems to a single set of standards.
- Ensure that all involved are aware that when instruction is aligned to assessment rather than to standards, there is the danger of the curriculum becoming overly narrow. When the depth, range, structure, and balance of the standards are reflected in the assessment system, the curriculum can, in turn, be designed to reflect the depth, range, structure and balance of the standards.

The whole instruction and learning system can be seen as a bridge that is vital in closing the gap between what the standards expect and where assessments show the majority of students to be.

The comprehensive picture that emerges from Webb's alignment document and Black and Wiliam's vision for effective teaching (discussed in Chapter 4) is a rounded picture of the components that are necessary to ensure opportunity to learn for all students. The vision created by these two important recent publications is an exciting one.

In classrooms informed by this vision, students would work on a variety of tasks whose range would extend well beyond the range of tasks that are likely to appear in the on-demand component of the assessment plan. There would be a balance of work on skills, conceptual understanding, and problem solving. Perhaps more students would develop their repertoire of basic mathematical skills as they worked on challenging non-routine problems. There would be fewer instances of tightly sequenced instruction and assessment, and more students who put their mathematics to work in ways that demonstrate an integrated and robust understanding.

In such classrooms, teachers would be freer from the constraints of a fragmented and cluttered curriculum and would be more committed to teach for understanding rather than coverage alone.

In a world informed by this vision, parents might not worry as much about grades and about how well their students compared to other students. Instead, they might worry whether the tests that were administered to their children were aligned to the standards, and whether their children had been given opportunities to learn and to perform.

The primary role of assessment would not be to compare students to one another but to enable students to see how they perform in relation to a balanced, publicly negotiated, and challenging set of standards. Rather than rank or sort, assessment could provide feedback to the students on how well they have learned what they are supposed to learn. As a result, students would have new understandings of what they need to do to learn the mathematics that the standards expect them to learn. It is worth remembering that the main function of alignment is not simply to improve assessment but to use assessment as a means of enhancing student learning.

## References

---

- Arcavi, A., Kessel, C., Meira, L., & Smith, J. (1998). Teaching mathematical problem solving: A microanalysis of an emergent classroom community. In A. Schoenfeld, E. Dubinsky, & J. Kaput (Eds.), *Research in Collegiate Mathematics Education III* (pp. 1-70). Providence, RI: American Mathematical Society.
- Beck, P. (1998). Alignment analysis: Assessment to standards. Unpublished manuscript.
- Bell, A. (1993). Some experiments in diagnostic teaching. *Educational Studies in Mathematics*, 24, 115-137.
- Bell, A.W., Swan, M., Onslow, B., Pratt, K., & Purdy, D. (1985). *Diagnostic teaching: Teaching for long term learning. Report of ESRC Project HR 8491/1*, Nottingham, UK: University of Nottingham, Shell Centre for Mathematical Education.
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, 14(4), 21-24.
- Borasi, R. (1996). *Reconceiving mathematics instruction: A focus on errors*. Norwood, NJ: Ablex.
- Close, G. (1996). *Developing criterion-referenced, open-response, national test items*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, San Diego, CA.
- Cobb, P., & Lampert, M. (1998). *Communication*. White paper prepared for the National Council of Teachers of Mathematics. Reston, VA.

- Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational Psychologist*, 23, 167-180.
- Floden, R.E. (1996). *Teachers' choices about content: The Standards in use*. Unpublished paper prepared for November 1996 symposium sponsored by the Board on International Comparative Studies in Education of the National Research Council. Washington, DC.
- Graeber, A.O., & Campbell, P.F. (1993). Misconceptions about multiplication and division. *Arithmetic Teacher*, 39, 408-411.
- Greeno, J., & Hall, R. (1997). Practicing Representation: Learning with and about representational forms. *Phi Delta Kappan*, 78, 361-367.
- Hartocollis, A. (1999, April 8). In Bronx, Regents coaching is extended to teachers, too. *New York Times*, pp. B1, B3.
- Heid, M.K. (1988). Resequencing skills and concepts in applied calculus using the computer as a tool. *Journal for Research in Mathematics Education*, 19(1), 3-25.
- Henningsen, M. & Stein, M.K. (1996). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning.
- Hiebert, J. (1999). Relationships between research and the NCTM *Standards*. *Journal for Research in Mathematics Education*, 30(1), 3-19.
- Hiebert, J. & Carpenter, T. (1992). Learning and teaching with understanding. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*, (p. 65-97). New York: Macmillan.
- Hiebert, J. & Wearne, D. (1996). Instruction, understanding and skill in multidigit addition and subtraction. *Cognition and Instruction*, 14(3), 251-283.
- Jakwerth, P.R., Stancavage, F.B., & Reed, E.D. (1999). *An investigation of why students do not respond to questions*. Report commissioned by the NAEP Validity Studies (NVS) Panel. Palo Alto, CA: American Institutes for Research.
- Learning First Alliance. (1998). *Every child mathematically proficient: An action plan*. Washington, DC: Author.
- Lesh, R., Lamon, S.J., Lester, F., & Behr, M. (1992). Future directions for mathematics assessment. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 379-425). Washington, DC: American Association for the Advancement of Science.
- Linn, R. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.
- Massell, D., Kirst, M., & Hoppe, M. (1997). *Persistence and change: Standards-based reform in nine states*. Brunswick, NJ: Consortium for Policy Research in Education.
- National Assessment Governing Board. (1995). *Mathematics framework for the 1996 National Assessment of Educational Progress*. Washington, DC: Author.

- National Center on Education and the Economy. (1998). *Core assignment: Volume*. Washington, DC: Author.
- National Center on Education and the Economy & University of Pittsburgh. (1997). *New Standards performance standards*. (Volumes 1-3). Washington, DC: Author.
- National Center on Education and the Economy & University of Pittsburgh. (1997-1999). *New Standards mathematics reference examinations*. Washington, DC: Author.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1998). *Principles and standards for school mathematics: Discussion draft*. Reston, VA: Author.
- National Council of Teachers of Mathematics & National Research Council. (1997). *Improving student learning in mathematics and science: The role of national standards and state policy*. Washington, DC: National Academy Press.
- National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.
- National Research Council. (1990). *Reshaping school mathematics: A philosophy and framework for curriculum*. Washington, DC: National Academy Press.
- National Research Council. (1991). *For good measure: Principles and goals for mathematics assessment*. Washington, DC: National Academy Press.
- National Research Council. (1993a). *Measuring up: Prototypes for mathematics assessment*. Washington, DC: National Academy Press.
- National Research Council. (1993b). *Measuring what counts: A conceptual guide for mathematics assessment*. Washington, DC: National Academy Press.
- National Research Council (1997). *Learning from TIMMS: Results of the Third International Mathematics and Science Study, summary of a symposium*. A. Beatty, (Ed.), Board on International Comparative Studies in Education, Board on Testing and Assessment, Committee on Science Education K-12, and Mathematical Sciences Education Board. Washington, DC: National Academy Press.
- National Research Council. (1998). *High school mathematics at work: Essays and examples for the education of all students*. Washington, DC: National Academy Press.

- Porter, A.C., Kirst, M.W., Osthoff, E.J., Smithson, J.L., & Schneider, S.A. (1993). *Reform up close: A classroom analysis*. Final report to the NSF on Grant No. SPA-8953446 to the Consortium for Policy Research in Education. Madison, WI: Wisconsin Center for Education Research.
- Resnick, L., & Nolan, K.J. (1995). From aptitude to effort: A new foundation for our schools. *Daedalus*, 124(4), 55-52.
- Romberg, T.A., Zarinnia, E.A., & Williams, S.R. (1990). Mandated school mathematics testing in the United States: A survey of state mathematical supervisors. Madison, WI: National Center for Research in Mathematical Sciences Education.
- Schmidt, W.H., & Cogan, L.S. (1999). The rest of the story: Putting the U.S. twelfth grade TIMSS mathematics achievement results in perspective. *Focus on Calculus: A Newsletter for the Calculus Consortium Based at Harvard University*, No. 16, 6-8.
- Schmidt, W.H., McKnight, C.C., & Raizen, S.A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer.
- Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in mathematics*. Dordrecht, The Netherlands: Kluwer.
- Schoen, H.L., & Ziebarth, S.W. (1998). Assessment of students' mathematical performance (A Core-Plus Mathematics Project Field Test Progress Report). Iowa City, IA: Core-Plus Mathematics Project Evaluation Site, University of Iowa.
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.
- Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 189-215). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Schoenfeld, A.H. (1988). When good teaching leads to bad results: The disasters of "well taught" mathematics classes. *Educational Psychologist*, 23, 145-166.
- Schoenfeld, A.H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for Research in Mathematics Education*, 20, 338-355.
- Schoenfeld, A.H., Burkhardt, H., Daro, P., & Stanley, R. (1993). *A framework for balance*. Unpublished manuscript, Balanced Assessment & New Standards projects.
- Schoenfeld, A., Burkhardt, H., Schwartz, J., & Wilcox, S.J. (1999). *Balanced assessment for the mathematics curriculum* (8 packages). Menlo Park, CA: Dale Seymour Publications.

- Shannon, A., & Zawojewski, J. (1995). Mathematics performance assessment: A new game for students. *Mathematics Teacher*, 88(9), 752-757.
- Shavelson, R.J., & Baxter, G.P. (1992). What we've learned about assessing hands-on science. *Educational Leadership*, 49(8), 21-25.
- Shavelson, R.J., Gao, X., & Baxter, G.R. (1993). *Sampling variability of performance assessments* (CSE Tech. Rep. No. 361). Santa Barbara, CA: National Center for Research in Evaluation, Standards and Student Testing.
- Shavelson, R.J., Webb, N.M., & Rowley, G. (1989). Generalizability theory. *American Psychologist*, 44, 922-932.
- Stein, M.K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50-80.
- Stigler, J., & Hiebert, J. (1997). Understanding and improving classroom mathematics instruction: An overview of the TIMMS Video Study. *Phi Delta Kappan*, 79(1), 14-21.
- Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers.
- Wilson, J.W., Fernandez, M.L., & Hadaway, N. (1993). Mathematical problem solving. In Wilson, P.S. (ed.), *Research ideas for the classroom: High school mathematics*. New York: Macmillan.
- Zucker, A.A., & Esty, E.T. (1993). Promoting discourse in mathematics classrooms using a new video series for middle schools. Paper presented as part of a symposium entitled "The Potential of Video-Based Materials to Promote Classroom Discourse in Mathematics," at the annual meeting of the American Educational Research Association, Atlanta, GA.