Assessing Mathematical Proficiency
MSRI Publications
Volume **53**, 2007

Preface

Mathematics tests — and more broadly, assessments of students' mathematical proficiency — play an extremely powerful role in the United States and other nations. California, for example, now has a High School Exit Examination, known as CAHSEE. If a student does not pass CAHSEE, he or she will not be awarded a diploma. Instead, the four years that the student has invested in high school will be recognized with a certificate of attendance. In many states, annual examinations in mathematics and English Language Arts are used to determine whether students at any grade level will advance to the next grade.

Tests with major consequences like those just described are called high-stakes tests. Such tests raise myriad questions. What kinds of understandings do they test? Do they capture the kinds of mathematical thinking that is important? Are they equitable? Is it fair to let a child's career hinge on his or her performance on one particular kind of assessment? Do such tests reinforce and perhaps exacerbate patterns of discrimination, further penalizing students who already suffer from attending "low-performing" schools? Will they increase dropout rates, because students who see themselves as having no chance of passing a high-stakes exam decide to leave school early? Or, can such tests be levers for positive change, compelling attention to mathematics instruction and helping to raise standards?

These are complex issues, all the more so because of the multiple roles that the assessment of students' mathematical proficiency plays in the U.S. educational system, and the multiple groups that have an interest in it. As the chapters in this volume reveal, many different groups — mathematicians, researchers in mathematical thinking and learning, students and teachers, administrators at various levels (mathematics specialists, principals, school-district and state-level superintendents of education), policy-makers (some of the above; state and federal officials and legislators), test-makers and test-consumers (parents, college admissions officers, and more) all have strong interests in the information that mathematics assessments can provide. At the same time, many of these groups have different needs. Is the result of interest a number (How is this school or this district doing? How does this student rank against others?), or is it a profile (This

is what the student knows and can do; these are things he or she needs to work on)? Does it matter whether the score assigned to a student or school is legally defensible? (In some contexts it does, and that imposes serious constraints on the kind of reporting that can be done.)

Beyond these complexities is the fact that many of the groups mentioned in the previous paragraph have little knowledge of the needs of the other groups. For example, what relevance do "reliability" or "validity" have for teachers and mathematicians? There is a good chance that they would not know the technical meanings of the terms — if they have heard them at all. Yet, to the test manufacturer, an assessment without good reliability and validity data is worthless — the makers of the SAT or GRE would no sooner release such a test than a pharmaceutical company would release a drug that had not undergone clinical trials. Similarly, the policy-maker who demands that "high standards" be reflected by an increase in average scores or the percentage of students attaining a one-dimensional passing score may not understand teachers' needs for "formative diagnostic assessments," or how disruptive "teaching to the test" can be if the test is not aligned with an intellectually robust curriculum.

What to assess in mathematics learning is not as simple as it might seem. In the 1970s and 1980s, research on mathematical thinking and learning resulted in a redefinition of thinking and learning in general [Gardner 1985] and of mathematical understanding in particular [DeCorte et al. 1996; Grouws 1992; Schoenfeld 1985]. This refined understanding of mathematical competency resulted in a new set of goals for mathematics instruction, for example those stimulated by the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics*, commonly known as the *Standards* [NCTM 1989]. This document, which delineated content and process desiderata for K–12 mathematics curricula, focused on four "process standards" at every grade level: problem solving, reasoning, making mathematical connections, and communicating mathematics orally and in writing. Such learning goals, which continue to play a central role in NCTM's refinement of the *Standards* [NCTM 2000], pose a significant challenge for assessment.

Although it may seem straightforward to measure how much algebra, or geometry, or probability a student understands, the issues involved in obtaining accurate measurements of students' content understandings are actually complex. Measuring students' abilities to solve problems, reason, and make mathematical connections is much more difficult. There are myriad technical and mathematical issues involved. Which mathematics will be assessed: Technical skills? Conceptual understanding? For purposes of reliable scoring, when are two problems equivalent? How can one compare test scores from year to year, when very different problems are used? (If similar problems are used year after

year, teachers and students learn what they are, and students practice them: problems become exercises, and the test no longer assesses problem solving.)

One measure of the complexities of the issue of assessment in general, and mathematics assessment in particular, is the degree of attention given to the topic by the National Research Council (NRC). The NRC's Board on Testing and Assessment has published a series of general reports on the issue; see, for example, *High Stakes: Testing for Tracking, Promotion, and Graduation* [NRC 1999]. Among the NRC publications focusing specifically on assessments that capture students' mathematical understandings in accurate ways are *Keeping Score* [Shannon 1999], *Measuring Up* [NRC 1993a], *Measuring What Counts* [NRC 1993b]. A more recent publication, *Adding It Up* [NRC 2001] provides a fine-grained portrait of what we might mean by mathematical proficiency. All of these volumes, alongside the large literature on mathematical thinking and problem solving, and volumes such as NCTM's *Principles and Standards* [NCTM 2000], point to the complexity of teaching for, and assessing, mathematical proficiency.

Added to the intellectual challenges are a series of social, political, and ethical challenges. A consequence of the "standards movement" has been the establishment of high-stakes accountability measures — tests designed to see whether students, schools, districts, and states are meeting the standards that have been defined. Under the impetus of the No Child Left Behind legislation [U.S. Congress 2001], all fifty states have had to define standards and construct assessments to measure progress toward them. The consequences of meeting or not meeting those standards have been enormous. For students, failing a test may mean being held back in grade or being denied a diploma. Consequences for schools are complex. As has been well documented, African Americans, Latinos, Native Americans, and students in poverty have tended to score much worse on mathematical assessments, and to have higher dropout rates, than Whites and some Asian groups. One mechanism for compelling schools to focus their attention on traditionally lower-performing groups has been to disaggregate scores by groups. No longer can a school declare that it is performing well because its average score is good; all subgroups must score well. This policy raises new equity issues: schools with diverse populations now face more stringent requirements than those with homogeneous populations, and dire consequences if they fail to meet the requirements. Even before the advent of the No Child Left Behind legislation, various professional societies cautioned about the use of a single test score as the sole determinant of student success or failure: see, for example, the position paper by the American Educational Research Association [AERA 2000] and the updated paper by the National Council of Teachers of Mathematics [NCTM 2006].

There are also issues of how policy drives curricula. With high-stakes testing in place, many schools and districts take the conservative route, and teach to the test. If the assessments are robust and represent high standards, this can be a good thing. However, if the standards are high, schools run the risk of not making "adequate yearly progress" toward proficiency for all students, and being penalized severely. If standards are set lower so that they are more easily attainable, then education is weakened — and "teaching to the (narrow) test" may effectively lower standards, and limit what students learn [Shepard 2001].

In short, the issues surrounding mathematics (and other) assessments are complex. It may be that tests are simply asked to do too much. As noted above, many of the relevant stakeholders tend to have particular needs or interests in the assessments:

- Mathematicians want students to experience the power, beauty, and utility of mathematics. Accordingly, the mathematics represented in the tests should be important and meaningful, and not be weakened by technical or legal concerns (e.g., psychometric issues such as reliability, validity, and legal defensibility).
- Researchers, teachers, principals, and superintendents want tests that provide meaningful information about the broad set of mathematical skills and processes students are supposed to learn, and that provide diagnostic information that helps to improve the system at various levels.
- Psychometricians (and more broadly, test developers) want tests to have the measurement properties they are supposed to have — that the problems and their scoring capture what is important, that different versions of a test measure the same underlying skills, that they are fair in a substantive and legal sense.
- Parents and teachers want information that can help them work with individual students. Teachers also want information that can help them shape their instruction in general.
- Policy-makers want data that says how well their constituents are moving toward important social and intellectual goals.
- Cost, in terms of dollars and time, is a factor for all stakeholders.

Some of the goals and needs of some constituencies tend to be in conflict with each other. For example, multiple-choice tests that focus largely on skills can be easily graded, tend to be easy to construct in ways that meet psychometric criteria, and to provide aggregate statistics for purposes of policy — but, they are unlikely to capture the desired spectrum of mathematics, and provide little or no useful diagnostic information. Moreover, if there is a large amount of "teaching to the test," such tests may distort the curriculum.

Assessments that focus on a broad range of skills turn out to be expensive to grade, and are much more difficult to construct in ways that meet psychometric criteria — but, they can provide diagnostic information that is much more useful for teachers, and may be used for purposes of professional development as well. In addition, the vast majority of stakeholders in assessment are unaware of the complexities involved or of the needs of the other constituencies. That can lead to difficulties and miscommunication.

In short, different groups with a stake in the assessment of students' mathematical thinking and performance need to understand the perspectives and imperatives of the others. They need to seek common ground where it exists, and to recognize irreconcilable differences where they exist as well. To this end, in March 2004 the Mathematical Sciences Research Institute brought together a diverse collection of stakeholders in mathematics assessment to examine the goals of assessment and the varied roles that it plays. Major aims of the conference were to:

- Articulate the different purposes of assessment of student performance in mathematics, and the sorts of information required for those purposes.
- Clarify the challenges of assessing student learning in ways that support instructional improvement.
- Examine ethical issues related to assessment, including how assessment interacts with concerns for equity, sensitivity to culture, and the severe pressures on urban and high-poverty schools.
- Investigate different frameworks, tools, and methods for assessment, comparing the kinds of information they offer about students' mathematical proficiency.
- Compile and distribute a list of useful and informative resources about assessment: references, position papers, and sources of assessments, etc.
- Enlarge the community of mathematicians who are well informed about assessment issues and interested in contributing to high-quality assessments in mathematics.
- Articulate a research and development agenda on mathematics assessment.

This book is the product of that conference. Here is what you will find within its pages.

Section 1 of the book provides an orientation to issues of assessment from varied perspectives. In Chapter 1, Alan H. Schoenfeld addresses three sets of issues: Who wants what from assessments? What are the tensions involved? What other issues do we confront? This sets the stage for the contributions that follow. In Chapter 2, Judith A. Ramaley frames the issues of assessment in a broad way, raising questions about the very purposes of education. The underlying philosophical issue is, "Just what are our goals for students?" Then, as a

corollary activity, how do we know if we are meeting them? In Chapter 3, Susan Sclafani describes the current political context and national goals following from the No Child Left Behind Act (NCLB). The goals of NCLB are that all students will develop "fundamental knowledge and skills in mathematics, as well as in all core subjects." The mechanism for achieving these goals is premised on four basic principles: accountability for results, local control and flexibility, choice, and research-based practice. By virtue of these three introductory chapters, Section 1 identifies some of the major constituencies and perspectives shaping the varied and sometimes contradictory approaches to mathematics assessment found in the United States today.

Different notions of mathematical proficiency may underlie different approaches to assessment. Section 2 offers two perspectives regarding the nature of mathematical proficiency. Chapter 4, by R. James Milgram, describes one mathematician's perspective on what it means to do mathematics, and the implications of that perspective for teaching and assessing mathematics and understanding mathematics learning. Chapter 5, by Alan H. Schoenfeld, provides a top-level view of the past quarter-century's findings of research in mathematics education on mathematical thinking and learning. This too has implications for what should be taught and assessed. A comparison of the two chapters reveals that even the "basics" — views of the nature of mathematics, and thus what should be taught and assessed — are hardly settled.

Section 3 gets down to the business of assessment itself. At the conference organizers' request, the chapters in this and the following section contain numerous illustrative examples of assessment tasks. The organizers believe that abstract generalizations about mathematical competencies can often be misunderstood, while specific examples can serve to demonstrate what one cares about and is trying to do. Chapter 6, by Hugh Burkhardt, gives the lay of the land from the point of view of an assessment developer. It offers some design principles for assessment, discusses values (What should we care about?), the dimensions of performance in mathematics, and issues of the quality and cost of assessments. As in the chapters that follow, the examples of assessment items contained in the chapter show what the author values. There are modeling problems, and tasks that take some time to think through. Burkhardt has argued that "What You Test Is What You Get," and that in consequence assessment should not only capture what is valued but should help focus instruction on what is valued. Chapter 7, by Jan de Lange, lays out some of the foundations of the International Organisation for Economic Co-operation and Development Program for International Student Assessment (PISA) mathematics assessments. There is an interesting non-correlation between national scores on the PISA exams and scores on exams given in the Trends in International Mathematics and Science

Study, precisely because of their different curricular and content foci. Thus, de Lange's chapter, like the others, underscores the point that curriculum and assessment choices are matters of values. In Chapter 8, Bernard Madison expands the scope of assessment yet further. Madison reflects on the kinds of reasoning that quantitatively literate citizens need for meaningful participation in our democratic society. His examples, like others in this section, will certainly challenge those who think of mathematics assessments as straightforward skills-oriented (often multiple-choice) tests. In Chapter 9, Richard Askey provides a variety of assessment tasks, dating all the way back to an 1875 examination for teachers of mathematics in California. Some of Askey's historical examples are remarkably current — and, Askey would argue, more demanding than what is required of teachers today. In Chapter 10, David Foster, Pendred Noyce, and Sara Spiegel address the issue of teacher professional development through the use of student assessment. They describe the work of the Silicon Valley Mathematics Initiative, which uses well-designed assessments as a mechanism for focusing teachers' attention on the mathematics that students are to learn, and the kinds of understandings (and misunderstandings) that students develop.

Section 4 focuses on algebra. In Chapter 11, William McCallum takes a broad view of the "assessment space." McCallum gives examples of more advanced tasks that can be used to assess different strands of proficiency such as conceptual understanding and strategic competence. In Chapter 12, David Foster identifies some of the core skills underlying algebraic understandings — examples of algebraic "habits of mind" such as abstraction from computation, rule-building, and constructing and inverting processes ("doing-undoing"). Foster also discusses foundational concepts for young students transitioning to algebra, such as understanding and representing the concepts of equality. And he provides examples of assessment items focusing on these critically important topics. Chapter 13, by Ann Shannon, looks at issues of problem context in ways consistent with the approaches of Burkhardt and de Lange. Shannon discusses tasks that focus on understanding linear functions. Her chapter shows how very different aspects of reasoning can be required by tasks that all ostensibly deal with the same concept. Taken together, these chapters (and some examples from Section 3 as well) provide rich illustrations of the broad range of algebraic competency — and thus of things to look for in assessing it.

Section 5 focuses on fractions, with a slightly different purpose. The idea here is to see what kinds of information different kinds of assessments can provide. In Chapter 14, Linda Fisher shows how a close look at student responses to somewhat complex tasks provides a structured way to find out what one's students are understanding (and not). This kind of approach provides substantially more information than simple test results. Yet, as we see in Chapter 15, there is

no substitute for a thoughtful conversation with a student about what he or she knows. Chapter 15 provides the transcript of an interview with a student conducted by Deborah Ball at the conference (a video of the interview is available online). In Chapter 16, Alan H. Schoenfeld reflects on the nature and content of the interview in Chapter 15. He considers the detailed information about student understanding that this kind of interview can reveal, and what this kind of information implies — about the nature of learning, and about how different assessments can reveal very different things about what students understand.

Section 6 takes a more distanced view of the assessment process. To this point (with the exception of Section 1) the volume has focused on issues of which mathematics is important to assess, and what kinds of tasks one might give students to assess their mathematical knowledge. Here we open up to the issue of societal context — the fact that all assessment takes place within a social, political, and institutional set of contexts and constraints, and that those constraints and contexts shape what one can look for and what one can see as a result. In Chapter 17, Michèle Artigue presents a picture of the assessment system in France — a system very different from that in the U.S. (the educational system is much more centralized) and in which attempts at providing information about student understanding seem much more systematic than in the U.S. As always, a look outside one's own borders helps to calibrate what happens within them. In Chapter 18, Mark Wilson and Claus Carstensen take us into the realm of the psychometric, demonstrating the issues that one confronts when trying to build assessment systems that are capable of drawing inferences about student competence in particular mathematical domains. In Chapter 19, Lily Wong Fillmore describes the complexities of assessment when English is not a student's native language. If a student is not fluent, is his or her failure to solve a problem a result of not understanding the problem (a linguistic issue) or of not understanding the mathematics? Judit Moschkovich addresses a similar issue in Chapter 20, with a close examination of bilingual students' classroom work. Here she shows that the students have a fair amount of conceptual competency, while not possessing the English vocabulary to appear (to an English speaker) that they are very competent. This too raises the question of how one can know what a student knows — the real purpose of assessment. In Chapter 21, Elizabeth Taleporos takes us outside the classroom and into the realm of politics. The issue: How does one move a system toward looking at the right things? Taleporos describes her experiences in New York City. In Chapter 22, Elizabeth Stage discusses the systemic impact of assessments in California. Stage examines the ways in which the publicly released items on the California assessments shaped what teachers taught — sometimes for good, when testing identified consistent and important weaknesses in students' knowledge; sometimes for ill, when teaching to the test

resulted in a focus on less than essential mathematics. This brings us full circle, in that these issues address some of the critical concerns in Section 1, such as concerns about assessments mandated by the No Child Left Behind legislation.

At the end of the conference, participants reflected on what they had experienced and what they (and the field) needed to know. Eight working groups at the conference were charged with formulating items for a research agenda on the topic of the conference. The product of their work, a series of issues the field needs to address, is presented as an Epilogue. There is great variety in the chapters of this book, reflecting the diverse perspectives and backgrounds of the conference participants. As mentioned before, that mixing was intentional. The conference participants learned a great deal from each other. The organizers hope that some of the "lessons learned" have made their way into this volume.

References

[AERA 2000] American Educational Research Association, Position statement concerning high-stakes testing in PreK-12 education, July 2000. Available at http://edtech.connect.msu.edu/aera/about/policy/stakes.htm. Retrieved 7 Feb 2006.

[DeCorte et al. 1996] E. DeCorte, B. Greer, and L. Verschaffel, "Mathematics teaching and learning", pp. 491–549 in *Handbook of educational psychology*, New York: Macmillan, 1996.

[Gardner 1985] H. Gardner, *The mind's new science: A history of the cognitive revolution*, New York: Basic Books, 1985.

[Grouws 1992] Grouws, D. (editor), *Handbook for research on mathematics teaching and learning*, New York: Macmillan, 1992.

[NCTM 1989] National Council of Teachers of Mathematics, *Curriculum and evaluation standards for school mathematics*, Reston, VA: Author, 1989.

[NCTM 2000] National Council of Teachers of Mathematics, *Principles and standards for school mathematics*, Reston, VA: Author, 2000.

[NCTM 2006] National Council of Teachers of Mathematics, Position statement on high stakes testing, January 2006. Available at http://www.nctm.org/about/position_statements/highstakes.htm. Retrieved 24 Jun 2006.

[NRC 1993a] National Research Council (Mathematical Sciences Education Board), *Measuring up: Prototypes for mathematics assessment*, Washington, DC: National Academy Press, 1993.

[NRC 1993b] National Research Council (Mathematical Sciences Education Board), *Measuring what counts: A conceptual guide for mathematics assessment*, Washington, DC: National Academy Press, 1993.

[NRC 1999] National Research Council (Committee on Appropriate Test Use: Board on Testing and Assessment — Commission on Behavioral and Social Sciences and

Education), *High stakes: Testing for tracking, promotion, and graduation*, edited by J. P. Heubert and R. M. Hauser, Washington, DC: National Academy Press, 1999.

[NRC 2001] National Research Council (Mathematics Learning Study: Center for Education, Division of Behavioral and Social Sciences and Education), *Adding it up: Helping children learn mathematics*, edited by J. Kilpatrick et al., Washington, DC: National Academy Press, 2001.

[Schoenfeld 1985] A. H. Schoenfeld, *Mathematical problem solving*, Orlando, FL: Academic Press, 1985.

[Shannon 1999] A. Shannon, *Keeping score*, National Research Council: Mathematical Sciences Education Board, Washington, DC: National Academy Press, 1999.

[Shepard 2001] L. Shepard, Protecting learning from the harmful effects of high-stakes testing, paper presented at the annual meeting of the American Educational Research Association, Seattle, April 2001.

[U.S. Congress 2001] U.S. Congress, H. Res. 1, 107th Congress, 334 Cong. Rec. 9773, 2001. Available at http://frwebgate.access.gpo.gov.