Section 1
The Big Picture

This introductory section establishes the context for what follows in this volume. In broad-brush terms, it does so by addressing three P's: pragmatics, philosophy, and policy.

Alan H. Schoenfeld opens with a discussion of pragmatics in Chapter 1. A simple question frames his contribution: Who wants what from mathematics assessments? As he shows, the issue is far from simple. It is true that at some level everyone involved in mathematics testing is interested in the same basic issue: What do students know? However, various groups have very different specific interests. For example, a teacher may want to know what his or her particular students need to know in order to improve, while a state superintendent may be most interested in trends and indications of whether various performance gaps are being closed. Other groups have other interests. Those interests are not always consistent or even compatible. Understanding who has a stake in assessment, and what different groups want to learn from it, is part of the big picture — the picture that is elaborated throughout the volume, as representatives of different stakeholder groups describe the kinds of information that they need and that carefully constructed assessments can provide. Schoenfeld also looks at the impact of assessment. Tests reflect the mathematical values of their makers and users. In the United States, tests are increasingly being used to drive educational systems — to measure performance aimed at particular educational goals. This is the case at the national level, where various international assessments show how one nation stacks up against another; at the state level, where individual states define their intended mathematical outcomes; and at the individual student level, where students who do not pass state-mandated assessments may be retained in grade or denied a diploma. Schoenfeld addresses both intended and unintended consequences of assessments.

Judith A. Ramaley's discussion in Chapter 2 addresses the second P, philosophy. As indicated in the previous paragraph, testing reflects one's values and goals. The issue at hand is not only "What is mathematics," but "Which aspects of mathematics do we want students to learn in school?" Is the purpose of schooling (and thus of mathematics instruction in school) to provide the skills needed for successful participation in the marketplace and in public affairs? Is

it to come to grips with fundamental issues of truth, beauty, and intellectual coherence? To use some common jargon, these are consequential decisions: the answers to questions of values shape what is taught, and how it is taught. As Ramaley notes, the Greeks' two-thousand-year-old philosophical debates lie at the heart of today's "math wars." But, as she also observes, science brings philosophy into the present: questions of "what counts" depend on one's understanding of thinking and learning, and of the technologies available for assessing it. Discussions of what can be examined, in what ways, bring us firmly into the twenty-first century.

Susan Sclafani's contribution in Chapter 3 brings us into the policy arena. Having certain goals is one thing; working to have a complex system move toward achieving those goals is something else. At the time of the MSRI conference Sclafani served as Assistant Secretary in the Office of Vocational and Adult Education of the U.S. Department of Education. One of her major concerns was the implementation of the No Child Left Behind Act (NCLB), federal legislation that mandates the development and implementation of mathematics assessments in each of the nation's fifty states. The creation of NCLB was a political balancing act, in which the traditional autonomy granted to states on a wide range of issues was weighed against a federal interest in policies intended to have a beneficial impact on students nationwide. How such issues are resolved is of great interest. In NCLB states are given particular kinds of autonomy (e.g., what is tested is, in large measure, up to the states) but they are subject to national norms in terms of implementation. Broadly speaking, NCLB mandates that scores be reported for all demographic groups, including poor students, English as a Second Language students, various ethnic groups, and more. In order for a school to meet its state standard, every demographic group with 30 or more students in the school must meet the standard. In this way, NCLB serves as a policy lever for making sure that under-represented minority groups cannot slip through the cracks.

In sum, then, the philosophical, pragmatic, and policy discussions in the three chapters that follow establish the overarching context for the more detailed discussions in the core of the volume.

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

Chapter 1
Issues and Tensions in the Assessment of Mathematical Proficiency

ALAN H. SCHOENFELD

Introduction

You'd think mathematics assessment — thought of as "testing" by most people — would be simple. If you want to know what a student knows, why not write (or get an expert to write) some questions about the content you want examined, give those questions to students, see if the answers are right or wrong, and add up a total score? Depending on your predilections (and how much time you have available) you might give a multiple-choice test. You might give an "open answer" or "constructed response" test in which students show their work. You could give partial credit if you wish.

This version of assessment fits with most people's experiences in school, and fits with descriptions of the National Assessment of Educational Progress (NAEP) as "the nation's report card." From this perspective, mathematics assessment — discovering what mathematics a person (typically, a student) knows — seems straightforward.

Would that things were so simple. As this essay and later contributions to this volume will indicate, different groups can have very different views of what "counts," or should count, in mathematics. Assessing some aspects of mathematical thinking can be very difficult — especially if there are constraints of time or money involved, or if the tests have to have certain "psychometric" properties (discussed further in this essay) in order to make sure that the test-makers stand on legally safe ground. Different groups may want different information from tests. And, the tests themselves are not neutral instruments, in the ways that people think of thermometers as being neutral measures of the temperature of a system. In many ways, tests can have a strong impact on the very system they measure.

This essay introduces and illustrates such issues. I begin by identifying a range of "stakeholder" audiences (groups who have an interest in the quality or outcomes) for mathematics assessments, and identifying some of the conflicts among their interests. I proceed with a discussion of some of the side effects of certain kinds of large-scale testing. These include: test score inflation and the illusion of competence; curriculum deformation; the stifling of innovation; the disenfranchising of students due to linguistic or other issues; and a possible impact on drop-out rates.

My purpose is to lay out some of the landscape, so that the varied groups with a stake in assessing mathematical proficiency (nearly all of us!) can begin to understand the perspectives held by others, and the kinds of issues we need to confront in order to make mathematics assessments serve the many purposes they need to serve.

Who Wants What from Mathematics Assessments?

Here, in brief, are some assertions about what various communities want from mathematics assessments. It goes without saying that these communities are not monolithic, and that my assertions are simple approximations to complex realities.

Mathematicians. Generically speaking: Mathematicians want assessments to be true to mathematics. From the mathematician's perspective, mathematics assessments should focus on revealing whether and what students understand about mathematically central ideas.

This does not happen automatically. For example, I served for many years on the committee that produced the Graduate Record Examination (GRE) advanced mathematics exam. In a sense, our task was simple: the audience (mathematics professors assessing the potential of applicants to their graduate programs) wants to judge applicants' promise for success in graduate school. This is usually understood as "How much do they know?" or "What problems can they solve?" The paper-and-pencil version of the GRE advanced mathematics exam had 65 multiple-choice questions, which students worked over a three-hour period. Student scores correlated reasonably well with the grades that those students earned during their first year of graduate school — but those grades themselves are not great predictors of future performance in graduate school or beyond, and there has been perennial dissatisfaction with the test because it reveals so little about actual student thinking. Indeed, the Educational Testing Service (ETS) spent some time trying to alter the exam, replacing it with an exam that focused on a deeper examination of what students know. ETS looked at the possibility of putting essay questions on the test to see if students could produce proofs,

explain concepts, etc. For a variety of reasons, including cost and the difficulty of creating essay tests with the right psychometric properties, some of which are examined below, ETS abandoned this approach.

Mathematics education researchers. Again, generically speaking: From the perspective of mathematics educators, mathematics assessments should reflect the broad spectrum of mathematical content and processes that comprise what might be called "thinking mathematically." (See my chapter "What is Mathematical Proficiency and How Can It Be Assessed?" later in this volume.) Thus, for example, the National Research Council document *Adding It Up* [NRC 2001, p. 5] describes five interwoven strands of mathematical proficiency:

- *conceptual understanding*: comprehension of mathematical concepts, operations, and relations
- *procedural fluency*: skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
- *strategic competence*: ability to formulate, represent, and solve mathematical problems
- *adaptive reasoning*: capacity for logical thought, reflection, explanation, and justification
- *productive disposition*: habitual inclination to see mathematics as sensible, useful and worthwhile, coupled with a belief in diligence and one's own efficacy.

The National Council of Teachers of Mathematics offers a more fine-grained characterization of desired proficiency in its *Principles and Standards for School Mathematics* [NCTM 2000]. This document argues for competency along these ten dimensions, clustered by content and process:

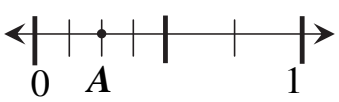
Content:

Number and operations
Algebra
Geometry
Measurement
Data analysis and probability

Process:

Problem solving
Reasoning and proof
Making connections
Oral and written communication
Uses of mathematical representation

To give a far too simple example, someone interested in students' ability to operate comfortably with fractions might ask a student to "reduce the fraction $2/8$ to lowest terms." Someone interested in whether students understand different representations of fractions, and operate on them, might ask a question such as this:



Write a fraction for point A. _____

Now write an equivalent fraction. _____

Part of the interest in this question is whether the student understands that the interval from 0 to 1 must be partitioned into n parts of equal length in order for each sub-interval to have length $1/n$. A student who merely counts parts is likely to say that point A is $2/6$ of the way from 0 to 1; that student may well reduce the fraction $2/6$ to $1/3$. To the mathematics educator, this is evidence that mastery of procedures in no way guarantees conceptual understanding. And, the problem above only examines a small part of conceptual understanding. With their broad interpretation of what it means to understand mathematics and to think mathematically, mathematics education researchers tend to want assessments to cover a very broad range of mathematical content and processes. Moreover, they would want to see students answer analogous questions with different representations — e.g., with a circle partitioned into parts that are not all congruent.

Researchers employ a wide range of assessment techniques. These range from extended interviews with and observations of individual children to the large-scale analysis of standardized test data.

Parents. From parents' point of view, a mathematics assessment should enable parents to understand their children's knowledge and progress (and to help their kids!). Thus, parents tend to want: (a) simple performance statistics that place the student on a continuum (e.g., a grade or a percentile score); and (b) perhaps enough information about what their child is doing well and doing poorly so that they can provide or arrange for help in the areas where it is needed. Homework plays this role to some degree, but assessments can as well.

Policy-makers. By policy-makers I mean those who have a direct say in the ways that schools are run. This includes principals, superintendents, and school boards at the local level; it includes state departments of education and their policy leaders (often elected or appointed state superintendents of education); it includes legislatures at the state and federal level; it includes governors and the president. For example, the No Child Left Behind (NCLB) Act [U.S. Congress 2001], passed by the Congress and signed by the president in 2002, mandated

mathematics testing at almost all grades K–12. State bureaucracies, under direction from their state legislatures, established curriculum frameworks, standards, and assessments: Local administrators are responsible for making sure that students meet the standards, as defined by performance on the assessments.

From the policy-maker's perspective, the primary use of assessments is to provide indicators of how well the system is going. The further away instruction the policy-maker is, the less details may matter. That is, a teacher is interested in detailed information about all of his or her students (and in some detail — see below). A principal may want to know about how the school as a whole is doing, how subgroups of students are doing, and perhaps how particular teachers are doing. By this level of the system, then, one number per student (the test score) is as much as can be handled, if not too much; one number per subgroup is more typically used. Statistical trends are more important than what the scores reveal about individuals. As one travels up the political food chain, the unit of analysis gets larger: what counts is how well a school or a district did, the typical question being "Did scores go up?" Test scores may be used to "drive" the system, as in the case of NCLB: each year scores must go up, or there will be serious consequences. Those making the policies may or may or may not know anything about the content of the assessments or what scores mean in terms of what students actually know. Such details can be (and usually are) left to experts.

Publishers and test developers. Commercially developed assessments play a significant and increasing role in schooling at all levels. In the U.S. students have faced the SAT and ACT (for college admission) and the GRE (for admission to graduate school) for decades, but with NCLB, students are assessed annually from grade 3 through 8 and often beyond. These tests are typically developed and marketed by major corporations — Harcourt Brace, CTB McGraw-Hill (CTB used to stand for "Comprehensive Test Bureau"), ETS, the College Board, etc.

What must be understood is that these corporations — indeed, every developer of "high-stakes" assessments for wide distribution and use — must design their tests subject to some very strong constraints. An easily understandable constraint is cost. School districts or other test consumers typically want tests that can be administered and graded at low cost. This is one reason multiple-choice questions are so common. Another constraint is security — the need to administer tests in ways that the potential for cheating is minimized. Thus tests are given under high security conditions. "Objective" machine grading is used not only to lower costs, but to lower the possibility of teachers acting, individually or collectively, to modify papers or scores to their students' advantage. (There is some individual scoring of tests such as the Advanced Placement ex-

ams that students take to get college credit for high school courses. However, the scoring is done at centralized locations where teachers do not have a stake in the scores, and the high costs of these exams are borne by the students who pay to take them.)

More important and more constraining, however, are the constraints imposed by the test design system itself. No major commercial test producer in the United States will market a test that fails to meet a series of technical criteria. These are called psychometric criteria. (Psychometrics is the field engaged in the quantification of the measurement of mental phenomena.) There are technical terms called "reliability," "construct validity," "predictive validity," and "test comparability" that play little or no formal role at the classroom level, but that are essential considerations for test designers. Indeed, the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education have jointly issued sets of criteria that standardized tests should meet, in their (1999) volume *Standards for Educational and Psychological Testing*. Relevant issues include: Will a student get the same score if he or she takes the test this month or next? Is the same content covered on two versions of a test, and does it reflect the content that the test says it covers? Will a score of 840 on this year's GRE exam mean the same thing as a score of 840 on next year's GRE? Such considerations are essential, and if a test fails to meet any of these or related criteria it will not be used. Among other things, test producers must produce tests that are legally defensible. If a test-taker can demonstrate that he or she suffered lost opportunities (e.g., failed to graduate, or was denied admission to college or graduate school) because of a flaw in a test or inconsistencies in grading, the test-maker can be sued, given that the test was used for purposes intended by the test-maker. Thus from the perspective of test-makers and publishers, the psychometric properties of a mathematics test are far more important than the actual mathematical content of the test.

Here is one way the tension between "testing that is affordable and meets the relevant psychometric criteria" and "testing that informs teaching and careful decision-making" plays out. Almost all teachers (from elementary school through graduate school) will say that "performance tasks" (asking the student to do something, which might be to build a mathematical model or write a proof, and then evaluating that work) provide one of the best contexts for understanding what students understand. Such tasks are found very rarely on high-stakes exams. This is partly because of the cost of grading the tasks, but also because it is very difficult to make the grading of student work on open-ended tasks consistent enough to be legally bullet-proof. The kinds of tasks that K–12 teachers and college and university faculty put on their in-class exams rarely meet the

psychometric criteria that are necessary to satisfy the technical departments of the publishers. Thus high-stakes exams tend to be machine graded. This has consequences. Machine-graded exams tend to focus largely on skills, and do not emphasize conceptual understanding and problem solving to the degree that many would like.

Teachers. From the teacher's perspective, assessment should help both student and teacher to understand what the student knows, and to identify areas in which the student needs improvement. In addition, assessment tasks should have curricular value. Otherwise they steal time away from the job of teaching.

Teachers are assessing students' proficiency all the time. They have access to and use multiple assessments. Some assessment is done informally, by observation in the classroom and by grading homework; some is done one-on-one, in conversations; some is done via quizzes and in-class tests; and some is done by formal assessments. But, if the formal assessments deliver just scores or percentile ratings, that information is of negligible use to teachers. If the results are returned weeks or months after the test is taken, as is often the case with high-stakes assessments, the results are of even less value. Moreover, if a teacher spends a significant amount of time on "test prep" to prepare students for a high-stakes test, then (depending on the nature of the mathematics that appears on the test) the exam — though perhaps low-cost in terms of dollars — may actually come at a significant cost in terms of classroom opportunities to learn. (See the issue of curriculum deformation, below.)

Professional development personnel. For the most part, professional development (PD) personnel are concerned with helping teachers to teach more effectively. One important aspect of such work involves helping teachers develop a better understanding of students' mathematical understanding, as can be revealed by appropriately constructed assessments (see the chapters by Foster and Fisher in this volume). Detailed information from assessments can help PD staff identify content and curricular areas that need attention. The staff can, therefore, make good use of rich and detailed assessment reports (or better, student work), but they, like teachers, will find the kinds of summary information typically available from high-stakes tests to be of limited value.

Students. Assessments should help students figure out what they know and what they don't know; they should be and feel fair. Thus tests that simply return a number — especially tests that return results weeks or months later — are of little use to students. Here is an interesting sidebar on scoring. A study by Butler [1987] indicates that placing grades on test papers can have a negative impact on performance. Butler's study had three groups: (1) students given feedback as grades only; (2) students given feedback as comments but with no grades written

on their papers; and (3) students given feedback as comments *and* grades. Not surprisingly, students from Group 2 outperformed students from Group 1, for the straightforward reason that feedback in the form of comments helped the students in the second group to learn more. More interesting, however, is that there were no significant differences in performance between Groups 1 and 3. Apparently the information about their grades focused the attention of students in Group 3 away from the content of the comments on their papers.

In any case, assessments can serve useful proposes for students. The challenge is to make them do so.

Discussion. The preceding narrative shows that many of the "stakeholders" in assessment, especially in standardized testing, have goals for and needs from assessments that can be complementary and even contradictory. If any conclusions should be drawn from this section of this chapter, they are that the goals of and mechanisms for testing are complex; that different testing constituencies have very different needs for the results of assessments; and that no one type of measure, much less any single assessment, will serve the needs of all those constituencies. To compound the problem, these constituencies do not often communicate with each other about such issues. These facts, among others, led MSRI to bring together the various groups identified in this part of this chapter to establish lines of communication between them.

Unintended Consequences of Assessment

Test score inflation and the illusion of competence. If you practice something a lot, you are likely to get good at it. But the question is, what have you gotten good at? Have you learned the underlying ideas, or are you only competent at things that are precisely like the ones you've practiced on? In the latter case, you may give the illusion of competence while actually not possessing the desired skills or understandings.

Suppose, say, that elementary grade students study subtraction. Typical "two-column" subtraction problems found on standardized tests are of the form

1. $\begin{array}{r} 87 \\ \underline{-24} \end{array}$

Problem 1 is, of course, equivalent to the following:

2. Subtract 24 from 87.

A priori one would expect students to do more or less as well on both problems — perhaps slightly worse on Problem 2 because it gives a verbal instruction, or because students who do the problem in their heads might find it slightly harder to keep track of the digits in the two numbers when they are not lined up vertically. But, if the students understand subtraction, Problems 1 and 2 are the same problem, and one would expect students to perform comparably on both.

Now, what would happen in a school district where students spent hour after hour being drilled on problems like Problem 1?

Roberta Flexer [1991] studied the test performance of students in a school district that had a "high-stakes" test at the end of the year. She compared test performance with an equating sample of students from another district that did not have a high-stakes test. She obtained the following statistics.

Sample	Problem 1 $\begin{array}{r} 87 \\ \underline{-24} \end{array}$	Problem 2 subtract 24 from 87
High-stakes district	83% correct	66% correct
Low-stakes equating sample	77% correct	73% correct

On problems that looked just like the ones they had practiced, such as Problem 1, students in the high-stakes district did substantially better than students from the low-stakes district. But their scores plummeted on the equivalent problem, and they did far worse than the "control" students. Thus the procedural skill of the students in the high-stakes district came at a significant cost.

Flexer's statistics demonstrated this kind of pattern on a range of tasks. In short, drilling students on tasks just like those known to be on the high-stakes exam resulted in the *illusion of competence*. (My mentor Fred Reif tells the story of his visit to a medical clinic. A technician asked him for the index finger of his left hand, in order to take a blood sample. Reif asked the technician if she could use his right index finger instead — he wanted the left hand untouched because he had a viola recital coming up soon after the blood test and did fingering with his left hand. The technician said no when he first asked. When he asked again, she allowed after some deliberation, and some misgivings, that "it might be OK" to use his right hand instead of the left for the blood sample. This is a case in point. Do we really want students, or professionals, who follow procedures without understanding?)

It is folk knowledge in the administrative community that an easy way for a new superintendent to appear effective is to mandate the use of a new test the first year he or she is in office. Partly because students are unfamiliar with the test format, scores are likely to be very low the first year. Then, because

students and teachers become more familiar with the test format, scores go up over the next two years. Whether the students have actually learned more is open to question — but the administrator can take credit for the increase in test scores.

Shepard and Dougherty [1991] highlight a case in point — actually, 50 cases in point. "A national study by Linn [1990] documented that indeed all 50 states and most districts claimed to be above average." (That is, compared to the average score established when the test was instituted.) However, "achievement gains on norm-referenced tests during the 1980s were not corroborated by gains on NAEP."

Curriculum deformation. In the years immediately before California instated high-stakes testing in reading and mathematics, California's students scored, on average, in the middle tier of the NAEP science examinations. In the 2000 administration of the NAEP science examinations, the first since the advent of high-stakes testing in California, California student performance dropped to the very bottom. The reason: teachers stopped teaching science, because their students were not being held accountable for their science knowledge. (NAEP is a "low-stakes" test, which affects neither the teacher nor the student in any way.) They devoted their classroom time to reading and mathematics instead.

This is an example of what has been called the WYTIWYG phenomenon — "What You Test Is What You Get." WYTIWYG can play out in various ways. For example, if the high-stakes assessment in mathematics focuses on procedural skills, teachers may drill their students for procedural fluency — and conceptual understanding and problem solving skills may be left unaddressed as a consequence. (Recall the subtraction example in the previous section.)

These negatives can be contrasted with some significant positives. Colleagues have reported that in some schools where little curriculum time had been devoted to mathematics or literacy prior to the advent of high-stakes tests, as much as two hours per day are now being devoted to each. In both subject areas, a substantial increase in instructional time can be seen as a significant plus. Thus, high-stakes testing can be seen as a powerful but double-edged sword.

Stifling innovation. In the spring of the year before my daughter entered middle school I visited a number of mathematics classrooms. A number of the teachers told me that I should return to their classrooms after the testing period was over. The state mathematics exams were coming up in a few weeks, and the teachers felt they had to focus on skills that were related to items on the test. Hence what they were teaching — in some cases for weeks or months — did not reflect the practices they wished to put in place.

In some cases, curricular innovators have faced the problem that without "proof of concept" (evidence that a non-standard approach will guarantee high enough test scores) school districts are reluctant to let people try new ideas. At the MSRI conference, civil rights leader and mathematical literacy advocate Robert Moses spoke of the testing regime he had to put into place, in order to reassure administrators that his students would do OK. This represented a significant deflection of time and energy from his larger educational goals. (See [Moses and Cobb 2001], for instance.)

Disenfranchisement due to linguistic or other issues. Here is a problem taken from the Arizona high-stakes math exam:

> If x is always positive and y is always negative, then xy is always negative. Based on the given information, which of the following conjectures is valid?
>
> A. $x^n y^n$, where n is an odd natural number, will always be negative.
>
> B. $x^n y^n$, where n is an even natural number, will always be negative.
>
> C. $x^n y^m$, where n and m are distinct odd natural numbers, will always be positive.
>
> D. $x^n y^m$, where n and m are distinct even natural numbers, will always be negative.

Imagine yourself a student for whom English is a second language. Just what is this question assessing? Lily Wong Fillmore [2002, p. 3] writes:

> What's difficult about it? Nothing, really, if you know about, can interpret and use—
>
> - exponents and multiplying signed numbers;
> - the language of logical reasoning;
> - the structure of conditional sentences;
> - technical terms such as negative, positive, natural, odd, and even for talking about numbers.
> - ordinary language words and phrases such as if, always, then, where, based on, given information, the following, conjecture, distinct, and valid.

I would also add that the problem statement is mathematically contorted and ambiguous. It is hardly clear what getting the question right (or wrong) would indicate, even for a native English speaker.

An impact on drop-out rates? California has instituted High School Exit Examinations (CAHSEE) in English in Mathematics. Students will have multiple

chances to take the exams, but the bottom line is this: As of June 2006, a senior who has not passed both the CAHSEE mathematics exam and the CAHSEE English language exam will not graduate from high school. Instead, the student will be given a certificate of attendance. The odds are that the certificate of attendance will not have a great deal of value in the job market.

It remains to be seen how this policy will play out. But, imagine yourself to be a somewhat disaffected high school sophomore who has taken the CAHSEE for the first time. Months later your scores arrive, and they are dismal. On the basis of your scores, you judge that the likelihood of your passing the test later on, without Herculean effort, is small. You don't really like school that much. What incentive is there for you to not drop out at this point? Some have argued, using this chain of reasoning, that the net effects of instituting the exams will be a sharp rise in early drop-outs.

Discussion

My purpose in this introductory chapter has been to provide readers a sense of the assessment landscape. Developing an understanding of the mathematics that someone knows (assessing that person's mathematical proficiency) is a complex art. Different stakeholder groups (mathematicians, mathematics education researchers, parents, policy-makers, test manufacturers and publishers, teachers, professional developers, and students) all have some related but some very different needs from assessments. Not only is it the case that one assessment size does not fit all, it may well be the case that one assessment size does not fit anyone's needs very well. Thus the issue of figuring out which kinds of assessments would provide the right kinds of information to the right stakeholders is a non-trivial and important enterprise. Moreover, the set of examples discussed in the second part of this chapter shows that assessment (especially high-stakes assessment) can often be a blunt instrument, with the tests perturbing the system they measure. My hope is that these observations help to open the doors to the insights and claims in the chapters that follow.

References

[Butler 1987] R. Butler, "Task-involving and ego-involving properties of evaluation", *Journal of Educational Psychology* **79** (1987), 474–482.

[Fillmore 2002] L. W. Fillmore, *Linguistic aspects of high-stakes mathematics examinations*, Berkeley, CA: University of California, 2002.

[Flexer 1991] R. Flexer, "Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts", paper presented at the

annual meeting of the American Educational Research Association, Chicago, April 1991.

[Linn et al. 1990] R. L. Linn, M. E. Graue, and N. M. Sanders, "Comparing state and district test results to national norms: Interpretations of scoring 'above the national average' ", Technical Report 308, Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, 1990.

[Moses and Cobb 2001] R. Moses and C. Cobb, *Radical equations: Math literacy and civil rights*, Boston: Beacon Press, 2001.

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

[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.

[Shepard and Dougherty 1991] L. A. Shepard and K. C. Dougherty, "Effects of high-stakes testing on instruction", paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1991.

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