

## 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.

# 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:

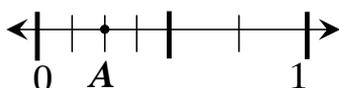
**Content:**

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  $\frac{2}{8}$  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  $\frac{1}{n}$ . A student who merely counts parts is likely to say that point  $A$  is  $\frac{2}{6}$  of the way from 0 to 1; that student may well reduce the fraction  $\frac{2}{6}$  to  $\frac{1}{3}$ . 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 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 purposes 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

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

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

2. Subtract 24 from 87.