The K-12 Mathematics Curriculum Center

Issues in Mathematics Education

Assessing Students' Mathematics Learning

Ilene Kantrov, Education Development Center, Inc.

The recommendations of the National Council of Teachers of Mathematics (NCTM) 1989 *Curriculum and Evaluation Standards for School Mathematics* have prompted changes in mathematics textbooks, teaching, and testing. While traditional mathematics education has emphasized memorization of facts and fluent application of procedures, the *Standards* call not only for fluency with facts and skills but also for sophisticated mathematical reasoning and problem solving.[1] Students are expected to apply mathematical procedures as well as to understand mathematical concepts.

The mathematics curricula created to reflect these recommendations—often called *Standards*-based curricula—look quite different from traditional mathematics textbooks. They tend to integrate several mathematics topics or skills in one lesson, extend lessons over several class periods, and embed skill mastery and practice within other activities. They call for students to work together to investigate problems, use concrete objects to model mathematical situations, and explain their mathematical ideas in speech and writing.[2]

The tests that accompany these curricula also look unfamiliar to most adults accustomed to traditional math tests. Flip through the pages of one of the new mathematics curricula, and you are likely to find pages headed "Quiz" or "Test," as you would in any mathematics text. However, much of what appears on these pages looks very different from traditional mathematics tests.

Consider, for example, the contrast between two test questions in Figure 1.[3]

In both examples, the student taking the test needs to be able to add in order to answer the question accurately. In the second example, however, in addition to adding correctly, the student needs to understand the concept of place value and be able to explain it in his or her own words.

At a time when student testing is debated on newspaper op-ed pages, mentioned in the president's *State of the Union* address, and even legislated in Congress, the new curricula and their unfamiliar assessments often provoke anxiety on the part of both teachers and parents:

Example 1

Add 3842
+ 4104

a. 7946
b. 7746
c. 7806
d. 7942
e. None of these

Example 2

Adam says that 4 + 52 is 452. Is he right or wrong? What would you tell Adam?

Figure 1

- How well do these curricula, and the teaching and testing approaches they incorporate, prepare students to succeed on "high-stakes" tests—that is, tests that determine, for example, whether students move to the next grade or graduate from high school?
- Do students in schools using *Standards*-based curricula fare poorly or well on national and state tests?
- Will the need to prepare students for these tests interfere with a teacher's ability to implement a new curriculum?
- Are high-stakes tests moving closer to measuring the kinds of learning emphasized by the new curricula?

"High-stakes" *tests have serious consequences; for example, they may determine whether students move to the next grade or graduate from high school.*

This paper explains the different kinds of tests used to assess mathematics learning. It also can help you answer questions about the compatibility of student assessments with the content and approaches to teaching embodied in *Standards*-based curricula. More specifically, this paper addresses:

- some of the terms used in debates about mathematics assessment
- the advantages and disadvantages of different kinds of assessments when used for different purposes
- the (limited) evidence about the impact of the new mathematics curricula on student achievement on high-stakes tests
- some criteria to apply to your own school's or district's assessments

Different Audiences and Purposes for Tests

All tests are intended to provide information, but the audiences for this information and the purposes for which it is used vary. In the case of mathematics tests, audiences include parents, teachers, students, school and district administrators, school boards, state departments of education, and policymakers at the local, state, and national levels. Sometimes the distinction is made between two broad purposes for testing: (1) to "improve the learning process in classrooms" and (2) to provide "reliable information to hold students and schools accountable for results."[4] It is difficult to find a single test that can equally well serve these very different purposes, as well as provide the information desired by all of these different audiences. In considering the kinds of assessments used in the *Standards*-based curricula, as well as the national and statewide tests students may be required to take, the critical questions to ask are, therefore: What kind(s) of information do these tests provide? and, For what purpose(s) is each test well-suited? Some of the criteria traditionally used to evaluate assessments help to answer these questions.

Criteria for Evaluating Assessments

Reliability refers to how consistently an assessment measures students' knowledge, skills, and understandings. To be reliable, the results of an assessment should be consistent across each of the following:

- different test items intended to measure the same knowledge or skill
- different administrations of the same test to the same student (for example, would a high school junior who takes the SAT twice score the same both times?)
- different raters (for example, would three different teachers scoring the same student's response to Example 2 on page 1 give it the same rating?)[5]

Another criterion traditionally used to evaluate assessments is validity, which refers to the extent to which an assessment measures what it is intended to measure, and the accuracy of inferences and decisions made on the basis of the assessment results.[6] For example, if a timed test of one-digit multiplication is used to determine how quickly students can recall their multiplication facts, the test is measuring what it was designed to measure. If the same test were employed to assess students' capacity to determine whether to use addition, subtraction, multiplication, or division to solve a variety of problems, the test would not meet the criterion. To the extent that a *Standards*-based mathematics test is valid, we should be confident that a student who does well on it is in fact competent in the mathematics skills and processes specified in the *Standards*. To be valid, an assessment should also be fair, or equitable; that is, it should enable students to demonstrate their mathematical competence, regardless of their language or cultural background, or physical disabilities.

Feasibility refers to the demands that a particular assessment places on a teacher's, school's, district's, or state's resources for example, how much the assessment will cost and how much time it will take to develop, administer, and score.

An additional criterion used to evaluate assessments which has earned increasing attention in the past decade is the extent to which a test is aligned with the curriculum students are using in the classroom. If a test includes items that require knowledge and skills not included in the curriculum, then it is not well aligned with what students are expected to learn in school. A particular test may be valid as a measure of a student's mastery of the content of the *Standards*; however, a student in a school in which the curriculum is not well aligned with the *Standards* could not be expected to perform well on that test.

To understand how these criteria operate in practice, think about the test you had to pass to secure your driver's license. If this test were *reliable*, you should have scored about the same regardless of which form of the written test you took, how many times you took it, or which driving examiner you had for the road test. (You should, of course, score higher if you learned more and improved your skills between tests.) If the test were *valid*, you should only pass it if you were a reasonably competent driver. If the test were *feasible*, that would mean most people should be willing to spend the time to take it and pay the taxes required to administer it. And if the test were aligned with the curriculum, then, for example, you should not be taught to drive a car with an automatic transmission but then be tested in a car with a standard transmission.

Testing to Guide Learning and Teaching

Many teachers use multiple-choice and other similar tests (for example, fill-in-the-blank, true/false) primarily for reasons of feasibility and reliability. These tests are easy and quick to administer and score, and they leave little room for discussion (by students or their parents) about the grades assigned based on the results.

Other than to determine grades, however, the primary purpose of classroom assessment should be to help teachers figure out what students do and do not understand, and thus to guide their instruction. When the goal is to determine how students think about a problem or task, as well as to elicit the results of their thought processes, then multiple-choice questions are less useful. Even if a student selects the correct answer to a problem, it is impossible to determine whether he or she made a lucky guess or actually figured it out. If the student selects an incorrect answer, there is no way to know whether he or she made a minor arithmetic error or fundamentally misunderstood the problem. The Bus Problem (Figure 2) provides an example of how responses to a multiple-choice question can misrepresent a student's understanding of a problem.

In contrast, the kinds of assessments included in the *Standards*-based curricula are often designed to help teachers determine what specific assistance students need in order to guide instruction. Some of these are performance assessments, which can take a variety of forms. Performance assessments ask students to demonstrate their knowledge, skills, and understanding by, for example, writing a response to an open-ended question, compiling a portfolio—a collection of their best work over a period of time—or completing a project that calls on them to apply what they have learned. An example of such a project appears at the end of a sixth grade unit on perimeter and area: it asks students to design a park,

An army bus holds 36 soldiers. If 1,128 soldiers are being bused to their training site, how many buses are needed?

This problem appeared on a National Assessment of Educational Progress exam.[7] The answers of most students who took the test suggested that they didn't understand that they had to round up the fractional answer (31.3 or 31 1/3) to the next largest number of whole buses (32). Later, a researcher gave the same problem to a group of students and then interviewed them. He found that many test takers did in fact understand that the answer had to correspond to reality. But they had in mind different alternatives: for example, that the remaining students would be transported in a minibus, or that the remainder of the partly-filled bus would be used to carry equipment. The multiple-choice test could not distinguish between those students who were mechanically applying the division algorithm and those who had given more careful thought to their response.[8] The teacher would presumably want to take different paths in instructing these two groups of students, but the results of a test composed only of such multiple-choice items would not provide the necessary information to guide the teacher's instruction.

Figure 2: Bus Problem

including a scale drawing, the dimensions of all the components of the design, a list of the quantities of materials needed, and a written description.[9] Such a project is intended to be both a learning activity *and* an occasion for assessing what students have learned. To successfully complete the project, students must accurately calculate perimeters and areas of various components of their park design, but they also must figure out which dimensions to measure and how their calculations apply to the practical challenge of determining what quantities of materials to purchase. Their written descriptions reveal how they went about the task, what decisions they made, and why.

In contrast with multiple-choice questions, performance assessments commonly require students to show, and sometimes explain, how they got to a result, in addition to what the result was. Thus, analysis of students' work can help teachers see the depth of students' thinking as well as pinpoint sources of error or misunderstanding. Often, the performance assessments included in the *Standards*-based curricula ask students to engage in tasks set in everyday contexts.[10] For example, in the Pizza Problem (Figure 3), students must figure out the best method for a pizza parlor manager to price different-sized pizzas: on the basis of their diameters, circumferences, or areas.[11]

Completing such tasks requires students to apply what they have learned and thus can tell more about the depth of their understanding of mathematics concepts and the flexibility of their skills. Depending on the contexts used, such assessments can also give students of differing backgrounds, interests, and strengths different ways to demonstrate their skills and understandings.

Performance assessments used in the classroom are closely aligned with the *Standards* and *Standards*-based curricula, which value student thinking and the process of arriving at solutions, as well as the solutions themselves. However, they do present challenges to students and teachers in terms of both reliability—achieving consistency in judging the quality of performances—and feasibility—finding the time to both administer and judge the assessments.

To address reliability concerns, scoring guides, called rubrics, are commonly used. Rubrics identify the particular skills, knowledge, and understandings that are being assessed, and describe different levels of quality for each.[12] Usually, rubrics are accompanied by one or more samples of student work that exemplify each of the ratings.

Rubrics can be more or less useful, depending on a number of factors. If a rubric is so specific that it applies only to a single task, it is not very useful to either teachers or students as a guide to what is important to learn and what constitutes superior performance. On the other hand, if the criteria in a rubric are so general that they can be interpreted in any number of ways, they likewise provide no useful guidance for learners or teachers. The most useful rubrics are those that give students (and their parents) a clear idea of what they should be striving to achieve, and that offer teachers guidance in helping students learn the particular mathematics being assessed.

Pizza shops often sell round pizzas in various prices. At Pizza Nook they sell the following sizes:

6 inch - $3.00
12 inch - $8.00
18 inch - $12.00

The new manager of the Pizza Nook is thinking about changing the prices of pizzas. It appears to him that he could think about the pricing in three ways:

I. The prices of the pizzas are influenced by comparing the diameters.
II. The prices of the pizzas are influenced by comparing the circumferences.
III. The prices of the pizzas are influenced by comparing the areas.
   a) If you were the manager, which method would be most appropriate for pricing the pizza?
   b) Explain your reasoning.

Figure 3: Pizza Problem

Student demonstrates proficiency – Score Point = 3

The student provides a satisfactory response with explanations that are plausible, reasonably clear, and reasonably correct, e.g., includes appropriate diagram(s), uses appropriate symbols or language to communicate effectively, exhibits an understanding of the mathematics of the problem, uses appropriate process and/or descriptions to answer the question, and presents sensible supporting arguments. Any flaws in the response are minor.

Student demonstrates minimal proficiency – Score Point = 2

The student provides a nearly satisfactory response which contains some flaws, e.g., begins to answer the question correctly but fails to answer all of its parts or omits appropriate explanation, draws diagram(s) with minor flaws, makes some errors in computation, misuses mathematical language, or uses inappropriate strategies to answer the question.

Student demonstrates a lack of proficiency – Score Point = 1

The student provides a less than satisfactory response that only begins to answer the question, but fails to answer it completely, e.g., provides little or no appropriate explanation, draws diagram(s) which are unclear, exhibits little or no understanding of the question being asked, or makes major computational errors.

Student demonstrates no proficiency – Score Point = 0

The student provides an unsatisfactory response that answers the question inappropriately, e.g., uses algorithms which do not reflect any understanding of the question, makes drawings which are inappropriate to the question, provides a copy of the question without an appropriate answer, fails to provide any information which is appropriate to the question, or fails to answer the question.

Figure 4: Sample Rubric[13]

The sample rubric included in Figure 4 is a generalized scoring guide from which appropriately specific rubrics can be developed.

The feasibility challenges of using performance assessments in the classroom are not much different from those of using a *Standards*-based curriculum. Preparing students to accomplish performance assessment tasks, engaging them in those tasks, and judging their performance, while time-consuming, can all be seen as elements of good teaching. Think again of the park design project mentioned earlier. Carrying out this assessment task contributes to students' learning and also gives teachers evidence of what students have learned.

Given the need for teachers to understand students' thinking in order to figure out how to teach effectively, most assessment experts, even those who criticize the use of performance assessments for high-stakes testing purposes, recognize the value of these assessments for classroom use. However, teachers need opportunities to learn how to use these assessments well in order to guide and individualize their classroom instruction. They can find it challenging to make accurate inferences about what students actually understand, based on their responses to performance assessment tasks. It is therefore wise for schools and districts that adopt *Standards*-based curricula to devote professional development time and resources to helping teachers learn how to analyze and respond to student work.

Testing to Hold Students and Schools Accountable

Nationally Used Standardized Tests

Standardized tests, usually in a multiple-choice format, are the tests most commonly used to hold students and schools accountable.[14] Many states and school districts use these tests not only to evaluate individual students' achievement, but also to evaluate their schools. The popularity of these tests over the past 75 years has much to do with their strengths in meeting the criteria of reliability and feasibility: they do not require scorers who must exercise their own judgment; in fact, they can be scored by machine.[15] Because human scorers aren't required (and don't need to be trained and paid for their time), such tests are less costly than tests that require human judgment. For the same reason, they are also perceived as more objective.

The standardized tests used for accountability purposes are usually achievement tests, such as the widely used Comprehensive Test of Basic Skills, Iowa Test of Basic Skills, and Stanford Achievement Test. These tests are designed to evaluate students' mastery of particular content in comparison to the performance of other students nationwide. This comparative feature of standardized tests is known as norm referencing, meaning that any individual student's performance is evaluated against the performance of a group identified as the "norm." Norms are determined by administering the test to large numbers (usually thousands) of students before it is released for general use. The performance of students who subsequently take the test is compared to the results from this initial administration.[16] In contrast, criterion-referenced tests compare an individual student's test results to a set of expectations, usually for the student's age or grade. In response to criticism of norm referencing, more states are reporting how students perform on standardized, multiple-choice tests as compared to a state benchmark—that is, a particular level of performance—as well as to other students.[17]

*To understand the difference between **norm** and **criterion referencing**, think again about the test you had to pass to secure a driver's license. Imagine that the road test were norm-referenced. And suppose that of all the drivers who took the test, those who scored at or above the 50th percentile achieved a passing grade and received their licenses. (A ranking in the 50th percentile means that the driver scored as high or higher than 50 percent of the drivers in the group used to set the norm. This percentile ranking is what makes a score norm referenced; it compares this person's performance to that of others who took the test.) All of these drivers would be "better than average." Yet some of them might not know how to parallel park or merge with traffic, and might ignore red lights and fail to signal when turning. In contrast, actual road tests are criterion-referenced; that is, in order to pass, prospective drivers need to demonstrate their competence on all the critical components of the test. Every driver who succeeds in doing this passes the test.*

Advantages and Disadvantages of Standardized Tests

Standardized, norm-referenced, multiple-choice tests are a fixture of the educational landscape, and there are purposes for which they can be quite useful. These tests can provide teachers and parents with a picture of a student's mastery of the content on the test relative to that of a sample of students nationwide—the norm group. In addition, standardized achievement tests can suggest students' relative strengths and weaknesses across different areas of mathematics.[19]

*A **benchmark** is "a detailed description of a specific level of . . . performance expected of students at particular ages, grades, or development levels. Benchmarks are often represented by samples of student work. A set of benchmarks can be used as 'checkpoints' to monitor progress toward meeting performance goals within and across grade levels, i.e., benchmarks for expected mathematics capabilities at grades 3, 7, 10, and [high school] graduation."[18]*

However, determining the significance of these tests requires more than just a score. For instance, what is the content of the test? How important is that content to your school and community? To what extent is that content part of the curriculum?

Moreover, although standardized, norm-referenced tests are designed to provide information about individual students in relation to a comparison group, these tests are often used to make inferences about the effectiveness of education, a purpose for which they are ill-suited.[20] These tests have flaws both as measures of individual students' achievement and as indicators of the success of teachers and schools. Most of these defects relate to the validity of the tests, including their fairness to students of different backgrounds, and to the alignment of the tests with standards and curriculum.

A key concern is that the kinds of multiple-choice items included in standardized tests do not adequately reflect the kinds of problem solving and mathematical thinking that are required for mathematical competence and that are emphasized in the *Standards*. Such items are also criticized because they are only "stand-ins" or "indicators" of what students are supposed to have learned. For example, to write well, students need a good grasp of grammar and vocabulary. However, students' performance on tests of grammar and vocabulary knowledge in isolation has been shown to bear little relationship to their capacity to write clearly and persuasively. The principle is the same in mathematics: students' knowledge of math facts, algorithms, and concepts in isolation does not ensure that they understand them, can call them up, and can use them appropriately in problem-solving contexts.[21]

There is also evidence that some of the content of standardized tests reflects knowledge and skills acquired by students at home as well as in school. Students from families of higher socioeconomic status are more likely

to acquire such knowledge and skills.[22] If elements of what is tested are not taught as part of the school curriculum, the test cannot fairly be used to assess students' learning in school or to evaluate the quality of teachers and schools.

Performance Assessment in Statewide Proficiency Testing

Because of concerns about the limitations of standardized, multiple-choice tests, a number of states have introduced high-stakes tests that are wholly or in part performance assessments. In the late 1980s, Vermont launched a statewide assessment system that used portfolios (compilations of student work). Around the same time, Kentucky and several other states began to introduce various kinds of performance assessments. By January 1999, 48 states had testing systems that were at least in part performance assessments.[23]

The main attraction of performance assessments for high-stakes testing has been their potential for greater validity and closer alignment with the *Standards* and *Standards*-based curricula. Such assessments can better match with what students are expected to learn, and the results can provide more accurate information about what students have actually learned. In addition, the rubrics used to score performance assessments make clear that individual students' work is being compared to particular benchmarks, rather than to other students' levels of achievement.[24]

Characteristics of classroom performance assessments:

- *Students apply knowledge, skills, and understanding.*
- *Students show and explain the process as well as the product.*
- *Problems are often set in everyday contexts.*
- *Problems offer students multiple ways to demonstrate skills and understanding.*
- *Results can help guide instruction.*
- *Assessments are aligned with the Standards and Standards-based curricula.*
- *Assessments present rater reliability and feasibility challenges.*

However, despite the claims made for the advantages of performance assessments for accountability testing, states that have begun using performance assessments for high-stakes purposes have encountered significant challenges. Designing large-scale performance assessments that actually assess the knowledge and skills the curriculum teaches and that correspond to a set of mathematics learning goals can be difficult, and can raise validity concerns. For example, does a complex task in fact allow students to demonstrate the particular kinds of learning the test is intended to assess? In order to determine the extent to which students understand a particular concept, such as

Several years ago, a researcher showed parents of third graders two sets of mathematics problems: a sample of multiple-choice items from standardized achievement tests (including Example 1 on page 1 and Example 1 below), and a sample of performance assessment questions (such as Example 2 on page 1 and Example 2 below). Parents who saw mathematics as a set of fixed knowledge favored the multiple-choice test items; however, most parents saw that the performance assessments asked kids to think, and preferred that approach. A typical comment from the parents in the latter group was, "I think it gives them a broader understanding of what they're doing, rather than just A + B = C. . . . [H]ow did you get it? Use logic rather than just being told this is the answer. . . It's not just memorization."[25]

Example 1

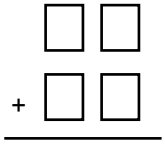
Multiply

6
x 9

a. 63
b. 48
c. 54
d. 69
e. None of these

Example 2

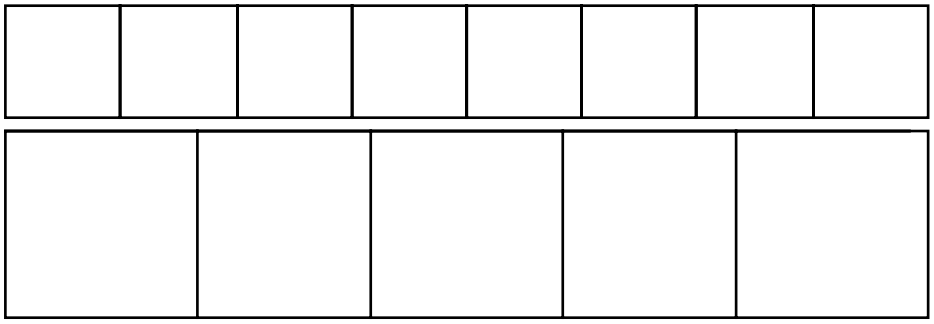
Put four different one-digit numbers in the boxes to make the largest possible answer.



How did you know what numbers to choose?

Task 1

The small squares are all identical. Their side length is 100 mm. The large squares are also all identical. Show how you can find the side length of the large squares without measuring. Give your answer to the nearest millimeter.



Task 2

Make a two-dimensional paper replica of yourself using measurements of lengths and widths of body parts that are half those of your own body.

Figure 5

ratio and proportion, asking them to solve a complex problem that requires other knowledge and skills may not be the best strategy.[26] Consider the two examples in Figure 5. Task 1 focuses directly on the concepts of ratio and proportion. Task 2 requires students to apply what they have learned about ratio and proportion; however, students' relative success in completing this more complex task may also depend on how well they can make plans, organize information, and draw.

To determine how well students can combine and apply several skills and concepts, a complex problem-solving task, like Task 2 in Figure 5, may be appropriate. A complex task, however, may reveal information about students' reading levels, drawing abilities, or familiarity with a particular context as well as their mathematical knowledge. Particularly in the use of "real-life" contexts, performance assessments can have defects in fairness for students of different cultural and economic backgrounds; for example, using a miniature golf setting to explore probability could put students unfamiliar with miniature golf at a disadvantage.

Another validity concern stems from the depth with which performance assessments measure a particular mathematical domain (such as area and perimeter of a circle). Because of the amount of time required for students to complete any particular performance assessment task, the total number of tasks is likely to be significantly lower than the number of items on a standardized multiple-choice test, which means that these assessments tend to "sample" (i.e., include items that test) a smaller range of student performance overall than do standardized multiple-choice tests. The results may therefore not be generalizable, which is especially problematic if the tests are being used for high-stakes assessments of students, teachers, or schools.[27]

Using performance assessments for high-stakes purposes also raises concerns about reliability, particularly the consistency of scoring by different raters. Increasing reliability in turn presents feasibility challenges, since development of accurate scoring guides and training of scorers are costly and time-consuming. Feasibility is also a factor in the administration of high-stakes performance assessments, which generally take more time out of classroom instruction than do multiple-choice tests.

Perhaps the most controversial aspect of using performance assessments for high-stakes purposes is the way decisions are made about what constitutes satisfactory performance. In most states, results of high-stakes tests are reported in three or four categories; for example, *advanced*, *proficient*, *basic* or *needs improvement*, and *failing*. Decisions about what scores mark the "cut-off" point for each category should not be arbitrary but should re-

flect actual benchmarks for performance, based on knowledge of how students develop mathematical understanding and skill over time. The scoring should also reflect the real demands of the next level: what students who pass the tests will face, whether in the next grade or course at school, in employment, or in college.[28] Even if the scoring criteria are appropriate, decisions about which category of performance qualifies a student for promotion or graduation can be highly political. A 1999 public controversy in Massachusetts over whether high school students must score in the "basic" or "proficient" range on the state's tenth grade test in order to graduate suggests how such decisions can move out of the realm of education and into the realm of politics.

It is important to note that classroom performance assessments and the rubrics accompanying them are less subject to some of the problems encountered in high-stakes testing. For example, because the assessed domain—what mathematics students are intended to learn—is particular to the curriculum, there is less likely to be a mismatch between the test and what is taught. A classroom teacher who knows what real-life contexts will be familiar to his or her students can also much more readily address fairness concerns about the use of those contexts. And, because teachers have the opportunity to assess students repeatedly over time, no single assessment carries the consequences of an assessment used for high-stakes purposes.

How Compatible Are Standards-based Curricula and High-Stakes Tests?

Individual teachers and schools do not generally control the content or format of the high-stakes assessments their students are required to take. Therefore, what happens if the curriculum focuses on learning goals such as mathematical problem solving and understanding of concepts, but students are held accountable for their scores on standardized, multiple-choice tests that focus on recall of isolated facts and routine application of algorithms?

Because the *Standards*-based curricula are so new, extensive research on this question has not been conducted, and the question cannot yet be answered definitively. The small number of studies that have been completed do provide some evidence to suggest that schools can successfully use a *Standards*-based curriculum in a context where students must perform well on standardized, multiple-choice tests. Virtually all of these studies come to the following conclusions:[29]

- Students using the new curricula generally perform at or above the levels of comparison groups taught by traditional methods on standardized, multiple-choice tests, such as the Iowa Test of Basic Skills.
- In the small number of cases where scores on standardized tests decline immediately after the introduction of a new mathematics curriculum, the decline is not dramatic, and scores recover within a couple of years.
- Students using the new curricula perform better on tests of problem solving and mathematical reasoning.
- Efforts to "teach to the test" by supplementing the new curricula with drill on basic skills are neither effective nor necessary.

In addition, some evidence suggests that students who learn basic facts in the context of the kinds of learning activities included in the *Standards*-based curricula are likely to remember the facts longer. For example, one study showed that students who learned computation facts and algorithms in the context of a variety of problem-solving activities were more likely to retain their knowledge beyond the time of the test than were students who learned isolated facts and algorithms.[30] The research evidence to date does not support the fears of some teachers that their students will suffer on standardized tests unless they supplement the *Standards*-based curricula with hours spent memorizing facts and practicing procedures.

As noted earlier, in an effort to bring their assessments more closely in line with the *Standards*, most states have begun to include some form of performance assessment in their high-stakes tests. There is some evidence that since testing influences curriculum and instruction, the move toward more high-stakes performance assessments is creating increased interest in *Standards*-based curricula and the kinds of teaching and learning these curricula promote. However, the challenges of large-scale performance assessments suggest that statewide performance assessments should be used with caution to determine students' future opportunities and to judge the quality of schools and teachers.

Making Decisions About What Tests to Use and What Test Results to Value

The relative advantages and shortcomings of performance assessments and standardized tests when used to hold students, teachers, and schools accountable suggest that

it is unwise to use just one assessment (of whatever kind) to make high-stakes decisions. Consider again the licensing requirements for driving a car, which usually require both a written test (generally in multiple-choice format) and a road test. Although it is undoubtedly expensive for states to maintain the road-testing capacity (which relies on potentially unreliable human judgments), states have decided that the written test alone is not sufficient to certify drivers' competence. The road test, despite its expense and potential for unreliability, is viewed as critical to determine the extent to which potential drivers can apply the knowledge and skills they have learned in real-life situations.

It is in the interest of school administrators, teachers, parents, and students to become as informed as possible about the tests used to evaluate mathematics achievement. Questions to ask include the following:

- To what extent do the tests, and the way they are scored, reflect what we want students to learn and what they are being taught in their classrooms?
- Do the tests value only the results of students' problem solving and computation, or do they also consider how well students understand and can apply the important mathematics?
- Are the tests fair to students of various cultures, ethnicities, levels of English language proficiency, and ranges of socioeconomic status? Are they fair to students of both genders?
- To what extent does students' success on the tests depend on teachers' ability to teach mathematical problem solving and develop students' understanding of important mathematical concepts?
- Are we relying on too small a range of assessments to make high-stakes decisions?

The more you know about what is on the national and statewide tests students are required to take, and how these tests are scored, the better able you will be to determine the significance and usefulness of the results. The higher the stakes, the more crucial it is to become a critical consumer of assessments, and to make your views known to policymakers. Within districts, schools, and classrooms, educators can also work to ensure that tests align with the *Standards* and the school's curriculum, that the tests are fair to everyone, and that teachers are prepared to effectively teach what the tests measure.

About the Author

Ilene Kantrov directs the Center for Educational Resources and Outreach at Education Development Center, Inc., in Newton, Massachusetts. She is a coauthor of *Choosing a Standards-Based Mathematics Curriculum* and *Facilitating Cases in Education*, and coeditor of *Casebook on School Reform*, all published by Heinemann.

Acknowledgements

I am most grateful for the excellent advice and support I received in writing this paper from Deborah Bryant Spencer, as well as June Mark, Kristin Winkler, and Paul Goldenberg. Sara Kennedy's formidable research skills, technical support, and cheerful approach to the work were invaluable. She was ably assisted for several months by Sarah Lee. All of these members of the K–12 Mathematics Curriculum Center issues paper group were extremely patient as I made my way through multiple drafts. I also appreciate the input provided by Kate Cress, Mark Driscoll, Lynn Goldsmith, and Dan Tobin. At various times Ki McClennan and Deborah Clark provided expert administrative support. The thoughtful readings by Jennifer Nichols, Ruth Gilbert Whitner, and Paul Maiorano were most helpful, and I especially appreciate the critique and suggestions of George Madaus. This work was supported in part by the National Science Foundation Grant No. ESI-9617783. Opinions and views remain my own, and not necessarily those of my colleagues, reviewers, or the Foundation.

Endnotes

[1] Goldsmith, L., Mark, J., & Kantrov, I. (2000). *Choosing a Standards-Based Mathematics Curriculum.* Portsmouth, NH: Heinemann.

[2] Ibid., 2; Goldsmith, L. & Mark, J. (1999). What Is a Standards-Based Mathematics Curriculum? *Educational Leadership, 57*(3), 40.

[3] Shepard, L. & Bleim, C. (1995). Parents' Thinking About Standardized Tests and Performance Assessments. *Educational Researcher, 24*(8), 28.

[4] Olson, L. (1999, Jan. 11). Making Every Test Count. *Quality Counts: Rewarding Results, Punishing Failure. Education Week* Special Report, 11.

[5] CRESST Assessment Glossary (1999). Available online at <http://cresst96.cse.ucla.edu/CRESST/pages/glossary.htm>.

[6] Ibid.

[7] Driscoll, M. & Bryant, D. (1998). *Learning About Assessment, Learning Through Assessment.* A report of the National Research Council, Mathematical Sciences Education Board. Washington, DC: National Academy Press, 21.

[8] An algorithm is a step-by-step procedure used to solve a mathematical problem, for example, the method of adding multi-digit numbers that involves "carrying."

[9] Lappan, G., Fey, J., Fitzgerald, W., Friel, S., & Phillips, E. (1996). *Getting to Know Connected Mathematics: A Guide to the Connected Mathematics Curriculum*. White Plains, NY: Dale Seymour Publications, 57.

[10] Such tasks are not necessarily real tasks but are set in real-life contexts to show students how mathematics might be used in a familiar setting.

[11] Lappan, G., et al, *Getting to Know Connected Mathematics*, 58.

[12] Popham, W. J. (1997). What's Wrong—and What's Right—with Rubrics. *Educational Leadership, 55*(2), 72.

[13] Mathematical Sciences Education Board and National Research Council (1993). *Measuring What Counts: A Conceptual Guide for Mathematical Assessment.* Washington, DC: National Academy Press.

[14] Standardized tests are those that "are given under the same conditions and ask the same questions across different populations in order to permit comparisons." Mitchell, R. (1992). *Testing For Learning: How New Approaches to Evaluation Can Improve American Schools.* New York: The Free Press, a division of MacMillan Publishers, 5.

[15] Madaus, G. & O'Dwyer, L. (1999). A Short History of Performance Assessment: Lessons Learned. *Phi Delta Kappan, 80*(9), 693.

[16] Stiggins, G. (1994). *Student-Centered Classroom Assessment*. New York and Toronto: Merrill/Macmillan College Publishing Co./Maxwell Macmillan Canada, 338, 346.

[17] Olson, Making Every Test Count, 18.

[18] Shepard & Bleim, Parents' Thinking, 28.

[19] Popham, W. J. (1999). Why Standardized Tests Don't Measure Educational Quality. *Educational Leadership, 56*(6), 8–16.

[20] Popham, Why Standardized Tests, 8–9.

[21] Boaler, J. (1999, March 31). Mathematics for the Moment, or the Millennium? *Education Week, 17*(29), 30, 34.

[22] Popham, Why Standardized Tests, 8–16.

[23] Olson, Making Every Test Count, 18.

[24] Parents often want to know how well their child is doing compared to others. It is possible to respond to this desire by publishing statistics about how many children in a particular grade in a school, district, state, or country have reached different levels of achievement in relation to particular benchmarks. Such reports provide parents with the information they want, without encouraging the kinds of ranking invited by parents who ask how much better, or worse, their child is doing than other children in the class, school, or district.

[25] Shepard & Bleim, Parents' Thinking, 28.

[26] Driscoll & Bryant, *Learning About Assessment*, 11.

[27] Madaus & O'Dwyer, A Short History, 693.

[28] Wiggins, G. (1992). Creating Tests Worth Taking. *Educational Leadership, 49*(8), 27.

[29] See, e.g., Briars, D. & Resnick, L. (forthcoming). *Assessments—and What Else? The Essential Elements of Standards-based School Improvement.* Pittsburgh: University of Pittsburgh; *Everyday Mathematics Student Achievement Studies* and *Everyday Mathematics Gets Results: Student Achievement Studies Volume 2.* Chicago: Everyday Learning Corporation; Hirstein, J. (1997). *The SIMMS Project: Student Assessment in the Pilot Study.* Missoula, MT: Montana Council of Teachers of Mathematics; Schoen, H. & Ziebarth, S. (1997). *A Progress Report on Student Achievement in the Core Plus Mathematics Project Field Test.* Iowa City: University of Iowa; Zawojewski, M. & Hoover, J. (1996). *Analysis of 6th-, 7th- and 8th-Grade Student Performance for the Connected Mathematics Project: Preliminary Analysis of Student Performances on ITBS Survey Battery and CMP Test Grades 6, 7, and 8, 1994–1995 School Year.* White Plains, NY: Dale Seymour Publications.

[30] Boaler, Mathematics for the Moment, 30, 34.