The Mathematics Assessment Collaborative: Performance Testing to Improve Instruction

David Foster and Pendred Noyce, M.D.
January 2004

American schools are awash in standardized testing. Students are spending more and more time on state-mandated, high-stakes tests linked to state standards. There are variations among the states in terms of how often they test students, what subjects they test, whether their exams are norm- or criterion-referenced, and whether multiple-choice questions are supplemented by other question types. But the passage of federal No Child Left Behind (NCLB) legislation makes it clear that, in the next few years, high-stakes standardized testing will only increase nationwide.

This rapid increase in high-stakes testing presents an educational dilemma. Measuring performance is not the same thing as improving it. And while the architects and supporters of statewide testing systems clearly mean for these systems to improve instruction and learning, most state testing systems are ill adapted for this purpose. Some critics argue that state accountability pressures actually lead school personnel to replace meaningful instruction with a narrow test-prep curriculum. Even proponents of high-stakes standardized testing acknowledge that such tests are generally summative in nature and provide little formative information for teachers to use with their current students. The complexities of scoring and reporting mean that months may pass between the administration and public reporting of the tests. Because of concerns about reliability, quality control, and test security, the tests themselves can become black boxes for teachers, administrators, and parents. The results, often reported after a school term ends, are usually limited to statistical measures, such as scaled scores or student percentile ranks on norm-referenced tests, and broad topic categories, such as "geometry and measurement" or "number sense." The reports do very little to inform instruction or to detail specific deficiencies of an individual student.

In this article, we describe a collaborative effort involving 30 school districts in California's Silicon Valley that are seeking to overcome this testing dilemma. These districts administer, score, and analyze a common set of performance assessments in mathematics in a way that guides professional development and leads to changes in teaching strategies. The assessment tasks place particular value on student understanding and knowledge transfer to new problems. Growth in students' test scores on statewide standardized tests occurs almost as an incidental effect of this approach.

The Mathematics Assessment Collaborative

Silicon Valley's Mathematics Assessment Collaborative (MAC) built on a number of existing partnerships and initiatives. In 1982 the California state legislature established the California Mathematics Project, which provided funding and guidance for a statewide network of professional development sites, usually led by mathematics education faculty

members at local universities. In 1996 the Noyce Foundation began to support local districts with mathematics professional development and quickly formed a partnership with the Santa Clara Valley Mathematics Project, which was led by Dianne Pors and Joanne Rossi Becker of San Jose State University. The new partnership was dubbed the Silicon Valley Mathematics Initiative (SVMI), and its early work focused on providing professional development, establishing content-focused coaching in schools, and collaboratively examining student work to inform teachers of pupils' understandings.

In 1996 the state of California was beginning a long and turbulent battle over the establishment of new state curriculum standards.[1] Following the state board's adoption of standards in mathematics, the governor pressed to establish a high-stakes accountability system. For the first time, California would require a test that produced an individual score for every student. Because developing a test to assess the state standards was expected to take several years, the state decided in the interim to administer an off-the-shelf, norm-referenced, multiple-choice test – Harcourt's Stanford Achievement Test, Ninth Edition (SAT-9). In the spring of 1998, students in grades 2 through 11 statewide took the test for the first time.

During the 1997-98 school year, the state department of education acknowledged some concerns about relying on a single measure to assess student performance, and indeed the department encouraged districts to develop performance assessment instruments to supplement the information gained from the SAT-9. In response, the SVMI invited regional school districts to join a discussion of the effects of a narrow accountability system.

The MAC grew out of those early district discussions. Twenty-four school districts joined the collaborative, paying annual membership dues of $3,500 per district. The Noyce Foundation matched district contributions, and the Santa Clara Valley Math Project became the mathematics advisory partner and fiscal agent. The MAC hired a director, Linda Fisher, and the collaborative began. Soon it became the cornerstone of the SVMI.

Selecting an Assessment

The MAC's first task was to create a framework of what was to be assessed. Keeping in mind William Schmidt's repeated refrain that the American curriculum is "a mile wide and an inch deep,"[2] the MAC decided to create a document that outlined a small number of core topics at each grade level. The goal was to choose topics that were worthy of teachers' efforts, that were of sufficient scope to allow for deep student thinking, and that could be reasonably assessed on a performance exam that lasted just a single class period. Using as reference materials standards developed by the National Council of Teachers of Mathematics (NCTM), by the state of California, and by the local districts, teacher representatives from the MAC districts met in grade-level groups to choose five core ideas at each grade level.

The resulting draft documents revealed some overlap and some inconsistencies. District teacher leaders, the MAC director, and Mathematics Project personnel refined the initial draft documents to eliminate the overlap and to create more consistent transitions between grade levels. Over the course of the intervening years, the assessment framework has been reviewed and revised to incorporate ideas from such important sources as NCTM's revised *Principles and Standards for School Mathematics*[3] and from feedback obtained through the MAC assessment process.

Once the core idea document was created, the next task was to develop a set of performance exams. Although several of the districts had been writing and using performance items on their own for a number of years, the collaborative quickly dismissed the notion of designing exam items internally, for the following reasons:

- Designing good assessments is a full-time job, not unlike designing curriculum materials. Teachers may have talent for task design, but they already have full-time jobs. When we ask teachers to do double duty, we ask them to divert their energy from their classrooms – the opposite of the message the collaborative wanted to send to teachers.

- Designing an exam means much more than writing good items. Field-testing, validity studies, rubric designs, reliability testing, and defining and addressing the sample space must all be dealt with to ensure a fair and consistent product.

- The education community, boards of education, and parents are interested not only in how students achieve on the exam but also in how results compare from one school system to another. If an exam is internal to one system only, such comparison is impossible.

The MAC considered two partners for its performance assessment work: the New Standards Project and the Mathematics Assessment Resource Service (MARS), creators of Balanced Assessment.[4] MARS is a project funded by the National Science Foundation, with roots in three universities: the University of California, Berkeley; Michigan State University; and the Shell Centre, Nottingham, England. MARS quickly became the collaborative's first choice. While New Standards offered an established exam specific to grade levels 4, 8, and 10, MARS was more flexible in allowing the MAC to design tests for multiple grades that fit the core ideas and other specifications. To supplement the state of California's two-day multiple-choice math exam, the MAC wanted a performance test that included only questions requiring extended open responses. The districts were also very interested in developing an internal scoring system that would bolster professional development and create teacher buy-in.[5] MARS had expertise in statistically measuring test outcomes and in developing reports. Finally, MARS could point to an impressive track record of providing performance assessment in England.

The MARS exam is a powerful tool. Each grade-level exam is made up of five tasks. The tasks assess math concepts and skills that correspond to the five core ideas taught at that grade. The exam also assesses the mathematical processes of problem solving,

reasoning, and communication. The tasks require students to evaluate, optimize, design, plan, model, transform, generalize, justify, interpret, represent, estimate, and calculate their solutions.

The MARS exams are scored using a point-scoring rubric. Each task is assigned a point total that corresponds to the complexity of the task and the proportional amount of time that the average student would spend on the task in relation to the entire exam. The points allocated to the task are then divided among its parts. Some points are assigned to how the students approach the problem, the majority to the core of the performance, and a few points to evidence that, beyond finding a correct answer, students demonstrate the ability to justify or generalize their solution. In practice, this approach usually means that points are assigned to different sections of a multi-part question.

The combination of open tasks and weighted rubrics provides a rich picture of student performance. Where the state's norm-referenced, multiple-choice exam asks a student merely to select from answers provided, the MARS exam requires the student to initiate a problem-solving approach to each task. Students may use a variety of strategies to find solutions. Most of the prompts require students to explain their thinking or justify their findings. This aspect of the performance assessment can never be duplicated by a multiple-choice exam. Details of the administration of the exam also differ from the state's approach. Teachers are encouraged to provide sufficient time for students to complete the exam *without rushing*. Students are allowed to select and use whatever mathematical tools they might need to solve a particular problem, such as rulers, protractors, calculators, link cubes, or compasses.

Performance Assessment in Practice

By the spring of 1999, the MAC was ready for its first test administration. Member districts voted to test grades 3, 5, 7, and 9 in the first year. Timing proved a more difficult decision. In order to separate the MARS exam from statewide exams given in April and May and to allow enough time following the exam for teachers to address any learning issues it uncovered, the collaborative chose to give the test in early March. While some teachers continue to worry that March is too early in the school year to measure accurately all five of the core ideas, the advantage of getting scored exams back in time to act on their diagnostic findings outweighs this concern.

Next came the challenge of collaboratively scoring 23,128 exams from four grade levels in 21 districts. The MAC director issued an application for scoring trainers. Ten teacher leaders, mostly district math coaches, were selected from across the region. Each pair of scoring trainers became expert on one task and corresponding rubric from each of the four exams. Initial scoring took place on two consecutive Saturdays. Each MAC district was required to provide one teacher to score for every 100 student papers sent. (This proved to be an underestimate of the number of scorers needed: today, we assign one scorer for every 60 student papers.)

The scoring trainers spent the first 90 minutes of the day training and calibrating the scorers on one task and rubric. After that initial training, the scorers began their work. After each problem was scored, the student paper was carried to the next room, where another task was scored. In that first year, scoring trainers checked random papers and rescored them as needed. The final step was a scoring audit. At the end of the initial scoring, a random sample of 5% of the student papers was pulled and rescored at San Jose State University. Reliability measures proved to be high. A final analysis across all grade showed that the mean difference between the original score and the audit score was 1/100 (0.01) of a point.

Along with checking for reliability, the 5% sample was used to develop performance standards for overall score reporting. The collaborative had voted to establish four performance levels in mathematics: Level 1, minimal success: Level 2, below standards: Level 3, meeting standards; and Level 4, consistently meeting standards at a high level

A committee of MARS staff members and SVMI leaders began the process of setting standards by analyzing each task to determine the core of the mathematical performance it required. Then actual student papers were examined to determine the degree to which students met the mathematical expectations of the task. The committee next examined the distribution of scores for each task and for the exam as a whole. Finally a cut score for each performance level for each test was established. The performance levels were reported to the member districts, teachers, and students.

Once the papers were scored, they were returned to the schools, along with a copy of the master scoring sheets, for teachers to review and use as a guide for further instruction. It was the responsibility of the districts to enter the student scores into a database and then merge them with each student's demographic information and score on the state-required exam. All the district databases were ultimately compiled into a central collaborative database. Using that database, the MAC director and a MARS statistician created an informative report that analyzed each of the tasks and provided comparative analysis and data disaggregated by student characteristics. This analysis, along with scored student papers, provided valuable information for instruction, professional development, and district policy.

In the MAC's second year, one district left the collaborative, and seven new districts joined. The departing district decided against joining for a second year because its students had such a high success rate on the state's norm-referenced, multiple-choice tests – although none of the district's third-graders had scored at the highest performance level on the MARS exam. The districts that continued, on the other hand, elected to increase the number of grade levels tested. In the spring of 2000, 35,061 students in grades 3 through 10 took the exam.

The most dramatic change in the second year was that the MAC transferred to the districts the responsibility for planning and conducting their own scoring sessions. To ensure the reliability of scoring, the MAC provided centralized training for the scoring leaders of the member districts and conducted the annual spring audit session of a random

sample of 5% of all scored papers. The reliability measures continued to remain extremely high, even though the tests were locally scored.

Currently, the cost for districts to administer, score, and report the MARS exam is approximately $7.50 per student. This cost, which is shared between the districts and the Noyce Foundation, includes the price charged by MARS for producing the test and creating score reports and the cost to the district for teachers' time spent scoring the exam.

How Assessment Informs Practice

Over time it has become clear that the tests, the scoring sessions, and the performance reports all contribute to the MAC's desired outcome: informing and improving instruction. The scoring sessions are powerful professional development activities for teachers. To be able to score a performance task accurately, teachers must fully explore the mathematics of the task. Analyzing different approaches that students might take to the content within each task helps the scorers assess and improve their own conceptual knowledge. The scoring process sheds light on students' thinking, as well as on common student errors and misconceptions. As one teacher said, "I have learned how to look at student work in a whole different way, to really say, 'What do these marks on this page tell me about [the student's] understanding?'"[6] Recognizing misconceptions is crucial if a teacher is to target instruction so that students can clarify their thinking and gain understanding. The emphasis on understanding core ideas helps teachers build a sound sequence of lessons, no matter what curriculum they are using. All of these effects on instruction grow out of the scoring process.

The scored tests themselves become valuable curriculum materials for teachers to use in their classes. MAC teachers are encouraged to review the tasks with their students. They share the scoring information with their students and build on the errors and approaches that students have demonstrated on the exams.

Each year, from the data collected and from a review of a random sample of student papers, the MARS and the MAC develop a document called "Tools for Teachers." Along with broad comparisons of performance across the collaborative's membership and analysis of the performance of different student groups, these reports provide a wealth of other information. A detailed portrait is compiled of how the students approached the different tasks, with a description of common misconceptions and evidence of what students understand. It becomes clear when certain practices in the classroom are contributing to errors, poor communication, or a genuine lack of understanding. Here are two examples.

During the first year, a fifth-grade task required students to look at a diagram that showed a linear relationship between the number of stones surrounding a flowerbed and the number of units making up the flower bed itself. (See Figure 1.) The students were given the first four stages of a functional pattern relating these two variables, and then asked to

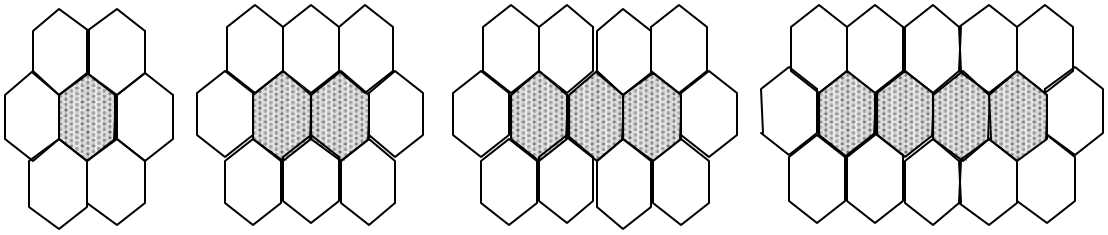
draw the fifth stage, complete a table of values, draw a graph, and determine why 28 stones was not the correct number to surround a flower bed made up of 13 units.

Figure 1.
Fifth-Grade Flower Bed Task

Flower Beds

In the diagram below, the shaded hexagons are flower beds, and the white hexagons are white paving stones.

Marco figures out how many white paving stones are needed around different numbers of flower beds.



1 flower bed
6 white stones

2 flower beds
8 white stones

3 flower beds
10 white stones

4 flower beds
12 white stones

1. Draw a diagram to show how many white stones are needed around 5 flower beds.

…

4. Marco says that 28 white stones are needed around 13 flower beds. Without drawing the flower beds, explain how you know that Marco is not correct. How many white stones are needed around 13 flower beds?

Based on a figure © 2000, MARS. Used with permission.

The students did very poorly on the flower bed task. The math leaders discovered that most of the collaborative's teachers were not teaching algebraic thinking or an introduction to patterns and functions in fifth grade, partly because many of the teachers themselves had only a superficial understanding of the mathematics involved.

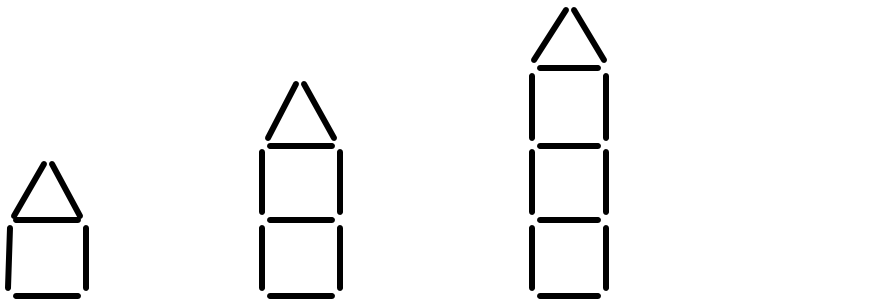
The following summer and school year, SVMI professional development for upper-elementary teachers focused on patterns and functions. The next spring, the 2000 MARS exam had a mathematically similar task – this time dealing with "toothpick shapes" – that once again assessed students' capacity to translate visible patterns into mathematical functions. (See Figure 2.) Year 2000 fifth-graders, who had never taken a MARS exam before, performed dramatically better on the algebraic thinking task than had fifth-graders the year before. In 2000 more than 50% of the students were awarded the maximum

number of points, compared to fewer than 10% in 1999. Clearly, the professional development had made a difference.

Figure 2.
Fifth-Grade Toothpick Task

Toothpick Shapes

Tom uses toothpicks to make the shapes in the diagram below.



shape 1
6 toothpicks

shape 2
9 toothpicks

shape 3

shape 4

1. How many toothpicks make shape 3? _____
2. Draw shape 4 next to shape 3 in the diagram above.

. . .

5. Tom says, "I need 36 toothpicks to make shape 12." Tom is *not* correct. Explain why he is not correct. How many toothpicks are needed to make shape 12?

Based on a figure © 1999, MARS. Used with permission.

This lesson on the value of using student results from the MARS test to tailor and inform professional development for the following year became a cornerstone of SVMI strategy.

The second example concerns an insight that came from an analysis of common errors occurring on student papers across multiple districts. The MAC has compiled a considerable list of such errors. One common early problem was the use of mathematical run-on sentences. Many students were creating mathematical run-on sentences when doing problems that involved multiple operations.

Consider the problem of how to calculate the number of feet in a picture of a girl walking three dogs. A typical (incorrect) student response reads: $4 \times 3 = 12 + 2 = 14$. This is a mathematical run-on sentence. Standard notation does not allow two equal signs in a single equation, and 4×3 does not equal $12 + 2$. The steps of the solution should have been written out as follows:

$$4 \times 3 = 12$$
$$12 + 2 = 14$$

At first glance, this correction may seem like nit-picking. But the problem with the notation is more than just sloppiness; a run-on sentence betrays a common mathematical misconception. Instead of understanding that the equal sign indicates that expressions on the two sides of the sign are mathematically equivalent, students using such run-on sentences take the equal sign to signal that an operation must be performed ("The answer is…").[7] This view contributes to the growth of further confusion as students learn to generalize and work with expressions containing variables in later grades.

We found that this error in notation occurred regularly throughout the tested population. On further investigation, we learned that teachers commonly allowed this notation to be used in classrooms – or even used it themselves when demonstrating solutions to multi-step problems. That year's assessment report pointed out the problem and clarified that solutions using run-on sentences would no longer receive full credit. Subsequent professional development showed teachers how such notation led to student misconceptions. Within a year, the collaborative noted a dramatic change in the way students in 27 districts communicated mathematical statements.

This matter of notation was just one example of how analyzing patterns of student error led to improvements in instructional practice. Other areas of improvement include differentiating between continuous and discrete graphs, noting and communicating the units in measurement problems, distinguishing between bar graphs and histograms, understanding correlation trends in scatterplots, and developing understanding of mathematical justifications. Examining MARS results has also led teachers to confront significant chunks of unfamiliar mathematical content. Discussing the tasks and student responses often uncovers the fact that, for many topics and concepts in algebra, geometry, probability, measurement, and statistics, teachers' understanding is weak. Uncovering these gaps in teachers' content knowledge is central to improving instruction.

Formative Assessment and the Link to Mathematics Coaching

Even before launching the assessment collaborative, the SVMI had built a network of district mathematics coaches who worked together to improve mathematics instruction in area schools. Mathematics coaches are accomplished teachers with records of leadership and strong understanding of mathematics content who are released from teaching duties to work with other teachers. Over the past five to six years, SVMI has built a strong and stable cadre of such coaches. Currently, the 28 MAC districts are served by 44 math coaches (or full time equivalents), working primarily with teachers in grades 2 through 7. Funding for the coaches is shared by the Noyce Foundation and the participating districts. The coaches spend 70% of their time supporting other teachers in the classroom and the remainder either offering professional development to groups of teachers or participating in further professional development of their own.

The school year for coaches begins in August, with a weeklong institute focusing on mathematics content, pedagogy, leadership skills, and coaching strategies. Each coach brings five teachers to the institute. Usually these are teachers who have been identified as potential leaders by their districts, but in some cases they are a group of teachers that the coach has chosen for more intensive coaching during the upcoming year. Coaches and the teachers who attend the institute meet as a MAC-wide group for five additional days each year.

Since the 2000-01 school year, the coaches have explicitly linked their work to the kind of performance assessment used in the MARS exam. During the summer, working with the MAC leadership, the coaches select two or three performance tasks to address each of the five core ideas at every level from grade 3 through grade 10. These tasks are drawn from among publicly released Balanced Assessment or New Standards tasks, from district-developed tasks, and, increasingly, from released items from past MARS tests.

Throughout the school year, the coaches encourage teachers to administer the tasks as they begin and complete instruction for each core idea. The coaches lead the teachers in discussing the mathematics that students need to know in order to complete the tasks and in analyzing patterns of student errors. These discussions often uncover weaknesses in the teachers' own mathematical understanding, which can be addressed in time to improve instruction for that year's class of students. As Melissa Adams, a district math coach explains, "Using the data from the previously administered MARS tasks gives our work focus. Teacher and coach have personal as well as collaborative-wide data to inform instructional planning. Through each unit of study our discussions are focused on the mathematics, on student understandings and misconceptions, and we have time to tailor instructional experiences."[8] The discussions and common assessments also lead to greater consistency in pacing and content coverage from teacher to teacher in a school. By tracking student performance on tasks given throughout the year, teacher teams are able to determine when their approaches to teaching for depth of understanding have been particularly effective and when they might have missed the mark.

Performance Assessment and Statewide Standardized Testing

The quality of information that the MAC has provided to its member districts has helped the districts maintain their commitment to coaching and to professional development that concentrates on improving teacher understanding. California offers significant incentives and sanctions for student achievement on the state exam (STAR: Standardized Testing and Reporting), and many districts across the state are thus tempted to embrace narrow quick-fix methods of test-prep and teaching to the exam.

To counter this temptation, the SVMI has been able to show that, even when a significant number of students are improving on the state test, their success may not translate into greater mathematical understanding as demonstrated by success on the more demanding performance assessments. The data also indicate that, as students move up the grades, the disparity increases: more and more students who appear to be doing well on the state exam fail to meet standards on the performance exam. Conversely, success on the

performance exam becomes an ever *better* predictor of success on the state's STAR exam. By grade 7, students whose teachers have prepared them to perform well on the MARS performance exam are extremely likely to perform above the 50th percentile on the STAR exam.

The SVMI has been able to demonstrate to the satisfaction of superintendents and school committees that high-quality professional development significantly enhances student achievement. District case studies show that students whose teachers participate in intensive SVMI professional development, including coaching and work with formative assessment, achieve higher averages on both the state math test and the MARS exam than students whose teachers who are less involved. As a result, districts have continued to invest in math professional development, coaching, and formative assessment. The number of students assessed and teachers and grade levels involved has grown every year, even as the SVMI has kept the number of member districts relatively constant. In 2003 more than 60,000 students of 1,200 teachers in 28 districts participated in the MARS exam.

The performance of MAC district students on the STAR exam has continued to rise. For example, while 53% of third-graders performed above the 50th percentile on the state test in 1998, 72% did so in 2002. There is similar growth in the other grades, with a minimum of 11% more students above the 50th percentile at each grade level.

While success on the state test is politically important, even more important as a measure of student learning is the growth that students demonstrate on the performance assessment exam. Over the first four years, grades 3 through 6 have shown considerable increases in the percentage of students meeting standard on the MARS exam. In 1999, 44% of the students met standards, while in 2002, 62% did so. In grades 7 through 10, student achievement on the MARS exam has remained steady. The distribution of improvement roughly parallels the distribution of SVMI professional development, since 80% of SVMI coaches provide services to grades 2 through 6. What's more, in grades 3 through 6, there has been a significant increase in the number of students achieving at the highest performance level.

These findings have convinced district leaders to embrace the theory of action central to our work. This theory states that, when teachers teach to the big ideas, participate in ongoing content-based professional development, receive support in the classroom from well-trained coaches, and use specific assessment information to inform instruction, their students will learn and achieve more.

The lessons learned in the Mathematics Assessment Collaborative are also institutional. Districts, schools, and teachers can more effectively balance the demands of learning and testing when they work together. The SVMI is first and foremost a community of learners. The interplay of intellects and the constant feedback provided by data are both humbling and informative. Knowing that we are always learning and improving creates a healthy, rich environment for change. Greater student learning cannot be mandated by

policy: it must be nurtured and supported by continual formative efforts that improve instruction.

---

[1] For a description of the history of the controversy over curriculum, standards, and testing in California, see Suzanne Wilson, *California Dreaming: Reforming Mathematics Education* (New Haven, Conn.: Yale University Press, 2003).

[2] William Schmidt, U.S. research director for the Third International Mathematics and Science Study, has made this statement in numerous places. See, for example, the press releases available on the Internet at http://ustimss/msu.edu.

[3] *Principles and Standards for School Mathematics* (Reston, Va.: National Council of Teachers of Mathematics, 2000).

[4] Balanced Assessment tasks can be found at http://balancedassessment.gse.harvard.edu.

[5] Asked to identify the ways that MARS assessment services are unique, users of the MARS test have cited the intellectually rigorous nature of the tasks and rubrics, the willingness of MARS to customize assessments to client needs, and the value of scoring tasks as professional development. See Mark St. John et al., "A Study of the MARS Project: The Contributions to Clients," December 2000, available from Inverness Research Associates, www.inverness-research.org.

[6] Quoted in St. John et al., p. 7.

[7] Robert S. Siegler discusses the problem of students misinterpreting the equal sign as "simply a signal to execute an arithmetic operation" in "Implications of Cognitive Science Research for Mathematics Education," in Jeremy Kilpatrick, W. Gary Martin, and Deborah Schifter, eds., *A Research Companion to PSSM* (Reston, Va.: National Council of Teachers of Mathematics, 2003).

[8] Personal communication.