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

Chapter 22
Perspectives on State Assessments in California: What You Release is What Teachers Get

ELIZABETH K. STAGE

In the 1970s, the California Legislature determined that California's curriculum was not well served by national standardized tests and developed its own testing program, the California Assessment Program (CAP). Folklore gives several rationales for this action: urban superintendents wanted to conceal their students' low achievement, which would be revealed with the publication of national norms; consciousness that the California population was more diverse than the U.S. population; or awareness that California's curriculum differed from the national composite used for national tests.

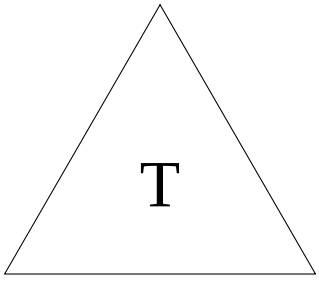
California has had state-wide frameworks to guide district curricula for many years. This story begins with the 1980 addendum to the 1975 framework [California 1982], a small volume that declared "problem solving" to be the umbrella for all of the framework's curricular strands (number, algebra, geometry, statistics, etc.). The task for CAP was to provide the state and districts with information about performance on these strands. It used matrix sampling and item response theory to provide detailed analysis, on the basis of which instructional improvements could be made. CAP was not designed to yield individual student scores; these could continue to come from standardized tests so that parents, teachers, principals, and superintendents could answer the question, "How does this student, this class, this school, or this district stack up in comparison with national norms?"

CAP was designed to provide scores on mathematical topics. It used matrix sampling to make sure that enough kids took items such as "whole number division" or "similar figures" to yield a reliable and valid score. That meant that people could see quite obviously that scores on the "number" component of the exam were relatively good, while other strands had weak performances.

This was similar to California results on the National Assessment of Educational Progress and many other measures. But, CAP showed information at that level of detail for districts and schools, as well for the state as a whole. That way teachers could focus on areas of poor performance because they were getting information in a form they could use. “I can work on division. That is something that I can target.”

The way in which I entered this conversation was with respect to gender differences. The pattern of CAP results, performance on the number strand that was strong relative to performance on the other strands, was even more pronounced for girls. (There wasn’t very good tracking of ethnic data at the time.) Professional development programs like EQUALS at the Lawrence Hall of Science focused attention on giving girls opportunities to learn geometry. Over time, the gap narrowed. In Lee Shulman’s formulation, CAP was a low-stakes, high-yield assessment. CAP yielded information that teachers could use to target instruction.

Work through this item before you read further.



A piece of cardboard shaped like an equilateral triangle with a side 6 cm is rolled to the right a number of times.

If the triangle stops so that the letter T is again in the upright position, which one of the following distances could it have rolled?

a. 24 cm b. 60 cm c. 30 cm d. 90 cm

A typical multiple-choice item on most standardized tests is allotted a response time of 45 seconds to a maximum of 1 minute. The item above, categorized “extended multiple choice,” was allotted as much as 3 minutes because, at the time, in 1988, it was unconventional and it required some thought. In other words, it doesn’t say, “What’s the perimeter of the triangle?” and “Which one of the numbers is divisible by 18?,” both of which suggest approaches to solving the problem. Another strategy is to mentally imagine the triangle rolling over, “Cabump, cabump, cabump, 18; cabump, cabump, cabump, 36”; until you get a match with one of the possible answers. Because CAP was low stakes, its designers could fool around with format and use the multiple-choice format more creatively. The idea was not to sort children, for which a speeded format is very helpful. If you want to sort people, make them run a race; if you want to see if kids can “do it,” them give them adequate time to “do it.”

At the same time, CAP had introduced something called “Direct Assessment of Writing.” Instead of asking where the comma goes in a sentence, which can

be done in a multiple-choice format, or about subject-verb agreements, the CAP test made students write — which was amazing and revolutionary at the time. In response, teachers started to ask students to write, using the seven CAP genres (persuasive, expository, etc.). And, student writing improved.

One could argue that the professional development programs like EQUALS or the California Writing Project (which was behind the direct assessment of writing) contributed to improved student performance. Or, one could argue that CAP scores improved due to a variety of other factors such as familiarity with the test, better instructional materials, improved nutrition, etc. Whether the focus for teachers on the CAP released items caused the improvement cannot be determined, but the fact is that many students spent considerable time improving their responses to tasks of this sort. That was not a bad way for them to spend their time.

The mathematics community was envious of what was taking place in writing and tried to take advantage of their success. We embraced the extended multiple-choice items and used them in our workshops with teachers, but behind the scenes we said to the CAP folks, "We'd really like to see how we can push the limits."

One of my favorite items from this era was this:

> James knows that half of the students from his high school were accepted to the public university nearby, half were accepted to the local private college. He thinks this adds up to 100% so he'll surely be accepted at one or the other. Explain why James might be wrong.

One can determine from the context that this was a high-school item, though the mathematics is clearly elementary. Therefore, it was amazing how many students proved that James was correct. The vast majority of the students wrote "one half plus one half equals one, which is the same as 100%." When this result was reported, even though it was an experimental item and didn't count for anything official, the teachers whose students had been happily adding probabilities without regard for context were horrified. And, assuming that problems like this would, once scoring issues were sorted out, count in the official assessment, they started to give students problems like this to solve. Thus, good assessments can have a beneficial curricular impact.

The extended multiple-choice items, like the triangle problem, and constructed-response items, like James being accepted at college, were released in a California Department of Education publication called *A Question of Thinking* [California 1989]. This was presented as simply a collection of problems that teachers might want to try — like CAP, low stakes, high yield.

Around the same time, the California Mathematics Council (CMC, the state affiliate of the National Council of Teachers of Mathematics) published *Alternative Assessment in Mathematics* [Stenmark 1989]. It showcased items from CAP, the Shell Centre, and other sources. Some 30,000 copies were distributed to CMC members, at conferences, and through other professional networks of teachers. This teacher-to-teacher communication, basically saying, "Look at all these cool ways that we can find out what is going on in our children's mathematical thinking," was — again — low (in fact, no) stakes, high yield.

California seems to prefer the rollercoaster model of educational policy. In the 1980s, under State Superintendent Bill Honig, the curriculum frameworks called for more authentic, extended work in every discipline. In addition to reading and writing in English Language Arts and problem solving in mathematics, the frameworks called for hands-on experimentation in science and working with primary source documents in history.

According to Honig, CAP had intentionally been designed as a low-level assessment because the state's urban superintendents wanted a measure that wouldn't show much differentiation. Honig used the bully pulpit of the state superintendency to get those urban superintendents to see that, in fact, using a low-level test perpetuated inequity; he argued that the test had to be ratcheted up to test what was really valued in order to expose the magnitude of current inequities. The revised state testing program, the California Learning Assessment System (CLAS), was designed to use the new item types, extended multiple-choice and constructed response, in addition to multiple-choice items, to assess the more demanding curricular goals that could not be assessed adequately with multiple-choice items.

Unfortunately, CLAS backfired. Tests that, like CLAS, use constructed-response formats depend on the training of the scorers. In the United States, teachers score only one large-scale test, the New York State Regents' exams. The Advanced Placement exams involve only a small group of elite teachers. Thus, the vast majority of California teachers had no experience evaluating constructed-response items outside of their own classrooms.

Teaching anybody to score with a rubric rather than their own personal standards or judgment is hard work. There is a whole technology to make sure that the scores are accurate, including scorer training and qualification procedures, table leaders and room leaders, and read-behinds. Installing that technology and training the leadership in that technology takes time; the constant monitoring of scorers' accuracy is a major culture change for teachers who are accustomed to grading papers in isolation.

Where would a California education story be if it didn't have some politics? The CLAS program was initially intended to be just like CAP, designed to

give curricular information to schools, districts, and the state. Midstream, the governor ordered the program to produce individual student scores. The teachers union had been arguing for individual student scores on the premise that, in exchange for the students' time, you owed them scores on the assessment. Because the time for constructed-response items exceeded the time needed for the multiple-choice tests, the union's increased demands for individual student scores dovetailed nicely with the governor's increased interest in accountability. Unfortunately, neither the governor nor the union understood a psychometric rule of thumb: You need 30 or more data points to get a reliable score. If your goal is to get the score at the school level, you can get lots of data points about lots of subscores. If your goal is an individual student score and you don't have unlimited testing time, then you have to narrow the domain of what you test in order to get a reliable and valid score. A test that was designed for one purpose — school-level scores — was asked suddenly to fulfill a different purpose: individual student scores. It didn't have the right design to accomplish the new purpose and the technology for the scoring wasn't yet robust enough to assign accurate scores to individual students.

To add insult to injury, the CLAS writing assessment asked students to write a brief essay based on their reading of a passage written by Alice Walker in which the protagonist wondered whether or not to get married. Despite the lack of personal freedom that she would experience in a conservative religious tradition, she decided to go ahead with the marriage for the sake of her children. The excerpt was published in the *Orange County Register* and the Eagle Forum became unglued: "Questioning marriage is not an appropriate topic for eighth graders!" (This was despite the fact that the protagonist decided to get married.)

The governor had a huge political problem. He used a psychometric argument concerning the accuracy of the prematurely released individual scores as an excuse for not defending CLAS. Bill Honig, the reform superintendent who had set the more demanding frameworks and assessments in motion, had stepped on the toes of the State Board of Education by getting ahead of them, and he had been indicted for the appearance of conflict of interest, so he was in no position to defend CLAS. Without further ado and with no rational discussion whatsoever, CLAS went down in flames.

When CLAS was in place, just as with the Direct Assessment of Writing, teachers were giving students opportunities to construct responses to challenging questions in mathematics, science, and social studies. For many students, particularly for some of our most neglected students, it was the first time anybody ever asked them, "What do you think?" (I think that question is the most profound assessment of a classroom; if you visit a classroom and don't hear

someone say to someone else, "What do you think?" in a whole class period, then those students are being shortchanged.)

Because of the political debacle, Governor Wilson, through his State Board of Education, asked the superintendent, "What shall we use as the state test?" Ninety percent of the local districts were using McGraw-Hill's CTB at that time. The state superintendent (a Democrat) recommended CTB, so the governor (a Republican) picked the Harcourt Assessment's Stanford Achievement Test, Ninth edition (SAT 9), making lots of people in San Antonio, Texas very happy.

Hardly anybody knows that Harcourt, Riverside Publishing, and McGraw-Hill produce practice exams that show the format of the items but not the level of difficulty or the range of mathematics that is assessed. They don't show the differences between what they call "skill problems," "concept problems," and "problem-solving problems." Because teachers don't know what's on the test and it doesn't get released, they drill on arithmetic. That is all they can be certain will be on the test. That is what they practice and the scores go up because practice pays off.

Using the SAT 9 was only an interim solution, because it was not aligned with California standards, so the plan was to add California standards items gradually to the test. There is a blueprint of the test available that shows how many items will be on which standards. But, in 2004, at the time of the MSRI assessment conference, there were no released items available, so teachers had to imagine what they might look like.

An interesting note is that after three years of administering California Standards Tests in mathematics, released items were posted recently [California 2006]. There are 65 items on the fifth grade test: 17 about operations on fractions and decimals, 17 about algebra and functions. An interesting example of the latter is:

What value for z makes this equation true?

$$8 \times 37 = (8 \times 30) + (8 \times z)$$

A. 7 B. 8 C. 30 D. 37

A student might recognize the equation as an instance of the distributive property. Or, the student might try the answer choices to see which number makes the equation true.

Until recently, the only clue that teachers had at their disposal to predict what would be on the test was the "key standards." Responding to initial reactions to the mathematics standards, that there were too many topics on the list to

teach in depth, about half of the standards were identified as "key" standards. The test blueprint was designed so that 80% of the test addressed these key standards. When school districts developed pacing plans that covered only the key standards, the percentage was dropped to 70% to encourage teachers to teach all of the standards. Key standards-only pacing plans persist in at least one large California district.

Released items from California state assessments, to the extent that they have been available, have been more influential than the standards or frameworks that they are supposed to exemplify. It's not what the framework or standards say or intend. It is the teachers' perceptions of what counts that afford the students the opportunity to learn.

References

[California 1982] California State Board of Education, *Mathematics framework and the 1980 addendum for California public schools, kindergarten through grade twelve*, Sacramento: Author, 1982.

[California 1989] California Department of Education, *A question of thinking: A first look at students' performance on open-ended questions in mathematics*, Sacramento: Author, 1989.

[California 2006] California Department of Education, 2003, 2004, and 2005 CST released test questions, January 2006. Available at http://www.cde.ca.gov/ta/tg/sr/css05rtq.asp. Retrieved 21 Apr 2006.

[Stenmark 1989] J. K. Stenmark, *Assessment alternatives in mathematics*, Berkeley, CA: University of California, 1989.