This is a story about a remarkable school, a school that has been labeled “underperforming” by the state of California. It is a true story, and it draws on a combination of research data, collected as part of a Stanford University research project on mathematics learning, and the lived realities of teachers and students working hard to achieve success in a low-income, urban school. At the heart of this story lies the conflict between learning and SAT-9 success, a conflict that has affected the lives of students and teachers in this school in profound ways.

As a new professor at Stanford University, recently arrived from England, I was considering schools to include in a study of mathematics teaching and learning. I soon learned that a school that I’ll refer to here as Railside High had a mathematics department that worked in very unusual ways. Some years ago the teachers “de-tracked” their classes in response to the low performance of some students. The mathematics department plans lessons collaboratively, and the teachers meet every week to discuss and improve their lessons. They visit one another’s classes frequently, and every new teacher is given the opportunity to watch every lesson that he or she will teach being taught by an experienced colleague first. The algebra curriculum that all students take on entry to the school was designed by the department, and it draws

Jo Boaler is an associate professor of mathematics education, Stanford University, Stanford, Calif., and the author of Experiencing School Mathematics (Erlbaum, 2002). She thanks Ed Haertel and Lani Horn, who both gave extremely helpful advice in the earlier stages of writing this article.
from a variety of different curriculum materials. Students from a range of ethnicities — Latino, African American, white, Filipino, and other Asian groups — work together in groups to solve complex mathematics problems. All the teachers in the department are mathematics specialists, and they all regularly attend professional conferences as a department.

These practices would be unusual for any school, but this is a school in a low-income area with few resources. Lessons are accompanied by the steady hum of cars zipping past on the two freeways that surround the school and are interrupted at frequent intervals by the sound of trains that pass just feet away from the school yard (thus our chosen pseudonym for the school). Financial resources are low — in the school and in the students’ homes. Yet qualified mathematics teachers are queuing up to join this department, and, after a year of studying and monitoring the mathematics teaching and learning in this school, we have discovered some unequivocally positive facts.

As part of a research project funded by the National Science Foundation, my research team and I gave incoming students at three high schools a mathematics examination. The students starting Railside High scored at a significantly lower level than the students starting the other — wealthier — schools in our study. However, after one year at Railside High, the students attained a higher average score on the end-of-year algebra examination than the students at the other schools in the study. By the end of the second year at the school, Railside students were significantly outperforming students at the other schools.

There is one other high school in Railside’s district — in a wealthier area. At the end of each course (algebra, geometry, trigonometry, etc.) both high schools give students the same final examinations, designed to carefully assess the competencies specified in the California mathematics standards. These exams are constructed and graded by the two departments and overseen by the district. In all three years that the exams have been given, the Railside students have significantly outperformed their wealthier counterparts at every level of mathematics.

In questionnaires, the Railside students are significantly more positive about mathematics than other students in our study. Indeed, there are many indicators that the mathematics teaching at Railside High is unusually effective. Moreover, the vast amounts of time teachers spend working together and preparing lessons to challenge and motivate students pay significant benefits in terms of student engagement and learning. The students at Railside appear to learn more mathematics than most, to develop more positive attitudes toward mathematics, and to take more mathematics courses.

But all is not well at Railside High. The hard-working teachers and students have been dealt a devastating blow, as the state has decided that they are “underperforming.” Despite outperforming the other district high school and the other schools in our study on varied mathematics assessments, the Railside students scored significantly lower on the SAT-9 than students at the other high schools. In addition, SAT-9 scores at Railside did not “improve” sufficiently over the one-year time period set by the state.

The “underperforming” label conferred upon this school bears no relation to the learning we see in its classrooms. This mismatch between the students’ achievements and the state’s label is unfortunate for many reasons, but it also provides us with an important opportunity to consider what is being assessed in the SAT-9 and what is not. Consider, for example, two of the questions from our assessment that directly assess the competencies in the California standards. On these questions, the Railside students performed at a significantly...
higher level than students from other schools.

1. Here is a rectangle. The sides are $2x + 4$ and 6 units.

   ![Rectangle Diagram]

   a. Find the perimeter of the rectangle. Simplify your answer if possible.
   b. Find the area of the rectangle. Simplify your answer if possible.
   c. Draw and label a rectangle with the same area that you found in part b, but with a different length and width.

2. Solve the following equations:
   a) $5x - 3 = 101$
   b) $3x - 1 = 2x + 5$

These questions differ from those in the SAT-9 in a number of ways. First, they are not set in contexts that are confusing to linguistic-minority and low-income students. Second, they reward all students who attain the correct answers, rather than only those who have answered the questions in the same form as the acceptable multiple-choice answer. Third, they do not use long and confusing sentences. By contrast, consider this question from the SAT-9 test for students of the same grade:

A cable crew had 120 feet of cable left on a 1,000-foot spool after wiring 4 identical new homes. If the spool was full before the homes were wired, which equation could be used to find the length of cable (x) used in each home?

   F. $4x + 120 = 1000$
   G. $4x - 120 = 1000$
   H. $4x = 1000$
   J. $4x - 1000 = 120$

The most obvious difference between our questions and this one is that the SAT-9 question is set in a context with which only some students will be familiar. In addition, it uses long sentences and words unknown to many students new to the country (e.g., spool, cable crew, wired). Note, too, that the expression that would sensibly be used to represent the length of cable used — $x = (1000 - 120) \div 4$ — does not appear as an acceptable answer. This question, as with many others in the SAT-9, assesses many things — confidence in the face of unfamiliar answers, knowledge of context, and language. However, none of these are indicators of mathematics knowledge, and they are all likely to stack the deck against language learners, students from low-income homes, students who are from minority ethnic and cultural groups, and girls.

In other SAT-9 questions, students are asked to consider a student’s bank balance and to calculate the possible values of combinations of nickels and dimes. Students who have a bank account will undoubtedly be advantaged by questions that refer to them. I arrived in the U.S. from England four years ago, and I am still thrown on the rare occasions when I come across the terms “nickel” and “dime,” because they are rarely used in modern-day American society, and I had no cause to learn them before I came to this country.

The publishers of the SAT-9 questions have used these contexts in response to recommendations from the National Council of Teachers of Mathematics and other groups that mathematics be taught through realistic problems and situations. But teaching situations, in which students are learning, are very different from standardized assessments, in which they are being tested. Contexts may be useful and motivating in classroom activities and questions, but their use is minimized in the standardized assessments used in most other countries because it is known that they present barriers to some groups of students and not to others and so they contribute to inequalities.

In interviews, the students at Railside reported that they found the SAT-9 totally confusing, mainly because of the language and contexts used in the mathematics questions. But our research has found that there was still another, more insidious, factor affecting the Railside students’ performance on the SAT-9. The students had been told by the state that their school was “underperforming,” so they did not expect to do very well on the tests.

Was this important? Claude Steele has shown the importance of “stereotype threat.” He found that, when students were told that the test they were about to take tended to produce achievement differences, with women and minority students scoring at a lower level, this is exactly what happened. In the control groups, where students took the same test but were not told about any expected performance differences, there were no performance differences among different groups of students. Educational research — a field that often produces contradictory results — shows remarkable con-
sistency on this issue. If you tell students they are low achievers, they achieve at a lower level than if you do not.

In one of our first visits to Railside, a young boy asked us why we were looking at the mathematics department. When we replied that it was interesting, he frowned quizzically and said, “But we are a 3.” He was referring to Railside’s score of 3 out of 10 on the Academic Performance Index (API), which is used to rank the state’s public schools based on student performance on the SAT-9. In recent interviews, the students told us that they go to a “ghetto” school. Students from other schools had told them so. The Railside students struggle to make sense of the label, as they believe that the teaching at Railside is good and that the teachers really care about them. But they have been seriously affected by these labels that have emanated from the state’s use of SAT-9 scores.

In addition to the labeling of the school overall, students received their own, individual labels when they took the SAT-9. Parents and students at Railside, as in all other schools in California, were sent the results of the students’ SAT-9 tests, displayed on a graph divided into three sections that are marked “above average,” “average,” and “below average.” Of course, such a label tells nothing of what students have learned over any period of time.

One of the students we interviewed, Simon, had arrived in the U.S. from Nicaragua as a young boy. He told us that elementary school was a time of constant failure for him, as he couldn’t understand what the teachers were saying. But he has since caught up and is now excelling in school. He told us that the teachers at Railside told him that he was smart and that he started to believe in himself and achieve. He now loves mathematics and is very appreciative of the teaching at the school. He performed extremely well on our assessments. Despite all of this, when the SAT-9 results arrived at his door, he started to question his ability: “My parents, they saw in the SAT-9 graph thing I was below average in the majority of the things and especially math. I was like — below average. Right there. The thing is like — below average — you want it to be a little bit above average.”

I asked Simon whether that SAT-9 report affected how he thought of his abilities as a mathematics learner. He said that it did. “You say — you tried so hard and then suddenly they give you a paper where it says you’re below average and you’re like, What? I did so much work.”

Simon had reasonably assumed that the result he was given should tell him something about how hard he had worked or about what he had learned in mathematics. But plainly it did not.

The teachers at Railside are concerned that the SAT-9 results that are sent home to parents will convey negative messages about the students’ mathematics learning. In response, they have started to organize portfolio days, when students show their parents their mathematics work and all that they have achieved. These are very positive, well-attended events, but Simon’s repetition of the term “below average” reminds us of the extent to which such labels are internalized and remembered.

Almost half of all students in California are told that they are “below average.” In the new system that will be used in California, students will not take the SAT-9. Instead, they will take a different test that also includes only multiple-choice questions and that will confer labels that seem even more damaging. In the new system, approximately half of all students will be told that their attainment is “basic,” “below basic,” or even “far below basic.” What impact, I wonder, do those who designed this system think that these labels will have on students’ confidence? On their future mathematics achievement? Research tells us that confidence in one’s ability to succeed in mathematics is an intrinsic part of success and motivation. The labels that the students at Railside received are working against the positive achievements that their teachers had brought about.

Now Railside students are being told by students in the other district school that they attend a “ghetto school” because their test scores are low. Of course, Railside students outscored the other school on our test, and the researchers I know who have spent time in the school agree that it has one of the most professional and dedicated mathematics departments they have ever seen. The hard-working mathematics teachers have,
suggests that the low performance of students in the SAT-9 at Railside is related less to mathematical understanding than to language, context interpretation (which relies heavily on language), and test-taking skills. Using such tests — and their associated labeling — as a tool to increase the performance of underachieving students, particularly those from low-income and ethnic minority homes, does not seem to be a wise decision on the part of policy makers.

What’s more, even if the SAT-9 were a good test of students’ mathematics achievement — and I strongly believe it is not — a single test can measure the students’ attainment only at a certain point in time. It does not track any increases in learning. When the British government moved from reporting test scores to reporting students’ improvement over their years in school, the list of schools — ranked by “success” — was completely rearranged. When “improvement” (i.e., learning) became the criterion, many of the excellent schools in low-income areas moved to the top of the lists.

The designers of the API system in California have tried to assess learning by focusing on the improvement of schools, but this is measured by comparing the scores of one cohort of students with the scores of a later cohort. Such a system not only compares different students, without capturing the learning of any particular students, but it is also open to considerable abuse. Schools that push needy students away and into other schools will report greater “improvements” than schools that keep their students with special needs and work to help them. And such exclusionary practices, aimed at higher improvement ratings, are already being reported.

Simon, the student from Nicaragua, may have worked harder, learned more, and improved more than anyone else in the country over the previous year, but a SAT-9 test score could never give any indication of that. The SAT-9 does not measure learning, and events in this school suggest that its main effects can be to lower students’ perceptions of what they can do and to demoralize teachers. The place of this and similar regimes of testing and labeling in an education system that purports to hold higher — and more equitable — achievement as its goal seems questionable at best. The students and teachers at Railside High have discovered this fact, at considerable personal cost.

