

# Crucial Issues in California Education 2000:



Are the Reform Pieces  
Fitting Together?

Policy  
Analysis for  
California  
Education  
PACE



# Crucial Issues in California Education 2000:



## Are the Reform Pieces Fitting Together?

Policy  
Analysis for  
California  
Education  
PACE

**Elizabeth Burr**  
*UC Berkeley*

**Gerald C. Hayward**  
*Sacramento*

**Bruce Fuller**  
*UC Berkeley*

**Michael W. Kirst**  
*Stanford University*

## Preface and Acknowledgments

*Crucial Issues in California Education 2000* is a successor to *Conditions of Education*, a PACE publication since 1984. *Conditions* combined up-to-date data and ongoing trends in a wide variety of indicators relevant to state education policy. In recent years, education in California has become more complex, undergoing both strident criticism and renewed support. To present a more analytical overview of California education, this year PACE has asked experts around the state to contribute chapters based on in-depth research projects. Their contributions allow PACE to offer the latest data analysis around a wider variety of issues, while continuing to provide overall strategic recommendations. This volume provides a unique function in policy analysis because it brings together numerous reports on components of California education in one source. Moreover, the scope of *Crucial Issues* is the largest in the history of the series, spanning child care to universities.

We have received generous financial support from several foundations. First, PACE could not survive without core support and advocacy from Ray Bacchetti at the William and Flora Hewlett Foundation. The Stuart Foundation has been helpful since 1998 when Jane Henderson came to an early planning meeting. Since then, Ted Lobman has been a strong supporter. Lisa Carlos and Joan Herman also attended early planning efforts. Robert Shireman at the James Irvine Foundation provided funding for the newly structured volume. The work reported in Chapter 7 was also supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The study described in Chapter 9 was commissioned by Stanford University's Bridge Project: Strengthening K-16 Transition Policies, a national study funded by the Pew Charitable Trusts and the U.S. Department of Education's Office of Educational Research and Improvement.

Warm thanks are extended to Sarah Baughn for early help in research and organization, to Tor Ormseth and Peter Scott for providing research assistance, and to Marsha Ing, who assisted with the analyses. Kay Cooperman provided editorial assistance and Rachel Montgomery supervised final production and tirelessly checked references. At the project's end, Judith Kafka provided speedy and meticulous copyediting.

Many people gave their time for careful review of chapters: Richard Duran, Eugene Garcia, Dave Jolly, Barbara Merino and Allan Odden. Lisa Carlos, Diane Hirshberg, Vicki Lavatos, Robert Manwaring and Rich Shavelson also offered help and guidance on several chapters.

We thank Jennifer Garner for her vision and talents in design, and for handling production so quickly and carefully. David Ruenzel brought patience and humor to the project, offering substantive and careful critiques, as well as original writing. He edited the entire manuscript.

Finally, we would like to show appreciation for our PACE team, especially Terry Alter, Regina Burley, Robert Dillman, and Diana Smith.

Elizabeth Burr  
*Berkeley*

Gerald C. Hayward  
*Sacramento*

Michael W. Kirst  
*Stanford*

Bruce Fuller  
*Berkeley*

# Crucial Issues in California Education 2000: Are the Reform Pieces Fitting Together?

## Table of Contents

Chapter 1	<b>California’s Ambitious Education Reform Agenda: Will It Energize Schools and Teachers?</b> <i>David Ruenzel</i>	<b>1</b>
Chapter 2	<b>Early Education and Family Poverty</b> <i>Elizabeth Burr and Bruce Fuller</i>	<b>9</b>
Chapter 3	<b>The Schooling of English Learners</b> <i>Russell Rumberger and Patricia Gandara</i>	<b>23</b>
Chapter 4	<b>School Finance</b> <i>Neal Finkelstein, William Furry and Luis Huerta</i>	<b>45</b>
Chapter 5	<b>Governance and Accountability</b> <i>Michael W. Kirst, Gerald C. Hayward and Bruce Fuller</i>	<b>79</b>
Chapter 6	<b>Teacher Quality</b> <i>The Center for the Future of Teaching and Learning</i>	<b>95</b>
Chapter 7	<b>Student Assessment and Student Achievement in the California Public School System</b> <i>Joan L. Herman, Richard S. Brown, and Eva L. Baker</i>	<b>113</b>
Chapter 8	<b>Connecting California’s K-12 and Higher Education Systems: Challenges and Opportunities</b> <i>Andrea Venezia</i>	<b>153</b>
Chapter 9	<b>Alignment Among Secondary and Post- Secondary Assessments in California</b> <i>Vi-Nhuan Le, Laura Hamilton and Abby Robyn</i>	<b>177</b>

## Chapter 7

# Student Assessment and Student Achievement in the California Public School System

Joan L. Herman, Richard S. Brown and Eva L. Baker  
*National Center for Research on Evaluation, Standards,  
and Student Testing (CRESST)*  
*University of California, Los Angeles*



### Introduction

**M**ore than 15 years ago, a prominent national commission declared us a nation at educational risk, noting *a rising tide of mediocrity that threatens our very future as a nation ...*<sup>1</sup> A decade later, California received its own special wake-up call when results from the 1990 and 1992 National Assessments of Educational Progress' state-by-state comparisons revealed that California students were scoring near the bottom nationally in eighth-grade mathematics and fourth-grade reading. California students surpassed only those in Mississippi, Washington, DC, and the Virgin Islands on the 1992 reading assessment. What of the situation today? How are California's students faring? Are our students making progress toward the rigorous standards that have been established for their performance? Are our schools improving? Are they better preparing our students for future success? As we strive toward excellence, who is being helped most and who the least by California's educational system?

Such seemingly simple, bottom line questions are foremost in the minds of the public and its policy-makers. Yet answers are more complex to formulate, made more so by the

history and current status of the state's assessment system, the nature of other available indicators of educational quality, and the imprecision of all assessments. Below, we first provide a context for examining the progress of students and schools by reviewing California's recent testing history and the state's progress in creating a sound, standards-based assessment system. We then review available data about student performance, examining how schools are doing and the factors which most influence assessment results. We close by returning to the goals of accountability and standards by which such systems should be judged.

### Where California's Assessment System Is Today and How It Got There

California, as the rest of the nation, is creating statewide assessment systems intended not only to measure student learning, but to leverage its improvement. The system itself is intended as part of the reform: It signals what is important to teach and learn by providing specific learning targets—i.e., the content of the test. The assessment also is intended to provide feedback

on how students are doing and thus enable school leaders to diagnose curriculum strengths and weaknesses. Coupled with sanctions and/or incentives, the assessment is expected to motivate educators, students and their parents to pay attention and act to improve their performance. As measurement experts have aptly put it, WYTIWYG—what you test is what you get, a phenomenon that any number of research studies have confirmed.<sup>2</sup>

### *How Does California's Current Assessment System Measure Up?*

Put simply, California's current system is still evolving toward a standards-based system, and the base requirements are not yet in place. As the result of the rocky and changing story of the state's plans over the last few years, the basic requirement for analyzing students' progress—a consistent measure used over time—is not yet available and the system's alignment with state standards remains problematic.

First, a short history. Beginning in 1993, the California Learning Assessment System (CLAS) was to be the primary measurement of student achievement in the state. CLAS was largely a performance-based assessment system, although it included both multiple choice and open-ended items. CLAS focused on the complex thinking and problem-solving aims of the state's curriculum frameworks in place at the time. CLAS came to an early demise after just two years because of both technical quality and public credibility concerns.

Following CLAS, instead of a common, statewide assessment, the state provided financial incentives to school districts to select and administer assessments that best reflected their local standards. The result was a plethora of

different standardized tests<sup>3</sup> being given across the state.

Meanwhile, as the state embarked anew on establishing statewide standards for student performance, the testing plan changed again the next year. Impatient to establish a baseline and hold schools accountable, the state (and particularly then-Governor Wilson) initiated in 1998 the new California Standardized Testing and Reporting (STAR) program. The centerpiece of STAR was and continues to be the Stanford Achievement Test Series, Ninth Edition, Form T (SAT-9) administered in grades 2 through 11. Thus, in contrast to an ideal scenario where a testing system would be selected or developed based on a state's standards, the initial STAR test pre-dated the state's standards. Unlike CLAS, the SAT-9, it should be noted, is a norm-referenced test, designed primarily to show how students or schools perform on basic skills relative to others—others in the state, others in similar schools and districts, others in the national norm group (or average).

California's adoption of the SAT-9 occurred at a time when most other states were making progress in meeting federal expectations for state standards and assessments. The federal plan, originally designed in 1991-92,<sup>4</sup> was given additional impetus by the Improving America School Act,<sup>5</sup> in which Title I, an act to support disadvantaged students, was lodged. Title I made the receipt of funds contingent on the development of standards and assessments that met criteria, including the use of multiple measures, assessments for children with different language backgrounds, and measures of progress.

With California's Board of Education's passing of state standards in December 1998, plans to retrofit the testing system to the standards

began. To provide a comparable measure, the plan featured the continued administration of the same SAT-9 that had been administered in Spring 1998, but in 1999, additional items were included to bring the test into better alignment with the state's standards. Thus was born the SAT-9 augmentation, additional items that the test publisher selected or developed to fill in some of the gaps between the existing SAT-9 and California's content and performance standards. With the augmented items, the SAT-9 would then eventually provide both norm-referenced and standards-referenced scores. The norm-referenced scores would communicate to parents, the public, students, and educators how students were performing relative to other students nationally. The standards-referenced scores would tell those stakeholders if students were meeting state-defined content standards at advanced, proficient, basic, or below-basic levels. As we describe later in this chapter, although a first set of augmented items was administered in spring 1999, there are some questions about their appropriateness. Performance standards have not yet been established for them, so results from the augmentation are not yet directly interpretable.

Additional components of STAR are in the works to bring California's assessment system into still closer alignment with the state standards. The California Assessment of Applied Academic Skills (CAAAS), the so-called "Matrix" test, is to be designed to focus on the disciplinary thinking and problem solving capabilities which are reflected in the standards, but not well assessed by the SAT-9. Since the SAT-9's multiple-choice items alone cannot assess the broad range of important thinking and

communication skills, the Matrix test is to include open-ended and performance-assessment tasks, such as asking students to explain their thinking or write an essay. This component will model the types of learning which are expected of students and preclude an exclusive focus on "drill and kill" formats in classroom instruction that often are encouraged by multiple choice test formats. The Matrix test employs a matrix sampling framework where the accuracy and comprehensiveness of the assessment are improved by having some students within a school respond to some assessment tasks while other samples of students respond to different tasks. Given that each open-ended and performance assessment takes substantially more time to administer than a multiple choice item, matrix sampling improves the overall coverage for the school as a whole while minimizing the time each student is required to spend taking an assessment. While it has not been designed to yield a score for each student, it does provide school-level results for judging the quality of a school's curriculum and instruction and students' collective achievement and progress at that school. CAAAS currently is scheduled to go operational in 2001.

A high school exit exam in language arts and mathematics is the most recent addition to the state's standards-based assessment arsenal. Enacted as part of Governor Davis' first 100 days education agenda, the exit exam will be required for high school graduation and is scheduled to go operational in 2004. An English Language Development Test also is under development (see the chapter on English Learners in this publication.).

### *Academic Performance Index*

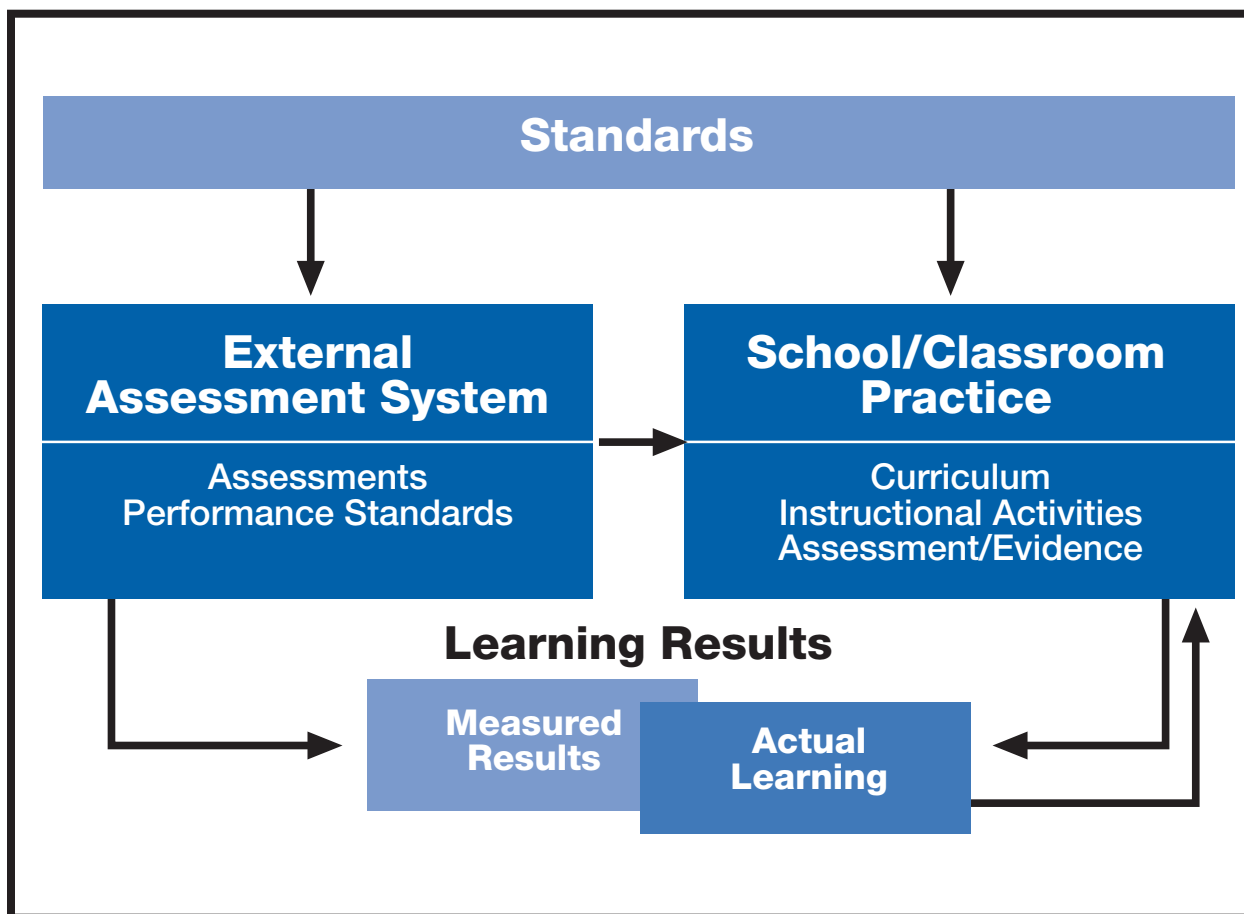
The components of the STAR program thus are abundant and in fact, there is continuing debate about whether the assessment load is too high and whether all planned components are necessary to achieve the system's goals. Even as the system components are under discussion and development, California already has developed a high stakes school performance index based on them. The Academic Performance Index (API) is being used to rank schools across the state based on their SAT-9 test scores. More will be described about the API in our concluding section.

### **The Assessment Context: Setting the Course with a System in Transition**

As we've noted, the ideal assessment system is developed after a state's content and performance standards are established in order to be in alignment with what students are learning. Since California's assessment system followed a different course, it remains a system in transition.

### *The Importance of Alignment*

The alignment between what is tested and what students are expected to learn is a critical



**Figure 1. Model of Standards-Based System**

criterion for any assessment or accountability system intended to promote the improvement of student learning and is the essence of current standards-based reform. As displayed in Figure 1, the idea is not really to teach to the test per se, but rather that both testing and teaching reflect the standards we hold for student performance. When standards, testing and instruction are in synchrony, the logic of the system works to leverage better performance. When not, then holding schools accountable and encouraging them to use the assessment results may not promote the standards we seek.

Consider, for example, the case where the assessment doesn't well reflect the standards. Under pressure to show improvement, schools

and teachers may use test results to modify their curriculum and instruction, but moving toward the test does not mean movement toward the standards. Minimally, the test and the standards are sending conflicting messages, which can cause confusion and dilute the focus of school efforts. Or consider where there is a poor match between what is taught and what is assessed. Here, while the results may tell us about gaps in the curriculum, they tell us little about the quality of instruction and teaching in that school. Even under the best scenario, as Figure 1 portrays, assessment results reflect only a portion of what students have learned and what they know and can do. In other words, the test is a reflection of standards and goals, it is not the goal itself.

#### Characteristics of Quality, Standards-Based Assessment Systems

- **Alignment.** Does the assessment reflect content and performance standards that have been established for students? Is the assessment content consistent with the best current understanding of the subject matter? Do it reflect the enduring themes and/or priority principles, concepts and topics of the discipline?
- **Instructional sensitivity.** Can the test detect differences in the quality of instruction? Does the test measure learnable and teachable knowledge, rather than simply general factors such as general ability or language background?
- **Technical quality.** To what extent are results reliable and consistent? Comparable over time and setting? Do the results enable accurate generalizations about student learning and achievement relative to standards?
- **Fairness.** Does the assessment enable students, regardless of race, ethnicity, gender or economic status, to show what they know and can do? Have students had the opportunity to learn what's being assessed?
- **Meaningfulness.** Do parents, teachers, students and the public find the assessment worthwhile and credible?
- **Consequences.** To what extent do the assessments model and encourage good teaching practice? Are intended positive consequences achieved? What are the unintended negative consequences?
- **Multiple Measures.** Does the mix of measures optimize alignment, technical quality, fairness, meaningfulness and consequences criteria?

\* Adapted from CRESST Criteria for Evaluating Assessment Quality (Linn, Baker and Dunbar, 1991) (National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, UCLA)

Figure 2. Characteristics of Quality Standards-Based Assessment Systems

### *Other Criteria for Quality Standards-Based Assessment Systems*

How well the results of an assessment system represent student learning is a complex validity issue and one which has driven traditional concerns for technical quality. One asks about the reliability, accuracy and consistency of measurement, at the same time acknowledging that there is error in any measure and that all tests are fallible—some more than others. But even alignment and indices of technical quality provide an inadequate base for evaluating the soundness of any assessment system. History shows that a number of other features of assessments are important to a quality system, the major ones of which are summarized in Figure 2.

Consider the importance of instructional sensitivity. If the assessment does not measure efforts made in the classroom—even if it nominally “matches” standards—it will be a poor device to provide feedback for improvement. Instead, scores will misrepresent the reality of serious educational reform. They may indicate improvement that might happen spontaneously with or without reform. Note also the final characteristic in our list—multiple measures—which is necessary to achieve the other listed criteria. It is unlikely that a single measure can adequately capture our goals for student performance or enable all students to show what they know. Some types of measures are efficient and cost effective for some purposes but have unintended consequences for other purposes. For example, multiple-choice tests can be highly efficient, cost effective, and reliable, but an over-reliance on such testing in the 1980s led to a narrowing of the curriculum to basic skills and an overemphasis on “drill and kill” types of instruction.<sup>6</sup> Different constituent-

cies, furthermore, find different types of information meaningful and useful. For example, basic skills are high among the public’s priorities and parents and the public often want to know how their children compare with others—nationally and internationally. Educational reformers and futurists, on the other hand, emphasize the importance of all children achieving high levels of skill in communication, problem solving and ability to learn and change—abilities which may not be well assessed through multiple choice testing.

### *Assessing Limited English Proficient Students*

While state code requires all students in grades 2 through 11—including those who are not fully proficient in English—to take the SAT-9,<sup>7</sup> it also provides that limited English proficient (LEP)<sup>8</sup> students who have been in school less than 12 months also be tested in their primary language. Students who have been in school more than 12 months but are still classified as LEP may also be administered a primary language test. The state has selected the Spanish Assessment of Basic Education, Second Edition, (SABE/2) as the statewide measure to be used for assessing students whose first language is Spanish, which begins this year. Currently, then, different districts are using different measures so it is not possible to know statewide how Spanish-language students are doing based on tests in their primary language.

The assessment of LEP students continues to be highly controversial. On the one hand, testing students in a language they do not understand does not allow them to show what they know and can do in content areas such as math and science, raising questions about the extent to which *fairness* criteria are being met in the state’s

system. On the other hand, it is important that LEP students' achievement and progress be monitored in publicly visible ways, and that schools be held accountable for all their students. The *consequences* of not testing and reporting LEP students' performance is that their progress and their needs may be ignored.

Testing in students' primary language at first glance might seem a better and fairer option. However, research shows that primary language testing only helps those students who have been instructed in their native language<sup>9</sup>—a circumstance which current education code prohibits for LEP students who have been in this country for more than a year. Statewide testing of English language proficiency will soon enable the state to at least monitor LEP students' progress in acquiring English, providing another measure that is potentially more sensitive to individual students' achievement and progress. Testing accommodations which attempt to reduce the language load of a test or otherwise compensate for students' reduced language skills (e.g., allowing students more time to take tests) also are currently being researched, but a solution that is equitable and fair for all students has not yet been found. Measurement experts, however, largely agree that test results of LEP students should be separated from those of English proficient students, and that the validity and utility of individual scores for LEP students on English language exams is limited.<sup>10</sup>

### ***Other Indicators of Quality***

Beyond the components of STAR, there are other statewide indicators that can be used to judge the quality of student performance.<sup>11</sup> As mentioned above, multiple indicators are

important to a balanced and valid view of any educational system. Some of these indicators act as counterbalances to others and are particularly relevant for different sub-populations. For example, the high school drop-out rate is of interest in itself, but also to assure that schools are not achieving higher test scores at the cost of more children leaving the system. Advanced placement exams, which are given to high school students who take college-level courses at their high schools through the College Board,<sup>12</sup> provide an indicator of how schools are serving their highest-ability students. As described further below, both the number of exams taken and the proportion passing are of interest. Similarly, college entrance exams, such as the SAT, provide an indicator of both students' expectations and preparation to attend college.

Other indicators are external to the K-12 system and provide a validity check of its academic quality. The National Assessment of Educational Progress (NAEP) periodically assesses national performance in the major subject areas—reading, mathematics, science, writing, etc. States participating in NAEP's state-by-state program are able to compare their performance to that of other states as well as nationally. College placement tests, which are used to decide whether entering college students have adequate mathematics and writing skills to handle college coursework or need remedial help, provide another external comparison point for judging the quality of the states' pre-collegiate systems.

### ***Alignment and Consistency***

The alignment of these various indicators of student performance is an issue under current

discussion. Some believe that college entry tests, such as the SAT, and college placement tests ought to be aligned with the state's standards and with the state's K-12 assessment system. Advocates believe that this would not only provide greater consistency and focus to California schools but would permit greater efficiency in testing. For example, they project scenarios where the state's graduation tests would serve a role in the college selection and placement process.

Consistency and alignment of each of these indicators with state standards aside, one looks for consistency in performance across various indicators to judge the quality of California's academic achievement. Although any individual indicator is flawed, when multiple indicators show consistent direction, we can be more confident of the breadth of our perspective and the validity of our conclusions. We now turn to a consideration of those indicators.

## **Student Achievement in California Public Schools**

A serious understanding of student performance in California requires in-depth knowledge of the wide variety of student achievement measures we've outlined thus far. In the next few pages, we'll describe those instruments, what they are intended to measure or monitor, and how well California schoolchildren are doing on them. We'll review data from both the most recent testing period and over a longer period of time to help the reader understand the status and progress of California performance.

We'll begin with a look at the state's standardized testing system, the program that applies to all students in the public educational system from elementary school through high school. Next we will analyze information regarding California's performance on NAEP. From there, we'll examine the results of a series of secondary school measures, including high school drop-out and graduation rates, advanced placement (AP) test results, course-taking patterns, and college entrance examination performance. To address the longer-term impact of public school, we will also present data on college attendance and preparedness by considering findings on reading remediation tests for college freshmen in the University of California system. Finally, we'll comment on some of the demographic trends for California students over the last decade and venture a summary judgment across this collection of information on what the state of academic achievement in our California public schools is and whether there is evidence it is headed in the right direction.

### ***STAR Results***

As we've noted, California began to implement STAR in 1998 with the SAT-9. In the sections below, we'll look at how well California students performed on the norm-referenced SAT-9 in reading, mathematics, language arts, spelling, science, and social studies in the 1997-'98 and 1998-'99 academic years, with some words of caution about the interpretation of the scores. We will follow with analyses of the performances of LEP students and students who are economically disadvantaged, and compare how

the performance gaps between these groups and others vary across different school contexts.

### **How are California’s students doing on the SAT-9?**

Before examining how California students are doing overall, it would do us well to review what the results from a norm-referenced test mean. The results tell only generally what students know and can do. The real information they provide is how California students’ performance compares with that of a national norming group. Results often are reported in terms of percentile scores, which reflect where students’ scores fall relative to the national distribution. For example, if a student scores at the 40<sup>th</sup> percentile, it means that the student’s performance equaled or exceeded 40% of the national norm group. A score at the 50<sup>th</sup> percentile—which the public often considers “average”—means that the student’s performance equaled or exceeded half the national norm group. Thus the nature of percentile scores means that some students will be above and some below the “average” relative to the norming group.

The national norming group is intended to represent students nationally. Ideally, for norms to be interpreted easily, the kinds of students tested in a particular state would be similar to those in the norming group. In the case of California, interpretation of the test results is difficult for a number of reasons. First, while no tested and norming groups are ever exactly alike, California’s student population differs substantially from the national norm group in its diversity and its urban concentrations. Plus, unlike other states, California assesses virtually all of its students using an English-language

examination, even though approximately a quarter of them are not fully proficient in English. It is not hard to predict that students who do not understand English are likely to fare poorly when compared to a national norm group consisting of only two percent of similarly non-English proficient students.

Thus, when we look on average at the results of all California students, it is not surprising to find that California students score below average (50th percentile) in practically all subject areas and in almost all grade levels compared to the national norm group. On the reading tests for grades 2 through 11, scores ranged from the 32nd to the 44th percentile in 1998, and from the 32nd to the 46th percentile in 1999. Average scores failed to exceed the 50th percentile at any grade level in either year, and performance shows a precipitous drop at the high school level.<sup>13</sup>

Observed scores were somewhat better in mathematics, where scores ranged from the 39th to the 50th percentile (grade 9) in 1998, and from the 44th to the 52nd percentile in 1999. For 1998, only grade 9 showed average scores above the 50th percentile. In 1999, grades 2, 6, and 9 showed average scores above the 50th percentile. In all other grades, average performance for California students was lower than the national average.

The subject areas of language arts and spelling showed similar levels of performance. In language arts, only one grade level (grade 7) exceeded the national average in 1999. None did so in 1998. For spelling, no grade levels surpassed the 50th percentile in either year.

Similarly, none of the three grade levels (grades 9-11) taking the science test demonstrated average performance above the 50th percentile in 1998 or 1999. In social studies, only

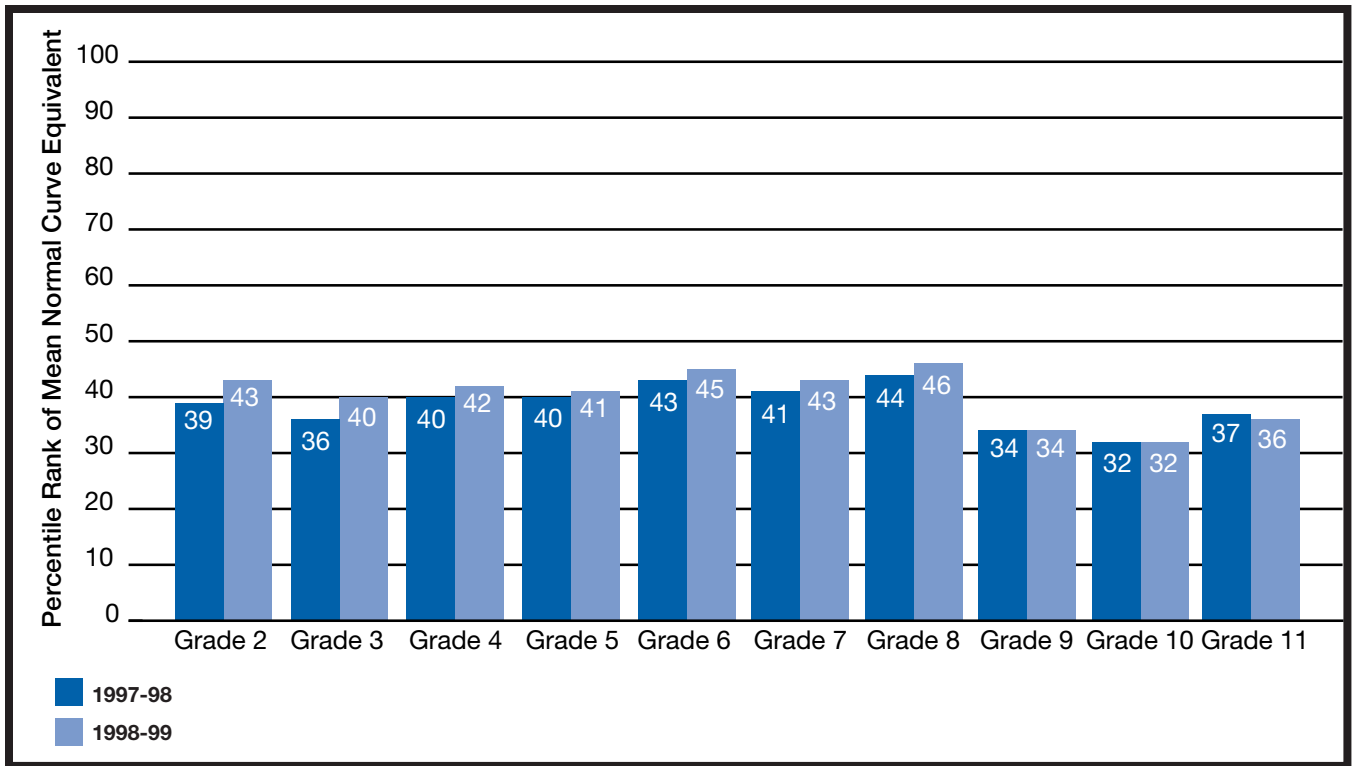


Figure 3. SAT-9 Reading Scores

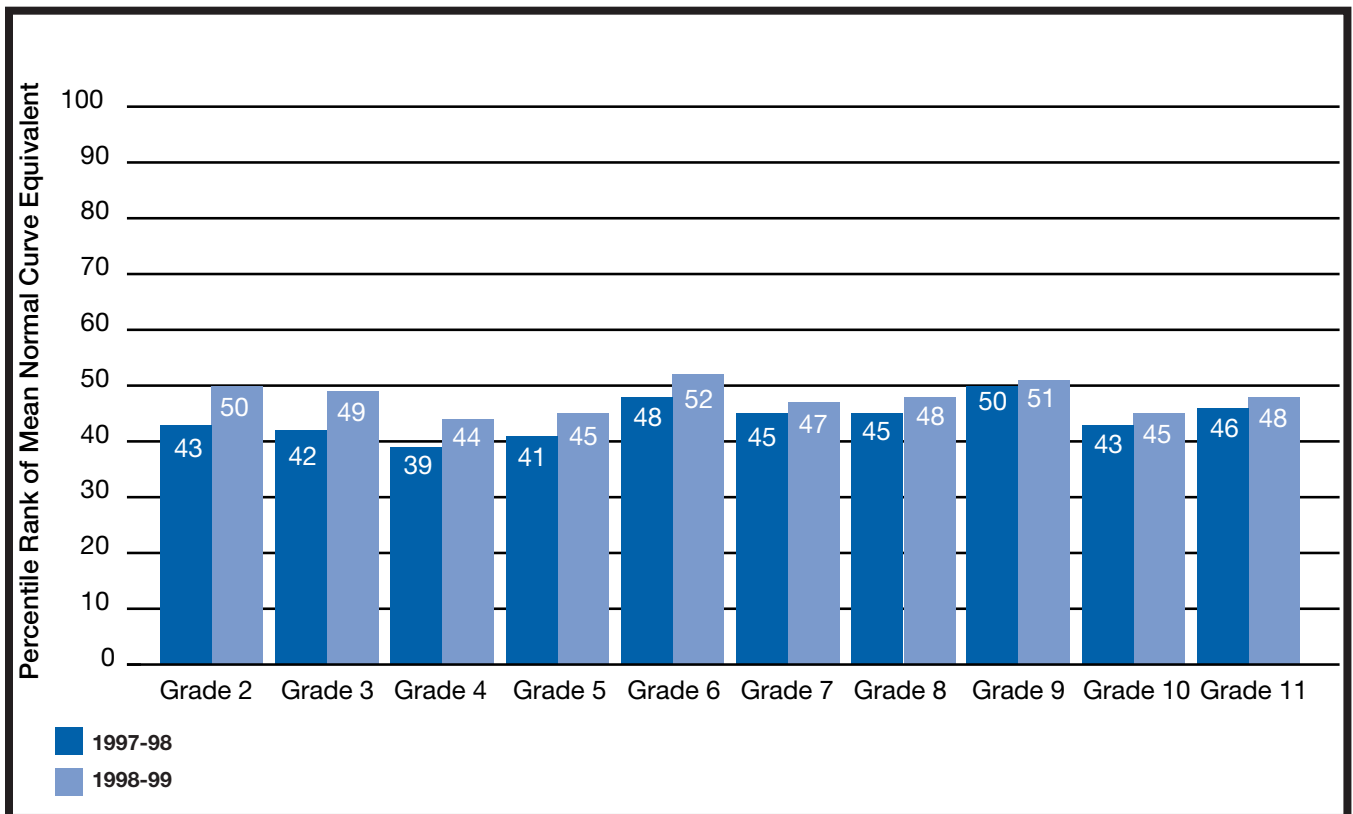


Figure 4. SAT-9 Math Scores

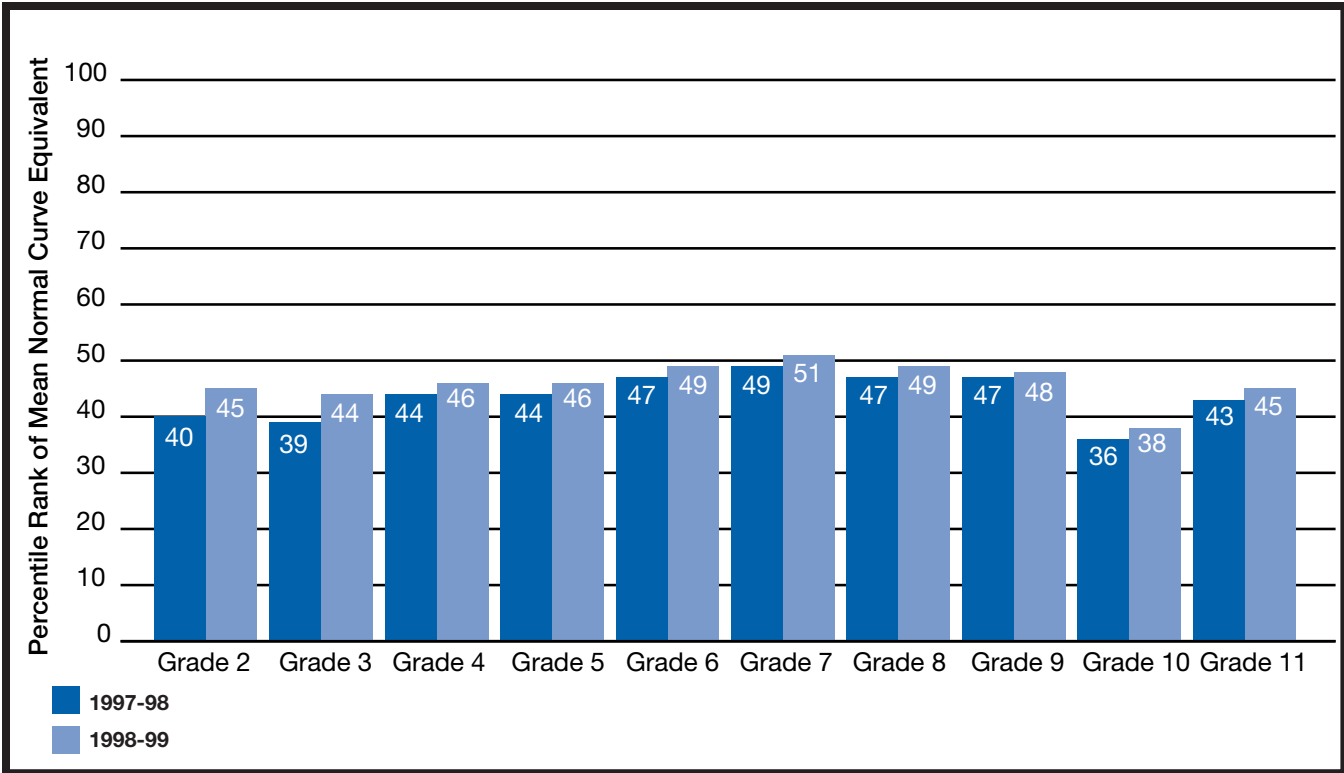


Figure 5. SAT-9 Language Scores

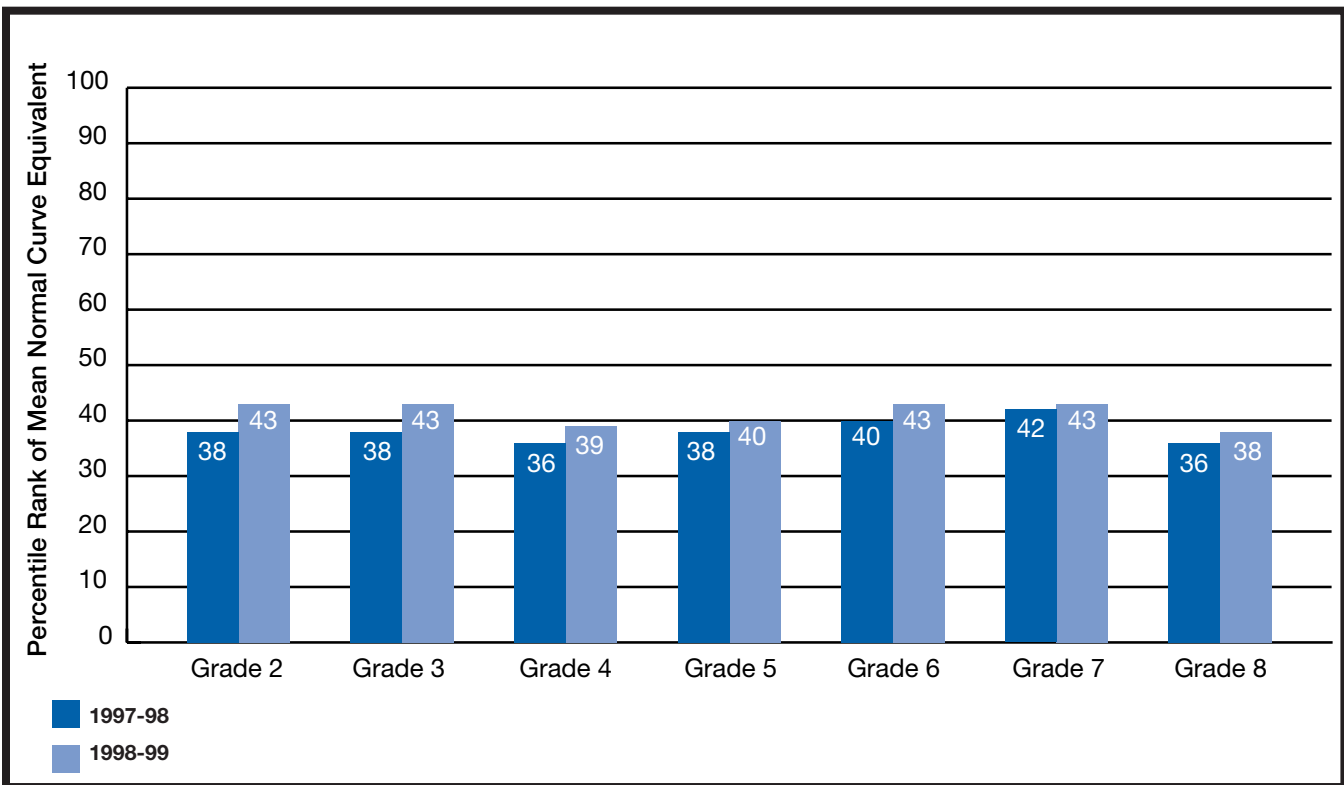


Figure 6. SAT-9 Spelling Scores

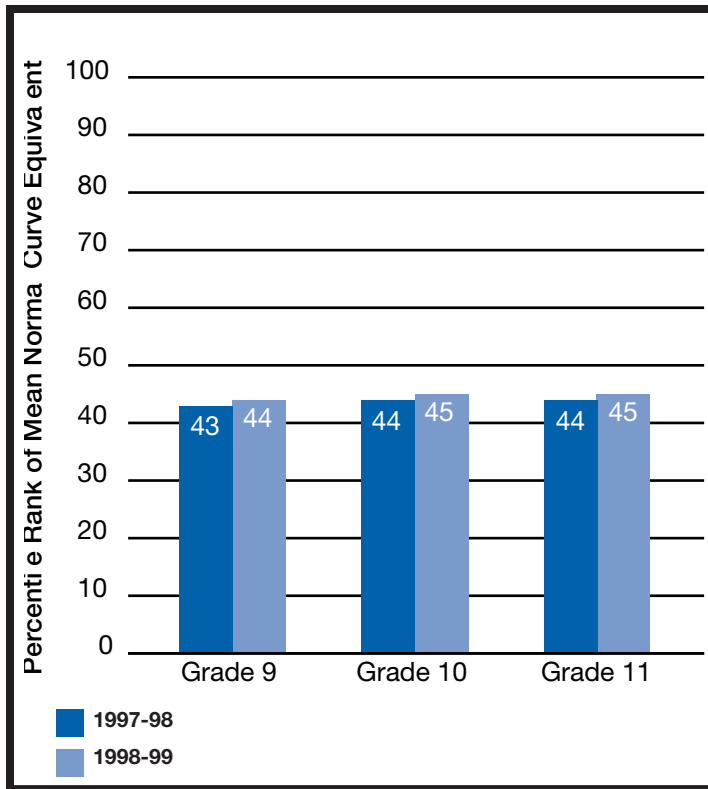


Figure 7. SAT-9 Science Scores

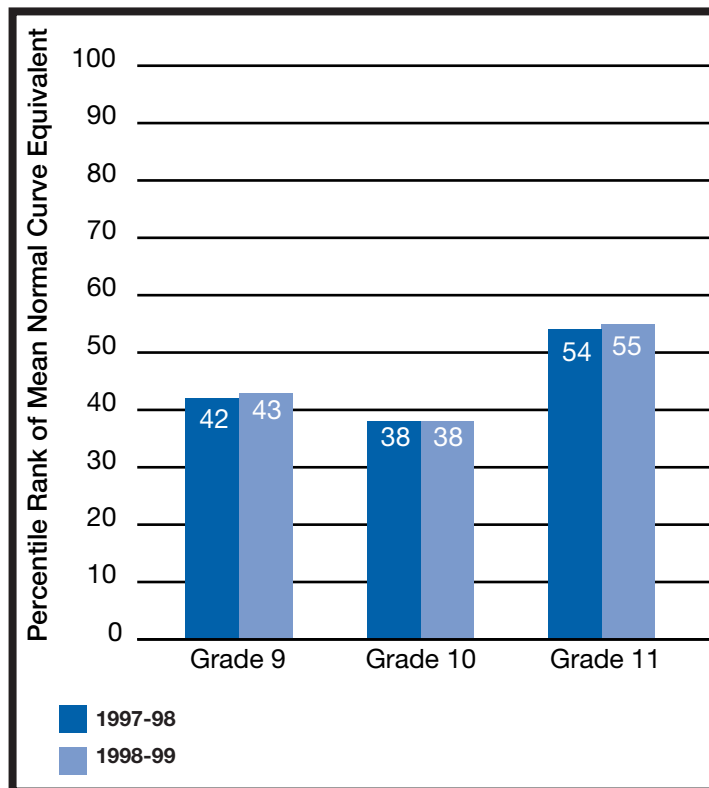


Figure 8. SAT-9 Social Studies Scores

grade 11 showed average performance above the national average, doing so in both 1998 and 1999. Average grade 9 performance in social studies came in at the 42nd and 43rd percentiles in 1998 and 1999, respectively. Average grade 10 performance was lower, reaching only the 38th percentile in both years.

### How are California's English-proficient students doing on the SAT-9?

One gets a slightly different picture, however, from looking solely at the results of California students who are fully proficient in English, a comparison that somewhat favors California students, since approximately two percent of the national norm group is not proficient. Here, the 1999 results show that California's English-proficient students are generally scoring at or above the national average. Differences between all students and English-only students are most pronounced in reading, as we might expect, at the elementary school level (grades 2-5). Yet student performance is still the best, relatively, in mathematics. And regardless of the comparison group, California students are performing the poorest, relatively, in spelling at the elementary school level and in science and social studies at the high school level.

### Are California's schools improving?

Comparisons between scores from the initial year (1998) of STAR and the most recent year (1999) are inevitable. Many claims of "improvement" or "progress" have been made based on such comparisons. However, a word or two of caution should be issued. First, observed test scores<sup>14</sup> are not without error. That is, observed scores—the score students receive when they take a test—do not exactly represent their actual or true capability, due to errors of measurement.

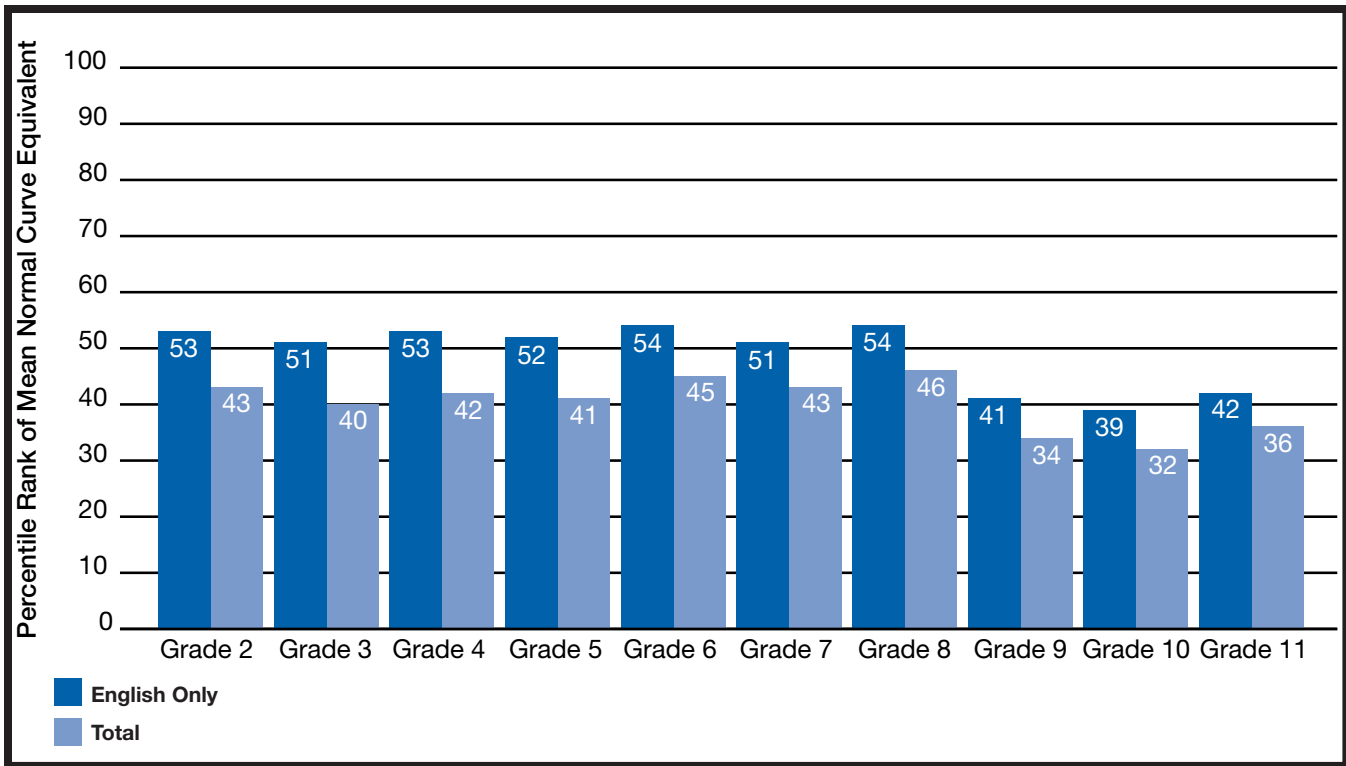


Figure 9. SAT-9 Reading—All Students versus English Proficient Students

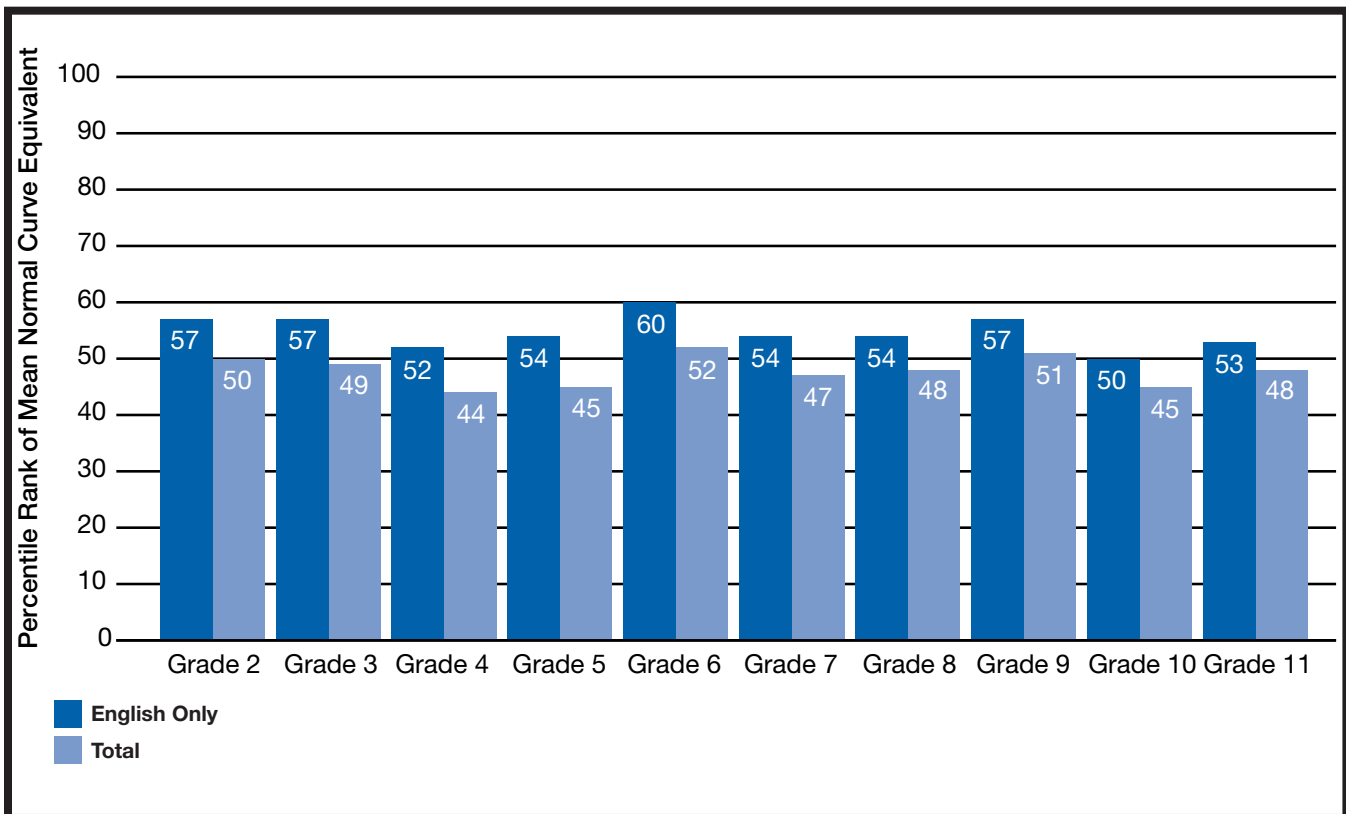


Figure 10. SAT-9 Math—All Students versus English Proficient Students

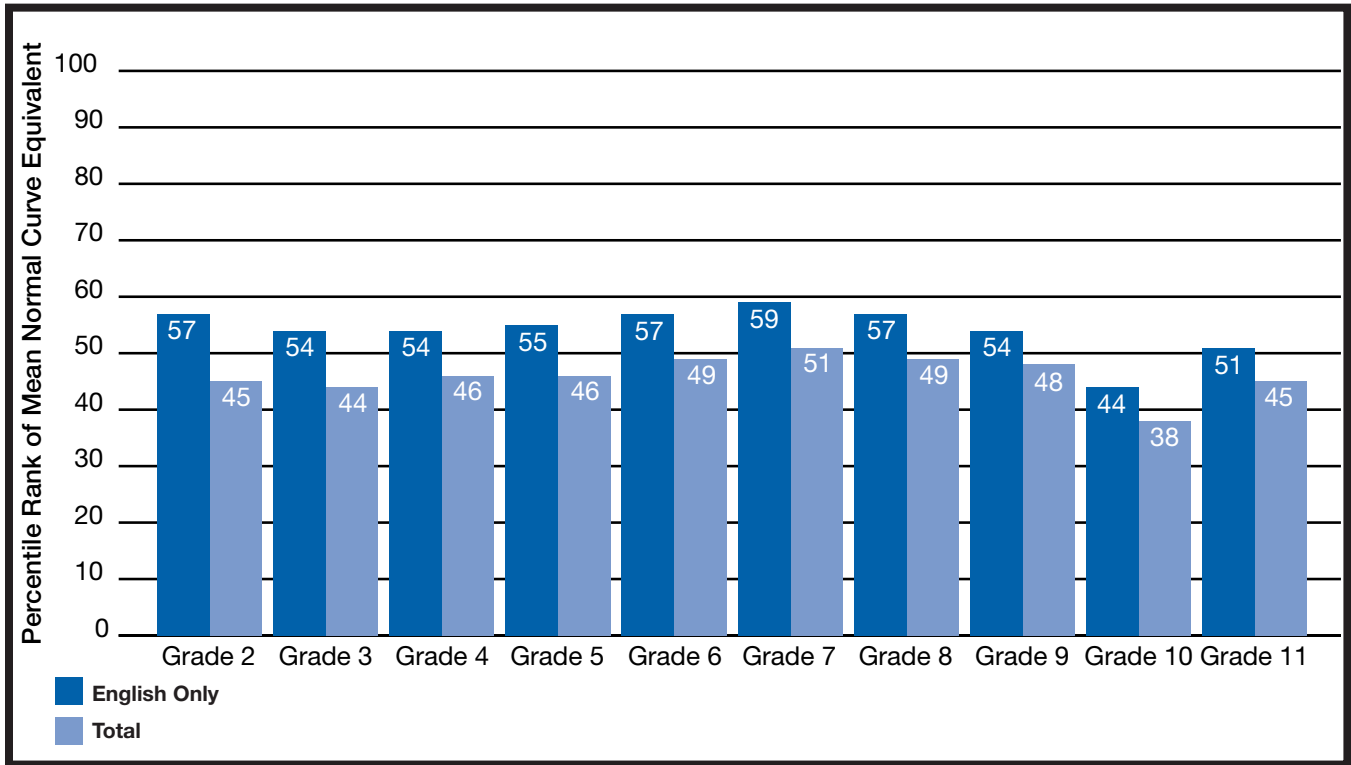


Figure 11. SAT-9 Language—All Students versus English Proficient Students

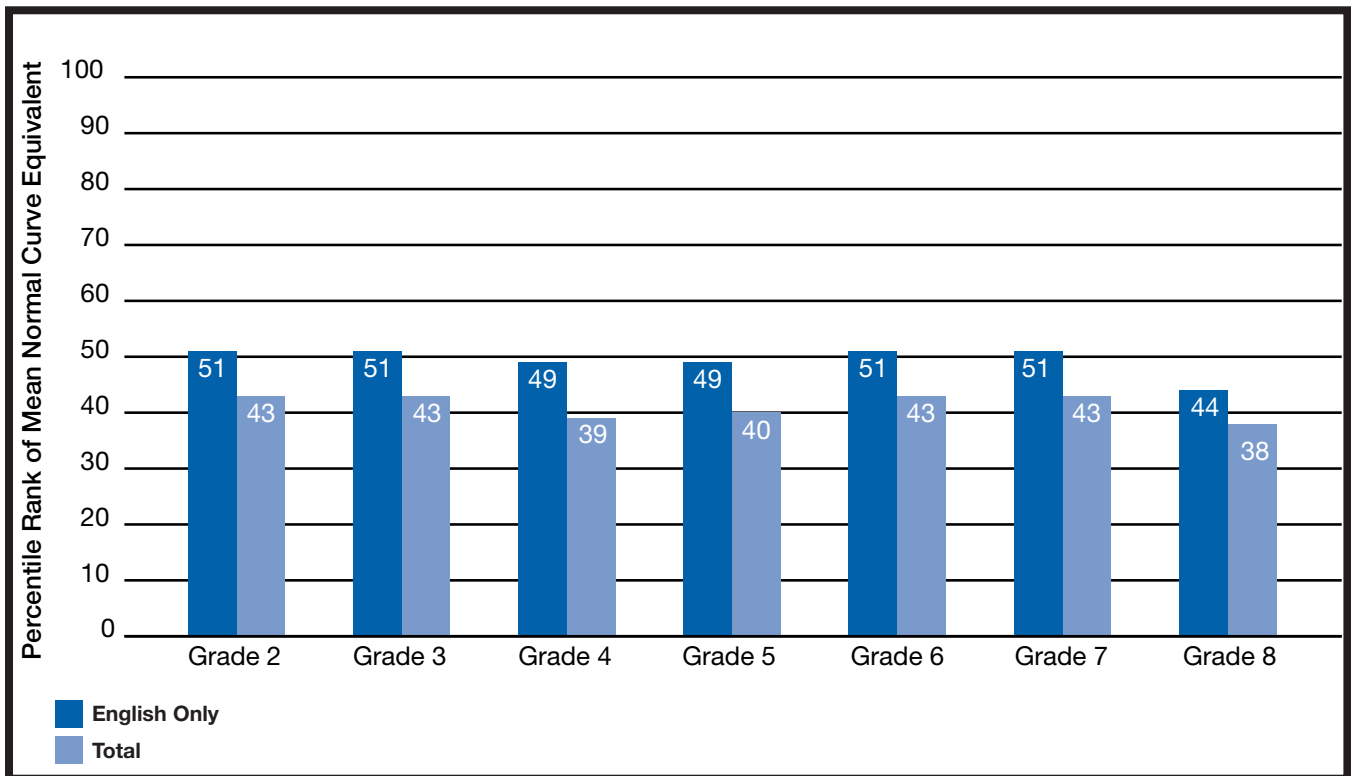
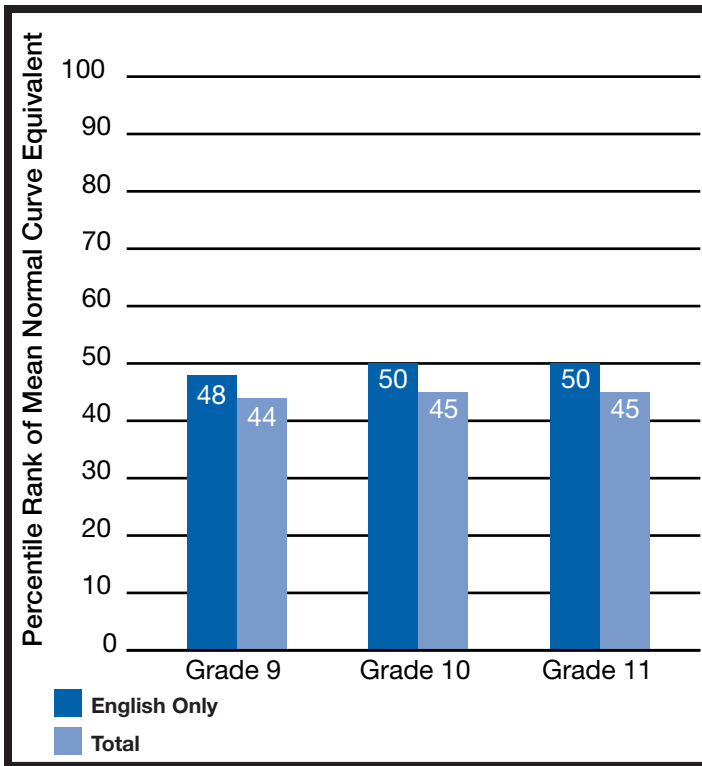
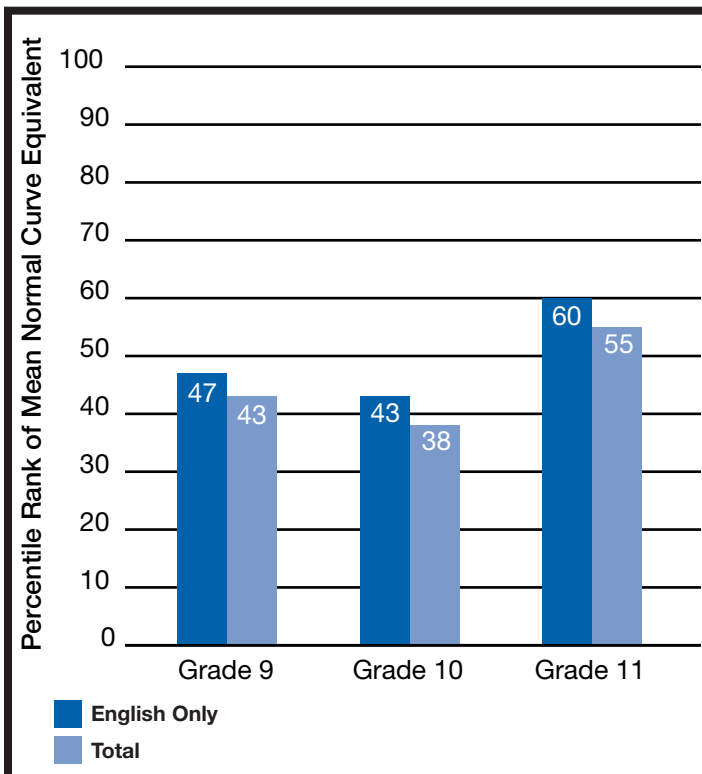


Figure 12. SAT-9 Spelling—All Students versus English Proficient Students



**Figure 13. SAT-9 Science—All Students versus English Proficient Students**



**Figure 14. SAT-9 Social Studies—All Students versus English Proficient Students**

The magnitude of this error varies, partly as a function of test reliability. One issue in interpreting these norm-referenced test scores as indicators of student or group achievement (or progress), thus, is how accurately the observed scores represent students’ true achievement.

Recent work by Stanford professor and CRESST researcher David Rogosa addresses this accuracy issue. In addition to technical reports that may be too complicated for the average citizen, Dr. Rogosa has created an easy-to-read guide titled, “How Accurate are the STAR National Percentile Rank Scores for Individual Students?—An Interpretive Guide.”<sup>15</sup> The results of this work will surprise many. Although most of the results are presented in the form of tables of data, the guide does provide a few samples in the form of responses to hypothetical questions. For example, the guide poses the question, “What are the chances that a ninth-grade math student whose actual capability or true score is at the 50th percentile of the norm group obtains a score more than five percentile points away from the 50th percentile?” The answer—70 percent! That is, there is only a 30 percent chance that the observed score is between the 45th and 55th percentile points.

With respect to interpreting progress, Rogosa’s guide also provides calculations for the probabilities of certain increases or decreases for students whose true percentile ranks remain constant from one year to the next. In one example, a ninth grade math student who is actually at the 60th percentile in both years has a greater than 50 percent chance of showing at least a ten percentile point change (up or down) in the second year! To state it differently, for this case it is more likely than not that student whose true score actually remains the same

from one year to the next will result in an observed score difference of more than ten percentile points. Given this level of imprecision in interpreting scores from one year to the next, it is advisable not to make too much of observed score differences, especially minor ones. While it is recognized that these analyses are based upon less precise student level scores and not state aggregates, it is nonetheless worthwhile to consider the issue of accuracy when utilizing standardized test scores to render important judgments.

Beyond the precision issues, there are also questions of the extent to which scores from one year to the next may be inflated by test preparation practices. That is, research suggests that under pressure to show improvement in test scores, teachers bring their curriculum more and more in line with just what's on the test and not the broader domain the test is intended to measure. They also are likely to spend substantial time on test preparation. Thus, the extent to which gains reflect real improvement in learning is an open question.<sup>16</sup>

Accuracy considerations aside, another issue to consider in comparing 1998 to 1999 observed scores is how progress is gauged. For assessing school-level progress, does it matter whether comparisons are made between mean-scaled scores or between the percentage of students scoring above a specified score point—two different ways of portraying “average” performance? And whose performance should be compared? What about comparing the performance of third graders in 1998 with the performance of third graders in 1999—commonly called cross sectional comparisons? Or should last year's third grade performance be compared with the performance of fourth graders

in 1999, an attempt to monitor the same group of students from one grade to the next? Does it make a difference?

A series of school-level analyses conducted by researchers at CRESST indicates there is rather low agreement between the rankings of schools using these two different methods of assessing change. Thus, it matters which method is used if schools are to be ranked as a result of their performance on those year-to-year comparisons. In other words, school rankings differed dramatically depending on whether average performance was compared from one year to the next based on grade level (e.g., the third grade in both 1998 and 1999) or on student cohort (e.g., second grade in 1998 and third grade in 1999). Quintile rankings across these two approaches agreed only about a third of the time (see Figures 15). This finding held across the different types of test scores (mean-scaled score, percentile rank of the mean normal curve equivalent, and percent scoring about the 50th percentile) and subject areas (reading, mathematics, language arts, and spelling), although only reading results are provided here.<sup>17</sup> Such inconsistency in rankings across methods advises thoughtful consideration of the method and what it purports to measure before placing much significance on the results.

### **How did students perform on the STAR augmentation?**

The rather poor showing by California students on the norm-referenced portion of the STAR system in 1998 has been attributed to many factors. As we've noted, one of the more widely discussed issues is the lack of alignment between the subject matter assessed by the

SAT-9 test and what was being taught in the public schools. In an effort to better align the assessment with what is outlined in the state content and performance standards, an augmented version of the SAT-9 was created for the subject areas of English and mathematics.<sup>18</sup> It is difficult to interpret student performance on the augmented test since the state has yet to identify what constitutes various performance levels. However the general consensus was that the tests—administered in the spring of 1999—sampled the more difficult elements of the state’s standards, and student performance was very low. In most grade levels, students on average correctly answered about half of the items on the English test (see Figure 16). Generally, the percentage of correct answers was lower on the math test at each grade level (see Figure 17), with better performance in the

lower grades.<sup>19</sup> Reports at open testimony at the October meeting of the California State Board of Education (1999) recounted anecdotes of students confronted with problems in mathematics far beyond their capability.

It is important to point out again that exactly what constitutes adequate or sufficient performance is undetermined at this time. Thus, not much should be made of student performance on the augmented tests until adequate performance standards are established and verified. Of more concern is the content sampling model used for the augmentation examinations, particularly since they are now termed the “standards-based” element of the STAR.<sup>20</sup> It will be important to follow the extent to which these particular tests are curriculum referenced and thus will reflect appropriate classroom instruction.

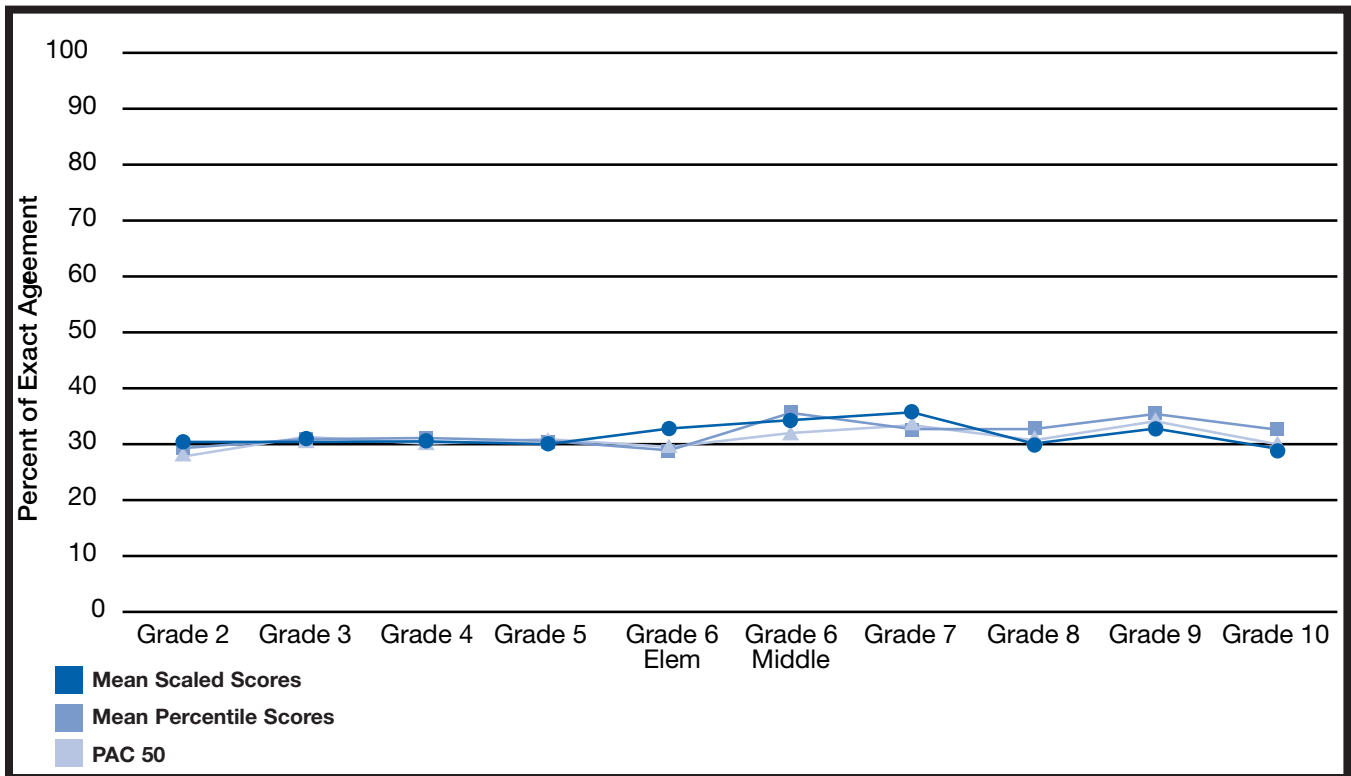


Figure 15. SAT-9 Reading Quintile Agreements

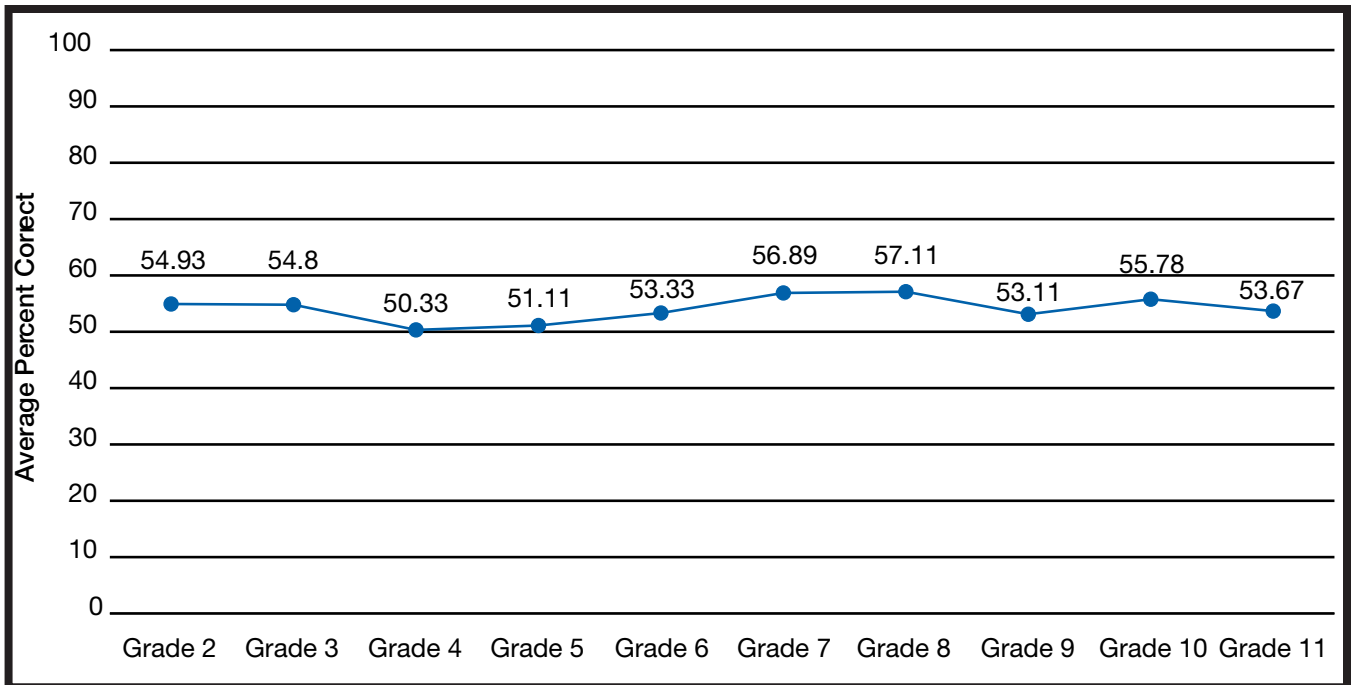


Figure 16. SAT-9 Augmented English

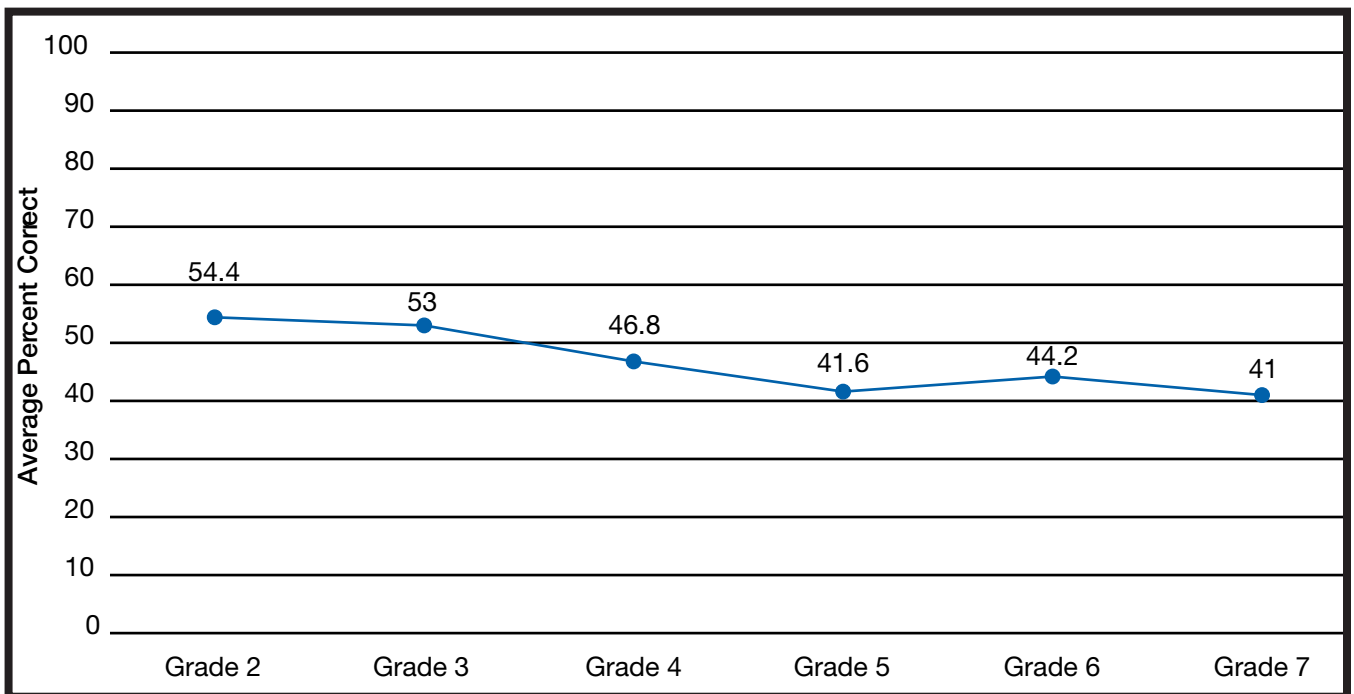


Figure 17. SAT-9 Augmented Math

**How is school composition related to a school's SAT-9 performance?**

On average, California students scored below the national average on the norm-referenced portion of the SAT-9. But clearly this finding does not imply that all students in the state are performing poorly. In fact, many schools and districts showed exceptionally high levels of average performance. Usually, these schools and districts are those that are challenged least with the forces of poverty and limited English proficiency. Simply stated, California school-children with limited English skills and those from economically disadvantaged backgrounds tend to score lower on the state's standardized test than students with English fluency or those from economically advantaged backgrounds.

This relationship is even greater where the concentration of disadvantaged students

increases. Schools with high proportions of students receiving free or reduced-price lunch score considerably lower than schools with lower proportions of such students.

Interestingly, the relationship holds for both economically disadvantaged and the more advantaged students at a particular school. That is, the average score for both groups of students tends to be lower in schools where there are high concentrations of poverty. Therefore, it appears the extent to which a school confronts the challenges of teaching impoverished children may affect not just the performance of poorer students, but of all students.

The same result was found for limited English proficient (LEP) students. The average performance of both LEP and non-LEP students is lower in schools with higher concentrations of LEP students. Thus, as in dealing

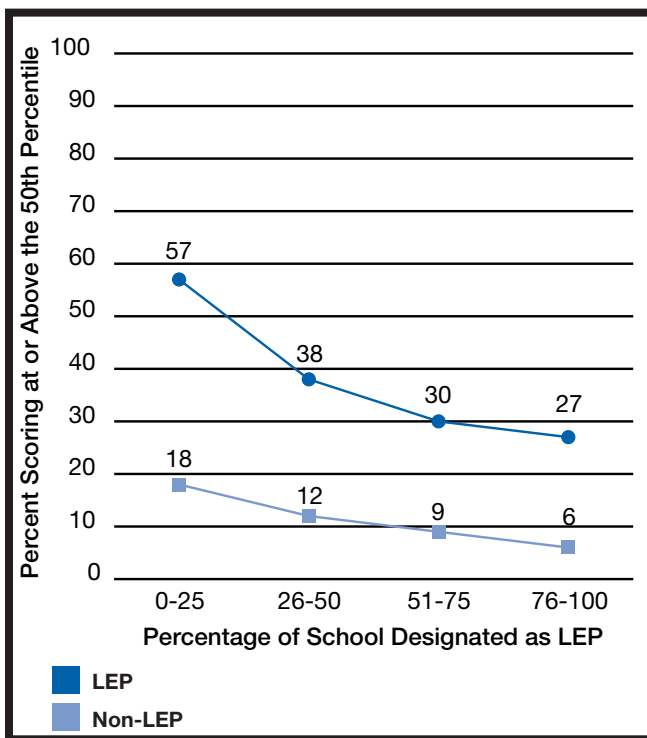


Figure 18. SAT-9 Grade 3 Reading LEP vs. Non-LEP

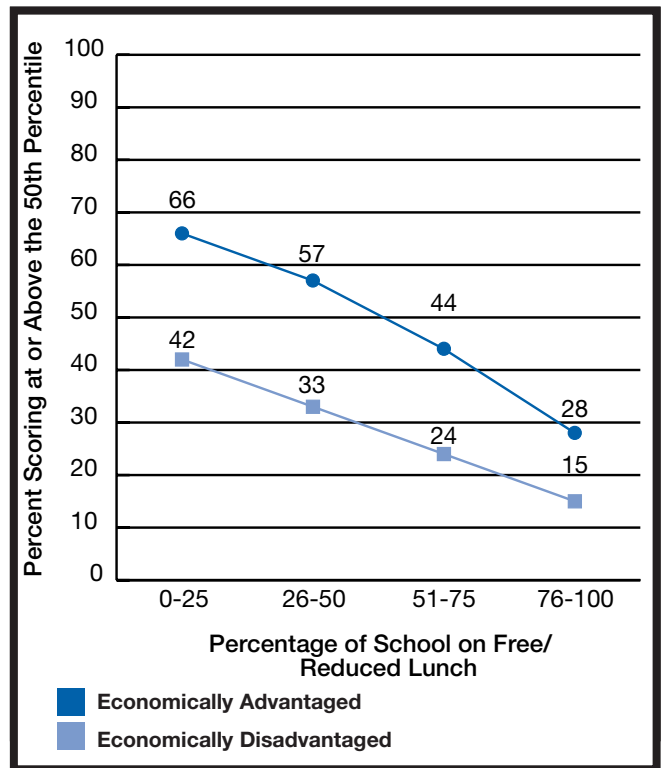


Figure 19. SAT-9 Grade 3 Reading Economically Disadvantaged vs. Economically Advantaged

with poverty, it appears the extent to which a school confronts the challenges of instructing children with limited English skills affects not only the performance of those students struggling to learn the language, but also the performance of students with sufficient English skills.

The observed relationship between language proficiency, poverty, and achievement on test scores is not surprising, and, as mentioned above, partly explains the relatively low overall average achievement of students in California. Since the SAT-9 norm group and the California student population differ dramatically on these key measures, lower average performance for California students as a whole relative to the normative group should be expected. Figure 18 graphically presents how the average performance varies for both those with and without sufficient language skills.<sup>21</sup> When there is a high proportion of LEP students at a local school site, all students perform at lower levels. Figure 19 illustrates similar findings for the problem of student poverty.<sup>22</sup>

The relationship between language proficiency, economic status, and test scores may not be as direct and clear, however, for students who are identified by our analyses as economically advantaged and/or fully English proficient because of the limits of the variables available to us. Clearly, those who are not eligible for free or reduced-price lunch (the advantaged or “non-disadvantaged” group in our analyses) represent a large range of socio-economic status (SES), from students whose families are just on the margin of qualification to those whose families reflect a very high level of SES. It may be the case that the relatively more advantaged students in schools that have high proportions of impoverished students are different from and

relatively less economically advantaged than those who are in schools with low proportions of children qualifying for free or reduced-price lunch. It may well be that these actual SES differences account for the differences in “non-economically disadvantaged” groups across the different types of schools.

Similar conclusions could be drawn for differences between the non-LEP population in schools serving a large proportion of LEP students compared to those that serve few or no LEP students. In the former case, a large proportion may be non-native English speakers who have relatively recently transitioned to English proficiency, but whose English language skills still are not totally secure; poverty may be another intervening variable. And it may be that it is these differences in the nature of the non-LEP group across the various types of schools that cause the observed performance differences. In any event, the relationship between school composition and the performance of different subgroups is vitally important and merits additional scrutiny.

### ***NAEP Results***

The National Assessment of Educational Progress (NAEP) is a federal effort at a nationwide assessment of educational achievement, conducted every few years nationally and including state-by-state comparisons in recent years for most states across the country. Generally, California students have performed poorly compared with the rest of the country. For instance, for the 1996 assessment of eighth-grade mathematics, California ranked 31st out of 41 states. The state did even worse in fourth-grade reading, coming in dead last out of 38 states. As indicated in Figures 20-23,

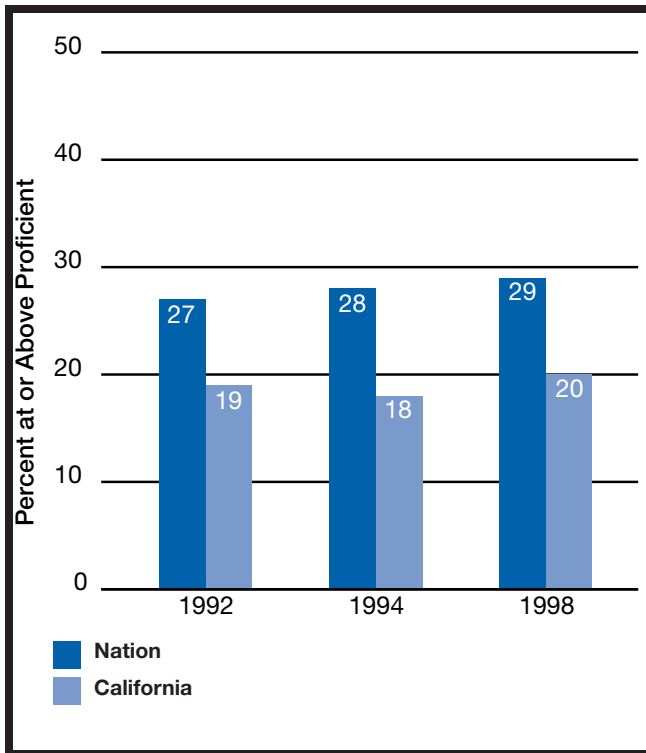


Figure 20. NAEP Grade 4 Reading 1992, 1994, and 1998

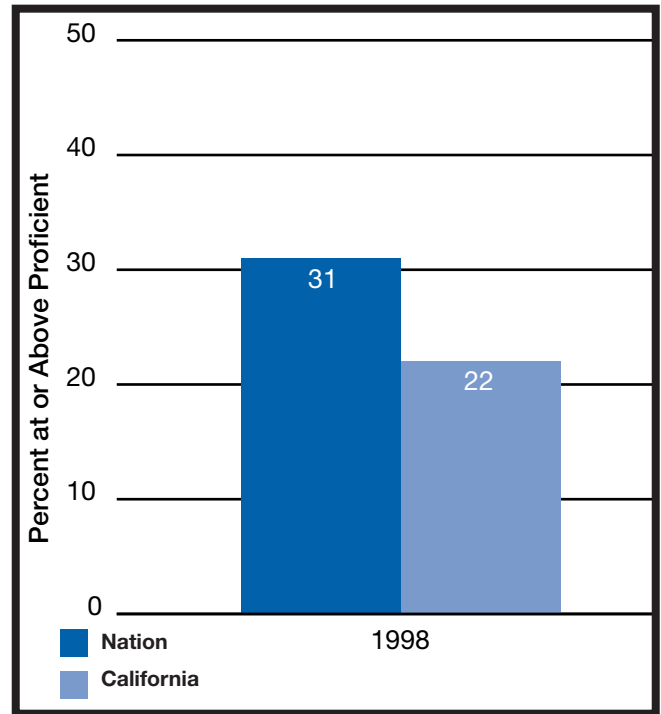


Figure 21. NAEP Grade 8 Reading 1998

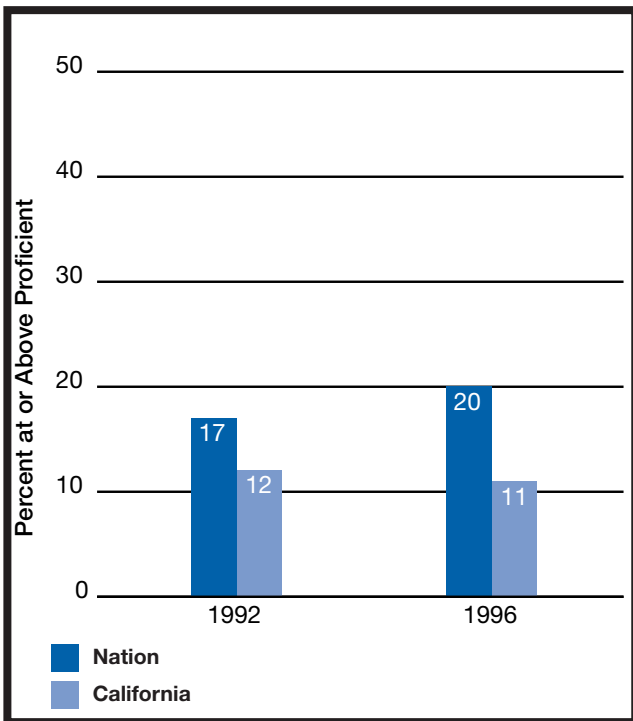


Figure 22. NAEP Grade 4 Math 1992 and 1996

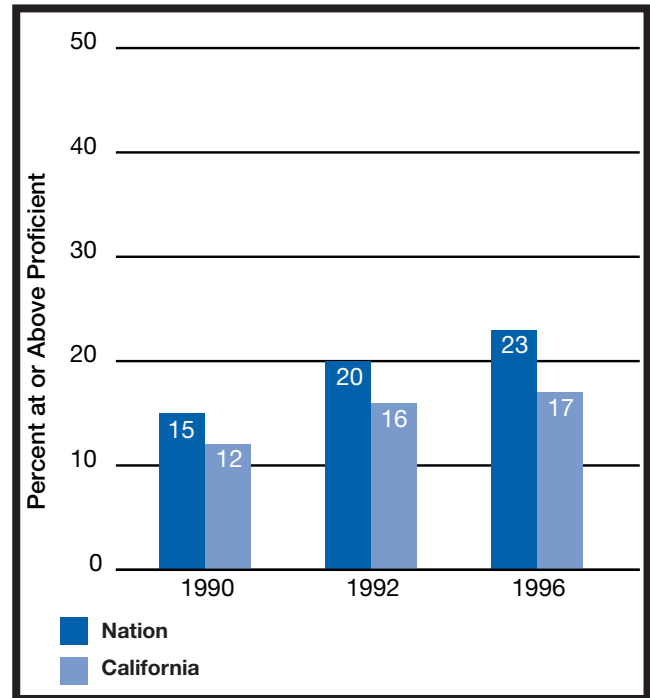
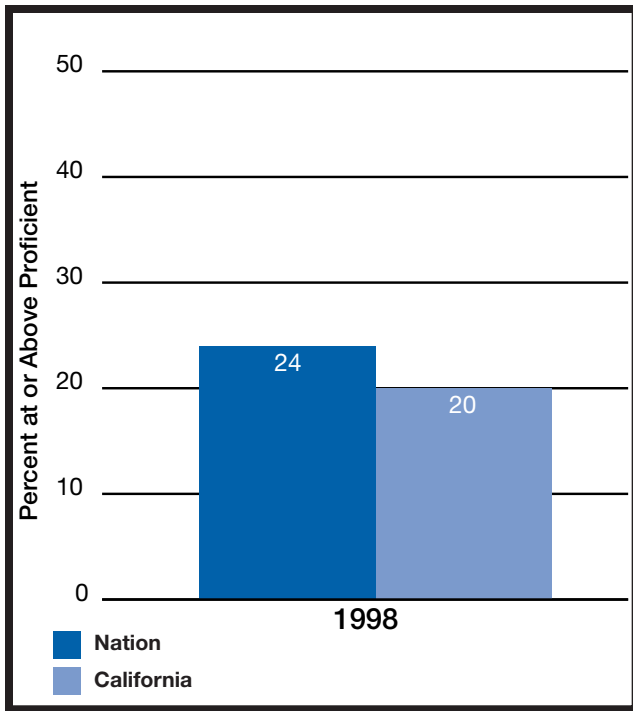


Figure 23. NAEP Grade 8 Math 1990, 1992, and 1996



**Figure 24. NAEP Grade 8 Writing 1998**

California lags the nation in grades 4 and 8 in both reading and mathematics achievement. Only 17 percent of California students performed at the proficient level in eighth-grade mathematics and 11 percent achieved that standard in fourth-grade mathematics—both of which are much lower than the national rates. Similarly, in eighth-grade writing, only one in five California students achieved at or above the proficient level, compared to one in four nationally (see Figure 24). Clearly, California students’ performance does not compare favorably to either the national sample or the standard of proficient performance.

Comparisons often provide a clear way to understand the meaning of performance. One way to understand California’s NAEP performance is to compare it to other states with similar characteristics. For example, with relation to poverty, 16.5 percent of California schools in the 1992 NAEP reading sample showed 75 per-

cent or more students on free or reduced-price lunch, and in the 1994 assessment, the figure was 16.6 percent of the California school sample.<sup>23</sup> In those two assessments, only 12.7 percent of those sampled in poverty-stricken schools scored at or above basic (the lowest level of achievement) in 1992, increasing to only 14.8 percent in 1994. Looking only at 1994 findings, ten states had higher percentages of schools in poverty than California. All of these states—Alabama, Arizona, Florida, Georgia, Louisiana, Mississippi, New York, New Mexico, South Carolina, and Texas—had higher proportions of disadvantaged students reaching the basic level than did California. In fact, some states with significantly higher proportions of schools in poverty, for example Georgia with 22.3 percent, Mississippi with 39 percent, and New Mexico with 26 percent, were substantially superior to California on this metric (Georgia and Mississippi with 29 percent scoring at basic and above, and New Mexico with 32 percent of students scoring at basic or above). Only one entity, the District of Columbia with about 62 percent of the schools meeting this poverty definition, scored below California, at 13.9 percent. Even so, the District of Columbia is doing a better job proportionally for its students when one looks at poverty and performance conjointly. These numbers show that the U.S. overall has a long way to go in educating its poor students, and California is clearly lagging behind the country.

In mathematics, the situation is comparable for the 1996 data. Twenty-one states have higher proportions of impoverished students than California and of these only the District of Columbia performed more poorly. For example, West Virginia with 29.7 percent poverty

had more than 60 percent of its students reaching or exceeding the basic level in mathematics.

Not all of the news from the NAEP assessment in California is bad. Other NAEP performance data<sup>24</sup> indicate the performance of low-income students in fourth-grade math is increasing. Education Watch reports a 7.8 percentage point increase in the number of these students scoring at or above the basic performance level from 1992 to 1996. In terms of cohort growth, furthermore, when one examines how fourth-graders performed on the 1992 mathematics assessment compared to the same cohort as eighth-graders in 1996, we find California in the top third of the states on this progress measure.<sup>25</sup> Clearly, California needs to continue to make progress and has a long way to go.

**Drop-Out/Graduation Rates<sup>26</sup>**

Despite the extensive focus placed on standardized test scores, other indicators of student performance have been collected and will be incorporated into the state’s accountability index at some undetermined future date. Two of them are the drop-out rates and completion rates for high school students. Definitions of dropouts often vary. California officially defines a dropout as a student at or above seventh grade who misses school for 45 consecutive days and does not enroll in another school. School completion rates tell us the proportion of high school seniors who graduate relative to those enrolled at the beginning of the year. Both of these indicators represent important ends in themselves, but also enable us to assure that improvements in test scores are not coming at the expense of more students being pushed out of school.

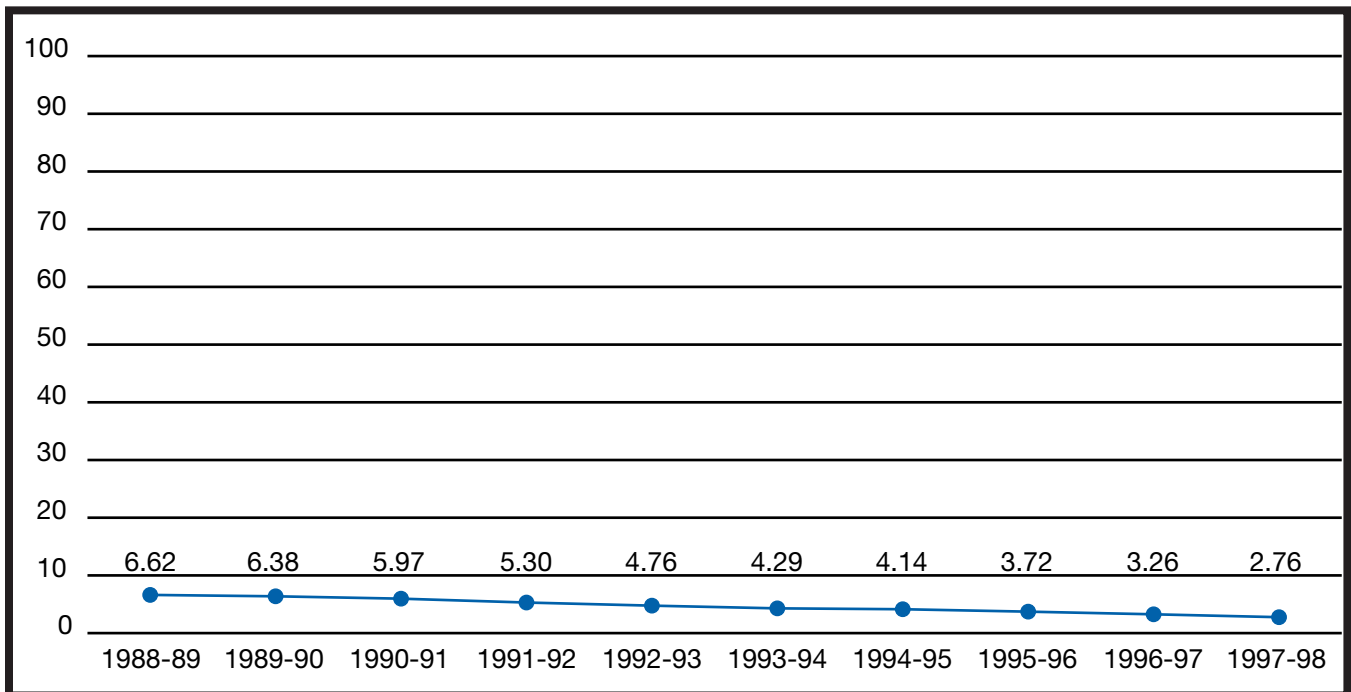
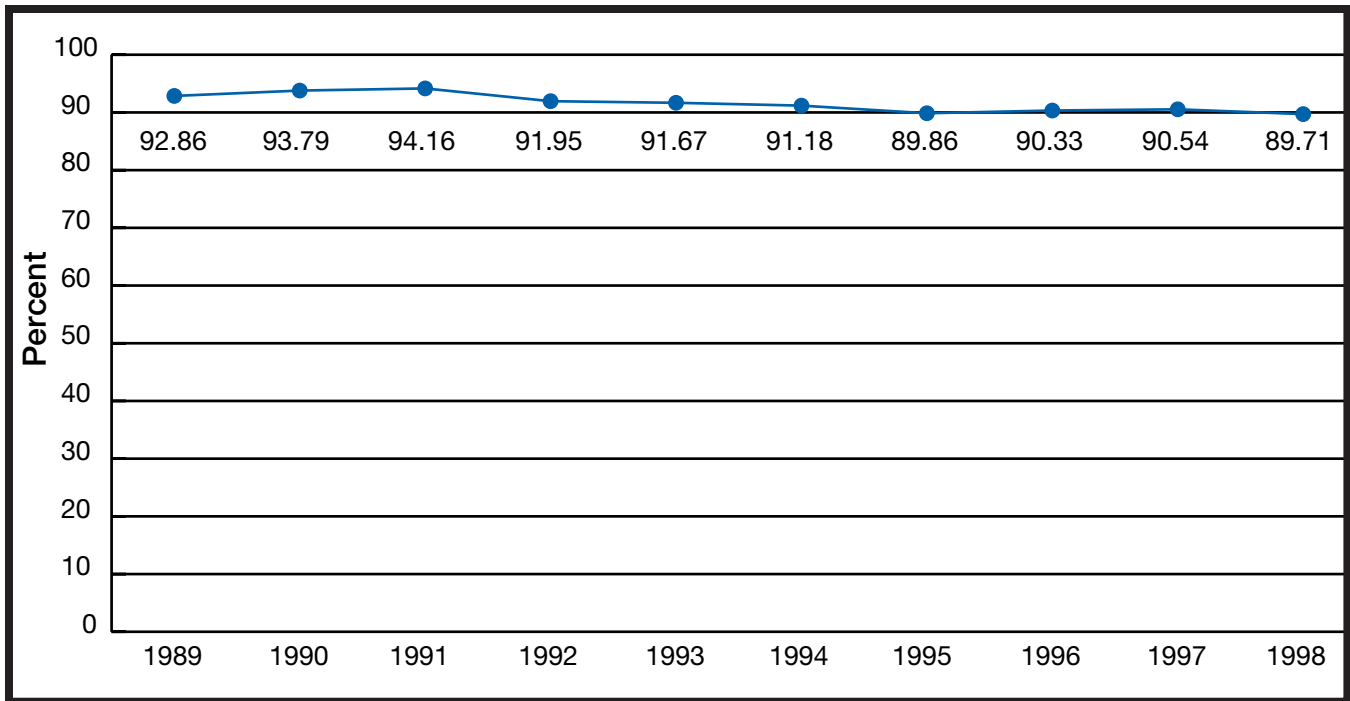


Figure 26. California High School Dropout Rates 1989-98



**Figure 26. California High School Graduation Rates 1989-1998**

Unfortunately, data regarding these two indicators are often unreliable or inaccurate because schools across the state do not use uniform definitions or share equally careful procedures for collecting the data. Poor data management may record students as dropouts when they have simply moved their home, or dropped out and then returned, after an extended hiatus. California is moving to a statewide student data system that will permit more precise understanding of these indicators. Nonetheless, in Figures 25 and 26, we use data from the California Department of Education to present ten-year trend lines of drop-out and graduation rates for California high school students. Drop-out rates are steadily declining and have done so in each year of the period. Graduation rates, on the other hand, have remained fairly stable, at around the 90 percent to 91 percent, though the rate was a few points higher at the beginning of the decade.

#### *High School Course-Taking Patterns* <sup>27</sup>

In California, high school students may choose to take a series of courses specifically defined to meet the University of California and California State University entrance requirements. These courses include the following:

- History/Social Science—two years required.
- English—four years required.
- Mathematics—three years required, four years recommended.
- Laboratory Science—two years required, three years recommended.
- Language Other than English—two years required, three years recommended.
- College Preparatory Electives—two years required. Two years (four semesters) in addition to those required in “A-E” above.

How many graduating students have actually completed this course series is, in some ways, a good indicator of how well the high schools in the state are preparing students for college in

the state's university system. It's also a marker for students' plans for college. Over the past ten years, the rate at which graduating seniors have met these course requirements has been consistently climbing. As shown in Figure 27, whereas fewer than 30 percent of graduates met the requirement in 1988, more than 38 percent did so in 1997.

Interpreting these changes depends upon how serious course titles match with actual course content. There is considerable evidence that actual topics covered and difficulty of content may vary in different courses with the same name. So at the least, increased college preparatory course-taking reflects better motivation if not always an increase in student performance.

*Advanced Placement Examinations*<sup>28</sup>

Another secondary school measure of interest is the availability of and participation in advanced

placement courses and examinations. Advanced Placement courses reflect college-level course work, and students passing advanced placement exams receive college credit. Thus, the percentage of students taking AP courses and passing the exams is an indicator of the extent to which students are being prepared for, pursuing, and are being successful in rigorous academic coursework. Because of the rigor of the courses, students receive extra points for their grades in these courses (5 for an A, 4 for a B), which in turn advantages their grade-point averages for college admissions. Recently civil litigation was brought against at least one California school district for allegedly providing disproportionate AP opportunities to students of varying ethnic backgrounds.<sup>29</sup> Though we do not present data on the availability of AP courses here, we do have some data on the frequency with which various ethnic groups take AP examinations, and contrast those numbers with the percent-

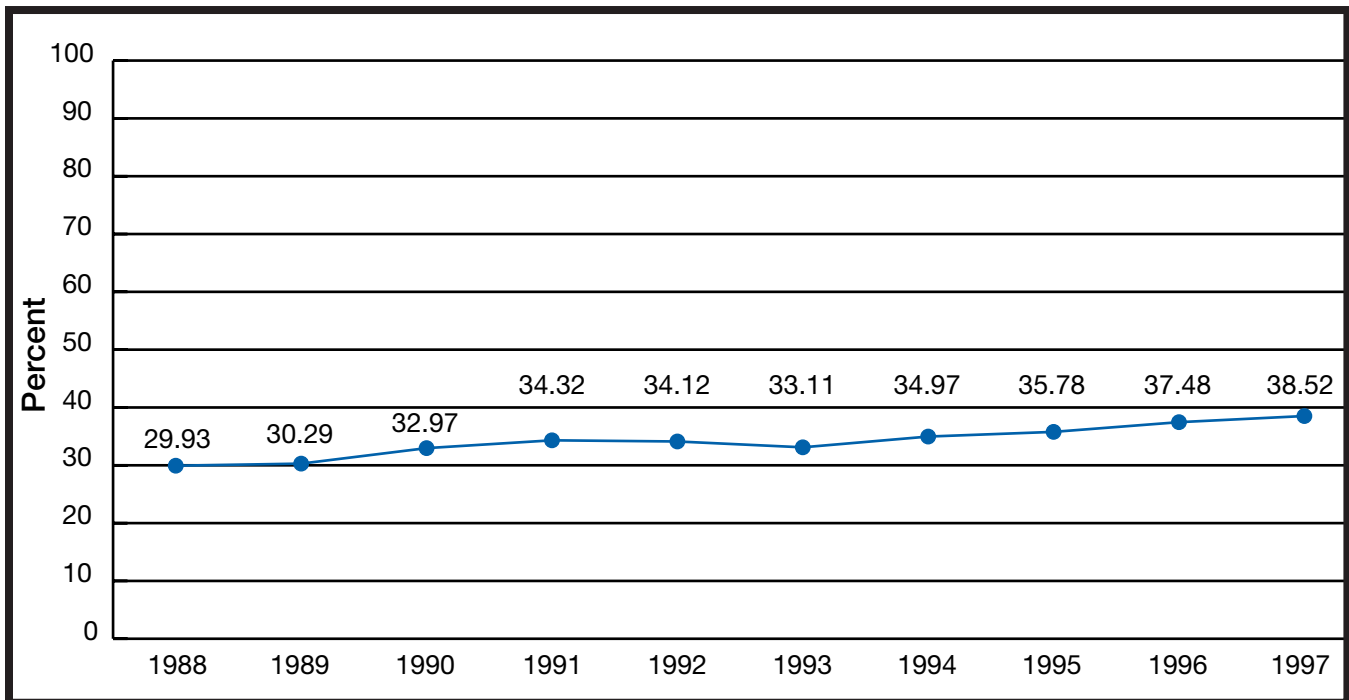


Figure 27. High School Graduates Meeting UC/CSU Course Requirements 1988-1997

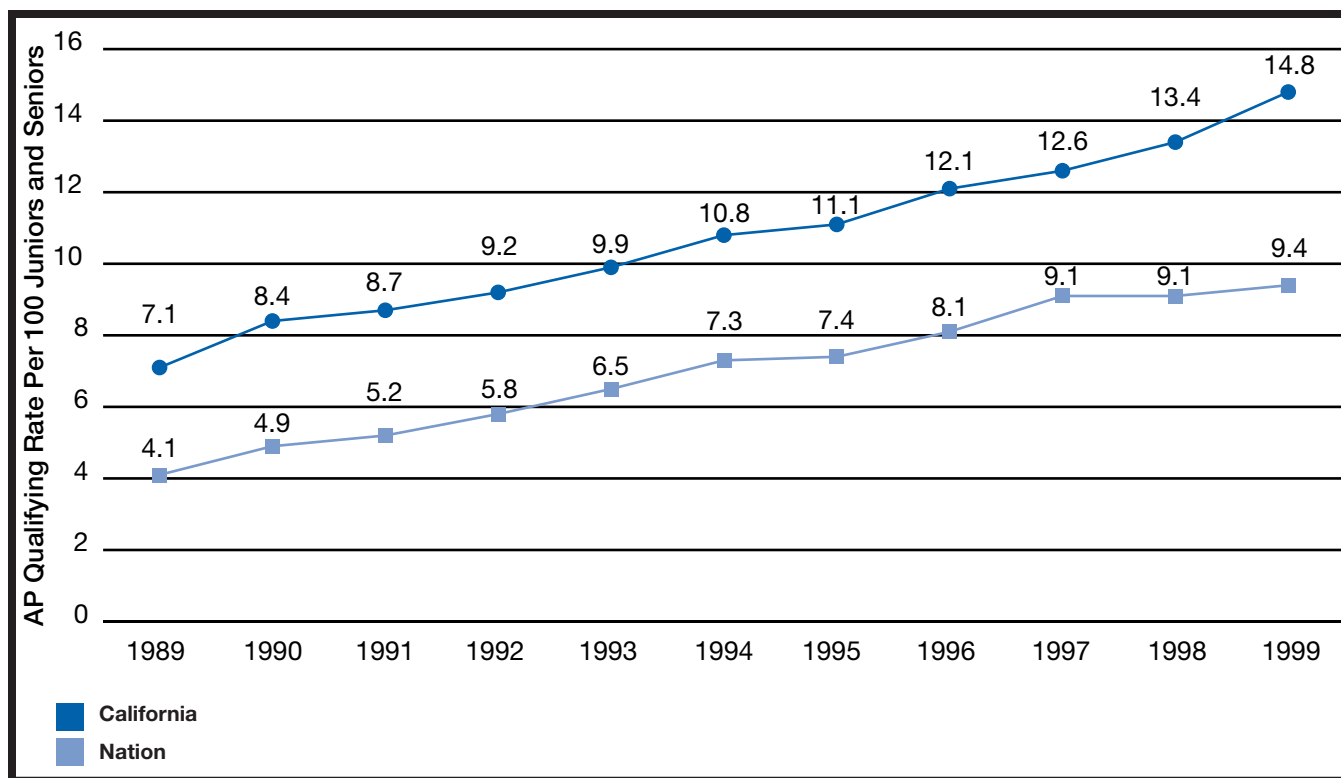
ages of each group in the student population. For instance, African-American students comprise 8.8 percent of the public school population, 3.5 percent of students taking the English Advanced Placement examination, and 2.5 percent of students taking the calculus test. In contrast, students of Asian ethnicity comprise 11.2 percent of the student population, but account for 28.1 percent of the English AP test takers and a whopping 42.8 percent of those sitting for the AP calculus examination.<sup>xxx</sup> Clearly, the ethnic makeup of students taking Advanced Placement examinations is not representative of the California student population as a whole.

One positive finding regarding the advanced placement data is the increased frequency with which California students are meeting the AP qualification standards. Since the 1991-'92 academic year, this rate has

steadily improved every year, going from 9.2 percent at the beginning of the decade to 14.8 percent last year (see Figure 28).

### *College Entrance Examinations*<sup>31</sup>

College entrance examination scores are another measure of who California high schools are preparing for college. As the figures below indicate, the performance of California's college-bound student population on the Scholastic Achievement Test (SAT) has been fairly stable over the last ten years. For both the math and verbal components of the test, statewide average scores dipped in the early part of the decade, but have steadily climbed back near the levels attained at the end of the last decade. For math, the achievement levels of the late 1980s have actually been surpassed in the last two years.



**Figure 28. Advanced Placement Qualifying Rate 1989-1999**

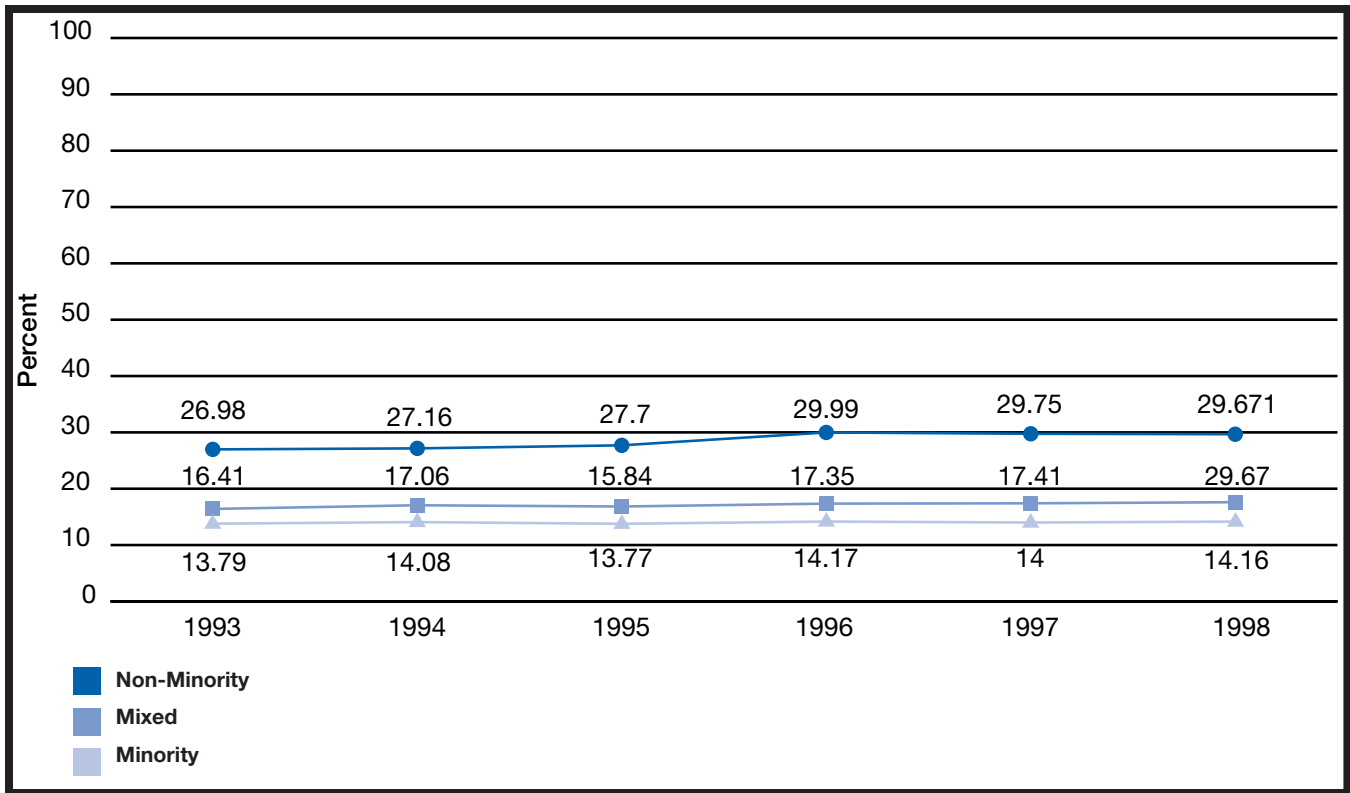


Figure 29. Students Meeting SAT Criteria in California 1993-1998

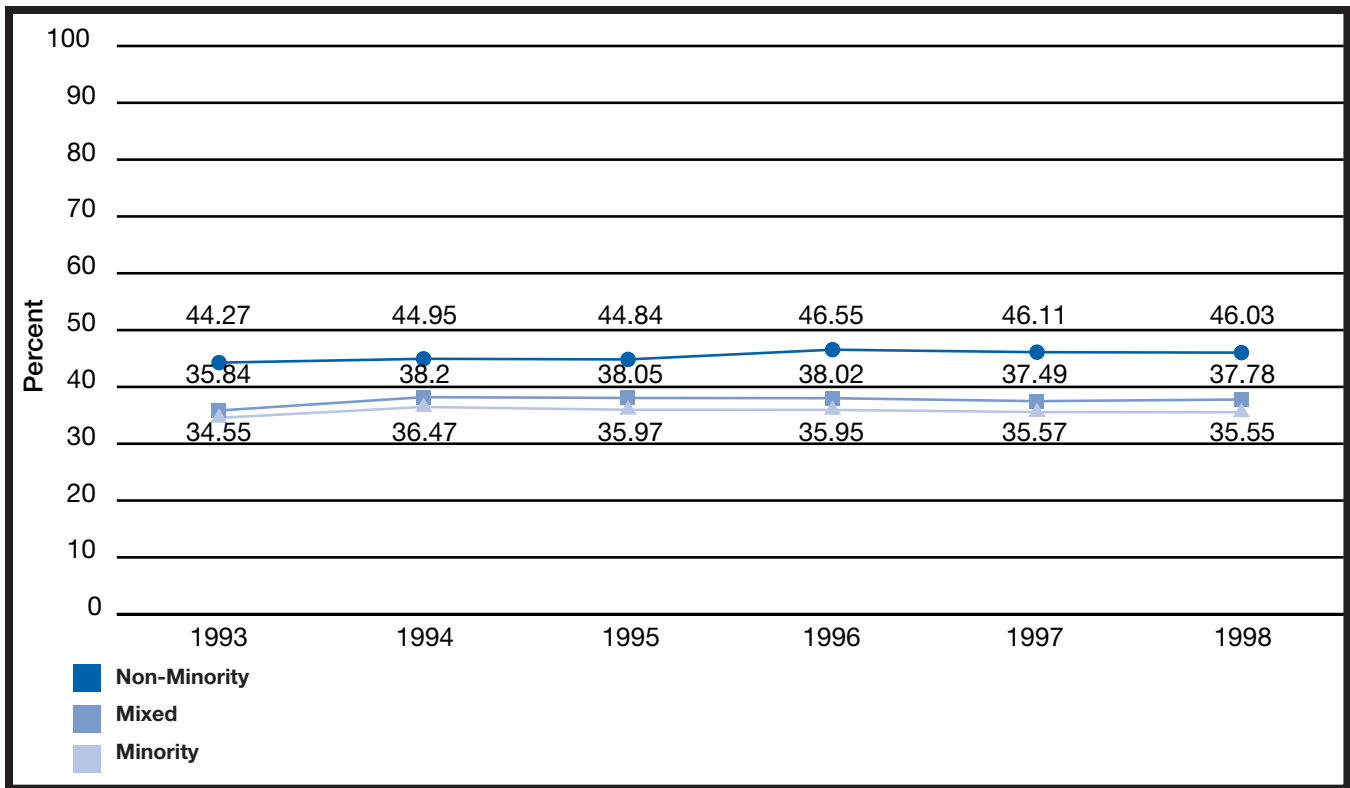


Figure 30. SAT Test Takers in California 1993-1998

As in the average scores for the math and verbal component of the SAT, there appears to be a rebounding trend in the percentage of test takers meeting or exceeding the combined 1000-point threshold. This measure, too, experienced a slight dip in the early 1990s but has inched up to remain between 18 and 19 percent over the last three years, levels comparable to the latter part of the 1980s. Of course, these rates vary by high school, with schools comprised of high minority populations achieving rates at roughly half those of low minority schools (see Figure 29). And as Figure 29 suggests, the differences in these rates do not appear to be decreasing over time.

Similarly, the rate at which high school seniors are taking the SAT has changed little over the last six years. As shown in Figure 30, the percentage of twelfth-graders taking the SAT in low minority schools has remained stable at around 46 percent, while that rate has hovered around 36 percent for high minority schools over the same time period.

### *College Attendance*

Data from *Education Watch* 1998 indicate that 66.4 percent of high school graduates in 1996 went on to enroll in college (full or part-time) by the time they were 19 years old. This rate ranked California fifth out of 50 states in providing students access and opportunity for college. However, college completion rates for minority students entering as freshmen—deemed the equity rate—is not so rosy. The equity rate for California is 58.4 percent, which is below the national average of 65 percent. California's four-year graduation rate is 41 percent, meaning that less than half of entering freshmen graduate within a four-year period.

### *College Remediation Rates*<sup>32</sup>

Part of the reason for the lower completion rates may be the fact that California has a high number of part-time community college students. Another reason may be the fact that many students enroll in college with severe limitations in their basic reading and mathematics skills. In the California State University system, more than 54 percent of incoming first-year students are required to take remedial math and more than 47 percent need remedial reading classes. In the state's elite University of California system, more than a third of the students fail to meet the minimal standards of writing proficiency.<sup>xxxiii</sup> This number has improved in the last year, from 38.9 percent in 1997 down to 33.3 percent in 1998. However, this indicator still suggests that although more high school graduates are completing the required sequence of high school courses, a great many are not at the basic levels of reading and mathematics ability that successful transition into a university education requires.

### *Summary of Achievement*

California student achievement is low compared to the rest of the nation. This is true based not only on SAT-9 scores but also on the NAEP. Average student performance in some schools is better than in others, and it is fairly easy to identify which schools these are by who is going to them. Although students now take an additional test designed to address their mastery of state-determined subject-matter standards, it is not ready for widespread implementation. Essentially across-the-board minimal gains on observed scores from the SAT-9 in 1999 compared with 1998 probably signify familiarity with the process more than "real" improvement.

Moreover, different measures of improvement greatly disagree with one another.

At the secondary level, we have seen improvements in reducing drop-out rates and maintaining graduation rates. Graduating students are taking more nominally challenging course loads, and greater numbers of them are meeting advanced placement requirements, although the rates of advanced placement test-taking vary markedly by ethnicity. Students are scoring higher on college entrance examinations, but the percentage of seniors taking the examinations is holding steady. On the positive side, California does a good job of providing college opportunities to high school graduates. Unfortunately, these students often are not prepared for the fundamental academic requirements for success in higher education.

On a more troubling note, the relationship between the socio-demographic complex of poverty, language skills and ethnicity and stan-

dardized student achievement measures is immense and getting stronger.<sup>34</sup> Over the past six years, this relationship has strengthened, not diminished (see Figure 31). These background measures relate to average school performance on the SAT at an extremely high level, accounting for greater than two-thirds of the variation in scores among schools. Similar evidence is found for the SAT-9 test, particularly at the lower grades, where background measures account for 60 to 80 percent of the variance in average school scores in reading, spelling, and language arts (see Figure 32).<sup>35</sup> A somewhat weaker relationship is found between background measures and mathematics, although a majority of the variance is still accounted for at each grade level, from a low of 56 percent in grade 2 to a high of 67 percent in grade 4.

The relationship is clear. More poverty relates to lower average scores. More limited

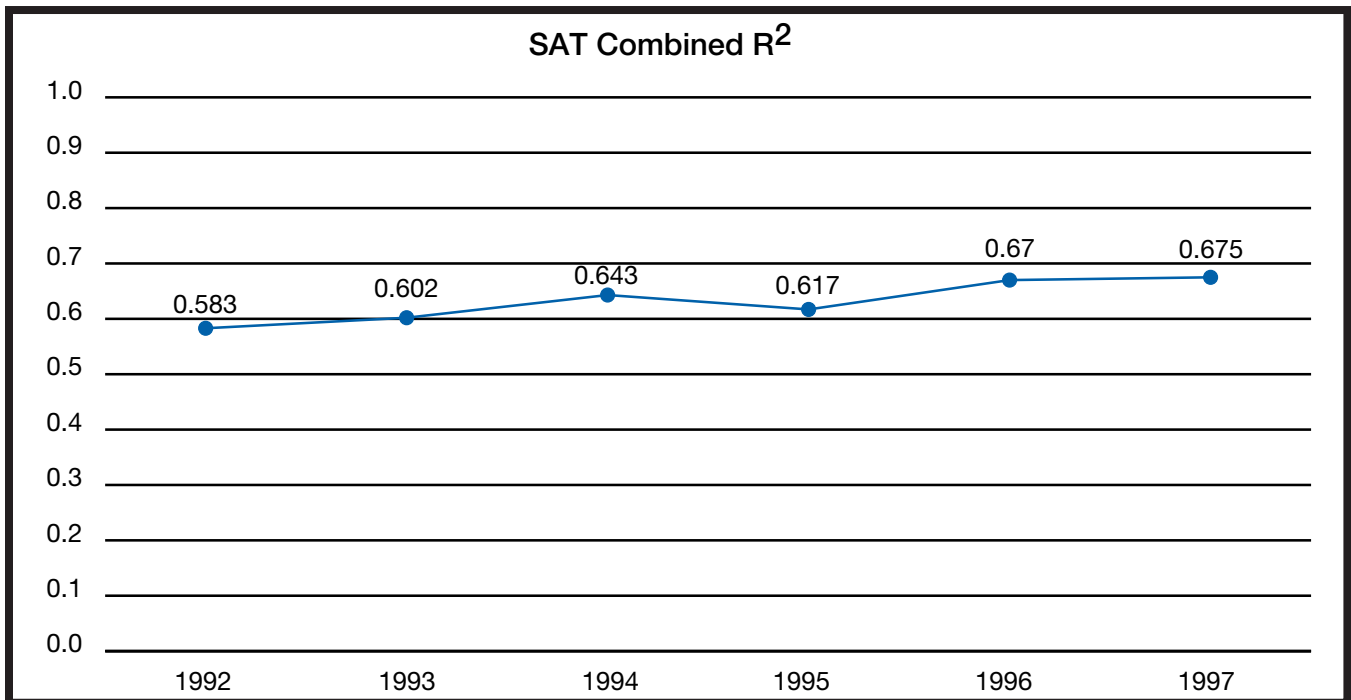
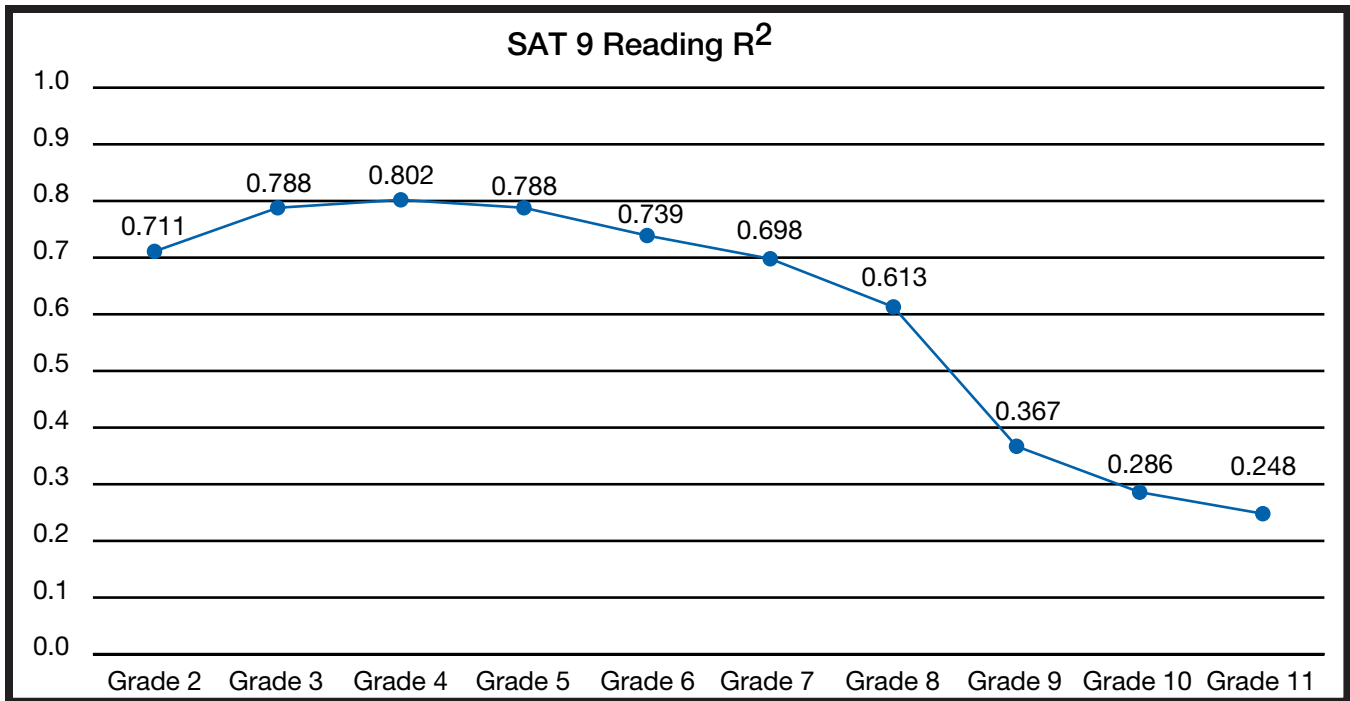


Figure 31. Relationship Between Socio-Economic Measures and SAT Scores 1992-1998



**Figure 32. Relationship Between Socio-Economic Measures and SAT-9 Reading Scores**

English skills relate to lower average scores. Greater minority representation in the student population, sadly, also relates to lower average scores. Poverty is increasing as the percentage of students qualifying for free or reduced-price lunch has risen from 32.19 percent in 1989 to 47.61 percent in 1999. Similarly, the percentage of California students with limited English proficiency has jumped from 16.29 percent in 1989 to 24.89 percent in 1999. Both of these increases represent about a 50 percent jump over the past decade (see Figures 33-34). Over the same period, the minority population has risen only slightly, but consistently (see Figure 35). And poor performance doesn't just affect those students who lack language skills or sufficient monetary resources. Students fully proficient in English and those not eligible for free or reduced-price lunch in schools with high con-

centrations of LEP and economically disadvantaged students perform more poorly than their counterparts in schools with lower numbers of these disadvantaged students.

The goal of California schools is to prepare all students to reach high academic standards. To do so, the educational system should seek to reduce the impact socio-demographic measures have on student achievement. Student achievement should relate more to what students learn in the classroom than to their background. Unfortunately, we currently are not seeing the desired effect.

Why? We have to consider the sensitivity of our measures to instructional change, to the capacity of the schools and school districts, to the motivation of students and parents, and to the period of time (fewer than two years) that California standards have been in place.

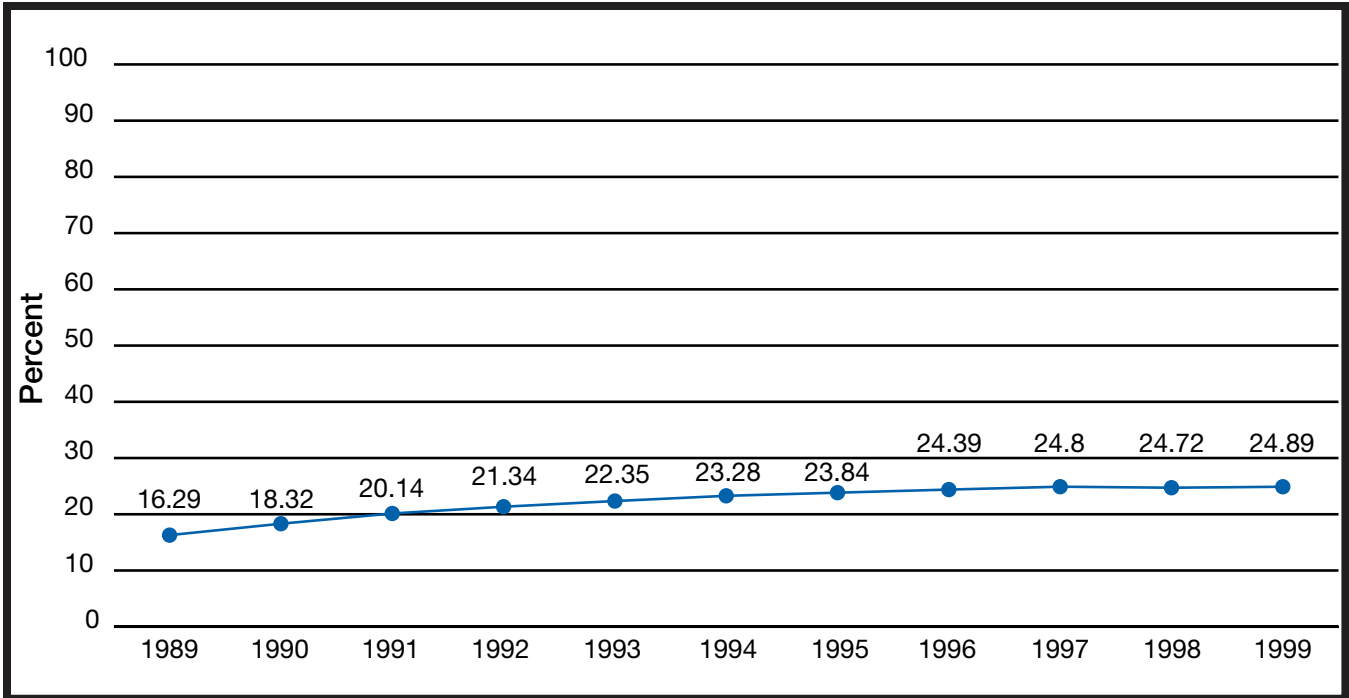


Figure 33. Limited English Proficient Students in California 1989-1999

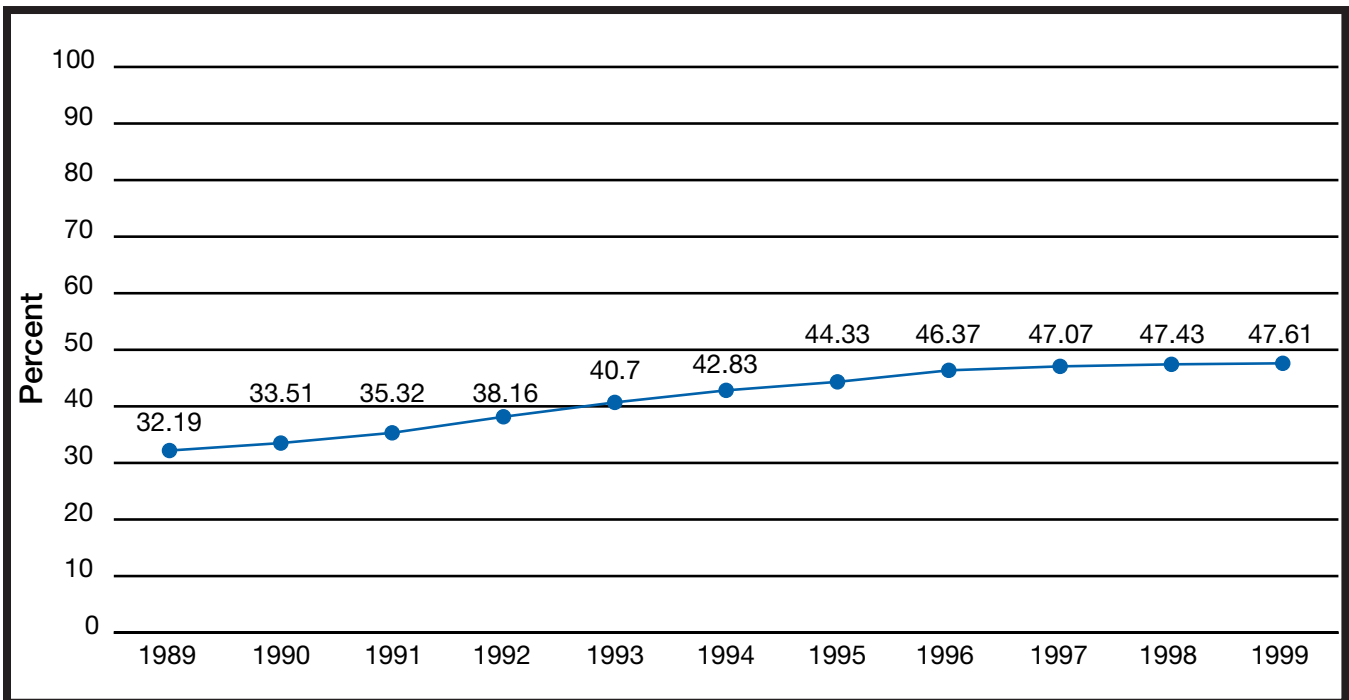


Figure 34. Students Receiving Free or Reduced Lunch in California 1989-99

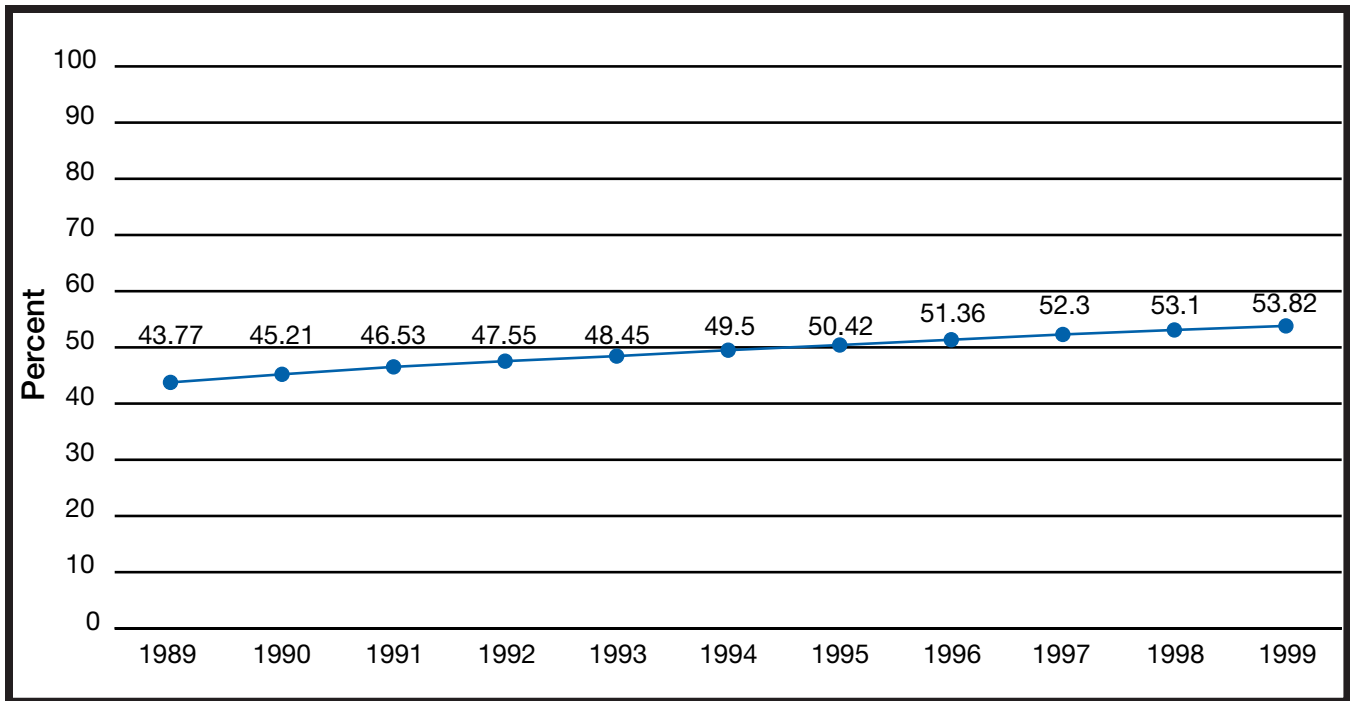


Figure 35. Minority Students in California 1988-1999

## The Future: Assessment and Accountability

With the adoption of the Academic Performance Index (API) in November 1999, California has moved into a new level of educational accountability. It has adopted a general plan to use assessment and other key school data, e.g., student absences and graduation rates, as part of a system to hold schools accountable. The plan is supposed to support *standards-based reform*. Over a six-month period, a committee of school policy-makers, academic experts, and practitioners met and prepared the requirements of the API. The details are available on the Department of Education Web page ([www.cde.ca.gov](http://www.cde.ca.gov)) and will eventually cover how growth targets are set (based on the distribution of performance of students at the school), how comparisons are made, the expectations for identifiable subgroups, and sanctions

and rewards. What is of most relevance here is the degree to which the API relies on assessments, and related to that, the degree to which the assessments represent and propel progress on the state's standards for student performance. The original plan for the API involved phasing in various assessments as they became available to bring the assessment into closer alignment with the standards. However, for the 1999-2000 year, only performance on the SAT-9 component of STAR enters into the accountability index.

Prior to adopting the API details, the California State Board of Education adopted a framework that enunciated principles to guide the use of the accountability system. The criteria comprising this framework are reproduced in Figure 36 below.

The relevance of these principles to concerns we have raised earlier about assessment and criteria for quality assessment systems is

### **Academic Performance Index Framework**

- The API Must Be Technically Sound
- The API Must Emphasize Student Performance, Not Educational Processes
- The API Must Strive to the Greatest Extent to Measure Content, Skills, and Competencies that Can Be Taught and Learned in School and that Reflect the State Standards (our emphasis).
- The API Must Allow for Fair Comparisons
- The API Should Include as Many Students as Possible In Each School and District
- The API Must Measure School Performance and Growth as Accurately as Possible
- The API Should Strive in the Long-Term to Measure Growth Based on Student-Level Longitudinal Data
- The API Should Be Flexible and Its Component Indicators Should Be Stable
- The API Should be Understandable, Particularly to Educators and Parents
- The API Is Part of an Overall Accountability System That Must Include Comprehensive Information Which Incorporates Contextual and Background Indicators Beyond Those Required by Law
- The API Should Minimize Burden
- The API Should Support Local Accountability Systems
- The API Must Conform to the Requirements and Intent of the Public Schools Accountability Act of 1999 as Well as Related Legislation

**Figure 36: Academic Performance Index Framework<sup>36</sup>**

(Adopted by the California State Department of Education at their July, 1999 meeting)

### **Relevant Standards for California from the Standards for Educational and Psychological Testing**

- State Purpose(s) and Minimize Negative Consequences of the test
- Give Evidence of Technical Quality of the test for Each Purpose
- Document Relationship to Content Standards
- High Stakes (Promotion) Requires Match Between Instruction and Test Content
- Give Evidence of Suitability of Test for Program and for Test Population
- When Use of a Test or System Implies a Specific Outcome, Provide Basis and Evidence for Expectation
- Minimize Possible Misinterpretation of Data with Appropriate Context
- No Student Decision Should Be Made on the Basis of One Test
- Test Preparation Should Not Adversely Impact Validity of Results
- Reports Should Include Classification Error and Error in Measurement of Change
- Public Interpretation Should Be Handled by Trained Personnel

**Figure 37: Relevant Standards for California from Standards for Educational and Psychological Testing**

(Adopted by the California State Department of Education at their July, 1999 meeting)

clear. In addition, the evolution of assessment for accountability in California calls for careful analysis. In general, California is starting with a measure—the SAT-9—that has only limited relationship to the state’s standards. While there are plans to add more elements down the line, the current accountability provisions may work to encourage a near exclusive focus on the SAT-9, since it was the first and most salient measure in use.

In adhering to the principles articulated by the state board, which in a preamble explicitly commit to continued studies of the validity of the state’s assessment system, it may be relevant to reference yet another set of guidelines for the design and use of assessments. From the recently published Standards for Educational and Psychological Testing<sup>37</sup>, the following paraphrased standards are applicable to California planning and evaluation.

As California moves forward with its assessment and accountability system, it will be important that it do so in line with its own principles and those of the testing profession.

## Conclusions

Starting with the available data, the story about California is mixed. When examining the overall performance on the SAT-9, we find that the state average, over all grades and all subject matters, is below the national average.

However, when we account for the state policy requiring that all students who have been in school for one year take the test—whatever their English proficiency—we find California students positioned around the national average. In fact, given the difference in the compo-

sition of the tested population and the norming groups, this result is somewhat better than we might expect.

However, when we move to standards-based measures, of which NAEP is a general example, California performance looks poor indeed. California especially falters when one addresses the performance of children in poverty. Also, it is important to recall that on the NAEP, only students who can comprehend the examination are tested. What will be important to watch in the future is whether California students, like those in many other states, at the outset have lower performance on new standards-based tests. We would expect lower performance if the tests are measuring and students are in fact attempting to meet more challenging goals. We would also expect to see test performance to rise over time as instruction becomes more relevant to the standards the assessments are measuring.

California has a number of important tasks to consider. We believe that there is direct action that can be taken to support the best possible development of the assessment system, of the accountability structure it supports, and of California education. First and foremost, it is desirable to focus on the appropriateness and validity of the assessments planned to be in the system, as they are under development. In simple terms, any test is usually not exchangeable for any other. For example, as we have seen, the SAT-9 is a general achievement test and not fully aligned with the state’s content and performance standards. It cannot simply be exchanged for a rigorous standards-based assessment system. Similarly, a high school graduation test presumably must make distinctions between those who are qualified and those who are not qualified, relative to explicit

standards, for a high school diploma, implying assessment items primarily focused on making that distinction. A college admission test, on the other hand, must make distinctions at a higher ability range, thus implying a different item focus and test-taker differentiation. A single test of limited duration probably cannot well serve both these purposes.

This is not to say, however, that these various measures themselves should not be consistent with the state's standards, albeit representing levels of performance and sophistication. Nor is it the case that a single test strategy for decision making is a good one or that all students necessarily should have to pass the same test. For example, a number of people have advocated using course-based exams for California's High School Exit Exam, as a direct way to align curriculum and testing and better assure that students have the opportunity to learn what is expected. Two examples of such course-based exams already exist – the Advanced Placement Exams, which were discussed earlier, and the Golden State Exams, a series of state-developed, academically rigorous, voluntary exams which are linked to specific high school courses. Both of these assessments probably represent a higher level of proficiency than can be expected from all high school students in the short run, but one might imagine a system where passing one or the other of these tests would count for the HSEE requirement, while still requiring students who were enrolled in other course to take the actual HSEE.

Assessments can be designed to serve various policy purposes, but there are times, such as we are seeing in other states, where policy imperatives have swamped technical capacity to deliver the assessments. Time frames have been

insufficient to assure a quality assessment or to prepare the educational system and its students for a new set of expectations. The result is usually some form of retrenchment. In California, we would hope to avoid this cycle.

## Recommendations

These recommendations will be brief and illustrative rather than exhaustive.

- Validity studies examining the extent to which California's assessment system is achieving intended purposes (school accountability, instructional improvement, consequences) must be undertaken immediately. These studies must address the impact of the assessment on various subgroups of students and schools.
- Evidence that the assessments detect instructional effects is needed.
- Efforts should be made to describe which standards are not measured by statewide programs (and are, therefore, appropriate for local scrutiny).
- Studies of side effects are needed, for example, to determine whether the developed form of accountability supports or interferes with the recruitment and retention of high-quality teachers for all children.
- Careful decisions need to be made about weighting of new measures as they become available for inclusion on the API. Modeling studies of potential volatile effects on API status by school and group will be required.
- Detailed studies of the relationship among all measures, those used for school report cards and the API, should be conducted to determine whether and how various out-

comes operate at cross purposes to one another.

- Smarter studies of alignment are necessary, including alignment of planned and enacted curriculum, resources, and preparation of teachers.
- Studies of the accuracy of the test are needed. In addition, strategies to help parents, the community, and the teaching force to understand the meaning of assessment—and what it does not mean—are essential.

Finally, well-designed assessments may tell us where we are and may communicate where we want to be. As we hope we have made clear, California's assessment and accountability system will need to continue to evolve to more fully achieve these goals and to support a standards-based system. We can all agree that the current status of student performance in California is insufficient, and that California schools need to improve. The real question is not where we are, but where we need to be and how we will get there. We should be looking for assessment results to show progress toward excellence—toward truly rigorous standards for student accomplishment—as well as progress toward equity. That is, we need to both raise our expectations for what children should know and be able to do, and assure that

as we move forward, we do not continue to leave some students—indeed a growing proportion—behind. We need to move all children ahead and reduce the gap between our least and most economically advantaged students. We need to find better ways to assure that poor students and students who start school without full English proficiency have effective opportunities to learn and are given what they need to make steady progress.

Certainly dramatic changes will not come overnight. Improvement will not come easily or quickly if we keep to high standards. It will take more than accountability and clear communication of expectations to change practice at a significant, meaningful level. It will take important and coordinated changes in capacity; in teacher quality; in curriculum, instruction, and assessment; in parent and community involvement; and in district and local capacity to support change—to name just a few, as the other chapters in this volume make clear. It also will require that we align and focus educational resources, policies and practices at the state, district, and local levels to assure all students achieve and learn what they need to be successful citizens of the future. We look to California's assessment system to be able to provide sound guideposts on how we are doing.

---

## Notes

1. National Commission on Excellence in Education. *A Nation at risk: The imperative for educational reform*. A report to the nation and the Secretary of Education. (Washington, DC: U.S. Government Printing Office, 1983).

2. See for example: Corbett, H. D., & Wilson, B. L. *Testing, reform, and rebellion*. (Norwood, NJ: Ablex Publishing, 1991); Dorr-Bremme, D. W., & Herman, J. L. *Assessing student achievement: A profile of classroom practices* (Los Angeles: UCLA Center for the Study of Evaluation, 1986); Kellaghan, T., & Madaus, G. F. "National testing: Lessons for America from Europe," *Educational Leadership*, 49.3(1991): 87-93; Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. "Can portfolios assess student performance and influence instruction?" in *The 1991-92 Vermont experience*, RAND (Santa Monica, CA: RAND, 1993; reprint from CSE Technical Report 371, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing, December.); Koretz, D. M., Barron, S. I., Mitchell, K. J., & Stecher, B. M. *Perceived effects of the Kentucky Instructional Results Information*

System (KIRIS). MR-792.PCT/FF. (Santa Monica: RAND, 1996); Koretz, D. M., Mitchell, K. J., Barron, S. I., & Keith, S. Final report: Perceived effects of the Maryland School Performance Assessment Program (CSE Tech. Report 409). (Los Angeles: UCLA National Center for Research on Evaluation, Standards and Student Testing, 1996); McDonnell, L. M., & Choisser, C. Testing and teaching: Local implementation of new state assessments (CSE Tech. Rep. No. 442). (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, 1997); Smith, M. L. Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (CSE Technical Report). (Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST), 1996); Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing (CSE Tech. Rep. No. 482). (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, 1998).

3. While most of these standardized tests were of the norm-referenced, multiple variety, some districts chose criterion-referenced tests that included some performance-oriented items. Selected tests had to meet technical quality criteria that were established by the state.

4. See NCEST report, 1992

5. IASA, 1994

6. See for example: Corbett, H. D., & Wilson, B. L. Testing, reform, and rebellion. (Norwood, NJ: Ablex Publishing, 1991); Dorr-Bremme, D. W., & Herman, J. L. Assessing student achievement: A profile of classroom practices (CSE Monograph Series in Evaluation, No. 11). (Los Angeles: UCLA Center for the Study of Evaluation, 1986); Kellaghan, T., & Madaus, G. F. "National testing: Lessons for America from Europe," *Educational Leadership*, 49.3(1991): 87-93. Shepard, L. Will national tests improve student learning? (CSE Technical Report 342). (Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing, 1991).

7. All students are required to take the test unless specifically exempted by an Individual Education Plan (IEP) or a written parent request.

8. Limited English Proficient (LEP) is the term used by California in its directives and reports. Many practitioners and researchers prefer the term English Language Learners (ELL) because of its accuracy and is more commonly found in the recent literature.

9. Abedi, J., Lord, C. & Hofstetter, C. "Impact of Selected Background Variables on Students' NAEP Math Performance." CSE Technical Report # 478, University of California. (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 1998).

10. The LEP designation applies to the full continuum of students from virtually no English proficiency to almost fully proficient. As students progress to the latter end of this continuum, scores from English language tests become more meaningful, though the point at which such meaning occurs is currently under investigation.

11. In addition to statewide measures, many districts have curriculum sensitive district assessments which are used to evaluate student achievement.

12. The Advanced Placement Program is conducted by the College Board in a total of 32 possible subjects.

13. Some believe that this drop may be an artifact of the norm group at this level, rather than representing an actual decrement in performance. Technical data that would more definitely determine the cause has not been available.

14. Measurement contrasts "observed" scores with "true" scores. "Observed" scores are the scores which students attain (and which are "observed") when students take a given test. Their true score is the score they would attain if the test were a perfect measure of their capability. We use "observed" scores to estimate what students "true" performance capability.

15. "How Accurate are the STAR National Percentile Rank Scores for Individual Students? – An Interpretive Guide" by David Rogosa is available to download at the CRESST Website, [www.cse.ucla.edu](http://www.cse.ucla.edu).

16. See, for example, Shepard, L. A. "Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?" *Educational Measurement: Issues and Practice*, 9 (1990): 5-22.
17. For reasons of space, only 3rd grade reading results are presented here. For analyses of other subjects, please see full technical report available through CRESST.
18. See CDE Stanford 9 Augmentation information at [www.cde.ca.gov](http://www.cde.ca.gov)
19. The SAT-9 mathematics augmentation was controversial at grades 8-11. Only students taking particular courses were required to take the test – e.g., students enrolled in Algebra at 8th grade, and critics raised serious questions about the technical and content appropriateness of the items. Because the test is still under development, we are not reporting results for these grades.
20. Hearing by the Joint Senate and Assembly Education Committee, November 1999.
21. Observed relationships were similar across grades and subject areas. Only 3rd grade results are displayed here. See full technical report for other subjects and grade levels. Herman, J., Brown, R, and Baker, E. "Student Assessment and Student Achievement in the California Public School System." CSE Technical Report (forthcoming). (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)).
22. Observed relationships were similar across grades and subject areas. Only 3rd grade results are displayed here. See full technical report for other subjects and grade levels. Herman, J., Brown, R, and Baker, E. "Student Assessment and Student Achievement in the California Public School System." CSE Technical Report (forthcoming). (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)).
23. United States Department of Education, 1998.
24. Education Watch 1998: The Education Trust State and National Data Book, Volume 2 (Washington, DC: Education Trust, 1998).
25. Barton, P.E. and Coley, R.J. *Growth In School: Achievement Gains from the Fourth to the Eighth Grade*. (Princeton, NJ: Policy Information Center, ETS, 1998).
26. Data from California DOE Website; 1988-89 – 1998-99 state summary numbers; One year dropout rate is calculated by the sum of the number of dropouts from grades 9-12, divided by the enrollment in grades 9-12 and un-graded secondary.
27. Data from California DOE Website; 1988-89–1997-98 state summary numbers.
28. Data from California DOE Website [www.cde.ca.gov](http://www.cde.ca.gov); 1991-92–1998-99 state summary numbers for public schools; rate of passing exams per 100 juniors and seniors in public high schools. See also, *The advanced placement program: California's 1997-98 experience*. (Sacramento, CA: California State University Institute for Education Reform, 1999).
29. David Hoff, "Inglewood ACLU Lawsuit," *Education Week*, 4 August 1999, Volume 2, 13.
30. Education Watch 1998: The Education Trust State and National Data Book, Volume 2. (Washington, DC, Education Trust, 1998).
31. Data from California DOE Website [www.cde.ca.gov](http://www.cde.ca.gov); 1988-89–1997-98 state summary numbers for public schools; SAT verbal, SAT math, percent meeting SAT criterion ( $\geq 1000$  on Verbal and Math sections), and percent of 12th graders taking the SAT. The percent of minority (American Indian + Black + Filipino + Hispanic + Pacific Islander/total enrollment) was calculated –and schools were designated as minority ( $>30\%$ ) or non-minority ( $<30\%$ ). The percent meeting the SAT criteria ( $\geq 1000$ ) and percent taking the SAT are presented.
32. Data from University of California for 1997 and 1998. Summary scores were created as the percent of students not meeting the requirements after taking the Subject A English examination.

33. The UC system does not have a consistent measure of mathematics preparation. Each campus uses its system for assessment/placement.

34. Regression analysis on SAT Combined scores (1993-97). The amount of explained variance from regressing average test scores onto the school measures of percent of students receiving free lunch and percent Limited English Proficient is plotted.

35. Results were similar across subject areas. For reasons of space, we only present figures for reading here. See full CRESST technical report for other subject areas. Herman, J., Brown, R, and Baker, E. "Student Assessment and Student Achievement in the California Public School System." CSE Technical Report (forthcoming). (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)).

36. Adopted by the California State Department of Education at their July 1999 meeting.

37. See Joint Committee on Testing (AERA, APA, NCME), Standards for Educational and Psychological Testing. (Washington, DC, 1999).



## About the Authors

*Elizabeth Burr* is Project Director for a child-care provider training and retention initiative at Policy Analysis for California Education (PACE). Her work focuses on early education policy research and analysis, with a focus on issues of access and equity. Recently she co-authored a Child Care Needs Assessment for Los Angeles County's Department of Public Social Services. She holds a B.A. degree from Brown University and an MPA degree from Columbia University.

*Bruce Fuller* is Associate Professor of Public Policy and Education at UC Berkeley. His current work focuses on family poverty and early education policy. As co-director of a Berkeley-Yale initiative called Growing Up in Poverty, Fuller is investigating how young children's lives are being affected by welfare reform. Fuller also writes in the area of decentralization and education policy, including school choice and charter schools. Before joining the Berkeley faculty, Fuller taught at Harvard University. He has also worked on education and family policy issues at the World Bank and in the California legislature. He received his Ph.D. in Sociology and Education from Stanford University.

Since 1969, *Michael Kirst* has been a Professor of Education and Business Administration at Stanford University. Before joining the Stanford faculty, Kirst held several positions with the federal government, including Staff Director of the U.S. Senate Subcommittee on Manpower, Employment and Poverty, and Director of Program Planning and Evaluation for the Bureau of Elementary and Secondary Education in the U.S. Office of Education (now the U.S. Department of Education). Kirst was a member of the California State Board of Education from 1975-1981 and its president from 1977-1981. He received his B.A. in Economics from Dartmouth College, his M.P.A. in Government and Economics from Harvard University, and his Ph.D. in Political Economy and Government from Harvard.

As director of PACE's Sacramento office, *Gerald Hayward* works primarily on issues in higher education, class size reduction, accountability and school finance. Until recently, Hayward served as Deputy Director of the National Center for Research in Vocational Education at UC Berkeley. He is also a founding partner of Management, Analysis and Planning, an educational consulting firm. From

1980-1985 Hayward served as Chancellor of the California Community Colleges, and prior to that served for a decade as Principal Consultant to the California State Senate Committees on Education and Finance. Hayward is a former teacher and administrator in California's public schools. He received his B.A. in Political Science from UC Berkeley and a master's degree in Education Administration from San Francisco State University.

*Richard S. Brown* obtained his Ph.D. in Social Research Methodology with a focus on advanced quantitative data analysis and measurement from the University of California, Los Angeles. His earlier research at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) involved developing assessments of teamwork skills in collaborative group settings using networked computers and investigations into the differential effects of question formats on meta-cognitive processes. His current research includes a project to develop an indicator system for California schools and an investigation into various methodological effects involved in setting performance standards on complex performance tasks using multi-level modeling, latent class analysis, and other advanced quantitative techniques.

*The Center for the Future of Teaching and Learning* is comprised of education professionals, scholars, and public policy experts who care deeply about improving education for California's children. The Center was founded in 1995 as a public, nonprofit organization with the purpose of strengthening the capacity of California's teachers for delivering rigorous,

well-rounded curriculum, and ensuring the continuing intellectual, ethical, and social development of children.

*Neal D. Finkelstein* is an education policy researcher, based in California. From 1997-99, he served as a senior program officer with the National Research Council's Committee on Education Finance which wrote *Making Money Matter: Financing America's Schools*. Prior to his work with the committee, Dr. Finkelstein was the assistant director of PACE and a research associate with the National Center for Research in Vocational Education, both located on the University of California, Berkeley campus. He has conducted research on numerous education policy issues, including public school finance, school governance, school-to-work programs and early childhood education. He holds a Ph.D. in Education Policy from UC Berkeley.

*Luis A. Huerta* is a research associate at PACE. His research focuses on issues of decentralization related to school reform and school choice, as well as the impact of school finance inequities on school reform. He is a contributing author to a PACE report titled *School Choice: Abundant Hopes, Scare Evidence of Results* (1999), and is also a contributing author to an upcoming book published by Harvard University Press, titled *Inside Charter Schools: The Paradox of Radical Decentralization* (Bruce Fuller, editor). Huerta received his master's degree in Education, and is presently a doctoral student at UC Berkeley. Before returning to graduate school, he served as a public school teacher for six years.

*William S. Furry* is an education consultant, engaged in policy research, political analysis, and strategic advocacy. Furry has held positions with The RAND Corporation, the California State Assembly, and the Governor's Office.

*Patricia Gándara* is Professor of Education at University of California at Davis and Associate Director of the University of California Linguistic Minority Research Institute. Among her recent publications are *The Dimensions of Time and the Challenge of School Reform* (1999) SUNY Press, and *Bilingual education programs: A cross national perspective*, co-authored with Fred Genesee, forthcoming in the *Journal of Social Issues*.

*Laura S. Hamilton* is an Associate Behavioral Scientist at RAND where she conducts research on educational assessment and the effectiveness of educational programs. Her areas of interest include the validity of large-scale achievement tests, the effects of high-stakes testing and accountability systems, and the relationships between classroom practices and student achievement in math and science. She received a M.S. in Statistics and a Ph.D. in Educational Psychology, both from Stanford University.

*Joan Herman* is associate director of the UCLA Center for the Study of Evaluation and co-director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Her research has explored the effects of testing on schools and the design of information systems to support school planning and instructional improvement. Her recent work emphasizes the validity and utility of per-

formance assessment and the measurement of students' opportunity to learn. She also has wide experience as an evaluator of school reform.

*Vi-Nhuan Le* is an Associate Behavioral Scientist at RAND where she conducts research on educational assessment. She is interested in group differences in achievement and exploring alternatives to multiple-choice tests. She earned a M.S. in Statistics and a Ph.D. in Educational Psychology, both at Stanford University.

*Abby Robyn* is an Associate Social Scientist at RAND. Her research focuses on the implementation of educational innovations. Her recent work includes studies of the relationship between teacher instructional practices and student achievement in math and science, a national conference on large-scale student assessment, and an evaluation of a high school reform initiative to improve the performance of students at risk of dropping out. She received a master's degree in English from UCLA.

*Russell W. Rumberger* is Professor of Education at the University of California, Santa Barbara and Director of the University of California Linguistic Minority Research Institute. He received his Ph.D. from Stanford University in 1978. A faculty member at UCSB since 1987, Rumberger has published widely in several areas of education: education and work; the education of disadvantaged students, particularly school dropouts; and education policy. He is currently working on a book on school dropouts and a study on the effects of school segregation on student achievement. He serves

on the editorial board of *Teachers College Record*, *Economics of Education Review*, and the *American Education Research Journal*.

*Andrea Venezia* is researcher at the Stanford Institute for Higher Education Research and the director of the Bridge Project: Strengthening K-16 Transition Policies. Venezia's work focuses on education policy research and analysis, particularly as related to the transition from K-12 to postsecondary education, with an emphasis on issues of

access, equity, and policy coherence. She has worked for a variety of state, federal, and not-for-profit education organizations including the National Education Goals Panel, the Texas Higher Education Coordinating Board, and the American Institutes for Research. Venezia earned a Ph.D. in Public Policy from the Lyndon B. Johnson School of Public Affairs at the University of Texas at Austin and a master's degree in Administration and Policy Analysis in Higher Education from Stanford University.



School of Education  
3653 Tolman Hall  
Berkeley, CA 94720-1670  
510-642-7223

URL: <http://pace.berkeley.edu>