Student Assessment and Student Achievement in the California Public School System

CSE Technical Report 519

Joan L. Herman, Richard S. Brown, & Eva L. Baker
CRESST / University of California, Los Angeles

April 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.5 Quality Education Forum Eva L. Baker, Project Director CRESST/UCLA

The work reported herein was supported in part by Policy Analysis for California Education (PACE) and 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 findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education; nor do they necessarily reflect the positions or policies of PACE.

STUDENT ASSESSMENT AND STUDENT ACHIEVEMENT IN THE CALIFORNIA PUBLIC SCHOOL SYSTEM[1]

Joan L. Herman, Richard S. Brown, and Eva L. Baker
CRESST/University of California, Los Angeles

More than fifteen 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" (National Commission on Excellence in Education, 1983). A decade later, California received its own special wake-up call when results from the 1990 and 1992 National Assessment of Educational Progress (NAEP) 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 not by California's educational system?

Such seemingly simple, bottom-line questions are foremost in the minds of the public and its policymakers. 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 that most influence assessment results. We close by returning to the goals of accountability and standards by which such systems should be judged.

The Assessment Context: Setting the Course With a System in Transition

California, as the rest of the nation, is creating statewide assessment systems intended not only to measure student learning, but to leverage its improvement.

[1] An edited version of this report will appear in *Conditions of Education* (Berkeley, CA: University of California, Policy Analysis for California Education [PACE], in press).

The system itself is intended as part of the reform: It signals what is important to teach and learn by providing specific learning targets—that is, the content of the test. The assessment also provides feedback on how students are doing and thus enables a diagnosis of curriculum strengths and weaknesses; coupled with sanctions and or incentives, it motivates 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 (see, for example, Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Kellaghan & Madaus, 1991; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Koretz, Stecher, Klein, McCaffrey & Deibert, 1993; McDonnell & Choisser, 1997; Smith, 1997; Stecher, Barron, Kaganoff, & Goodwin, 1998).

The Importance of Alignment

It is not only that teachers and schools tend to emphasize the content of important statewide tests in their day-to-day curriculum—and, as a corollary, to ignore or neglect what is not tested—but they also tend to mimic *how* the content is tested. Under pressure to show improved performance, schools tend to prepare students by spending lots of instructional time on exercises that look just like the test items. In the extreme, an overreliance on multiple-choice testing thus can become a narrow, multiple-choice curriculum, devoid of the complex thinking and communication skills that are essential for students' future success. Designers of the now defunct California Learning Assessment System (CLAS), with some of it sister assessment systems across the country, recognized the power and limitations of traditional testing to promote change and sought, through the use of performance assessments, to provide teachers with models of *how* to teach.[2] That is, the performance assessment tasks themselves were intended to exemplify the kinds of curriculum and teaching that were expected to stem the "rising tide of mediocrity" and to be "tests worth teaching to" (see Resnick & Resnick, 1992). On the other hand, some of these systems lost credibility because of decisions made about content and implementation. But the lesson they falteringly tried to teach remains a challenge.

The alignment between what is tested and what students are expected to learn thus is a critical criterion for any assessment or accountability system intended to

[2] CLAS is but one example of how the state testing system has changed over the last five years, a subject to which we return below.

promote the improvement of student learning and 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, then 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 the case of a school 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, not the goal itself.

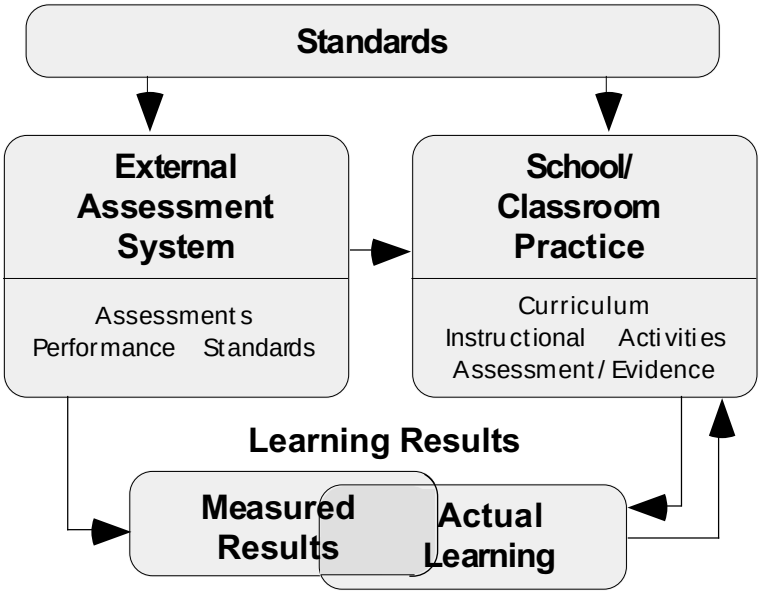

Figure 1. Standards-based assessment framework.

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 the one that 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 measure does not respond to 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

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? Does it reflect the enduring themes and/or priority principles, concepts and topics of the discipline?

Instructional sensitivity. Can the assessment detect differences in the quality of instruction? Does the assessment 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 does 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 Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

(National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, UCLA)

Figure 2. Characteristics of quality standards-based assessment systems.

misrepresent the reality of serious educational reform. Note also the final characteristic, 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 overreliance 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 (see, e.g., Corbett & Wilson, 1991; Dorr-Bremme & Herman, 1986; Kellaghan & Madaus, 1991; Shepard, 1991). Different constituencies, furthermore, find different types of information meaningful and useful. For example, basic skills are high in the public's priorities, and parents and the public often want to know how their children compare with others—nationally and internationally. Education 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 that may not be well assessed through multiple-choice testing.

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. The rocky and changing story of the state's plans over the last few years, furthermore, means that the basic requirement for analyzing students' progress—a consistent measure used over time—is not available. Five years ago, the state's system featured CLAS, a largely performance-based assessment system that focused on the complex thinking and problem-solving aims of the state's curriculum frameworks at the time. CLAS, however, 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 districts to select and administer assessments that best reflected their local standards. The result was a plethora of different standardized tests[3] 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

[3] Although most of these standardized tests were of the norm-referenced, multiple-choice 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.

establish a baseline and hold schools accountable, the state (and particularly then Governor Pete Wilson) initiated in 1998 the new California Standardized Testing and Reporting (STAR) program, using the Stanford Achievement Test Series, 9th Edition, Form T (SAT-9) in Grades 2-11. 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 predated the state's standards. The SAT-9, it should be noted, is a norm-referenced test, designed primarily to show how students or schools perform relative to others—others in the state, others in similar schools and districts, and others in the national norm group (or average).

All of this occurred at a time when most other states were making progress in meeting federal expectations for state standards and assessments. The plan, originally designed in 1991-92 (see National Council on Education Standards and Testing, 1992), was given additional impetus by the Improving America's Schools Act (1994), 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 the California State Board of Education's passing of state standards this past December, 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 the previous year, but added additional items 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 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, whereas the standards-referenced scores would tell those stakeholders the extent to which students were performing at advanced, proficient, basic, or below basic levels relative to the state standards. As we describe below, although a first set of augmented items has been administered, there are some questions about their appropriateness, and performance standards have not yet been established for them; so results from the augmentation are not yet directly interpretable.

Additional components are in the works, furthermore, to bring STAR into still closer alignment with the 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 that are reflected in the standards, but not well assessed by the SAT-9. Because, as mentioned above, 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. This component will particularly address the *instructional sensitivity* and *consequences* criteria by modeling the types of teaching and learning that are expected of students and precluding an exclusive focus on drill-and-kill formats in classroom instruction. The Matrix test is called so because it 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 task 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. Although it has not been designed to yield an individual-level student score, 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.

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 high school exit exam will be required for high school graduation and is scheduled to go operational in 2004. An English Language Development Test is also under development.

Assessing Limited English Proficient Students

While state code requires that all students Grades 2-11, including those who are not fully proficient in English, take the SAT-9,[4] it also provides that limited English proficient (LEP)[5] students who have been in school less than 12 months also be

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

[5] Limited English proficient (LEP) is the term used by California in its directives and reports. Many practitioners and researchers prefer the term English language learner (ELL) because of its accuracy. It is more commonly found in the recent literature.

tested in their primary language. Students who have been in school more than 12 months, but who are still classified as LEP, may also be administered a primary language test. The state has selected the SABE/2 as the statewide measure to be used for assessing students whose first language is Spanish, but this common measure will not be implemented statewide until 2000. 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 fully 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 *consequence* of not testing and reporting LEP students' performance is that their progress 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 helps only those students who have been instructed in their native language (Abedi, Lord, & Hofstetter, 1998)—a circumstance that 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 that attempt to reduce the language load of a test or otherwise compensate for students' reduced language skills (e.g., by providing students more time) are also currently being researched, but answers that are equitable and fair for all students have 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.[6]

[6] The LEP designation applies to the full continuum of students, from those with virtually no English proficiency to those who are 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.

Academic Performance Index

The components of the California Standardized Testing and Reporting (STAR) program thus are abundant, and 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 has developed a high-stakes, school performance index based on them. The Academic Performance Index (API) will be used to rank schools across the state predominately on their test scores. More will be described about the API in our concluding section.

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. 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 subpopulations. For example, the dropout 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 under the College Board,[7] 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 of students passing are of interest. Similarly, college entrance exams, such as the SAT, provide an indicator of both students' expectations and their preparation to attend college.

Other indicators are external to the K-12 system and provide a validity check of its 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 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 work, provide another external comparison point for judging the quality of the state's precollegiate system.

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

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 not only would provide greater consistency and focus to California schools but also 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 these indicators with state standards aside, one looks for consistency in performance across them to judge the quality of California's educational system. 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 knowledge of the wide variety of student achievement measures used in the state. In the next few pages, we'll describe those instruments, what they are intended to measure or monitor, and how well California school children are doing on them. We'll review data both from the most recent testing period and over a longer 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. In this section we will begin to discuss the evolution of assessment measures in the state. Next will follow information regarding California's performance on the National Assessment of Educational Progress (NAEP). From there, we'll examine the results of a series of secondary school measures, including high school dropout 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, 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 education in our California public schools is and whether there is evidence it is headed in the right direction.

STAR Results

California began to implement its Standardized Testing and Reporting (STAR) system with the Stanford Achievement Test, Version 9, Form T (SAT-9) in 1998. This norm-referenced standardized achievement test is given annually to over 5 million California students during the spring of the school year. In addition, Spanish-speaking students with limited English proficiency (LEP) who have been enrolled in the public school system for less than 12 months also are required to take a Spanish language test called the Spanish Assessment of Basic Education, 2nd Edition (SABE/2). Spanish-speaking LEP students who had been enrolled in the public school system for more than 12 months may also take the SABE/2 at the individual district's discretion.

Since the norm-referenced SAT-9 was not designed to align with California learning standards, students in 1999 also received an additional test, an augmented form of the SAT-9, developed as a partial attempt to address a broader range of the California standards in mathematics and language arts. Additional augmented forms have been proposed for science and social studies.

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 1997-1998 and 1998-1999 academic years, with some cautions about the interpretation of the scores. We will follow with analyses of the performances of limited-English-speaking 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 is well to recall that the SAT-9 is a norm-referenced test and the meaning of results from that type of test. 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. For norms to be interpreted easily, one would wish that the kinds of students tested in a state and in the norming group were similar. While no tested and norming groups are ever exactly alike, in California interpretation is made difficult for a

number of reasons. First, California's student population differs substantially from the national norm group in its diversity, its urban concentrations, and because unlike other states, California assesses virtually all its students on an English language examination, and approximately a quarter of them are not fully proficient in English. Just using this example, it is not hard to predict that students who do not understand English are less likely to perform well on an English language test, and thus that California students are likely to fare poorly when compared to a national norm group where only two percent are similarly non-English proficient.

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. In reading for Grades 2 through 11 (see Figure A1; Figures A1-A67 are presented in Appendix A), scores ranged from the 32nd to the 44th percentile in 1998, and from the 32nd to the 46th percentile in 1999. In no grade level did the percentile rank of the mean normal curve equivalent exceed the 50th percentile in either year, and performance shows a precipitous drop at the high school level.[8]

Observed scores were somewhat better in mathematics (see Figure A2), 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 (see Figure A3) and spelling (see Figure A4) 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, in science (see Figure A5) 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 (see Figure A6), only 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

[8] 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 have not been available.

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 at the results of California students who are fully proficient in English, a comparison that somewhat favors California students, since approximately 2% of the national norm group are not proficient. Here, the 1999 results (see Figures A7-A12) 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 students' performance is still relatively the best in mathematics. And regardless of comparison group, students are performing relatively the poorest in spelling at the elementary school level and in science and social studies at the high school level.

How are California's non-English-proficient students doing on the SAT-9? As expected, limited English proficient students score lower than English-proficient students on the SAT-9 across all subject areas and grade levels. However, when students were tested in their primary language, a different perspective is presented.

In addition to the SAT-9, students enrolled in the public school system for less than 12 months were required to take a test in their primary language (if available). Students enrolled in the system more than 12 months had the option of taking this additional test. The only primary language test available is the SABE/2, which is the primary language test for Spanish-speaking students. Figures A13-A16 represent the number of students who took the SABE/2 for reading, math, language and spelling in the 1999 testing period.

By comparing student performance on the SABE/2 with the performance of students with limited English proficiency on the SAT-9, we get a glimpse of how testing in an unfamiliar language may affect student achievement. With the exception of Grade 11 math, scores on the SABE/2 are higher than scores of LEP students taking the SAT-9 (see Figures A17-A20). Differences between groups diminish in the upper grade levels for math. However, differences are prevalent across all grade levels in reading, language and spelling. This suggests that Spanish-speaking students may perform better when tested in Spanish than when tested in English in these subject areas. However, several important distinctions between these two comparison groups should be made. First, the students designated as LEP

represent a range of languages other than Spanish. Since these analyses are not based on common students taking both measures, the extent to which scores of Spanish-speaking LEP students are related on the SABE/2 and SAT-9 is not identifiable, and therefore, definitive conclusions cannot be drawn. Second, comparisons between these measures involve different normative samples as referent groups, which may or may not be equivalent in terms of their subject proficiency. While the metric for the LEP students is the percentile rank of the mean normal curve equivalent for a representative national sample, the metric for the SABE/2 scores compares the average student's score with a sample of Spanish-speaking students in bilingual programs. Thus, the results of such comparisons should be interpreted with a great deal of caution.

Are California's schools improving? With 1999 being the second year of the statewide STAR testing system in California, comparisons between scores from the initial year (1998) 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 are not without error. That is, observed scores do not exactly represent true scores for individuals or for groups due to errors of measurement. 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) is how accurately do the observed scores represent 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, Rogosa has created an easy-to-read guide titled *How Accurate Are the STAR National Percentile Rank Scores for Individual Students?—An Interpretive Guide.*[9] 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 with a true score at the 50th percentile of the norm group obtains a score more than 5 percentile points away from the 50th percentile?" The answer—70%! That is, there is only a 30% chance that the observed score is between the 45th and 55th percentile points.

[9] *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 Web site, www.cse.ucla.edu.

With respect to interpreting progress, the accuracy 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% chance of showing at least a 10 percentile point change (up or down) in the second year! To state it differently, for this case it is more likely than not that a true score that actually remains the same from one year to the next will result in an observed score difference of more than 10 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.

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 (see, e.g., Shepard, 1990).

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 percent 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. The rankings of schools resulting from taking the difference between average performance in 1999 for a given grade level (e.g., third grade) and average performance in 1998 for that same grade level (e.g., third grade) differed from the rankings of schools resulting from taking the difference between average

performance in 1999 for a given grade level (e.g., third grade) and the average performance for that cohort in 1998 (e.g., second grade). Quintile rankings across these two approaches agreed only about a third of the time (see Figures A21-A24). 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). 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. One of the more widely discussed issues was the lack of alignment between the subject matter assessed by the SAT-9 norm-referenced test and the California state standards being taught in the public schools. In an effort to better align the assessment with what is outlined in the state standards, an augmented version of the SAT-9 was created for the subject areas of English and mathematics (see CDE Stanford 9 Augmentation information at www.cde.ca.gov). Though it is difficult to interpret student performance on the augmented test since the state has yet to identify what constitute various performance levels, the general consensus was that the tests sampled difficult elements of the 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 A25). Generally, a lower percentage of the items was answered correctly on the math test at each grade level (see Figure A26), with better performance at the lower grades.[10] Reports at open testimony at the October 1999 meeting of the California State Board of Education recounted anecdotes of students confronted with problems in mathematics far beyond their capability.

It is important to point out 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

[10] The SAT-9 mathematics augmentation was controversial at Grades 8-11. Only students taking particular courses were required to take the test (for example, students enrolled in Algebra at eighth 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.

"standards-based" element of the STAR.[11] It will be important to follow the extent to which these particular tests are curriculum referenced and thus will reflect appropriate classroom instruction.

How is school composition related to school's SAT-9 performance? We've seen that, 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 burdened 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 (see Tables B1-B6 in Appendix B) 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 higher proportions of students receiving free or reduced lunch score considerably lower than schools with lower proportions of such students. Interestingly, the relationship holds for both economically disadvantaged and advantaged students. That is, the average score for both groups of students tends to be lower in schools where there are higher concentrations of poverty. Therefore, it appears the extent to which a school confronts the difficulties 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 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 test scores is not surprising, and, as mentioned above, partly explains the relatively low overall average achievement for students in California. Since the SAT-9 norm group and the California student population differ dramatically on

[11] Hearing by the Joint Senate and Assembly Education Committee, November 1999.

these key measures (e.g., the norm group was comprised of 1.8% limited English proficient students whereas the California population consists of more than 25% limited English proficient students), lower average performance for California students as a whole relative to the normative group should be expected. Figures A27-A35 graphically present how the average performance varies for both those with and those without sufficient language skills. When there are more students at a local school site without English language proficiency, performance for all students is lower. Figures A36-A44 do the same for the problem of student poverty. Most clearly, poor students and students of limited English proficiency do less well as their proportions increase at a school.

The relationship may not be as direct and clear, however, for students who are identified by our analyses as economically advantaged and/or for students who are fully English proficient because of the limits of the variables available to us. Clearly, those who are not eligible for free or reduced lunch (the advantaged or "non disadvantaged" group in our analyses) represent a large range of 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 students in poverty are different from and relatively less economically advantaged than those in schools with low proportions of children qualifying for free lunch and reduced 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. And similarly, 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 are still not totally secure; poverty may also 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 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 to students in the rest of the country. For instance, for the 1996 assessment of eighth-grade mathematics, California ranked 31st out of 41 states and did even worse in fourth-grade reading, coming in dead last out of 38 states. As indicated in Figures A45-A48, California lags the nation in Grades 4 and 8 in both reading and mathematics achievement. Only 17% of California students performed at the proficient level in eighth-grade mathematics, and 11% achieved that standard in fourth-grade mathematics—both of which are much lower than the national rates. Similarly, in eighth-grade writing, only 1 in 5 California students achieved at or above the proficient level, compared to 1 in 4 nationally (Figure A49). 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, on poverty, 16.5% of California schools in the 1992 NAEP reading sample showed 75% or more students eligible for free or reduced lunch, and in the 1994 assessment, the figure was 16.6% of the California school sample (U.S. Department of Education, 1998). In those two assessments, only 12.7% of those sampled in poverty schools scored at or above Basic (the lowest level of achievement) in 1992 and only 14.8% in 1994. Looking at 1994 only, 10 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%, Mississippi with 39%, and New Mexico with 26%) were substantially superior to California on this metric (Georgia with 29% scoring at Basic and above, Mississippi with 29%, and New Mexico with 32% of students scoring at Basic or above). Only one entity, the District of Columbia, with about 62% of the schools meeting this poverty definition, scored below California, at 13.9%. Even so, the District of Columbia is doing a better job proportionally for its students when one looks at poverty and performance conjointly. Overall, these numbers show the U.S. has a long way to go in educating its poor students, but California is clearly lagging.

In mathematics, the situation is comparable for the 1996 data. Twenty-one states had higher proportions of poverty than California, and of these only the District of Columbia performed more poorly. Some of these states, like West Virginia with 29.7% poverty, had over 60% of their students reaching to or above the Basic level in mathematics.

However, not all data from the NAEP assessment in California are bad. Other NAEP performance data (*Education Watch 1998*) 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 perform 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 measures (Barton & Coley, 1998). Clearly, California needs to continue to make progress and has a long way to go.

Dropout/Graduation Rates[12]

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. Two of them are the dropout rates and completion rates for students in high school. Definition of dropouts often varies. California 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.

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 back, after an extended hiatus. California is moving to a state-level

[12] Data from the California DOE Web site; 1988-89–1998-99 state summary numbers; 1-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 ungraded secondary.

student data system that will permit more precise understanding of these indicators. Nonetheless, using data from the California Department of Education, we present 10-year trend lines of dropout and graduation rates for California high school students in Figures A50 and A51. As seen in these figures, dropout 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 90-91% rate, though the rate was a few points higher at the beginning of the decade.

High School Course Taking Patterns[13]

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:

A. History / Social Science—2 years required. (Two years of history / social science, including one year of U.S. history or one-half year of U.S. history and one-half year of civics or American government; and one year of world history, cultures, and geography.

B. English—4 years required. Four years of college preparatory English that include frequent and regular writing, and reading of classic and modern literature. Not more than two semesters of ninth-grade English can be used to meet this requirement.

C. Mathematics—3 years required, 4 years recommended. Three years of college preparatory mathematics that include the topics covered in elementary and advanced algebra and two- and three-dimensional geometry. Approved integrated math courses may be used to fulfill part or all of this requirement, as may math courses taken in the seventh and eighth grades that your high school accepts as equivalent to its own courses.

D. Laboratory Science—2 years required, 3 years recommended. Two years of laboratory science providing fundamental knowledge in at least two of these three disciplines: biology (which includes anatomy, physiology, marine biology, aquatic biology, etc.), chemistry, and physics. Laboratory courses in earth / space sciences are acceptable if they have as prerequisites or provide basic knowledge in biology, chemistry, or physics. The appropriate two years of an approved integrated science program may be used to fulfill this requirement. Not more than one year of ninth-grade laboratory science can be used to meet this requirement.

E. Language Other than English—2 years required, 3 years recommended. Two years of the same language other than English. Courses should emphasize speaking and understanding, and include instruction in

[13] Data from the California DOE Web site; 1988-89–1997-98 state summary numbers.

grammar, vocabulary, reading, and composition. Courses in languages other than English taken in the seventh and eighth grades may be used to fulfill part of this requirement if your high school accepts them as equivalent to its own courses.

F. College Preparatory Electives—2 years required. Two years (four semesters), in addition to those required in "A-E" above, chosen from the following areas: visual and performing arts, history, social science, English, advanced mathematics, laboratory science, and language other than English (a third year in the language used for the "E" requirement or two years of another language).

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, and also is 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 A52, whereas fewer than 30% of graduates met the requirement in 1988, more than 38% 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 in courses may vary in 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[14]

Another secondary school measure of interest is the availability of and participation in Advanced Placement (AP) 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 the percentage of those passing the exam are indicators 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

[14] Data from the California DOE Web site; 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 California State University Institute for Education Reform, *The Advanced Placement Program: California's 1997-98 Experience*, Sacramento, CA, 1999.

litigation has been brought against at least one California school district for allegedly providing disproportionate AP opportunities to students of varying ethnic backgrounds (Hoff, 1999). 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 percentages of each group in the student population. For instance, African American students comprise 8.8% of the public school population and 3.5% of students taking the English Advanced Placement examination, and 2.5% for the Calculus test. By contrast, Asian students comprise 11.2% of the student population, but account for 28.1% of the English AP test takers and a whopping 42.8% of those sitting for the AP Calculus examination (*Education Watch 1998*). Clearly, the ethnic makeup of the group 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 1991-1992 academic year, this rate has steadily improved every year, going from 9.2% at the beginning of the decade to 14.8% last year (see Figure A53).

College Entrance Examinations[15]

College entrance examination scores are another measure of how California high schools are preparing students for college. 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.

As with the average scores for the math and verbal components of the SAT, there appears to be a rebounding trend in the percent of test takers meeting or exceeding the combined 1000 point threshold. This measure, too, experienced a

[15] Data from the California DOE Web site; 1988-89–1997-98 state summary numbers for public schools; SAT verbal, SAT math, percent meeting SAT criterion (>=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 (>=1000) and percent taking the SAT are presented.

slight dip in the early 1990s but has inched up to remain between 18% and 19% 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 A54). And as Figure A54 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 A55, the percent of 12th graders taking the SAT in low minority schools has remained stable at around 46%, while that rate has hovered around 36% for high minority schools over the same time period.

College Attendance

Data from *Education Watch 1998* indicate that 66.4% of high school graduates in 1996 enrolled in college (full- or part-time) by the time they were 19 years old. This rate ranked California 5th out of 50 states in providing students access and opportunity for college. However, the college completion rate for minority students entering as freshman—deemed the equity rate—is not so rosy. The equity rate for California is 58.4%, which is below the national average (65.0%). California's 4-year graduation rate is 41%, which means that less that half of entering freshman graduate within a 4-year period.

College Remediation Rates[16]

Part of the reason for the lower completion rates may be that California has a high number of part-time community college students. Another reason may be 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% of incoming first-year students are required to take remedial math and over 47% need remedial reading classes. In the state's elite University of California system, over a third of the students fail to meet the minimal standards of writing proficiency.[17] This number has improved in the last year, from 38.9% in 1997 down to 33.3% in 1998. However, this indicator still suggests that although more high school graduates are

[16] Data from the 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.

[17] The University of California system does not have a consistent measure of mathematics preparation. Each campus uses its own system for assessment / placement.

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

Compared to the rest of the nation, California student achievement is low—not only on the SAT-9, but also on the National Assessment of Educational Progress. Average student performance in some schools is better than in others, but it is fairly easy to identify which schools these are—not by what's going on in the schools but by who is going to them. Although students take an additional test designed to address their mastery of state-determined subject matter standards, it is not ready for widespread implementation. Minimal observed scores gains, essentially across the board, from last year to this year on the SAT-9 probably do not signify "real" improvement, but familiarity with the process. Moreover, different measures of improvement greatly disagree with one another.

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

More troubling, the relationship between the socio-demographic complex of poverty, language skills and ethnicity and standardized student achievement measures is immense and getting stronger.[18] Over the past six years, this relationship has strengthened, not diminished (see Figure A56). 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 60-80% of the variance in average school scores in

[18] 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.

reading, language arts and spelling (see Figures A57, A59, and A60). A somewhat weaker relationship is found for the background measures and mathematics (see Figure A58), although a majority of the variance is still accounted for at each grade level, from a low of 56% in Grade 2 to a high of 67% in Grade 4.

The relationship is clear. More poverty relates to lower average scores. More students with limited English skills relates to lower average scores. Greater minority representation in the student population, sadly, also relates to lower average scores. Poverty is increasing as the percent of students qualifying for free or reduced lunch has risen from 32.19% in 1989 to 47.61% in 1999. Similarly, the percent of California students with limited English proficiency has jumped from 16.29% in 1989 to 24.89% in 1999. Both of these increases represent about a 50% jump over the past decade (see Figures A61, A62). Over the same period, the minority population has risen only slightly, but consistently (see Figure A63). 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 lunch in schools with high concentrations 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 what their background is. Unfortunately, we currently are not seeing the desired effect.

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

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 (for example, absences and graduation rates) as part of a system to hold schools accountable. The plan is supposed to support *standards-based* reform. Over a 6-month period, a committee of school policymakers, academic experts, and practitioners met and

prepared the requirements of the API. The details are available on the California Department of Education web page (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 relatedly, the degree to which the assessments represent and forward 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 of criteria (Figure 3) that enunciated principles to guide the use of the accountability system.

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
- 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 3. Academic Performance Index Framework. (Adopted by the California State Department of Education at their July, 1999 meeting.)

The relevance of these principles to concerns we have raised earlier about assessment and criteria for quality assessment systems is 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 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* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) the following paraphrased standards (see Figure 4) are applicable to California planning and evaluation.

Relevant Standards for California from the
***Standards for Educational and Psychological Testing* (1999)**

- 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 4. From *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). (Adopted by the California State Department of Education at their July 1999 meeting.)

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 around the national average. In fact, given the difference in the composition 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. Independent of our personal endorsement or skepticism about the program as it now stands, we believe that there is direct action that can be taken to support the best possible development of the assessment system, 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 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.

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 immediately undertaken. These studies must address the impact of the assessment for 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 outcomes 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 easy, 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. It will also require that we align and focus educational resources, policies, and practices at state, district, and local levels to assure that all students achieve and learn what they need to be successful citizens of the future. We look to California's assessment system to provide sound guideposts.

Appendix A

List of Figures

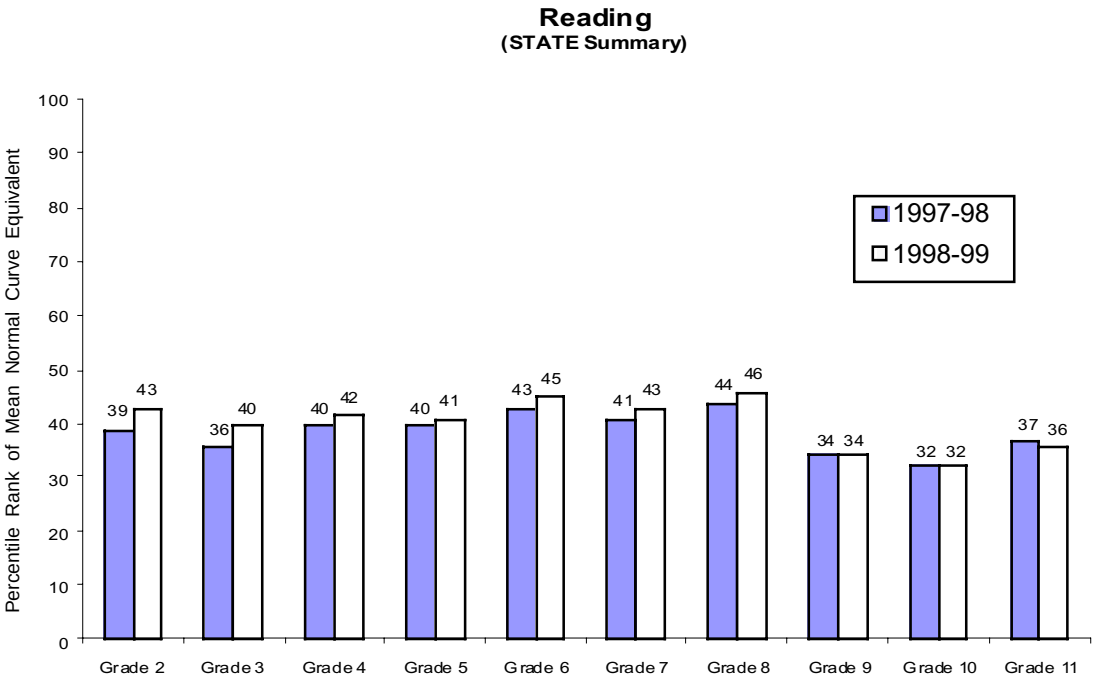

Figure A1. SAT-9 Reading scores.

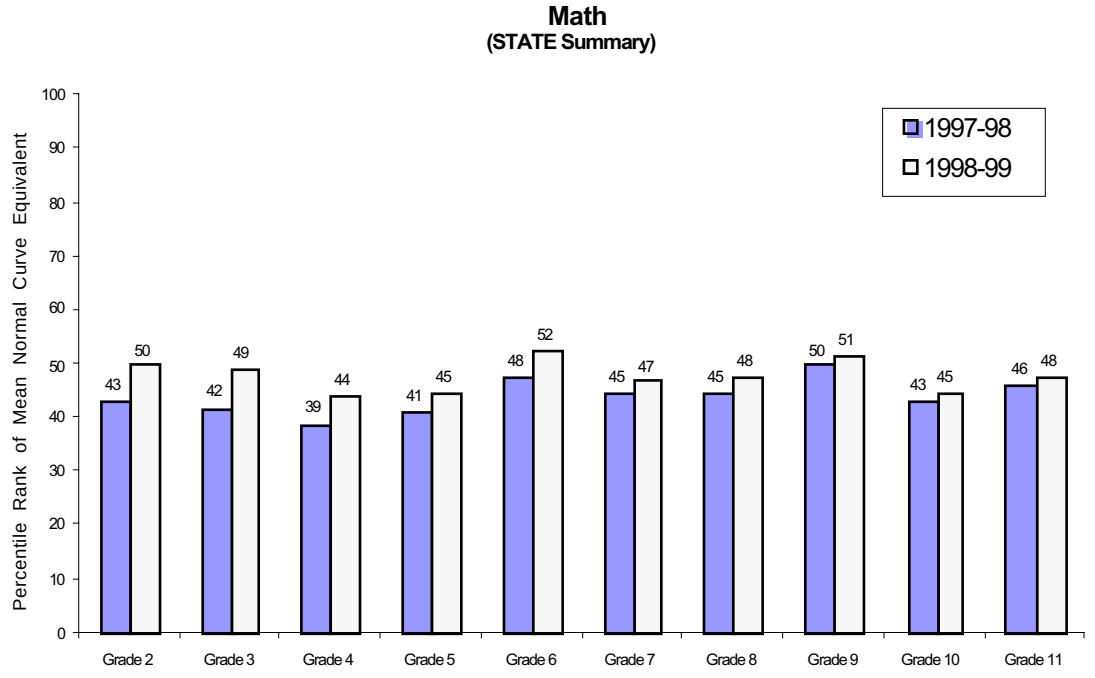

Figure A2. AT-9 Math scores.

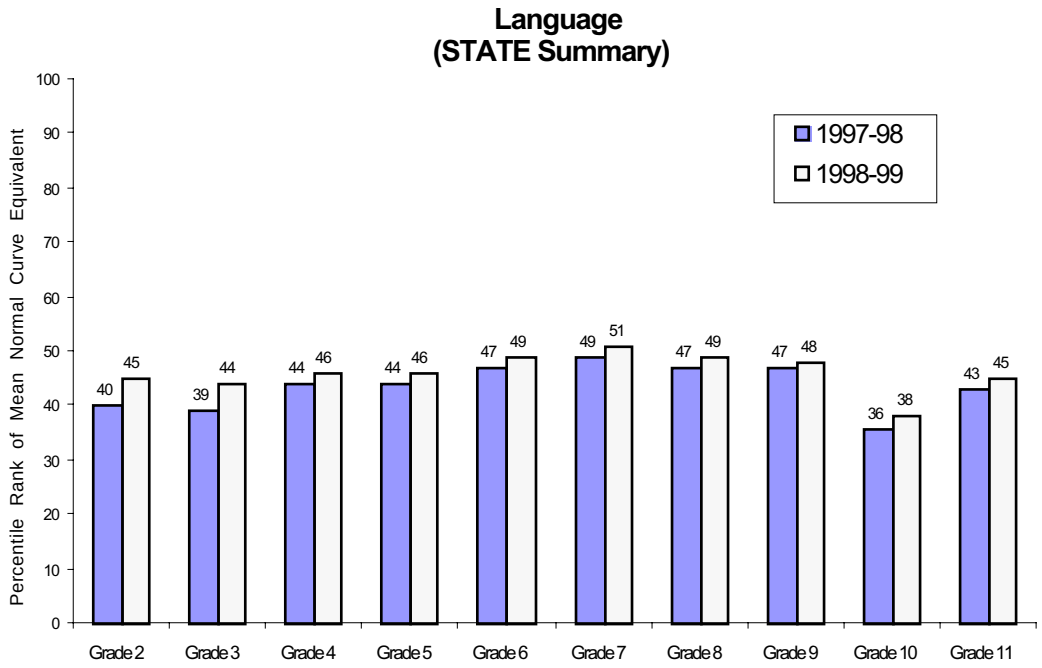

Figure A3. SAT-9 Language scores.

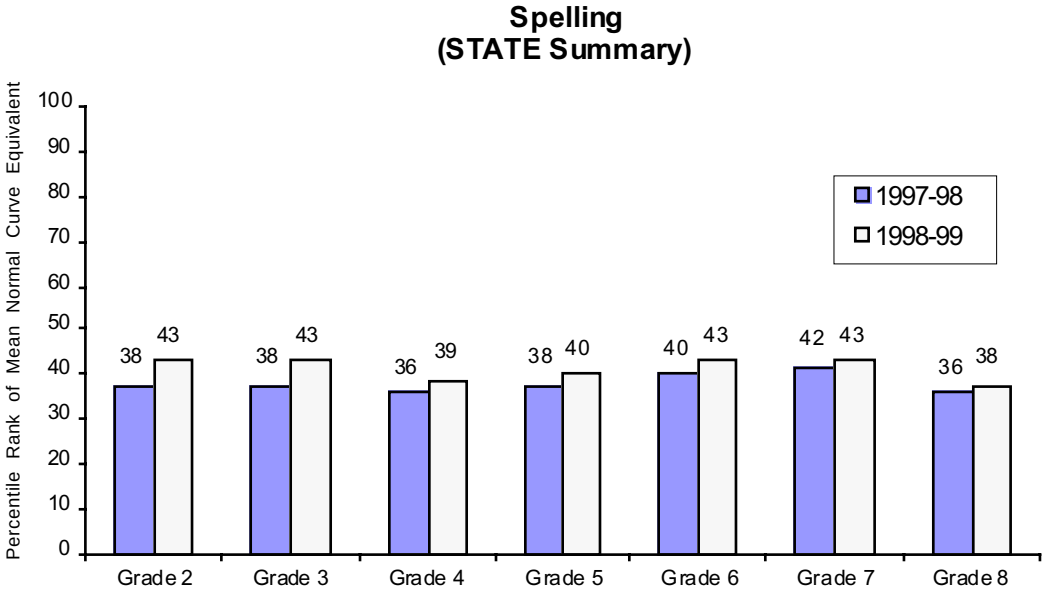

Figure A4. SAT-9 Spelling scores.

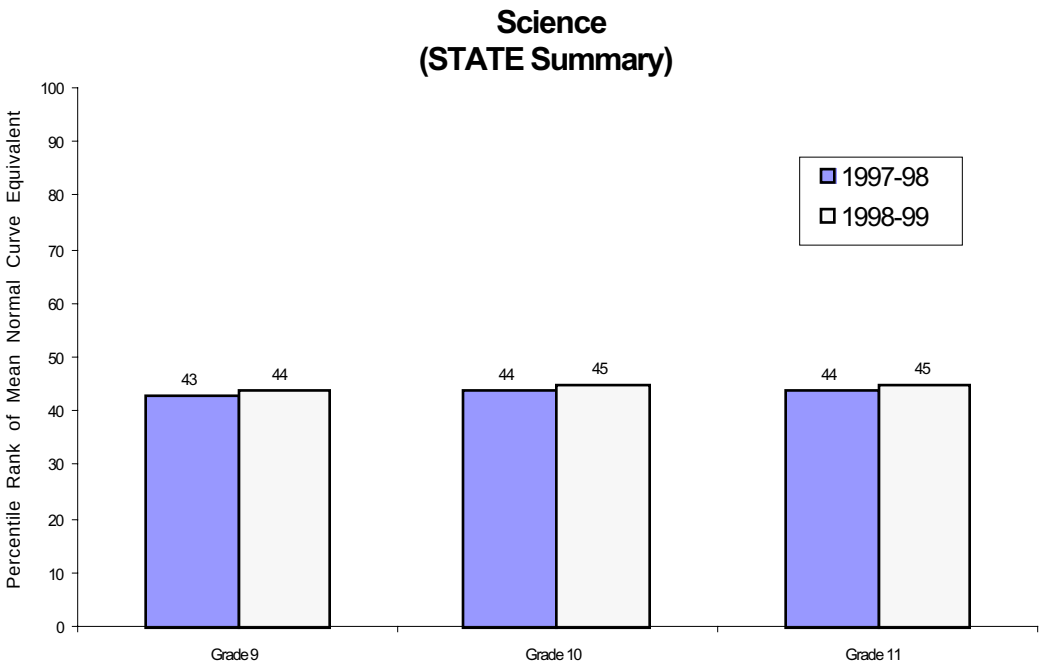

Figure A5. SAT-9 Science scores.

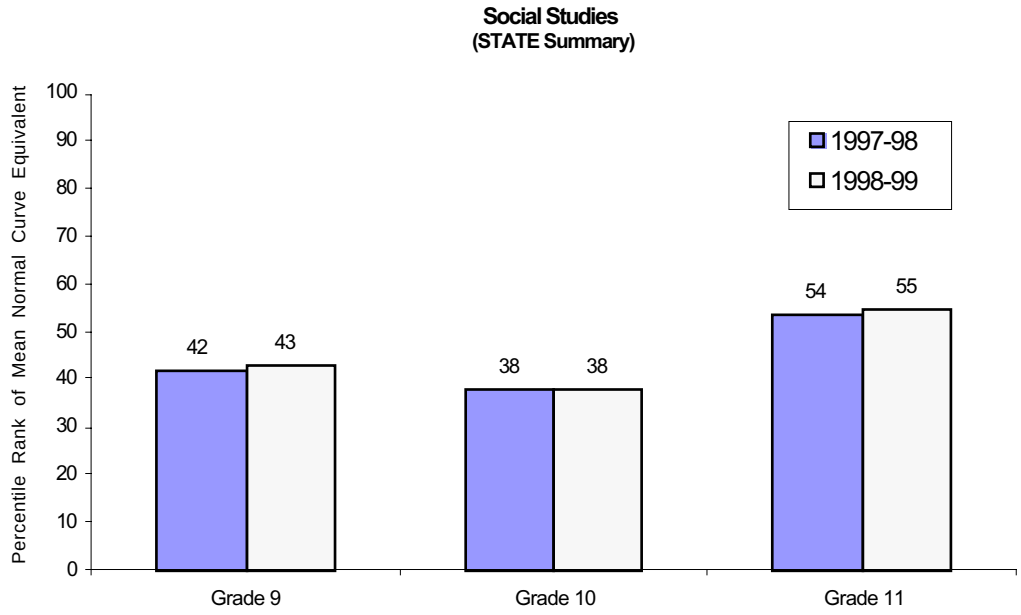

Figure A6. SAT-9 Social Studies scores.

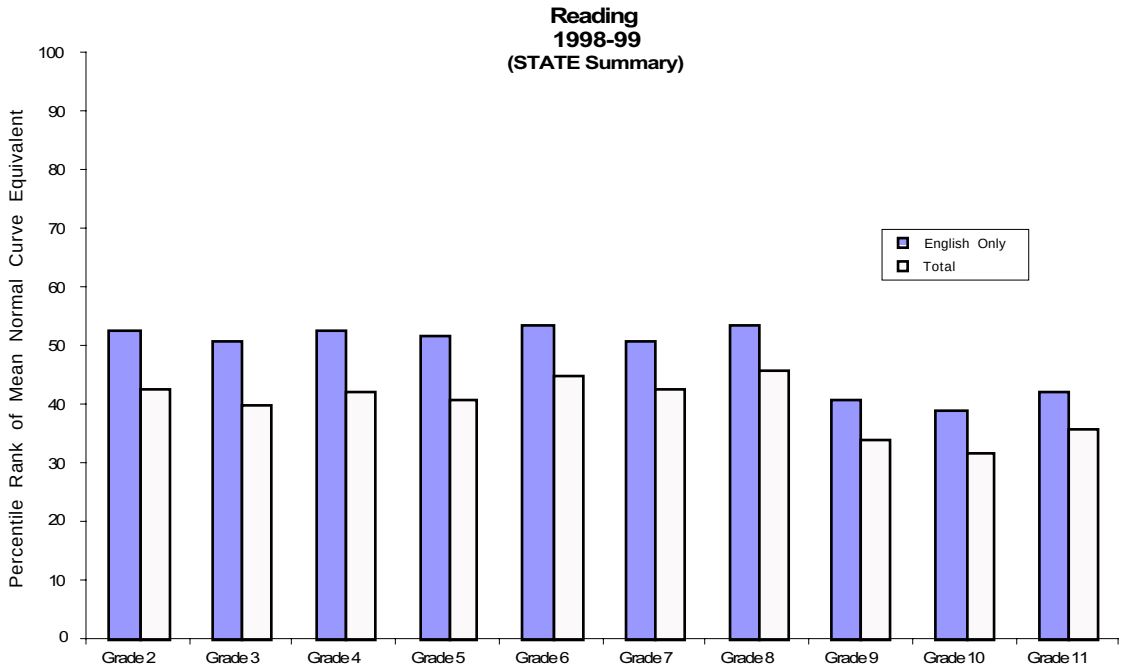

Figure A7. SAT-9 Reading – All students versus English proficient students.

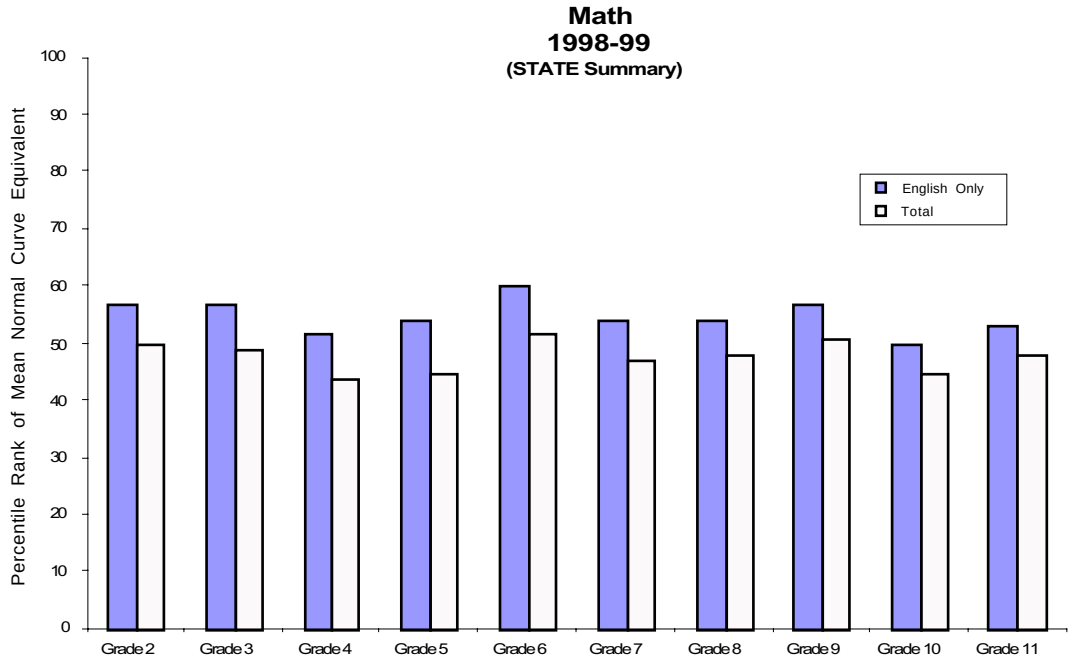

Figure A8. SAT-9 Math – All students versus English proficient students.

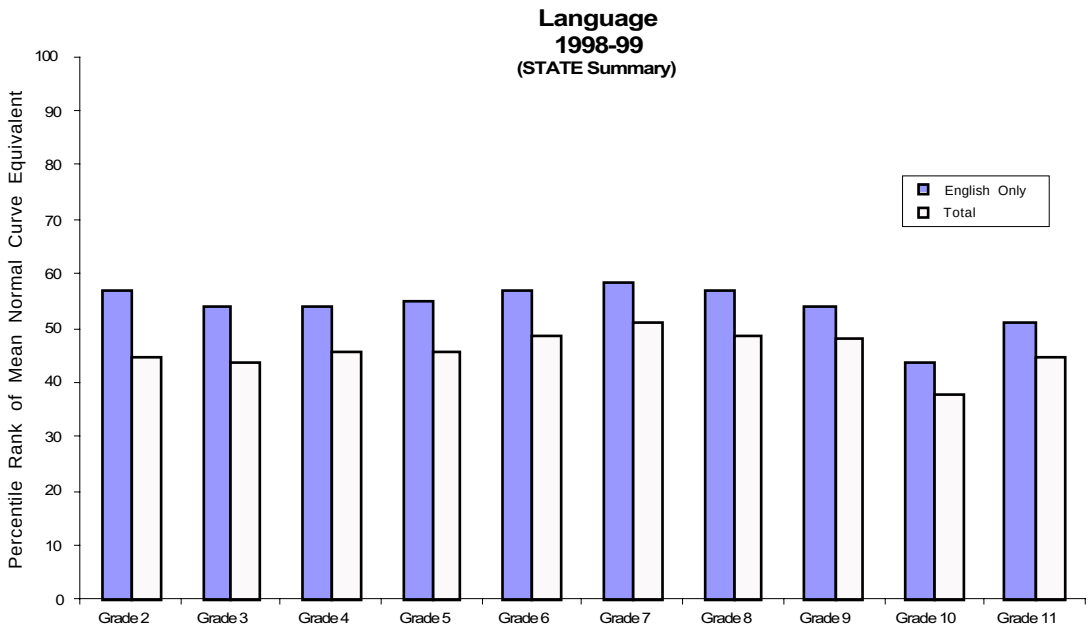

Figure A9. SAT-9 Language – All students versus English proficient students.

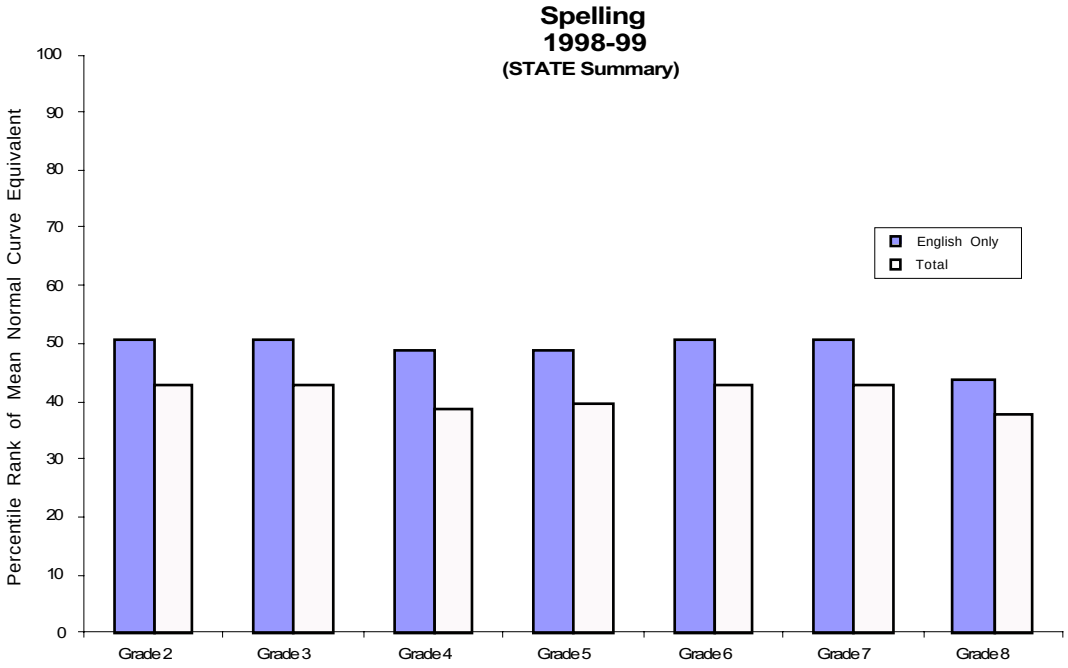

Figure A10. SAT-9 Spelling – All students versus English proficient students.

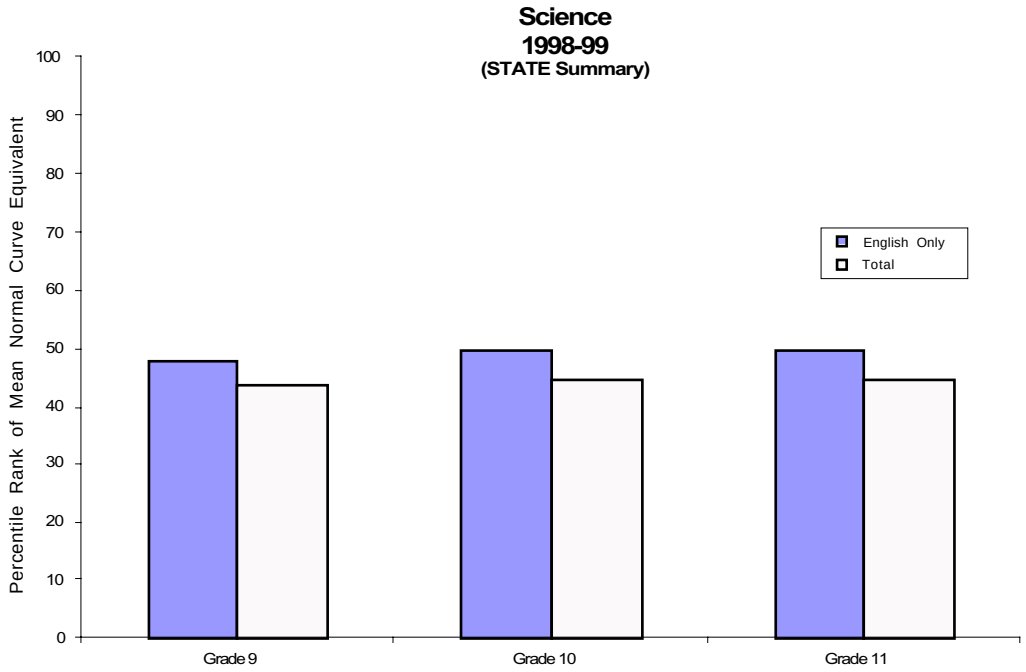

Figure A11. SAT-9 Science – All students versus English proficient students.

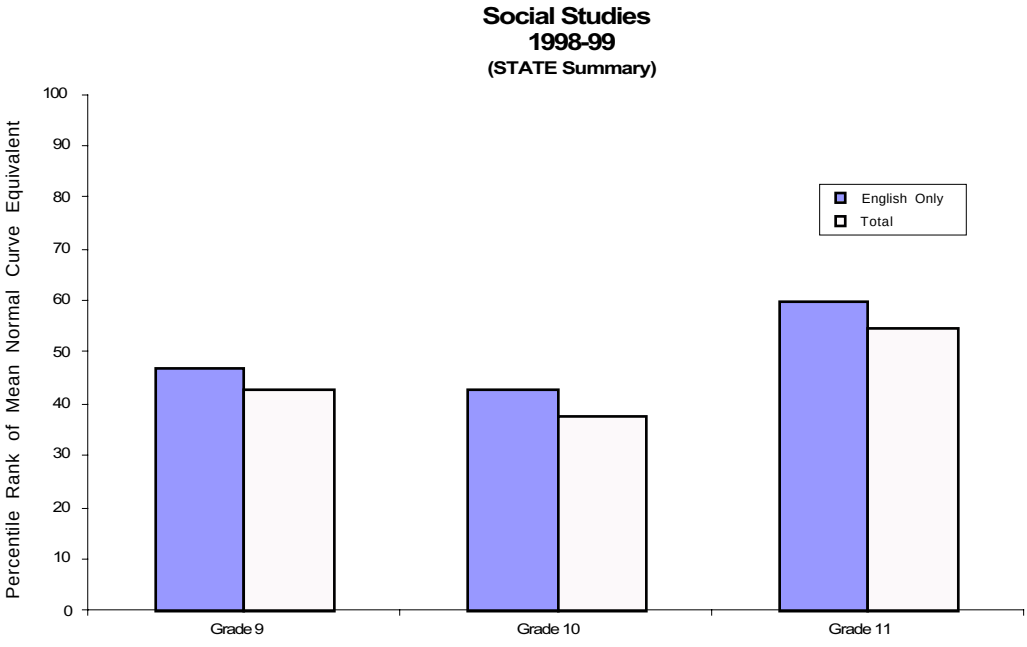

Figure A12. SAT-9 Social Studies – All students versus English proficient students.

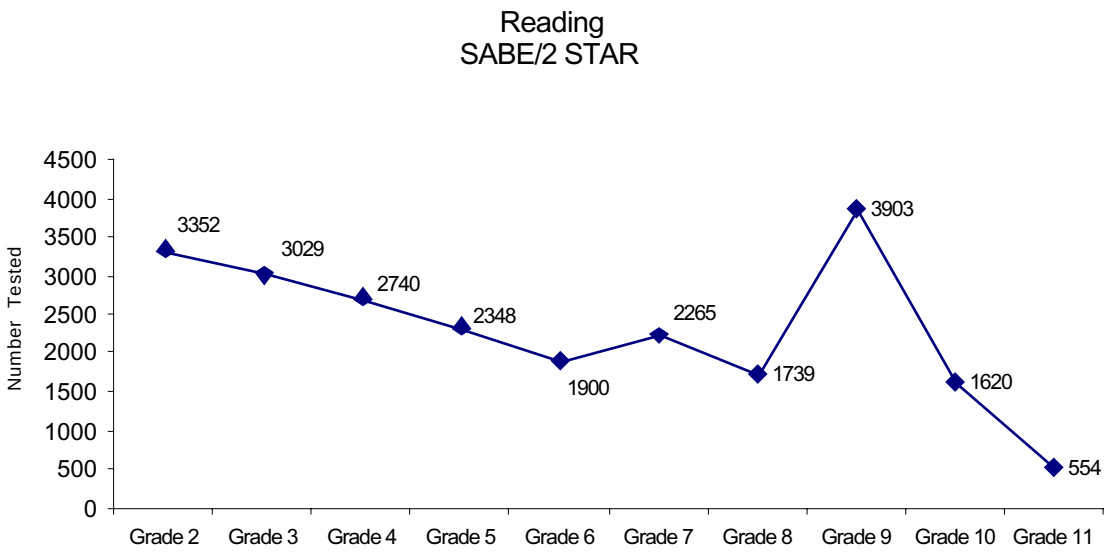

Figure A13. SABE / 2 Reading test takers.

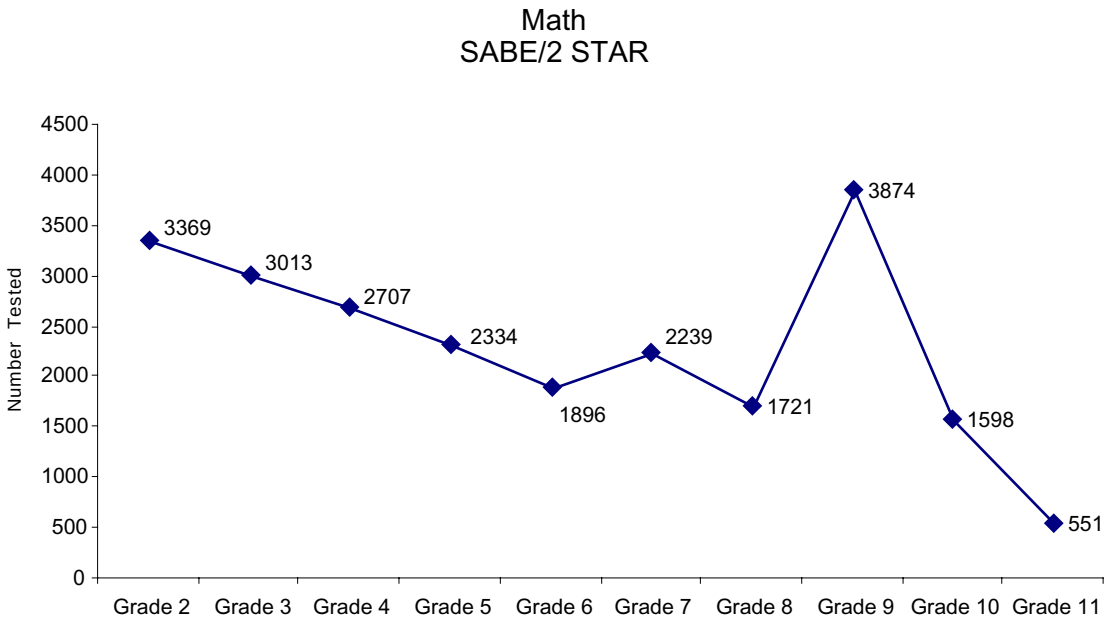

Figure A14. SABE / 2 Math test takers.

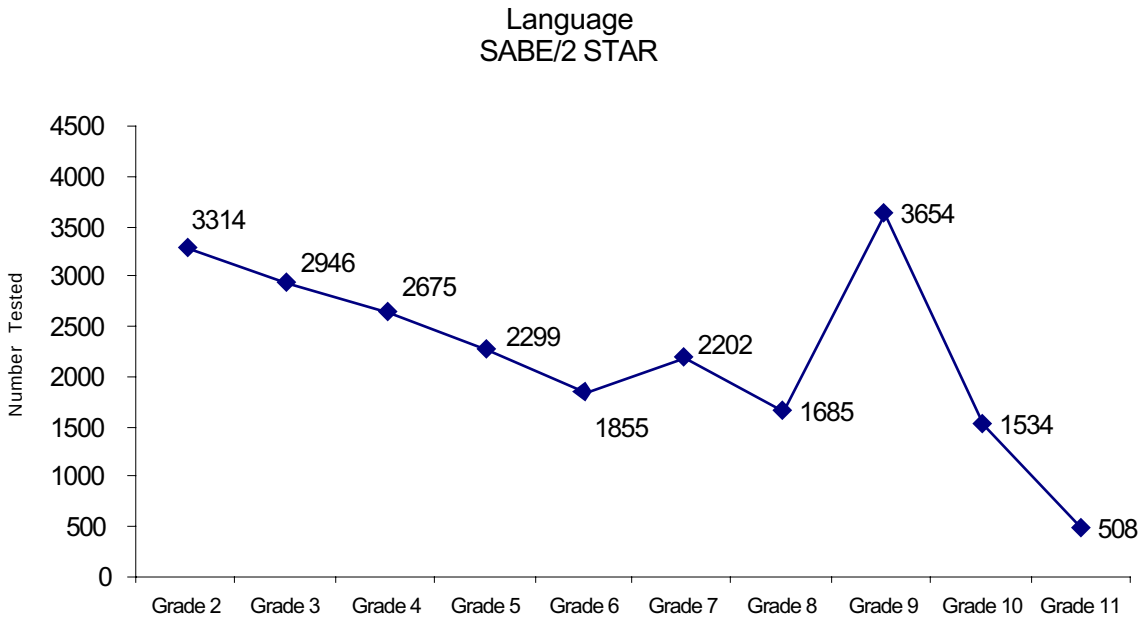

Figure A15. SABE/2 Language test takers.

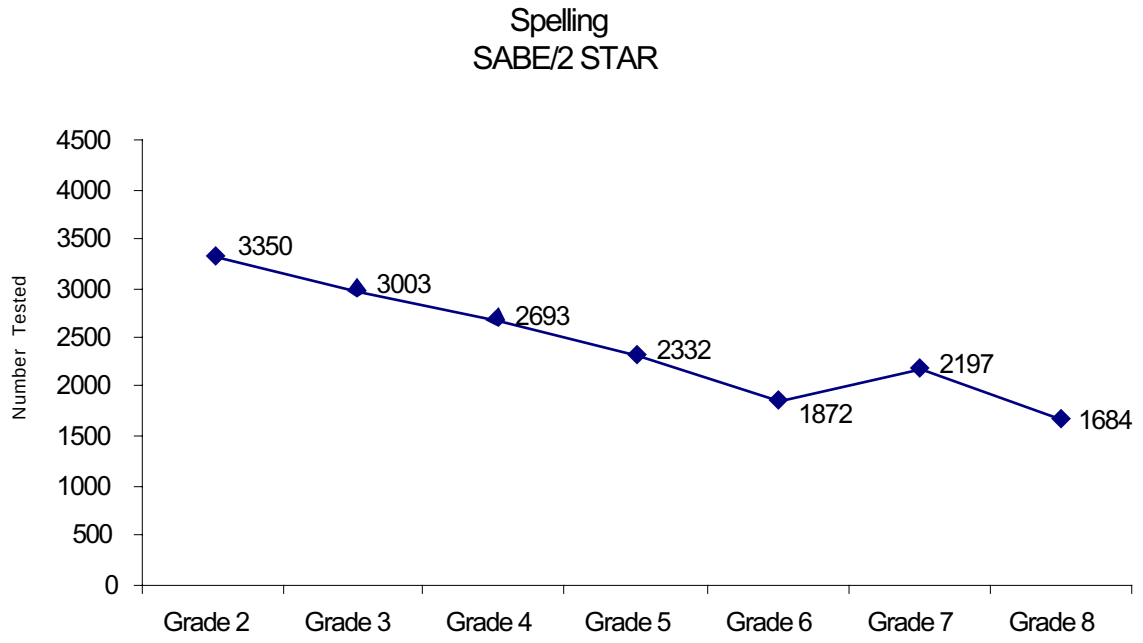

Figure A16. SABE/2 Spelling test takers.

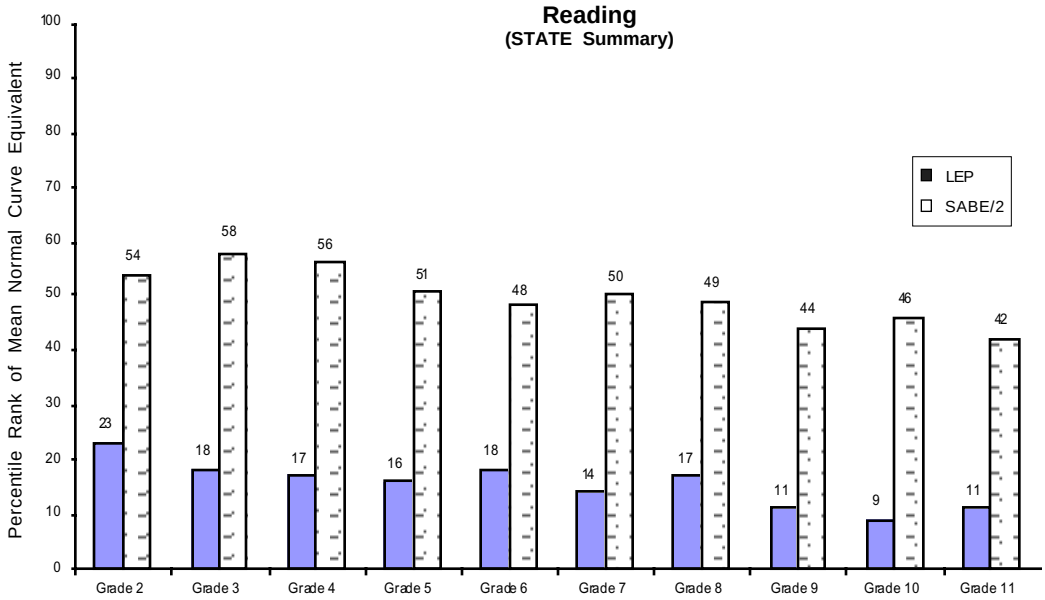

Figure A17. SAT-9 LEP Reading scores versus SABE/2 Reading scores.

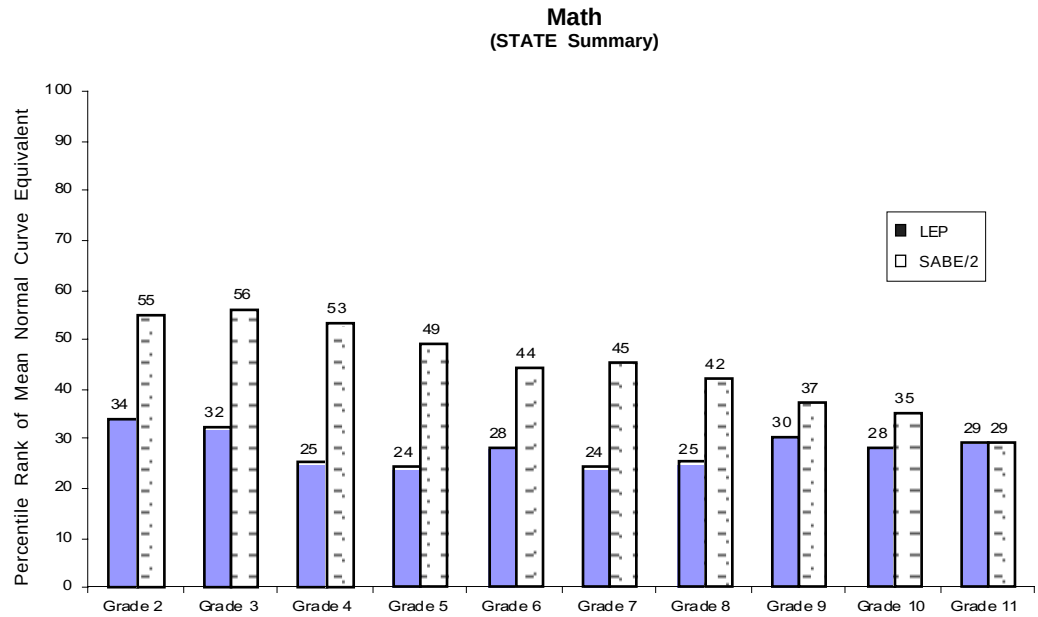

Figure A18. SAT-9 LEP Math scores versus SABE/2 Math scores.

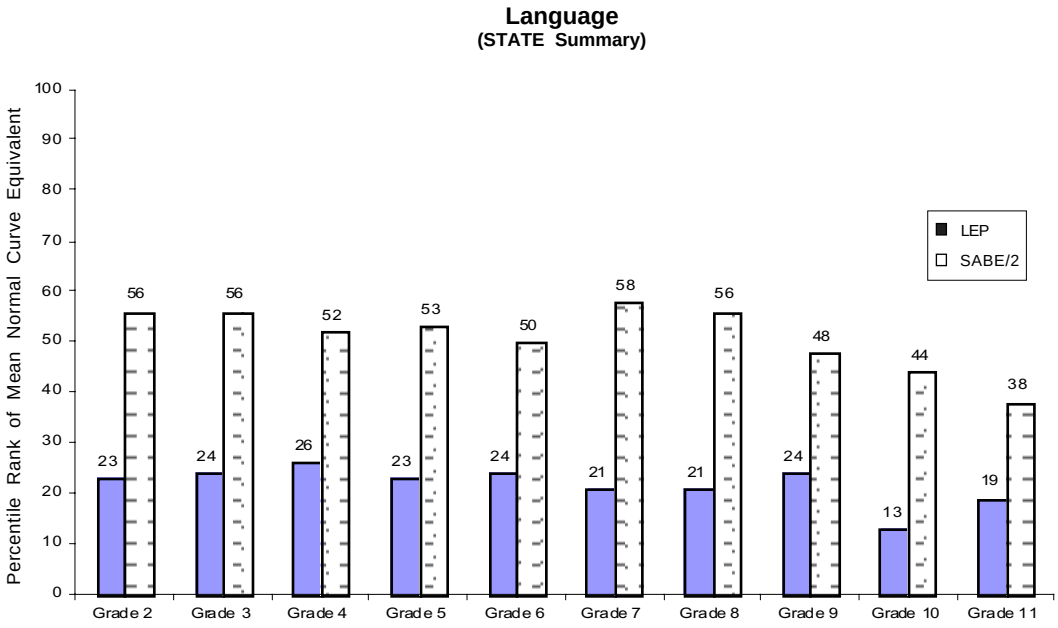

Figure A19. SAT-9 Language scores versus SABE/2 Language scores.

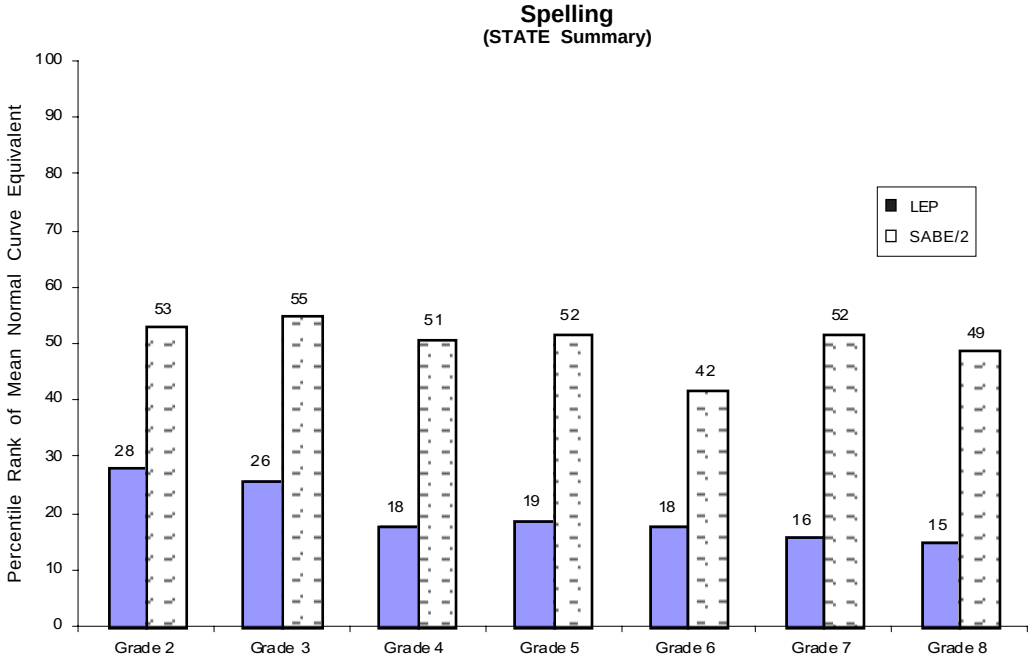

Figure A20. SAT-9 Spelling scores versus SABE/2 Spelling scores.

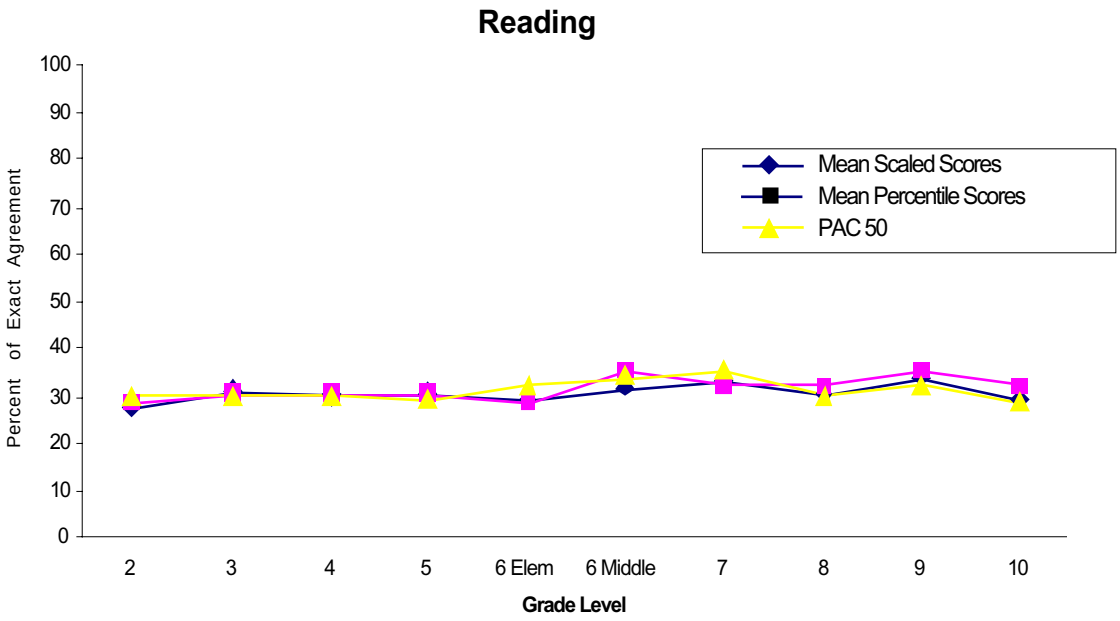

Figure A21. SAT-9 Reading quintile agreements.

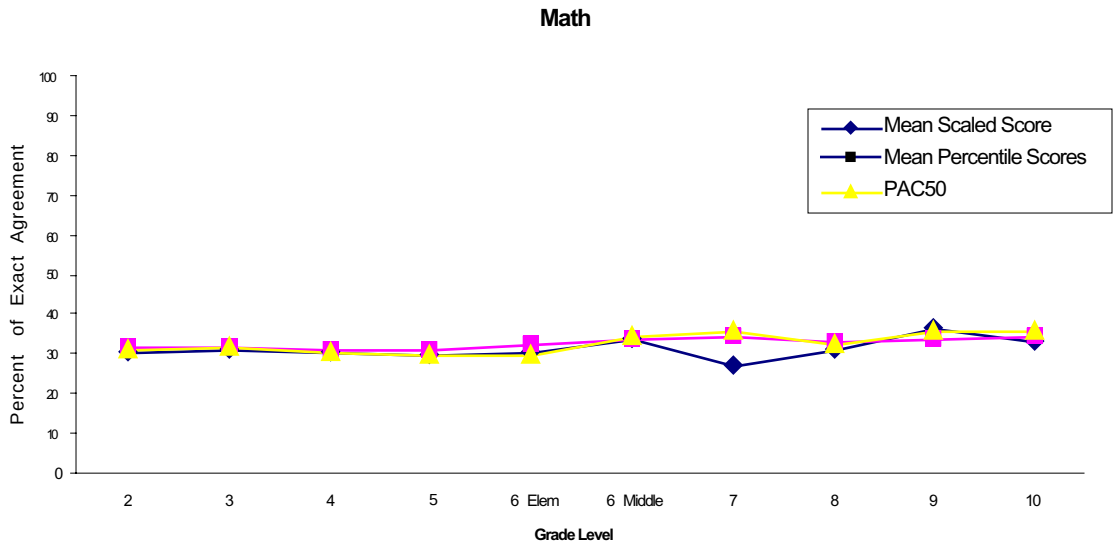

Figure A22. SAT-9 Math quintile agreements.

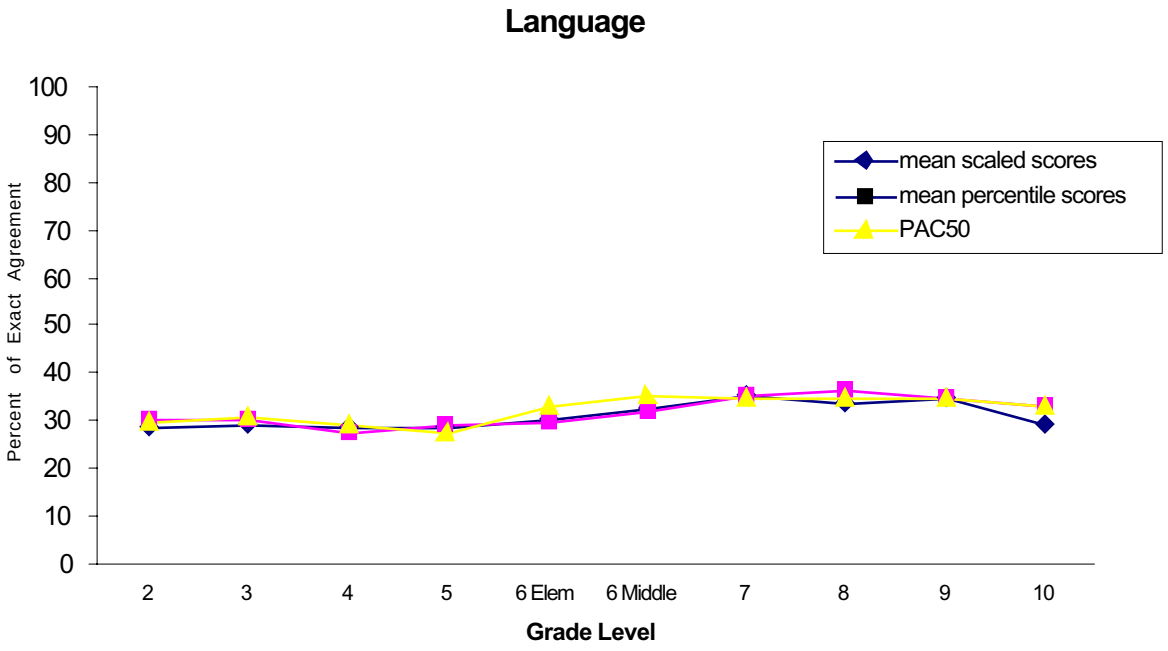

Figure A23. SAT-9 Language quintile agreements.

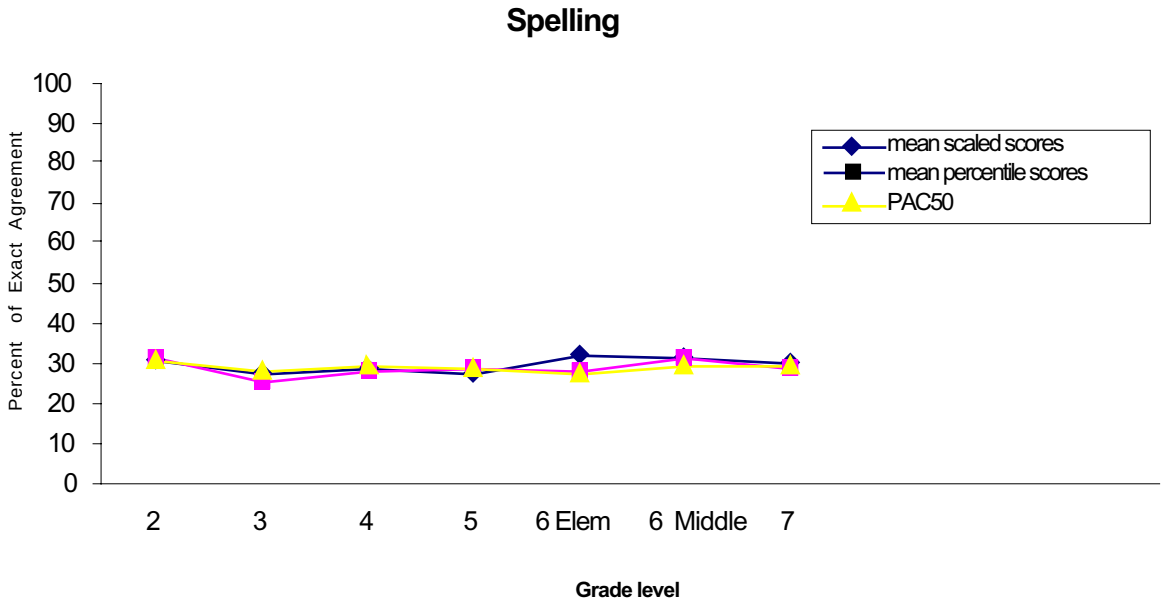

Figure A24. SAT-9 Spelling quintile agreements.

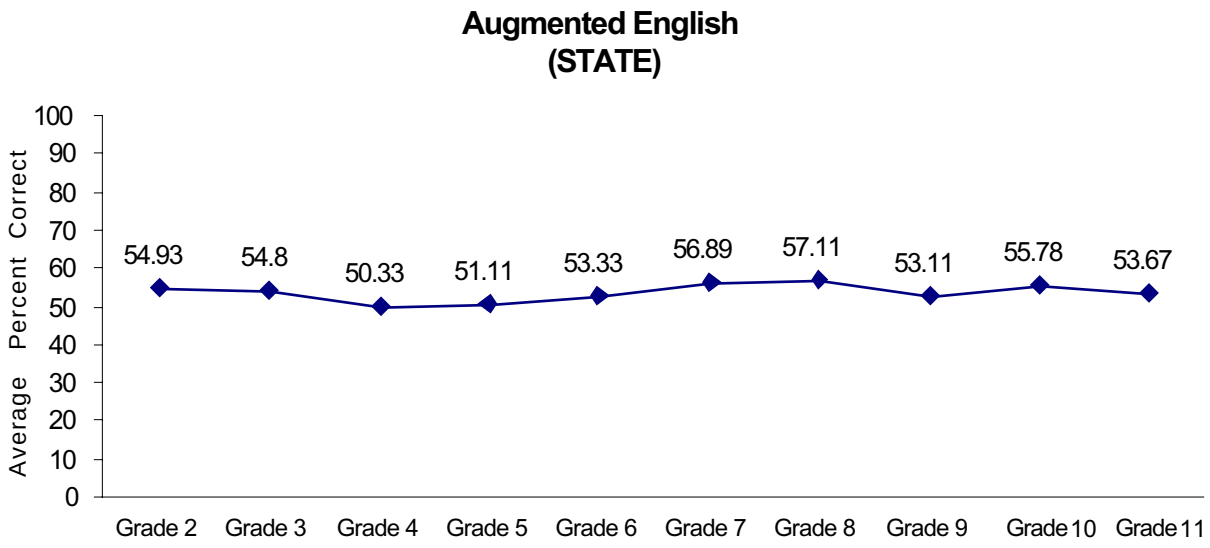

Figure A25. SAT-9 Augmented English.

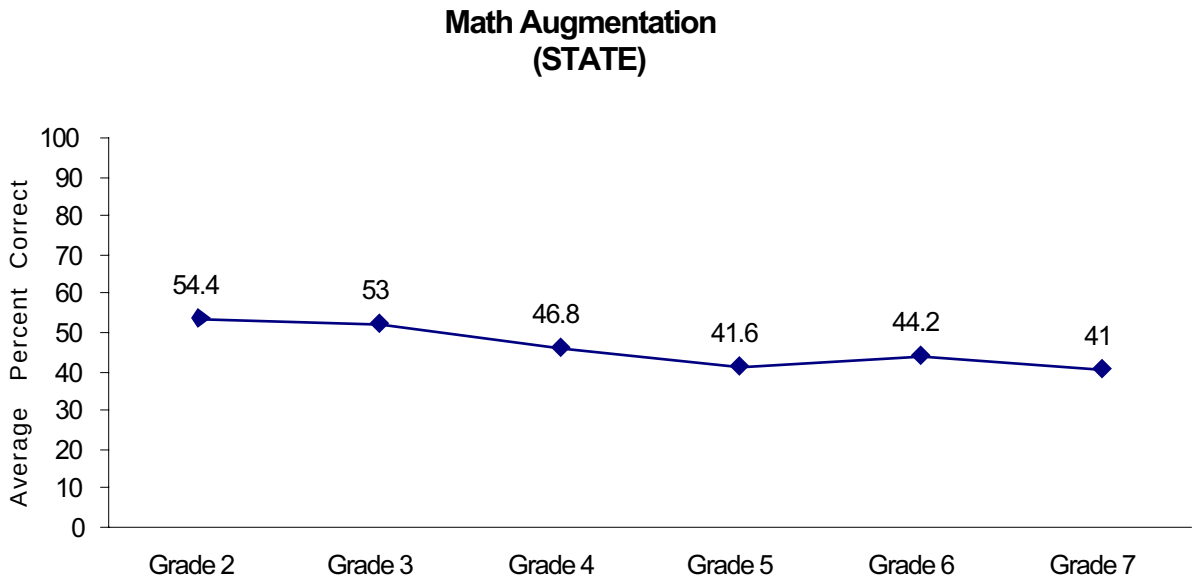

Figure A26. SAT-9 Augmented Math.

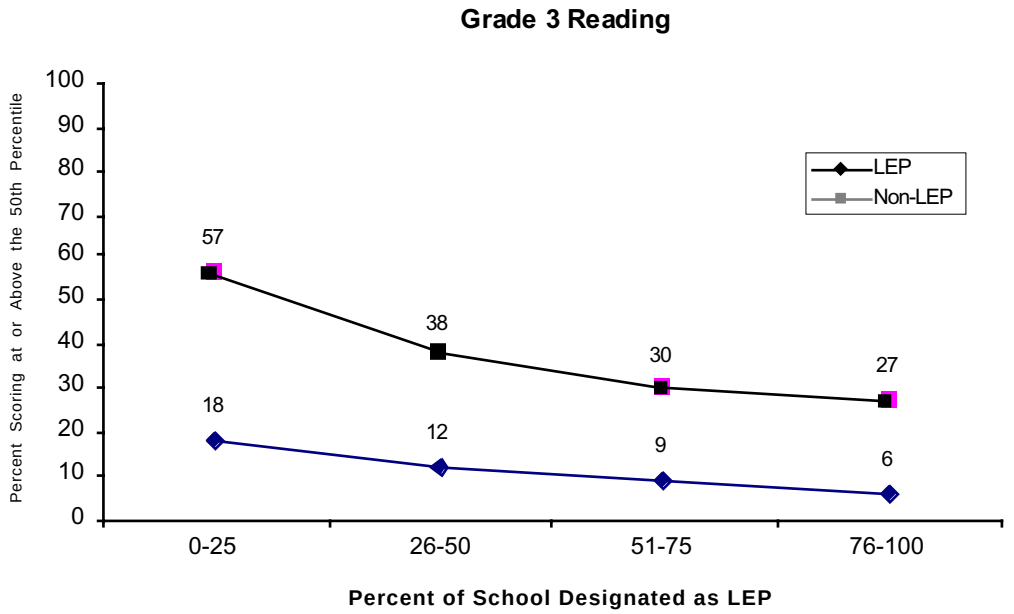

Figure A27. SAT-9 Grade 3 Reading LEP vs. non-LEP.

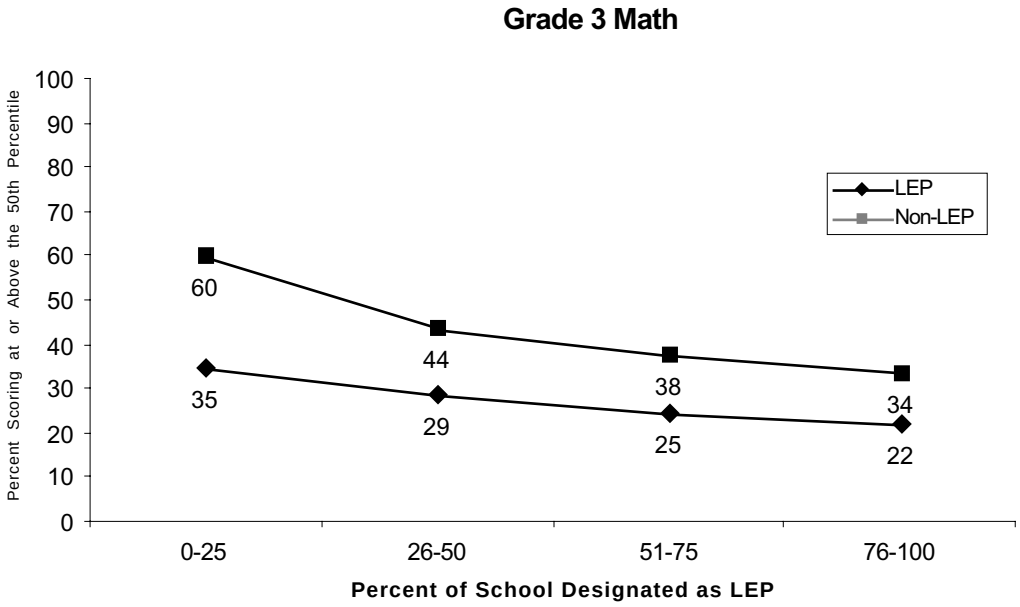

Figure A28. SAT-9 Grade 3 Math LEP vs. non-LEP.

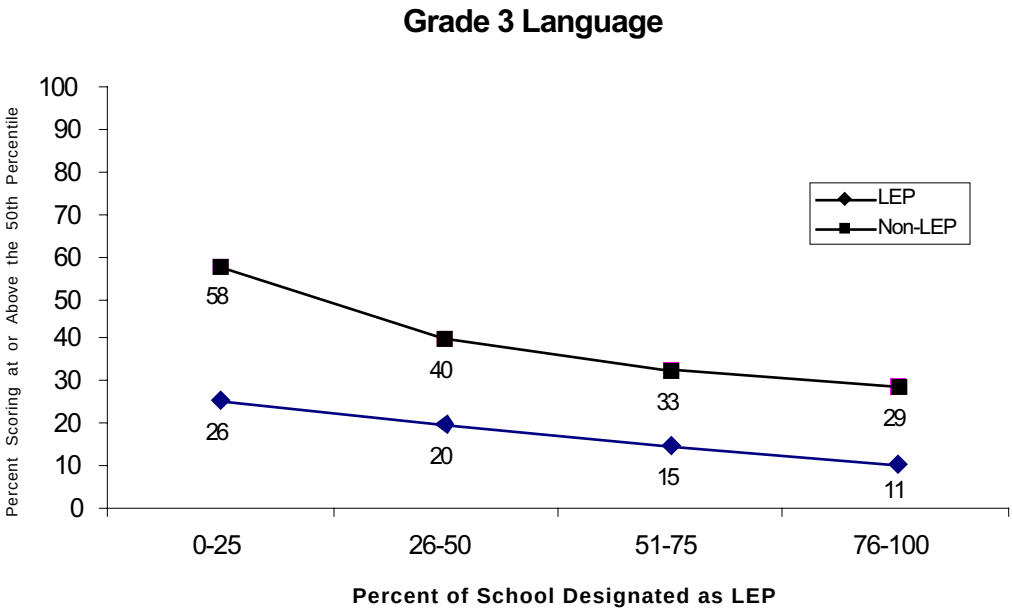

Figure A29. SAT-9 Grade 3 Language LEP vs. non-LEP.

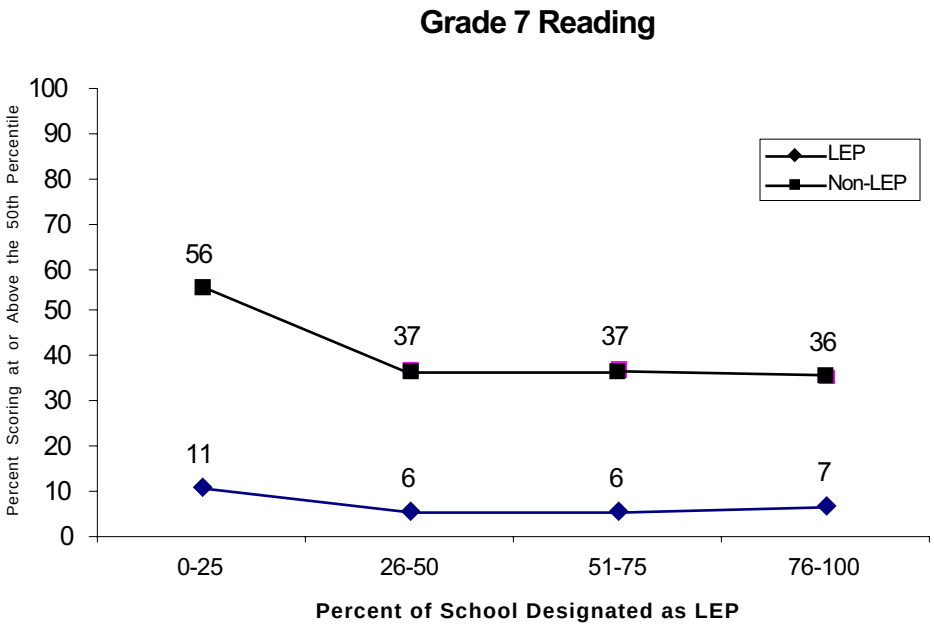

Figure A30. SAT-9 Grade 7 Reading LEP vs. non-LEP.

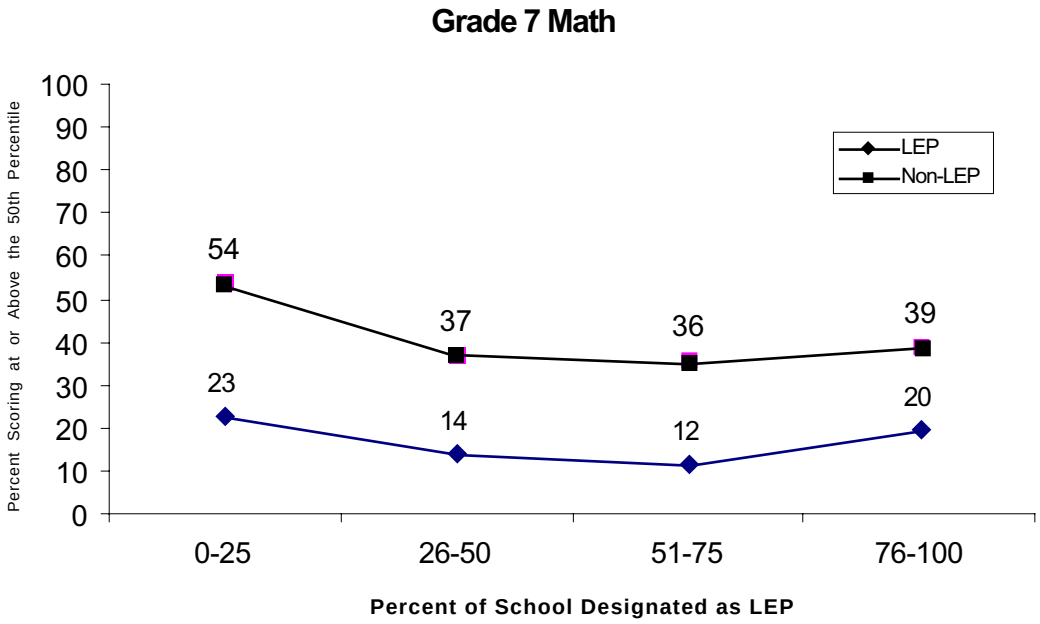

Figure A31. SAT-9 Grade 7 Math LEP vs. non-LEP.

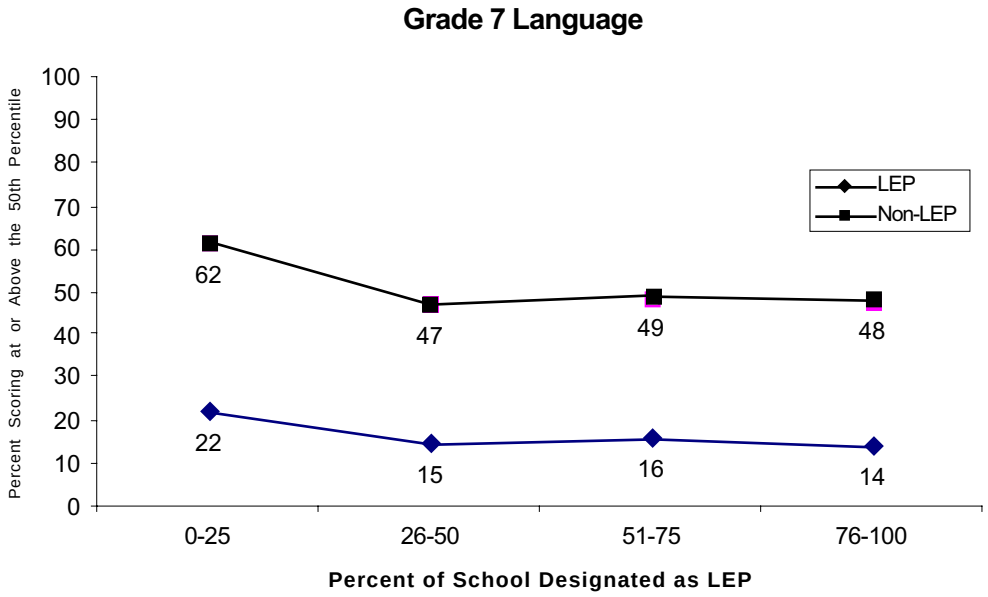

Figure A32. SAT-9 Grade 7 Language LEP vs. non-LEP.

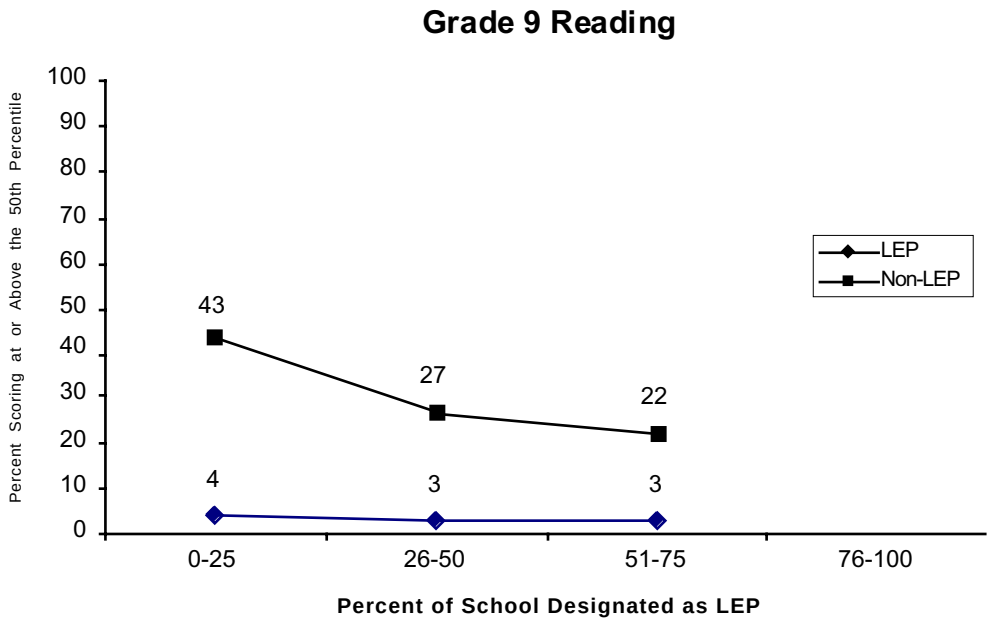

Figure A33. SAT-9 Grade 9 Reading LEP vs. non-LEP.

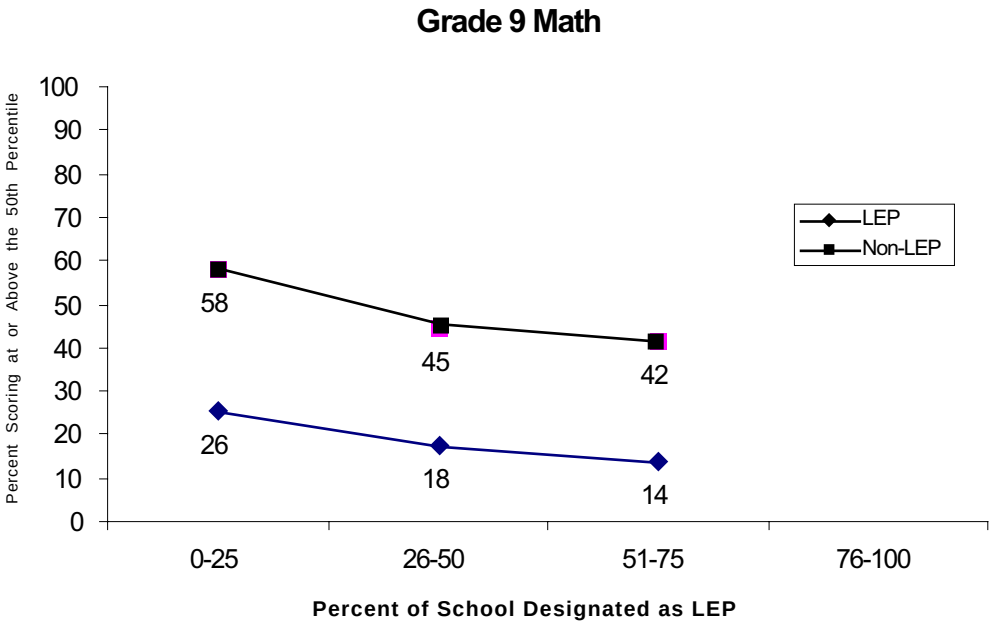

Figure A34. SAT-9 Grade 9 Math LEP vs. non-LEP.

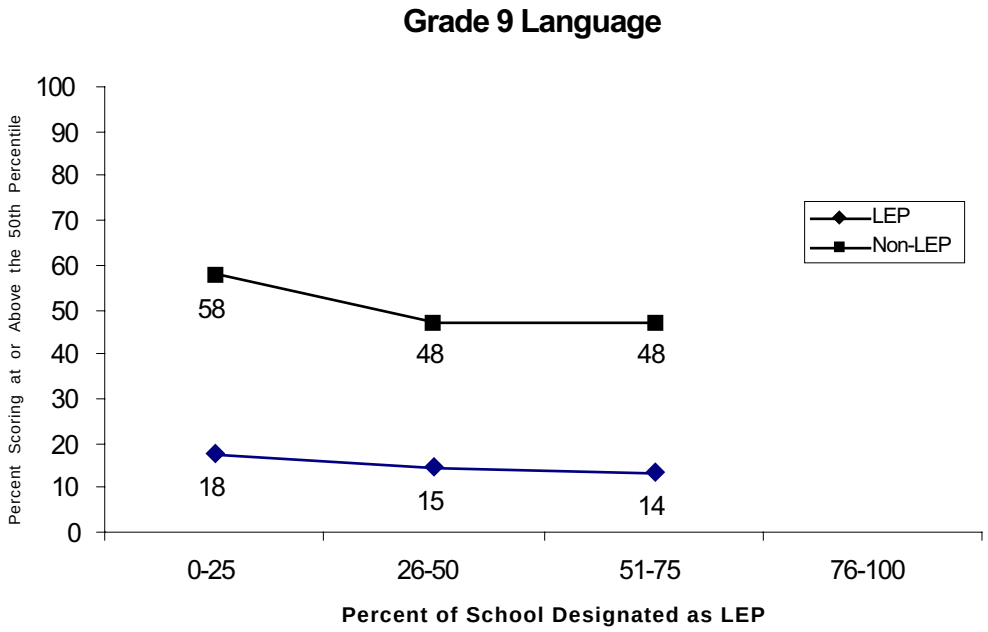

Figure A35. SAT-9 Grade 9 Language LEP vs. non-LEP.

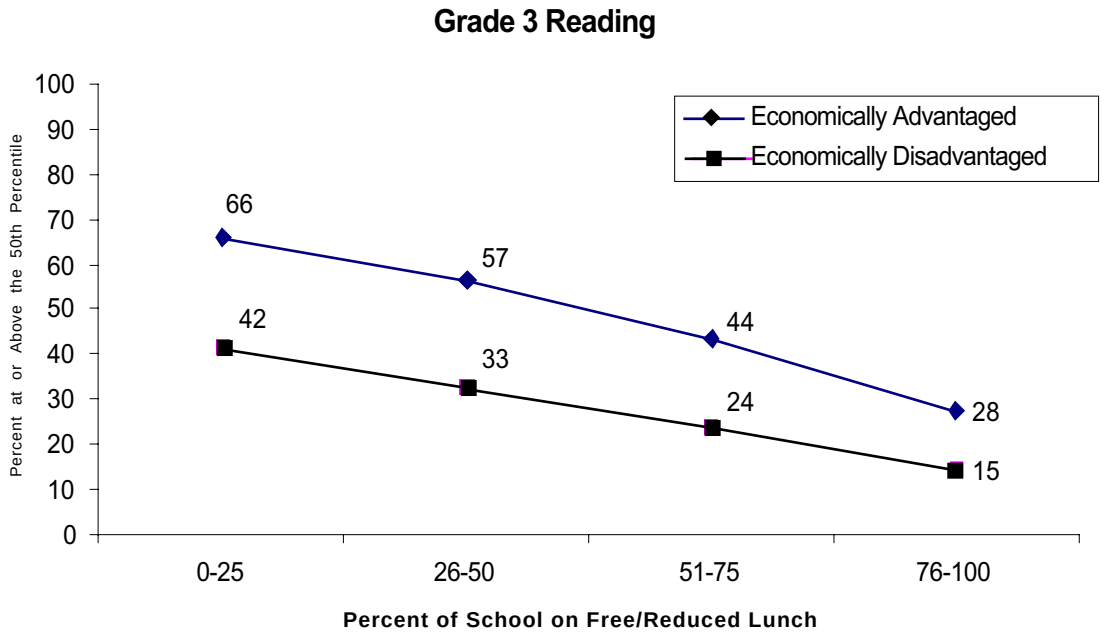

Figure A36. SAT-9 Grade 3 Reading economically disadvantaged vs. economically advantaged.

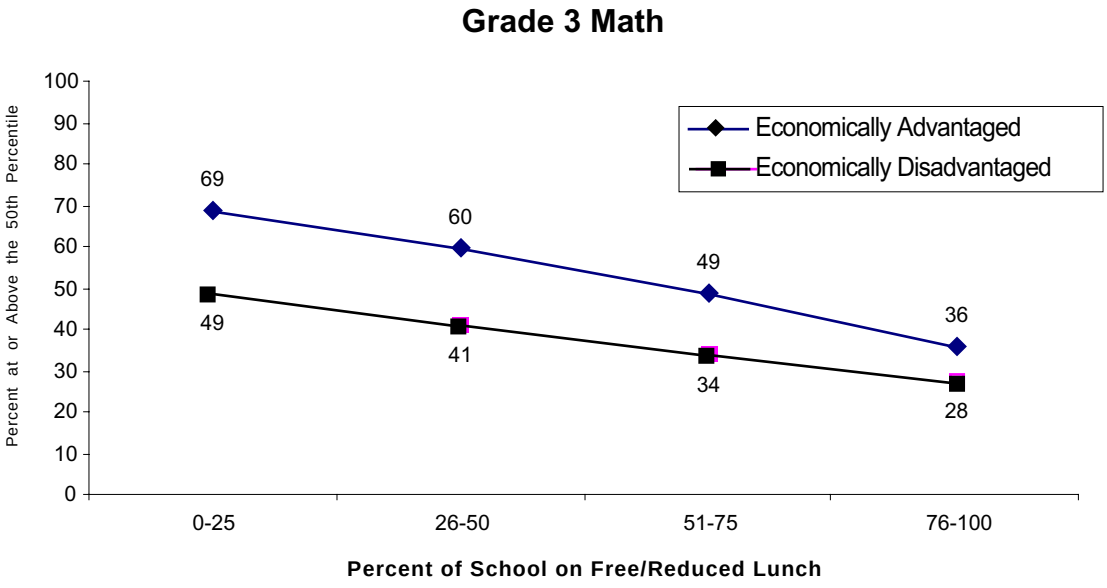

Figure A37. SAT-9 Grade 3 Math economically disadvantaged vs. economically advantaged.

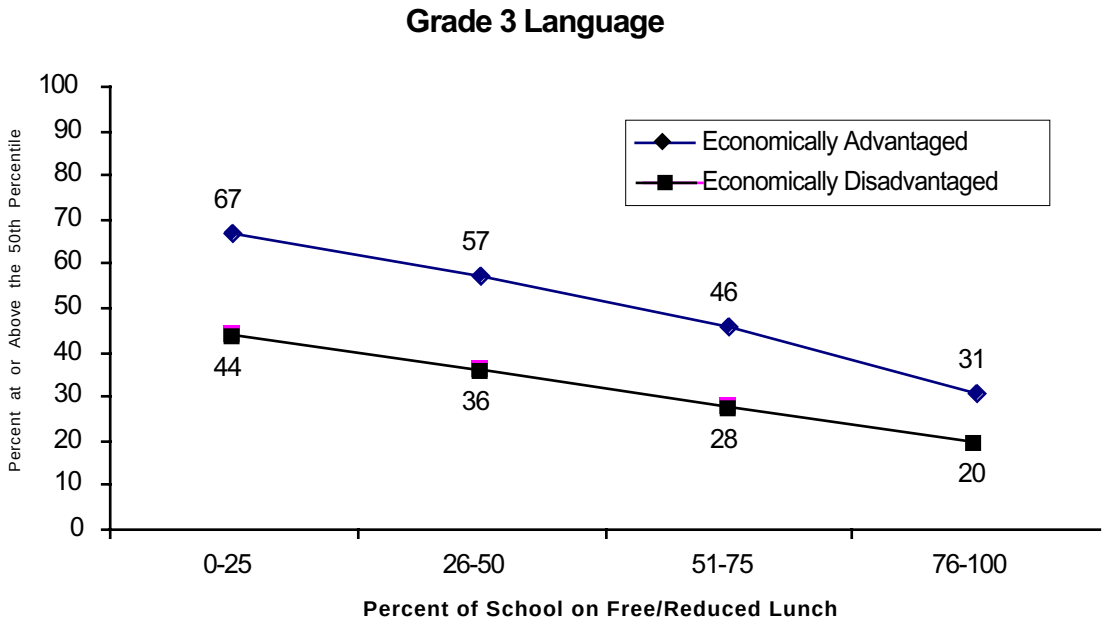

Figure A38. SAT-9 Grade 3 Language economically disadvantaged vs. economically advantaged.

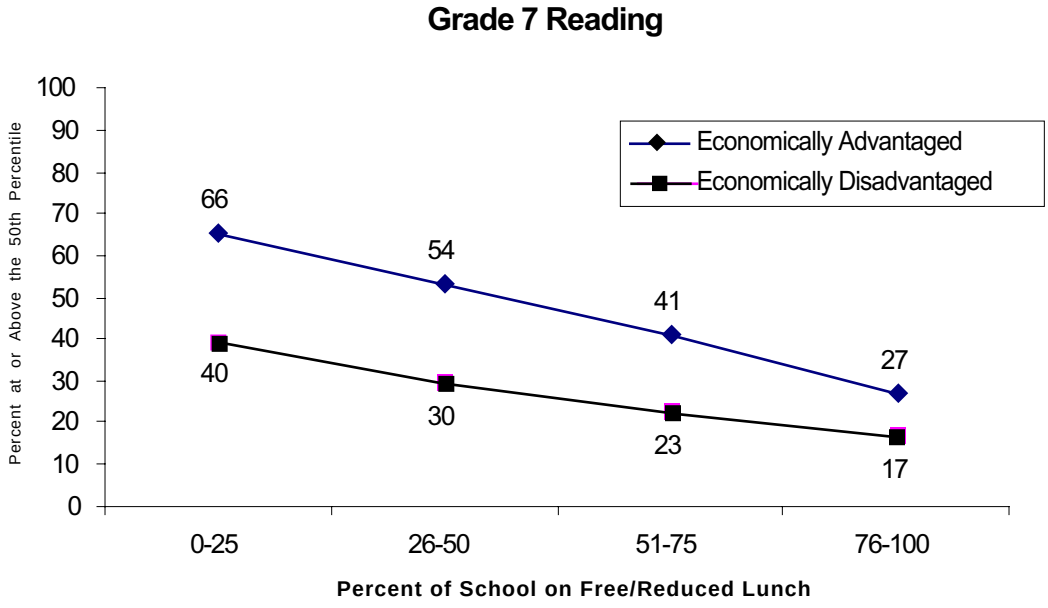

Figure A39. SAT-9 Grade 7 Reading economically disadvantaged vs. economically advantaged.

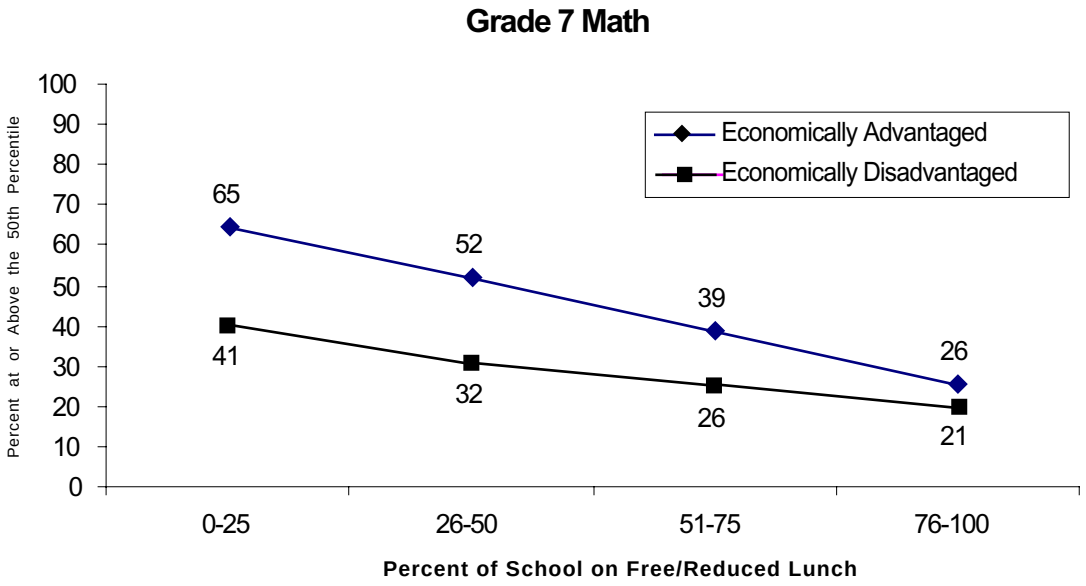

Figure A40. SAT-9 Grade 7 Math economically disadvantaged vs. economically advantaged.

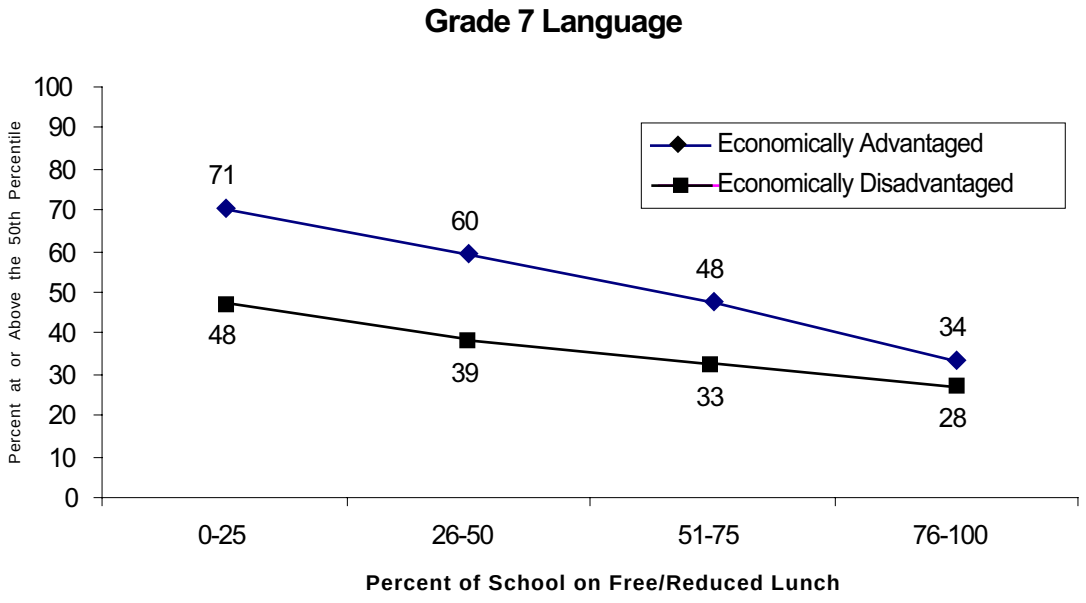

Figure A41. SAT-9 Grade 7 Language economically disadvantaged vs. economically advantaged.

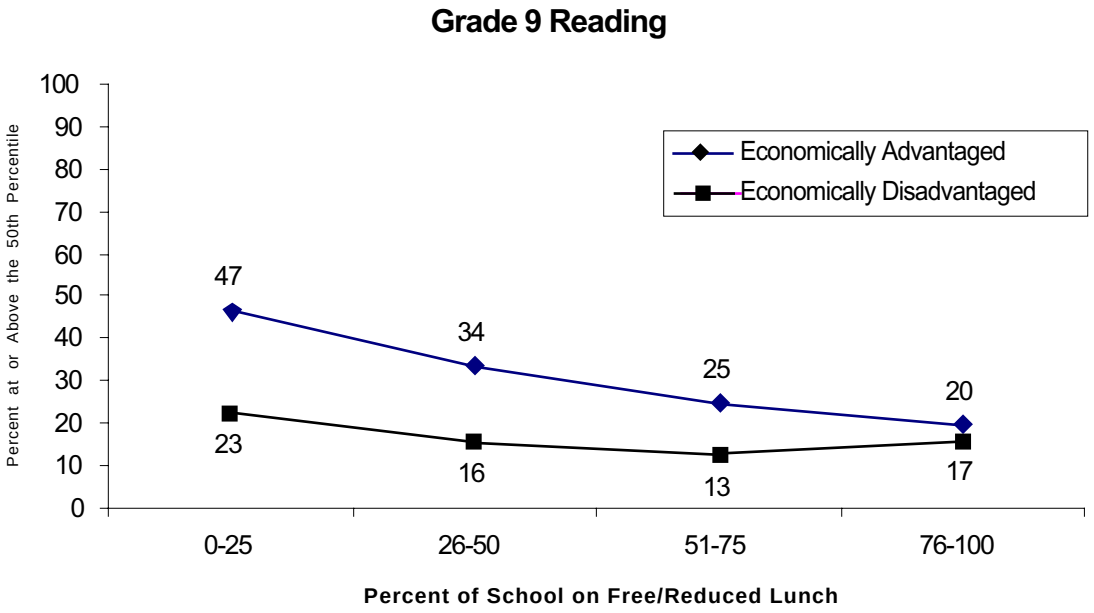

Figure A42. SAT-9 Grade 9 Reading economically disadvantaged vs. economically advantaged.

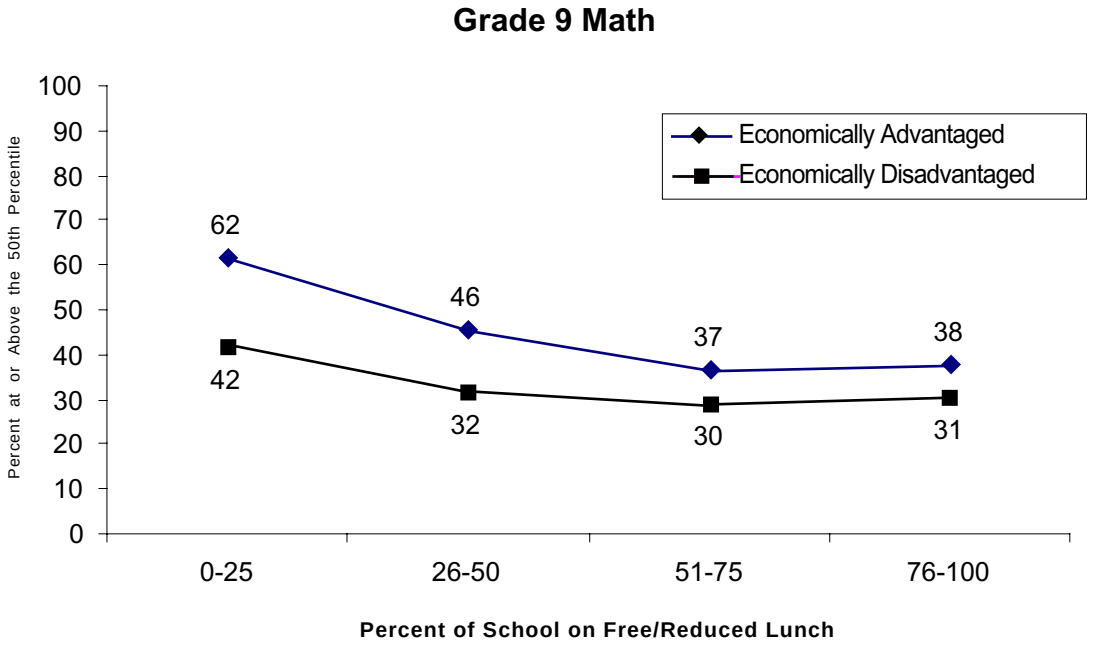

Figure A43. SAT-9 Grade 9 Math economically disadvantaged vs. economically advantaged.

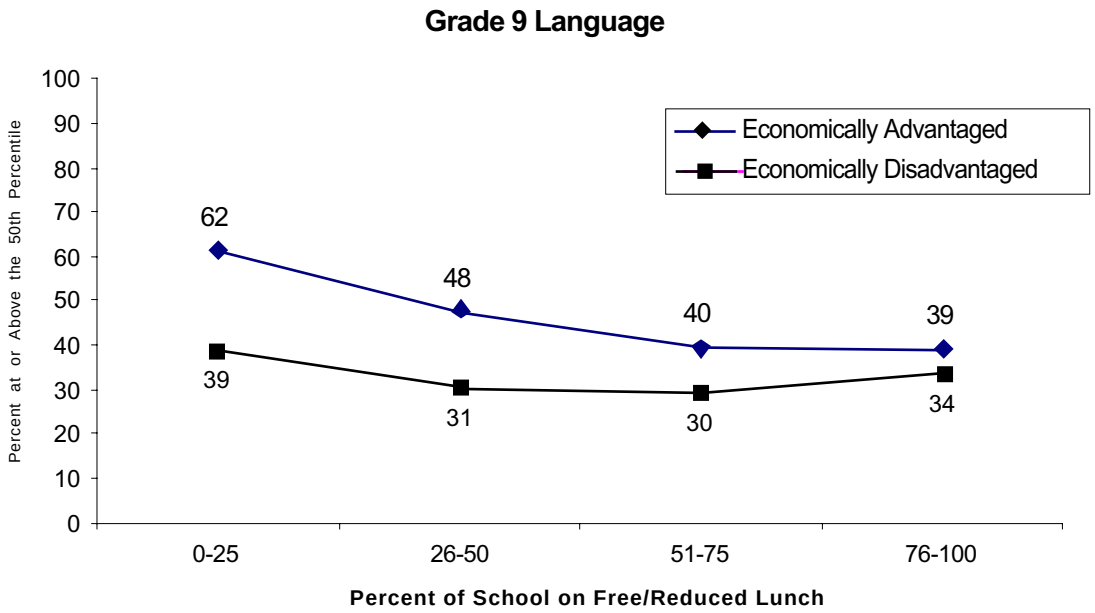

Figure A44. SAT-9 Grade 9 Language economically disadvantaged vs. economically advantaged.

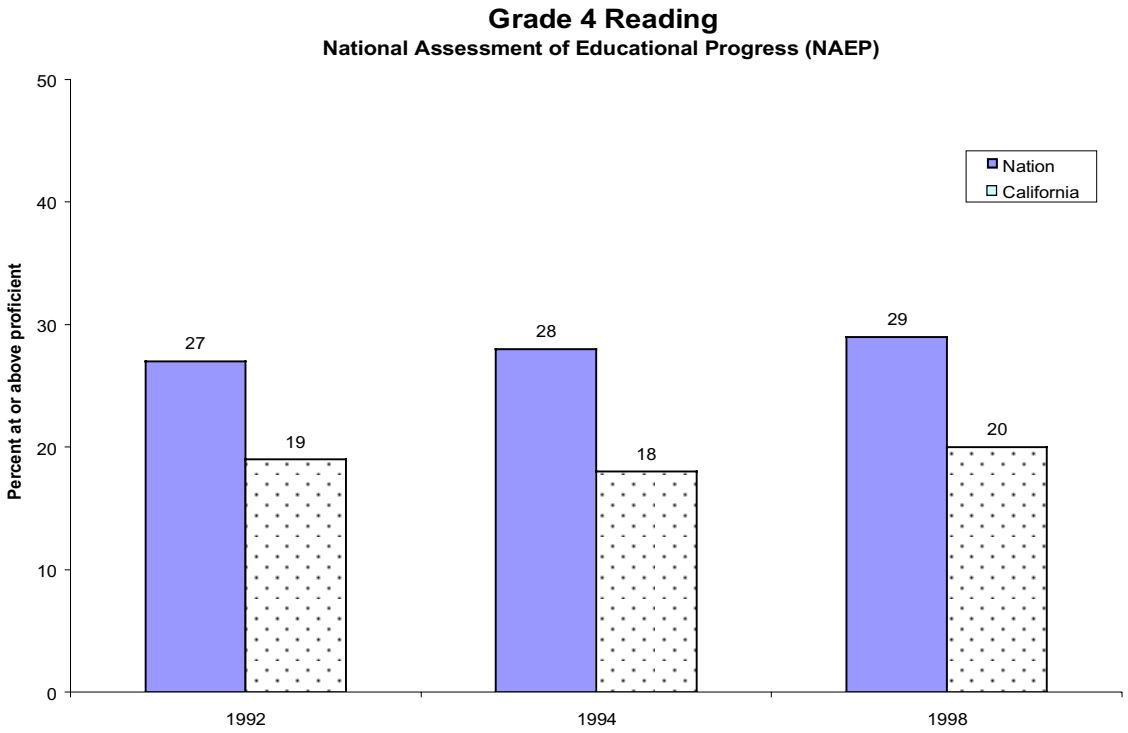

Figure A45. NAEP Grade 4 Reading 1992, 1994, and 1998.

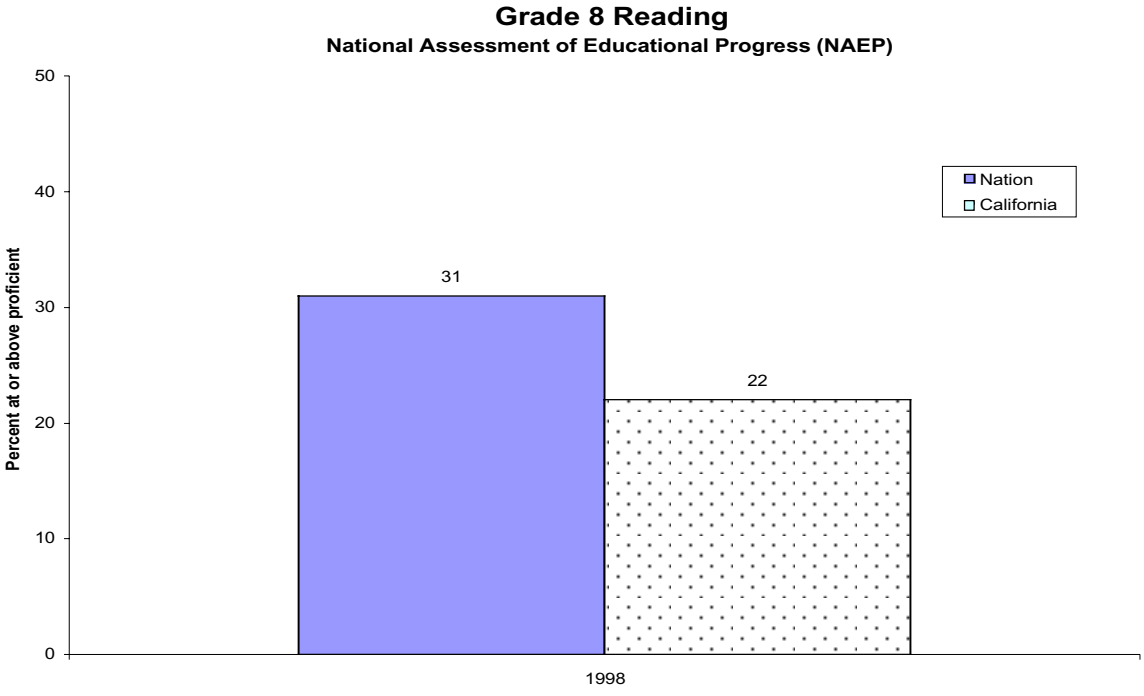

Figure A46. NAEP Grade 8 Reading 1998.

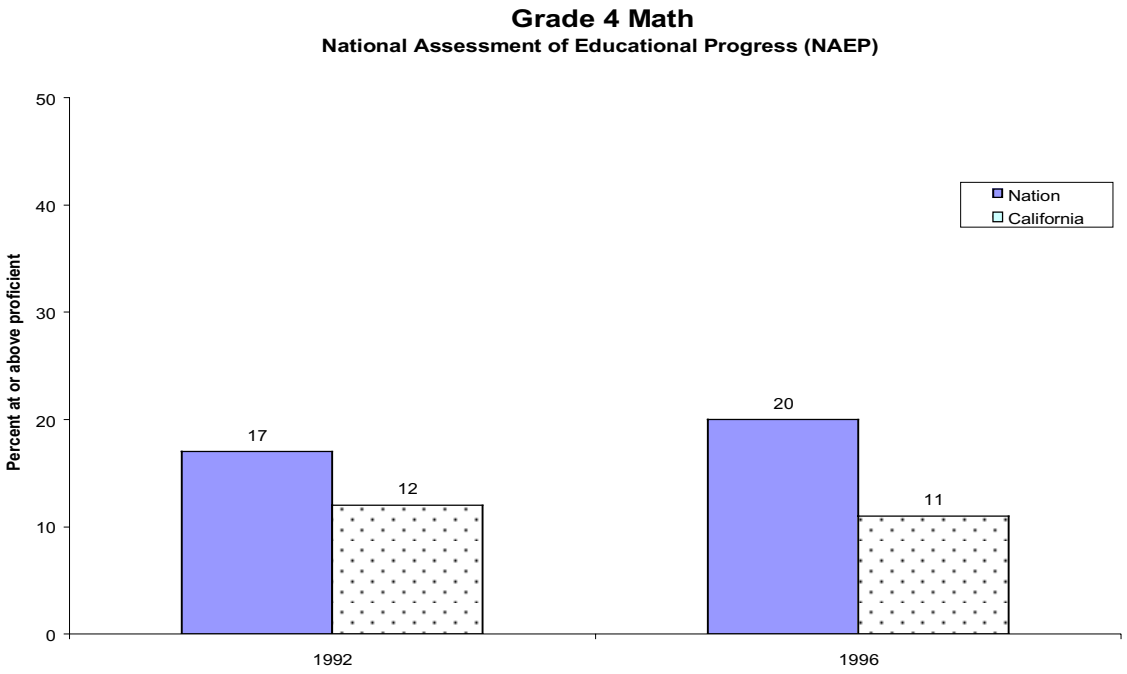

Figure A47. NAEP Grade 4 Math 1992 and 1996.

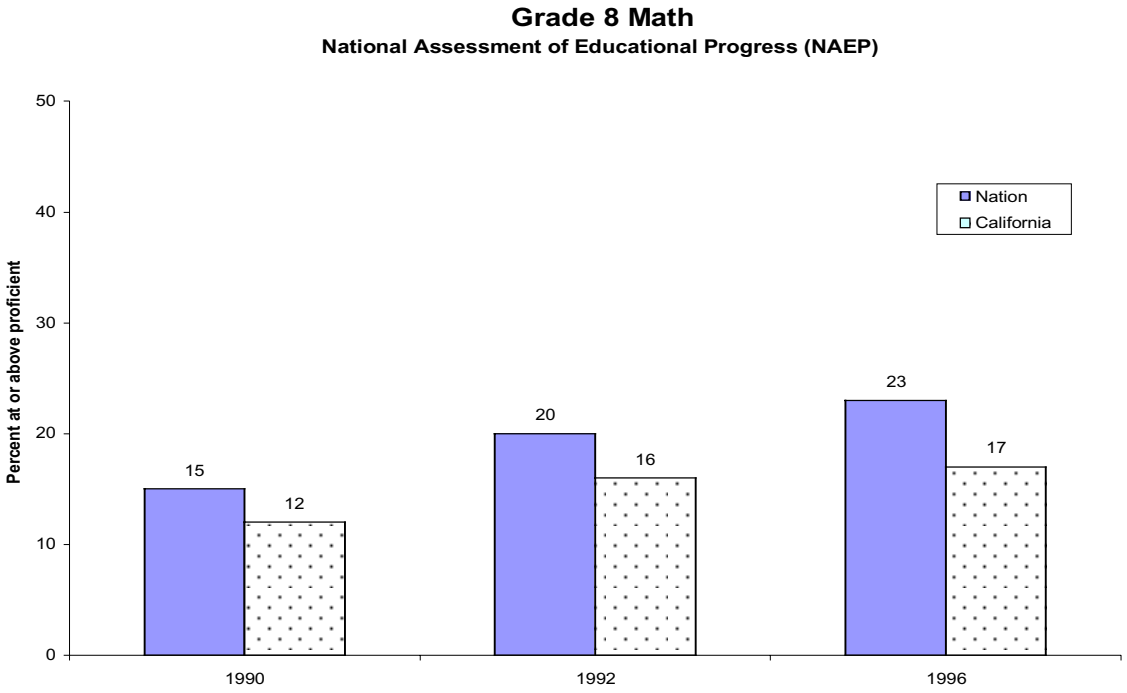

Figure A48. NAEP Grade 8 Math 1990, 1992, and 1996.

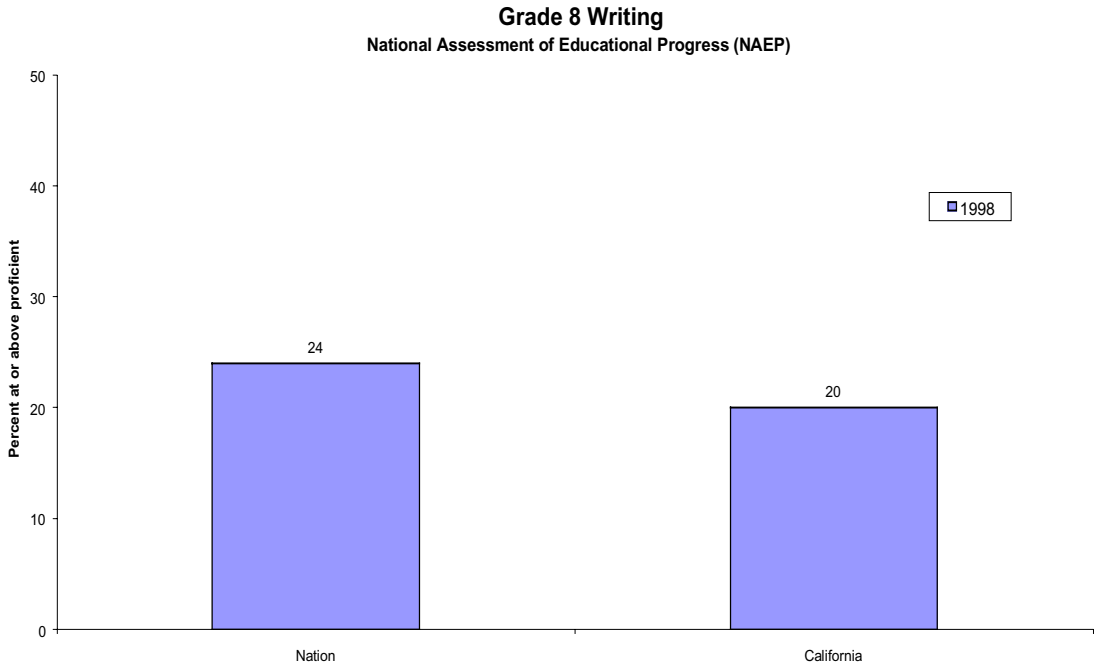

Figure A49. NAEP Grade 8 Writing 1998.

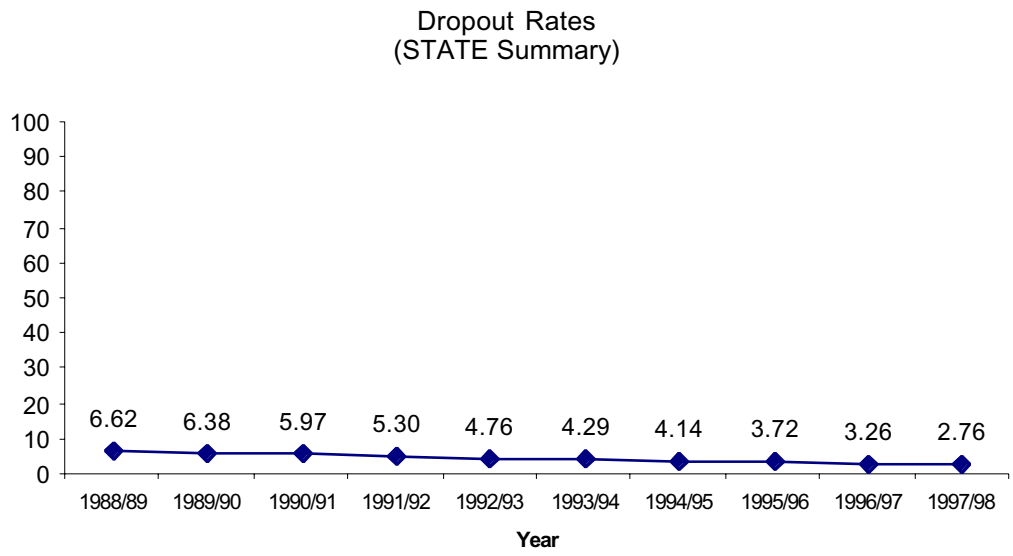

Figure A50. California high school dropout rates 1989-98 .

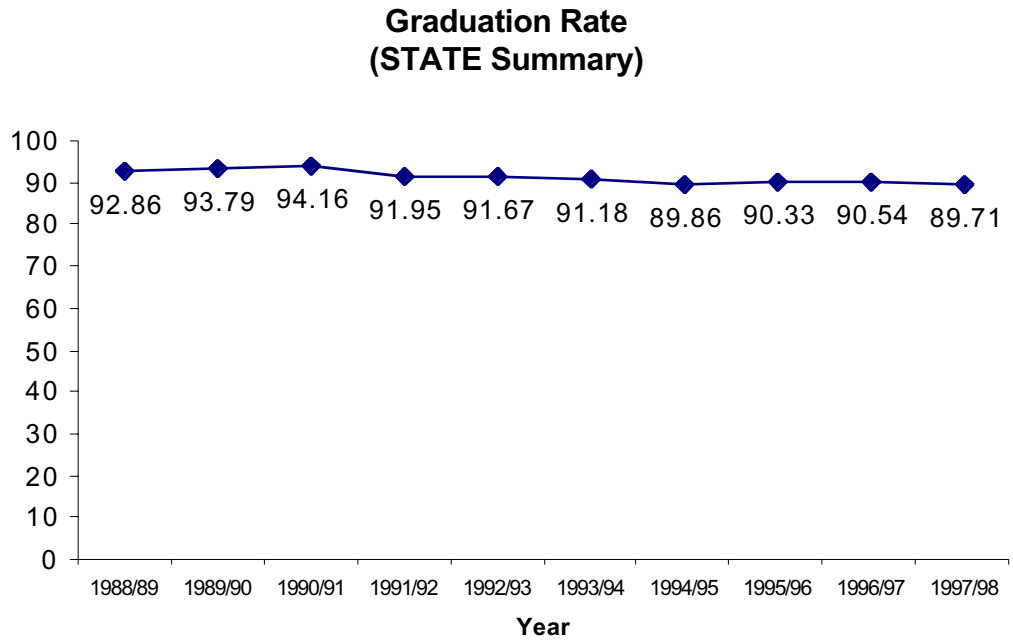

Figure A51. California high school graduation rates 1989-1998.

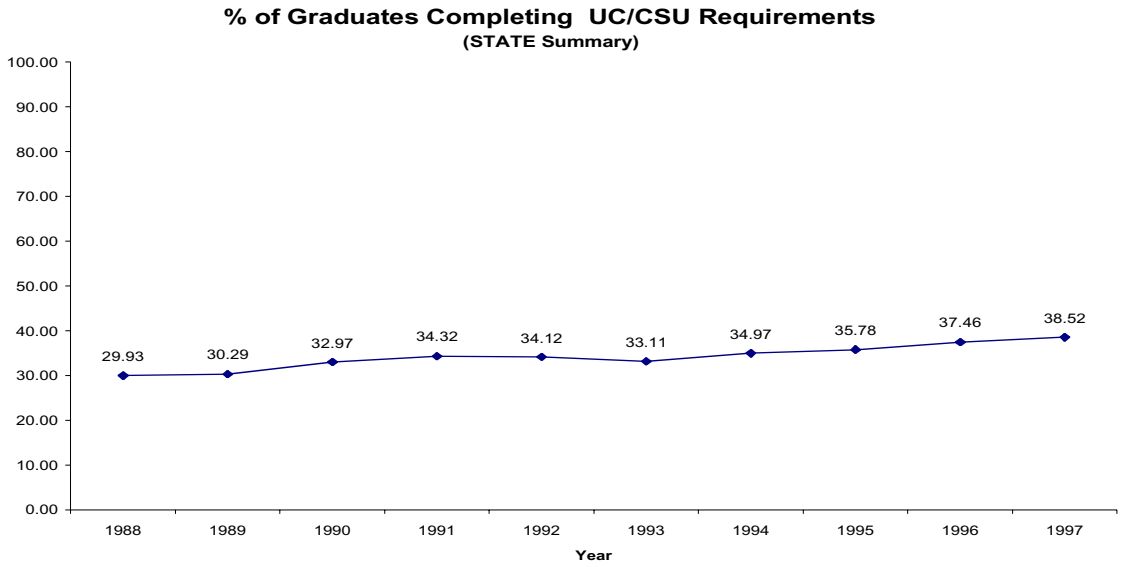

Figure A52. High school graduates meeting UC / CSU course requirements 1998-97.

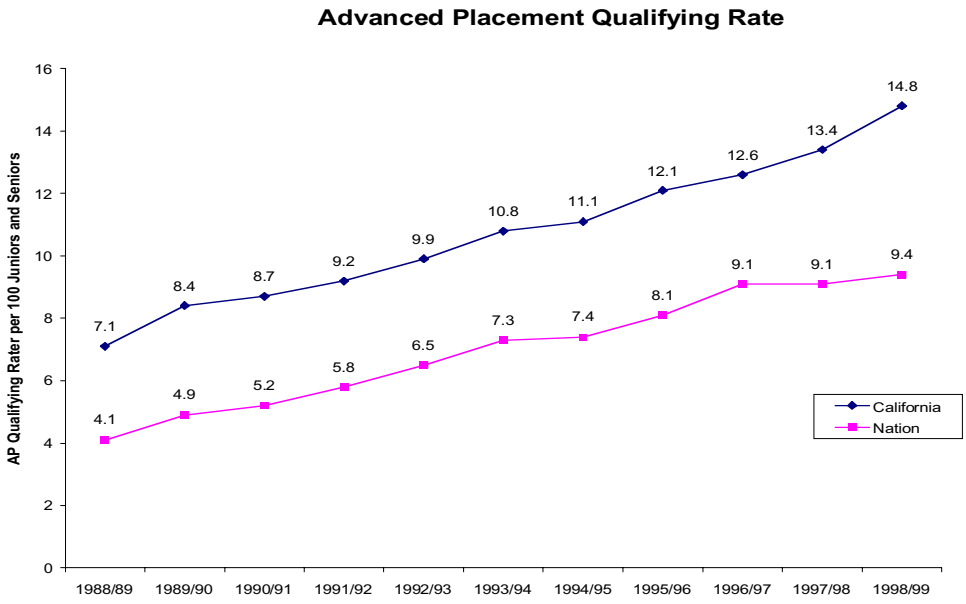

Figure A53. Advanced Placement qualifying rate 1989-1999.

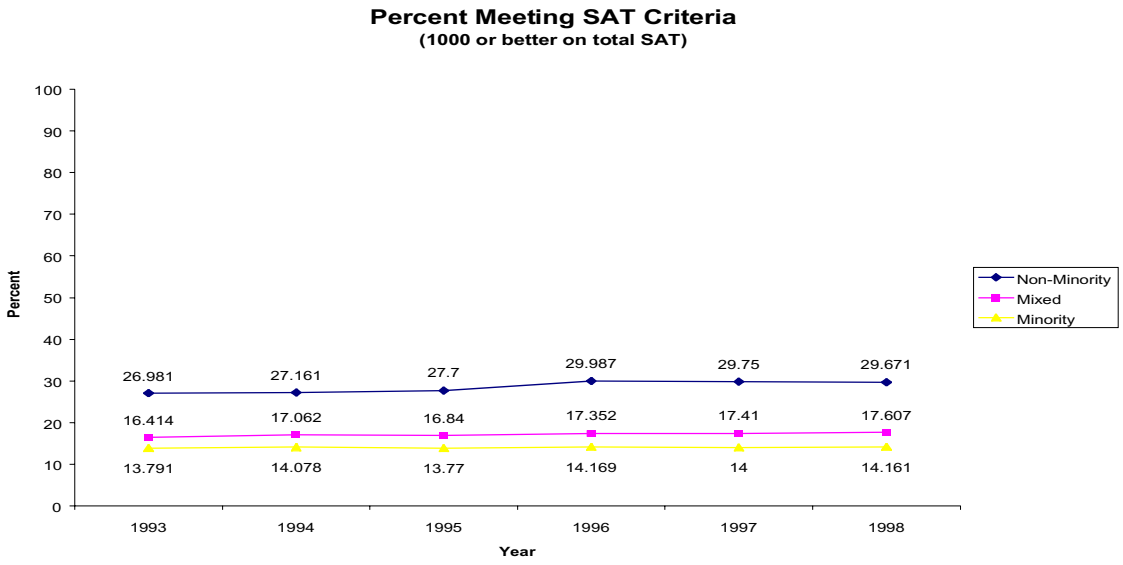

Figure A54. Students meeting SAT criteria in California 1993-1998.

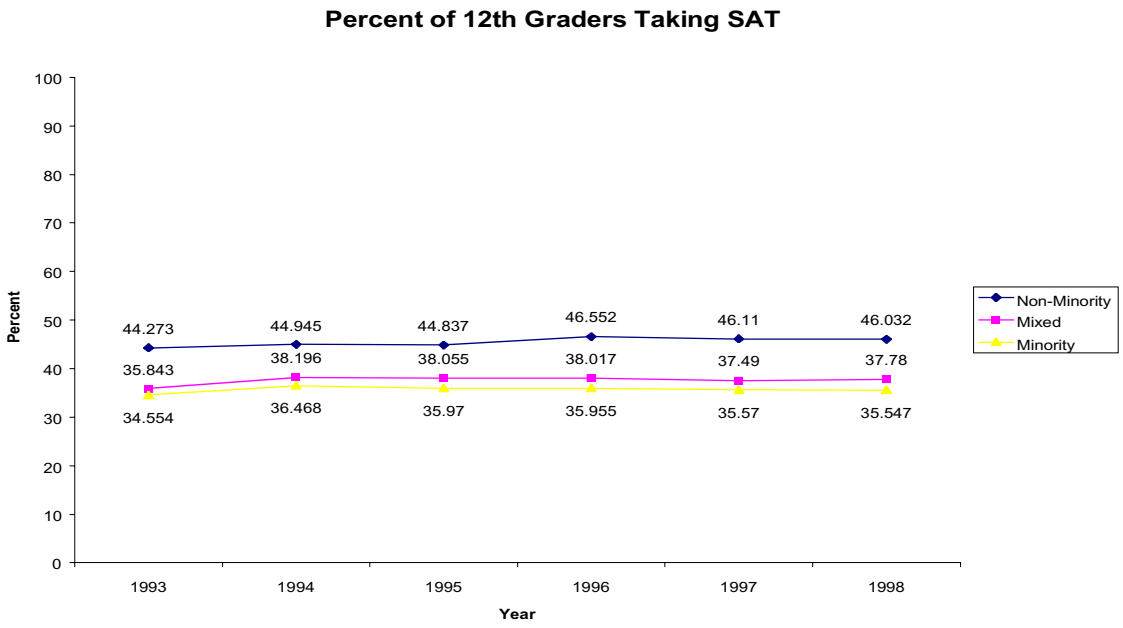

Figure A55. SAT test takers in California 1993-98.

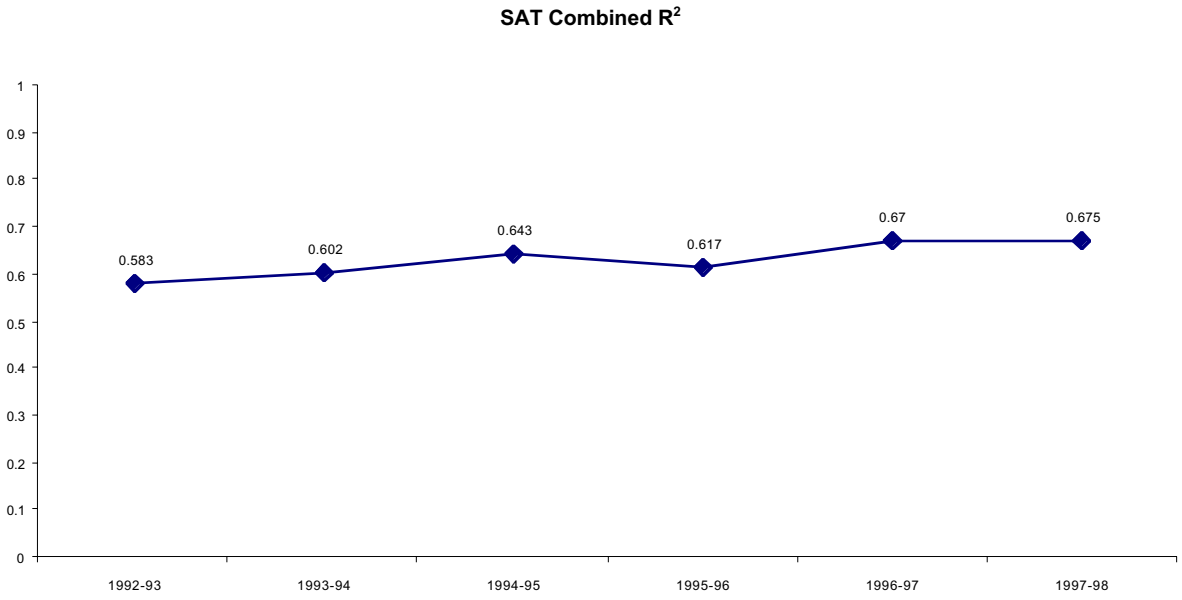

Figure A56. Relationship between socio-economic measures and SAT scores.

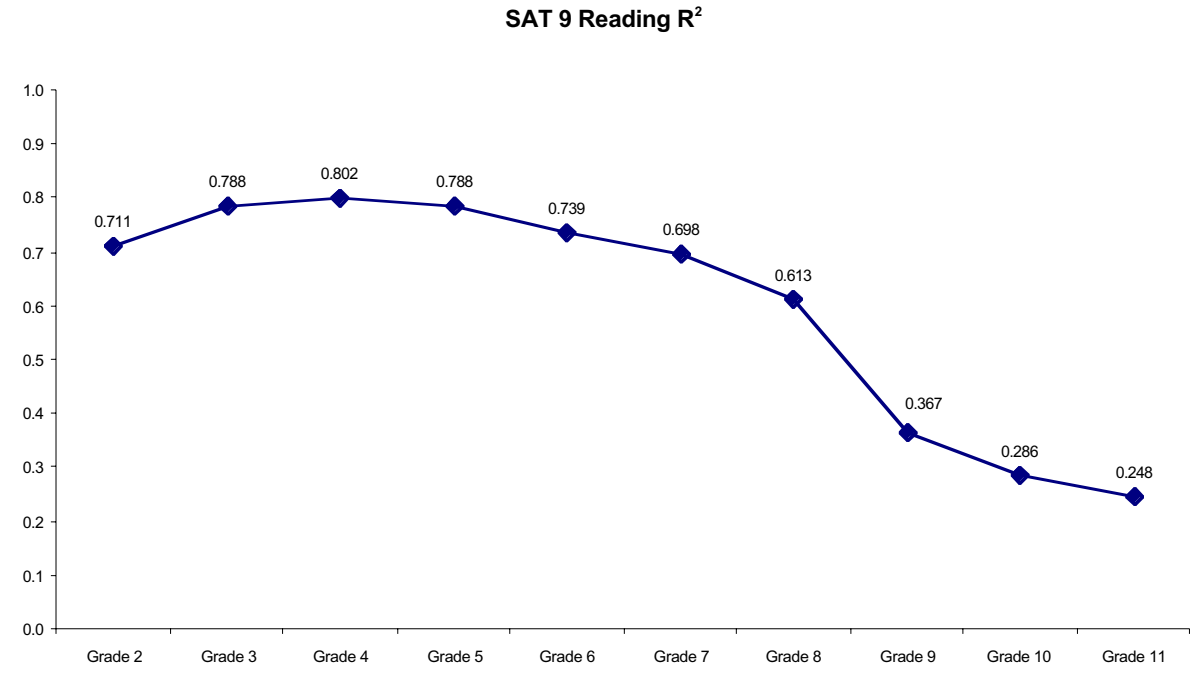

Figure A57. Relationship between socio-economic measures and SAT-9 Reading scores.

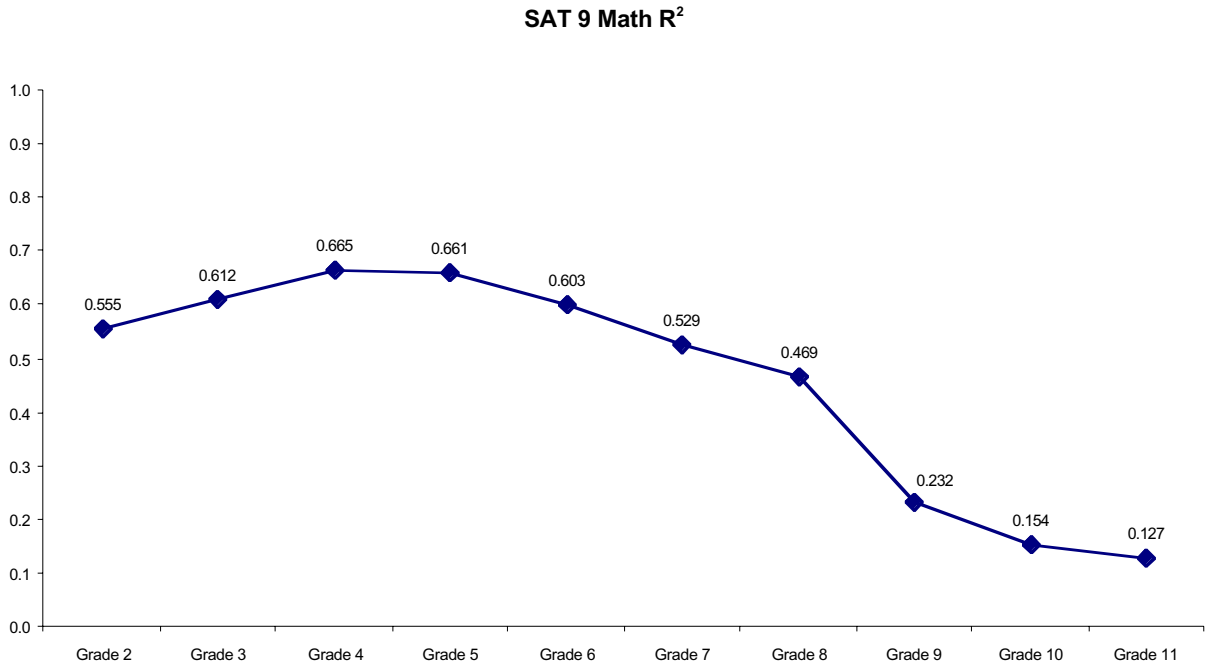

Figure A58. Relationship between socio-economic measures and SAT-9 Math scores.

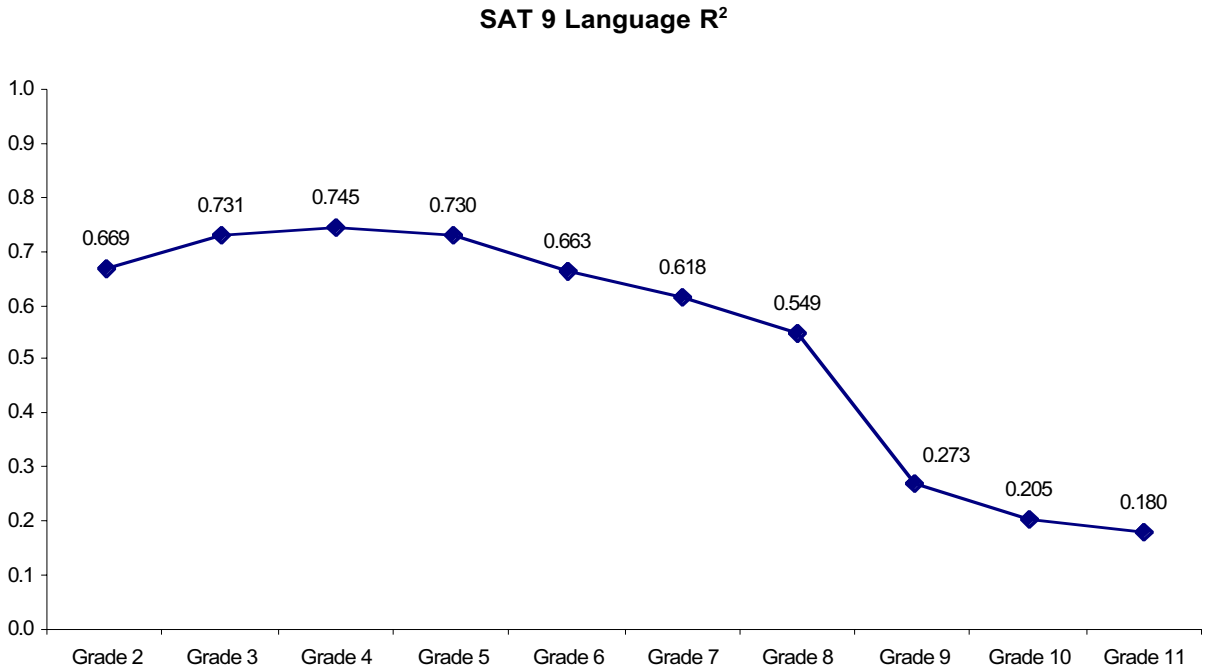

Figure A59. Relationship between socio-economic measures and SAT-9 Language scores.

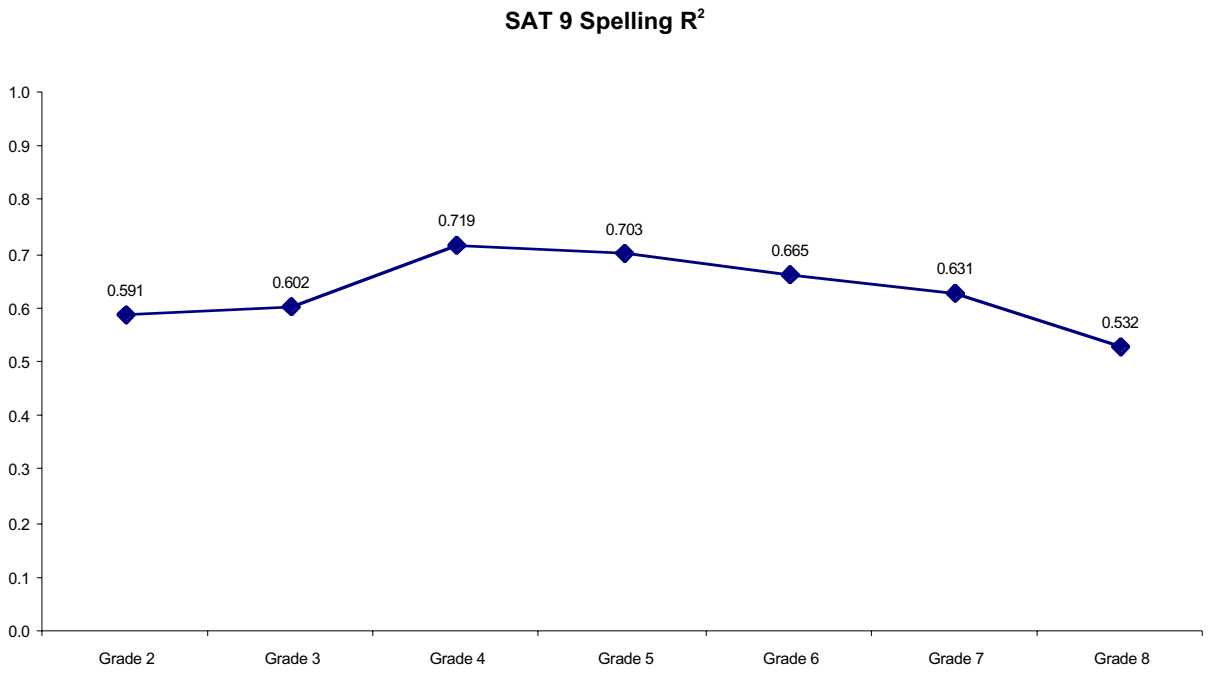

Figure A60. Relationship between socio-economic measures and SAT-9 Spelling scores.

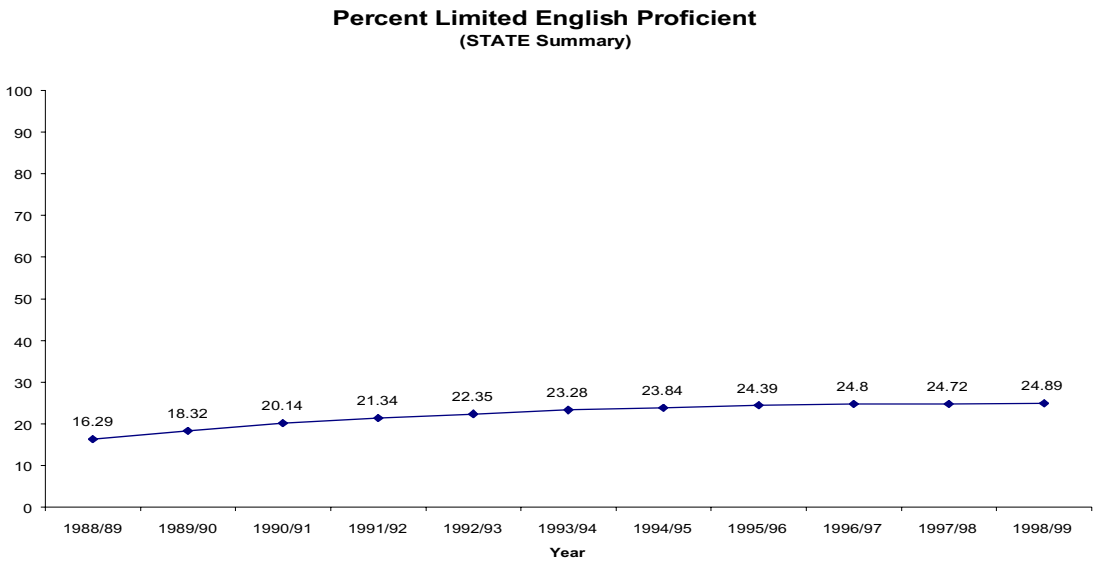

Figure A61. Limited English proficient students in California 1989-1999.

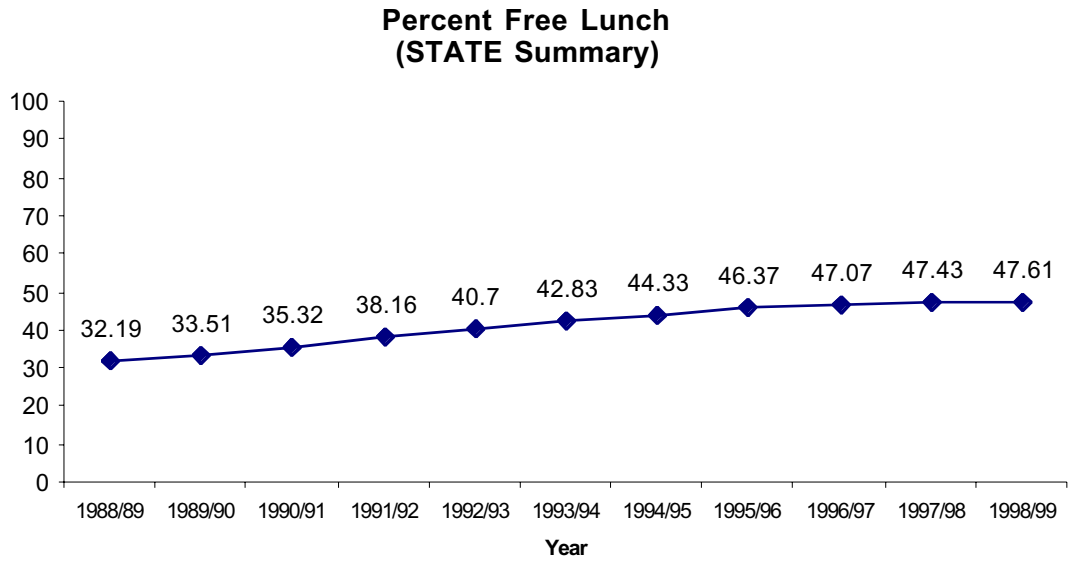

Figure A62. Students receiving free or reduced lunch in California 1989-1999.

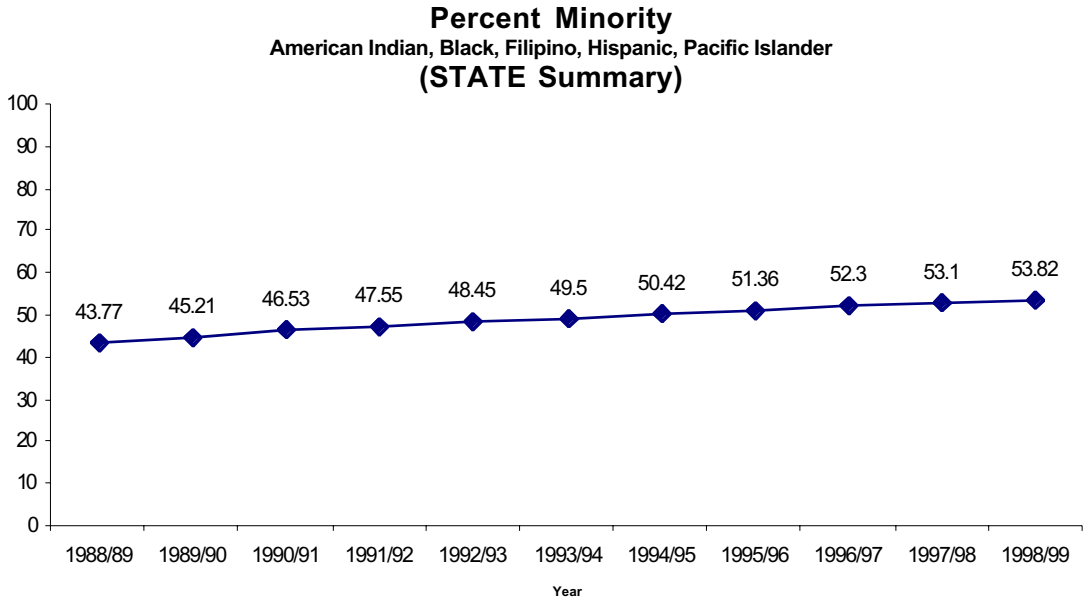

Figure A63. Minority students in California 1988-1999.

Appendix B

List of Tables

Table B1

Grade 3 LEP versus Non-LEP

			(98-99)3rd-LEP Stu.:Reading: % Student >50%ile	(98-99)3rd-LEP Stu.:Math: % Student >50%ile	(98-99)3rd-LEP Stu.:Language: % Student >50%ile	(98-99)3rd-non_LEP Stu.:Reading: % Student >50%ile	(98-99)3rd-non_LEP Stu.:Math: % Student >50%ile	(98-99)3rd-non_LEP Stu.:Language: % Student >50%ile
	low % LEP students	Mean	18	35	26	57	60	58
		Std Deviation	17	22	20	20	20	20
		Valid N	N=990	N=1007	N=986	N=2690	N=2691	N=2690
	low medium % LEP students	Mean	12	29	20	38	44	40
		Std Deviation	11	17	15	17	18	17
		Valid N	N=968	N=971	N=966	N=966	N=967	N=966
	medium high % LEP students	Mean	9	25	15	30	38	33
		Std Deviation	7	13	10	14	17	16
		Valid N	N=510	N=509	N=509	N=479	N=480	N=473
	high % LEP students	Mean	6	22	11	27	34	29
		Std Deviation	6	11	7	14	19	15
		Valid N	N=64	N=64	N=64	N=48	N=48	N=48
Group Total	Mean		14	31	21	49	54	50
	Std Deviation		14	19	17	21	21	21
	Valid N		N=2532	N=2551	N=2525	N=4183	N=4186	N=4177

Table B2

Grade 7 LEP versus Non-LEP

			(98-99)7th-LEP Stu.:Reading: % Student >50%ile	(98-99)7th-LEP Stu.:Math: % Student >50%ile	(98-99)7th-LEP Stu.:Language: % Student >50%ile	(98-99)7th-non_LEP Stu.:Reading: % Student >50%ile	(98-99)7th-non_LEP Stu.:Math: % Student >50%ile	(98-99)7th-non_LEP Stu.:Language: % Student >50%ile
	low % LEP students	Mean	11	24	22	56	54	63
		Std Deviation	11	19	15	17	18	16
		Valid N	N=525	N=528	N=524	N=730	N=729	N=729
	low medium % LEP students	Mean	6	15	15	37	37	47
		Std Deviation	5	10	9	14	15	15
		Valid N	N=217	N=217	N=217	N=218	N=218	N=218
	medium high % LEP students	Mean	6	12	16	36	36	49
		Std Deviation	4	8	8	11	13	16
		Valid N	N=40	N=40	N=40	N=40	N=40	N=40
	high % LEP students	Mean	7	20	14	36	39	48
		Std Deviation	2	8	7	18	21	18
		Valid N	N=4	N=4	N=4	N=2	N=2	N=2
Group Total	Mean		10	21	20	51	50	59
	Std Deviation		9	17	13	18	19	17
	Valid N		N=786	N=789	N=785	N=990	N=989	N=989

Table B3

Grade 9 LEP versus Non-LEP

			(98-99)9th-LEP Stu.:Reading: % Student >50%ile	(98-99)9th-LEP Stu.:Math: % Student >50%ile	(98-99)9th-LEP Stu.:Language: % Student >50%ile	(98-99)9th-non_LEP Stu.:Reading: % Student >50%ile	(98-99)9th-non_LEP Stu.:Math: % Student >50%ile	(98-99)9th-non_LEP Stu.:Language: % Student >50%ile
	low % LEP students	Mean	5	26	18	43	58	58
		Std Deviation	6	19	15	17	17	16
		Valid N	N=462	N=460	N=462	N=606	N=605	N=603
	low medium % LEP students	Mean	3	19	15	28	45	49
		Std Deviation	3	11	10	12	15	14
		Valid N	N=109	N=109	N=109	N=110	N=110	N=110
	medium high % LEP students	Mean	3	15	15	24	45	50
		Std Deviation	3	3	5	6	18	12
		Valid N	N=7	N=7	N=7	N=7	N=7	N=7
Group Total	Mean		4	24	18	40	56	57
	Std Deviation		6	18	14	17	17	16
	Valid N		N=578	N=576	N=578	N=723	N=722	N=720

Note. There were no schools with a high percentage of limited English proficient students in the sample for Grade 9.

Table B4

Grade 3 Economically Disadvantaged versus Economically Advantaged

			(98-99)3rd-Eco Disadv Stu.:Reading: % Student >50%ile	(98-99)3rd-Eco Disadv Stu.:Math: % Student >50%ile	(98-99)3rd-Eco Disadv Stu.:Language: % Student >50%ile	(98-99)3rd-Eco Adv Stu.:Reading: % Student >50%ile	(98-99)3rd-Eco Adv Stu.:Math: % Student >50%ile	(98-99)3rd-Eco Adv Stu.:Language: % Student >50%ile
	low % free or reduced lunch students	Mean	42	49	44	66	69	67
		Std Deviation	17	19	18	19	19	19
		Valid N	N=508	N=512	N=511	N=1184	N=1185	N=1185
	low medium % free or reduced lunch students	Mean	33	41	36	57	60	57
		Std Deviation	14	17	15	17	17	17
		Valid N	N=908	N=909	N=905	N=919	N=919	N=916
	medium high % free or reduced lunch students	Mean	24	34	28	44	49	46
		Std Deviation	12	15	13	17	18	17
		Valid N	N=1017	N=1018	N=1018	N=898	N=899	N=896
	high % free or reduced lunch students	Mean	15	28	20	28	36	31
		Std Deviation	8	13	11	18	20	19
		Valid N	N=740	N=741	N=741	N=336	N=342	N=332
Group Total	Mean		27	37	31	54	58	55
	Std Deviation		16	17	16	22	21	21
	Valid N		N=3173	N=3180	N=3175	N=3337	N=3345	N=3329

Table B5

Grade 7 Economically Disadvantaged versus Economically Advantaged

			(98-99)7th-Eco Disadv Stu.:Reading: % Student >50%ile	(98-99)7th-Eco Disadv Stu.:Math: % Student >50%ile	(98-99)7th-Eco Disadv Stu.:Language: % Student >50%ile	(98-99)7th-Eco Adv Stu.:Reading: % Student >50%ile	(98-99)7th-Eco Adv Stu.:Math: % Student >50%ile	(98-99)7th-Eco Adv Stu.:Language: % Student >50%ile
	low % free or reduced lunch students	Mean	40	41	48	66	65	71
		Std Deviation	14	16	15	15	17	14
		Valid N	N=220	N=220	N=218	N=273	N=273	N=273
	low medium % free or reduced lunch students	Mean	30	32	39	54	52	60
		Std Deviation	10	13	12	14	15	13
		Valid N	N=274	N=274	N=274	N=274	N=274	N=274
	medium high % free or reduced lunch students	Mean	23	26	33	41	39	48
		Std Deviation	8	11	11	16	16	16
		Valid N	N=201	N=201	N=201	N=196	N=196	N=196
	high % free or reduced lunch students	Mean	17	21	28	27	26	34
		Std Deviation	7	9	9	14	14	16
		Valid N	N=95	N=95	N=95	N=83	N=83	N=82
Group Total	Mean		30	32	38	52	51	58
	Std Deviation		13	15	14	19	20	19
	Valid N		N=790	N=790	N=788	N=826	N=826	N=825

Table B6

Grade 9 Economically Disadvantaged versus Economically Advantaged

			(98-99)9th-Eco Disadv Stu.:Reading: % Student >50%ile	(98-99)9th-Eco Disadv Stu.:Math: % Student >50%ile	(98-99)9th-Eco Disadv Stu.:Language: % Student >50%ile	(98-99)9th-Eco Adv Stu.:Reading: % Student >50%ile	(98-99)9th-Eco Adv Stu.:Math: % Student >50%ile	(98-99)9th-Eco Adv Stu.:Language: % Student >50%ile
Recode of % Free Lunch	low % free or reduced lunch students	Mean	23	42	39	47	62	62
		Std Deviation	13	15	14	16	16	15
		Valid N	N=285	N=284	N=283	N=356	N=356	N=355
	low medium % free or reduced lunch students	Mean	16	32	31	34	46	48
		Std Deviation	8	11	10	13	15	14
		Valid N	N=158	N=158	N=158	N=158	N=158	N=158
	medium high % free or reduced lunch students	Mean	13	30	30	25	37	40
		Std Deviation	7	12	11	16	19	19
		Valid N	N=76	N=76	N=76	N=75	N=75	N=75
	high % free or reduced lunch students	Mean	17	31	34	20	38	39
		Std Deviation	11	18	13	13	24	19
		Valid N	N=10	N=10	N=10	N=6	N=6	N=6
Group Total	Mean		19	37	35	41	54	55
	Std Deviation		11	15	13	18	19	17
	Valid N		N=529	N=528	N=527	N=595	N=595	N=594

References

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade*. Princeton, NJ: Educational Testing Service, Policy Information Center.

California State University Institute for Education Reform. (1999). *The Advanced Placement program: California's 1997-98 experience.* Sacramento, CA: Author

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices* (CSE Monograph Series in Evaluation, No. 11). Los Angeles: University of California, Center for the Study of Evaluation.

Education Watch 1998: The Education Trust State and National Data Book (Vol. II). Education Trust, Washington, DC. (http://www.EdTrust.org)

Hoff, D. (1999, August 4). Inglewood ACLU lawsuit. *Education Week, 2,* 13.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Kellaghan, T., & Madaus, G. F. (1991). National testing: Lessons for America from Europe. *Educational Leadership, 49*(3), 87-93.

Koretz, D. M., Barron, S. I., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)* (MR-792.PCT/FF). Santa Monica, CA: RAND.

Koretz, D. M., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). *Perceived effects of the Maryland School Performance Assessment Program* (CSE Tech. Rep. No. 409). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Koretz, D. M., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience* (RAND RP-259). Santa Monica, CA: RAND. (Reprinted from CSE

Tech. Rep. No. 371, Los Angeles, University of California, Center for Research on Evaluation, Standards, and Student Testing.)

McDonnell, L. M., & Choisser, C. (1997). *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.

National Commission on Excellence in Education. (1983). *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.

National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9,* 5-22.

Shepard, L. A. (1991). *Will national tests improve student learning?* (CSE Tech. Rep. No. 342). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Smith, M. L. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program* (CSE Tech. Rep. No. 425). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *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, Center for Research on Evaluation, Standards, and Student Testing.

U.S. Department of Education. (1998). *The Nation's Report Card. National Assessment of Educational Progress.* Washington, DC: National Center for Education Statistics. Available 28 April 2000 http://nces.ed.gov/nationsreportcard/site/home.asp.