Expert Report

Submitted for:

Eliezer Williams

vs.

State of California

By

Michael Russell

October, 2002

I. Introduction

In 1994, I joined the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College as a Research Associate. I am currently a Senior Research Associate at CSTEEP and a professor in the Lynch School of Education at Boston College. My areas of expertise include student assessment, school accountability, test theory and construction, and applications of technology to K-12 education. I have both participated in and directed several research and development efforts related to educational testing and accountability. These efforts have included:

- Developing and implementing the Co-NECT School Accountability Model used by approximately 25 schools operating in 5 states that were

implementing the Co-NECT School Design. This accountability model was a multiple measure system that included students test scores, surveys of students and, in some cases, parents, student drawings, and active reflection and goal setting by the school community;

- Conducting a three year study in 22 schools in Rhode Island to examine impacts of standards-based reform and standards-based accountability;
- Assisting districts supported by the Edna McConnell-Clark Foundation in developing accountability reports. These districts included Minneapolis, Corpus Christi, San Diego, Long Beach, and Chattanooga;
- Examining technical issues related to Massachusetts Comprehensive Assessment System including scaling, equating, scoring, and standard-setting procedures;
- Working with Department of Defense Schools in Aviano, Italy, in integrating technology program evaluations into their School Improvement Process;
- Working with Department of Defense Schools in Aviano, Italy in developing local accountability system that included multiple-measures;
- Conducting a series of randomized experiments that examined the validity of paper-based state and national writing tests for students accustomed to writing with computers;

- Developing an alternative accountability system that employed multiple measures and supported active reflection and accounting by schools in Massachusetts;

- Exploring applications of computer-based technology to the technology of testing with the aim of increasing the validity of inferences related to higher-order cognition;

- Collaborating with several educational and political leaders in Massachusetts on a proposal for the Bill and Melinda Gates Foundation to develop a comprehensive accountability system that capitalized on the powers of computer-based technologies

I have authored several articles that focus on the validity of paper-based tests for students accustomed to writing with computers, multiple-measure accountability systems, use of test scores for college admission decisions in California's University System, and methods for evaluating impacts of technology in K-12 schools.

I have previously assisted others in preparing testimony in litigation. This is the first time that I have personally provided testimony in litigation.

When preparing this report, I have considered previously published scholarly work, unpublished reports prepared by other scholars, data accessible via the California Department of Education's web-site, documents and minutes from meetings held by various parties within or associated with the Department of Education and/or the State Board of Education, a conversation with a member of the API Technical Advisory Committee, and depositions of a few personnel within the Department of Education.

Most of the material was obtained by myself and two very helpful assistants, namely Anastasia Raczek and Jennifer Cowan. The depositions were provided by the litigation team.

II. Nature of Assignment

The overarching assumption implicit in much of my opinion is that states (California included) provide funding and leadership for public education in order to provide all students with opportunities to develop academic, social, and work-related skills and knowledge so that they will be prepared to be productive, thoughtful, and active members of society. Given this role, I assume that state-level accountability systems should be designed to assist school systems in assessing the extent to which they provide an environment in which these academic, social, and work-related skills and knowledge develop. Thus, an effective and educationally beneficial accountability system would encourage schools to focus on inputs, outputs, and the relationship between the two-that is, the extent to which inputs impact outputs in desired ways. With that in mind, the plaintiffs have asked me to answer two questions:

1) Does California's current output-based accountability system accurately and sufficiently notify the State of whether students receive essentials required for learning?

2) If not, are there alternatives to California's current accountability system?

III. Opinions and Conclusions

Over the past nine years, my research activities have required me to become familiar with educational assessment and accountability systems in at least ten states, including Massachusetts, Tennessee, Texas, Florida, Maryland, Ohio, Michigan, Florida, Rhode Island, Kentucky, Alaska, and most recently California. While none of these states have established what I consider to be exemplary accountability systems, some are much better than others. If asked to rank the quality and utility of the systems in place in each of these states, the system currently in place in California (codified in the 1999 Public School Accountability Act ("PSAA")) would be near the bottom of the list. The Academic Performance Index (API) it employs is simply incapable of providing the type of information that State policy-makers need to make rational decisions as to which schools need help and how to help them.

As I describe in detail below, California's attempt to establish an educational accountability system over the past decade has been tumultuous. Setting aside the several proposed and implemented versions of the current PSAA, California has put into place five distinct systems within a ten year period. The current PSAA itself keeps changing. Recently, one of the "key components" of the PSAA system, Teacher Bonuses, was targeted for elimination by the Governor due to a budget shortfall and the State Board of Education is changing vendors for the state testing program. This change is expected to result in the replacement of the SAT-9 with California Achievement Test 6th Edition ("CAT 6").

Even if the technical shortcomings of California's accountability system were fixed and/or prior decisions were altered to make expectations for most schools more reasonable, the PSAA's single-minded focus on student outcomes as measured by standardized tests fails to adequately prevent, detect or deter gross disparities in education.[*] A system that focuses solely on student learning outcomes, no matter how broadly defined, cannot provide schools and their constituents with information that allows them to identify why students succeed or fail to succeed. Without placing outcomes (or outputs) in the context of inputs, schools, their constituents, researchers, and policy makers are limited as to what they can learn about how to improve the quality of education. Moreover, a narrow focus on outcomes ignores the roles played by quality teachers; quality resources (such as books, manipulatives, labs, and computer-based technologies); and a quality environment in developing students' academic, social, and work-related skills and knowledge.

A system like California's, which ranks, rewards, and punishes schools based on outcomes, without also requiring and assisting them to provide quality inputs, is not only extremely limited in terms of its ability to direct positive change, it is damaging in and of itself. Such a system promotes practices that are of poor educational value. As I describe below, these questionable practices include reducing or denying students' exposure to computer-based technologies; investing time and resources in test preparation while decreasing or eliminating investments in non-tested, but standards-based, areas of education; increasing retention without exposure to supplemental or alternative learning opportunities; aggravating school drop-out rates; and increasing (often without sound

[*] The non-beneficial effects of using standardized test scores for school accountability purposes are

reason) the number of students classified as having special educational needs. While many of these problems are more likely to occur in low performing schools than in high performing schools, a narrow focus on outcomes may also be harmful for students in high performing schools.

In most cases, high performing schools serve students with a high socio-economic status (SES) whose parents are generally well educated. These two factors are consistently correlated with high academic performance, which suggests that some of students' learning is influenced by factors outside of a school's control. Because these external factors play a role in high test scores, they may overcome poor educational practices employed within a high performing school. That is, a school could be high-scoring on tests and meet performance targets in the accountability system, despite a low quality of educational practices. Requiring all schools to place their performance in the context of the practices they employ, the resources they provide, and the performance of schools that serve similar students would more likely lead to improvements in the quality of education. By requiring all schools to consider the relationship between inputs and outputs, improvements are more likely to occur in low -- as well as high -- performing schools.

Counting may be a component of an accountability system, but absent an account of how those counts came to be, the seemingly precise final figures can be deceptive. (The recent collapse of Enron provides a good example, outside of education, of how a narrow focus on outcomes absent a solid understanding of how the outcomes are produced can be extremely deceptive and harmful.) Without a full understanding of the

---

discussed at length by Kohn (2000) and Meier (2002).

factors that influenced the final figures – whether they be the financial bottom line, a tally of judges' scores, or a summary of school test scores – desirable high performance numbers can be the product of undesirable practices. And little light is shed on the causes of undesirable outcomes.

EXECUTIVE SUMMARY OF KEY POINTS

THE API CANNOT ACHIEVE THE GOALS OF A STATE ACCOUNTABILITY SYSTEM

Counting Rather than Accounting: Because the system does not require schools to be accountable for adequately providing the inputs that will allow children to succeed (quality teachers, adequate facilities, textbooks, etc.), and because the State has no means to assess, let alone address, schools' input needs, many students in California go without the factors that matter most in improving their academic, social and work-related skills and knowledge.

Consider, for example, the fact that research consistently indicates that quality teaching matters. Wenglinsky (2002) uses data from the National Assessment of Educational Progress (NAEP) to examine the role teachers and their instructional practices play in impacting student achievement. Following extensive analyses, he summarizes his findings: "The effects of classroom practices, when added to those of other teacher characteristics, are comparable in size to those of student background, suggesting that teachers can contribute as much to student learning as the students themselves." Several other studies have also shown that the quality of instruction that students experience impacts their learning. When students are repeatedly exposed to low-quality teaching, their learning suffers. The State knows this. The State also knows which schools have the highest percentages of uncredentialed teachers, i.e. schools where

low-quality teaching is likely prevalent. Fully aware that the low test scores returned by students in those schools are likely linked directly to the quality of teaching they receive, the State has not taken any action that is likely to resolve the problem. In this instance, the State knows that a particular factor contributes to failure. It even knows where that factor exists. But the system is set up so that this information is not valued and is effectively ignored. Furthermore, unlike uncredentialed teachers, most factors that impair student learning are not even measured by the State (inadequate facilities, textbook unavailability, etc.).

California's educational accountability system is based solely on counting percentages of students in different performance bands on an assortment of tests, counting the percentage of schools performing above or below an arbitrary and unrealistic target, and counting the number of points each school's API changes each year. Based on these counts, schools and teachers may be eligible for monetary awards (although the Certified Staff Performance Incentive program was eliminated earlier this year) or to apply for assistance. If selected to receive additional assistance, an external evaluator is employed to examine various aspects of the school and its practices. This evaluation process is the closest California's accountability system comes to requiring schools to provide an account of their practices. Implied in this evaluation process is a desire for schools to take corrective action to improve problematic practices (whether they be curriculum misalignment, instructional practices, resource allocation, quality of teachers, quality of facilities and related educational materials, leadership, etc.).

However, evidence of success is based solely on the counting of students performing within specific performance bands and changes in API scores. Even in the

relatively few instances where the State does make an effort to identify potential input problems at low performing schools, the schools are deemed successful only if the desired outcome is reached – regardless of whether the problematic inputs are addressed or how the outcome is reached.

Higher APIs Don't Necessarily Indicate Improved Student Achievement: The state has specified that tests employed as part of the accountability system should measure skills and knowledge specified in the curriculum frameworks from which schools are expected to teach. However, the state has acknowledged that the SAT-9 is poorly aligned with state frameworks. Despite the misalignment, and in spite of the state's effort to develop tests that are better aligned with standards, SAT-9 scores were the sole outcome indicator used in 1999 and 2000 and constituted 64% of the outcome indicator for elementary and middle schools and 76% for high schools in 2001. And a recent plan issued by the State Board of Education (SBE) indicates that the final version of the API, due to be established in 2006, will still be based, in part, on a poorly aligned test.

Research consistently indicates that when high-stakes decisions are made based on test scores, teachers modify their instruction so that it focuses on the skills and knowledge included on the test, de-emphasizing skills and knowledge not on the test. It is reasonable to expect that teachers will "teach to the test" more often in schools that are performing poorly on tests used for accountability purposes given the close scrutiny such schools face. Therefore, increases in the SAT-9-based API scores over the past few years may very well be the result of inferior, test-centered teaching practices as opposed to student improvement in terms of state standards. Teachers must choose whether to focus

instruction on the skills and knowledge emphasized in the standards or on the misaligned content of the SAT-9.[1]

A Low API Score Says Nothing About Why It Is Low: Many schools in California serve large numbers of students who are English language learners ("ELL") who have limited proficiency in English. As a result, scores for these schools are currently well below the national average. Even if these schools were extraordinarily successful in improving test scores of ELL students, each year a new cohort enters – a cohort that cannot reasonably be expected to perform any better on the test than ELL students have historically performed in California (approximately 25-30 percentile ranks below the national mean). Unless the State changes its system to provide information as to *why* schools perform as they do, it will never be able to target assistance in a rational way.

The API System Is Not An Adequate Outcome Measure: My main argument in this report is that California's accountability system, because it fails to measure the inputs that determine the outputs it does measure, cannot provide information that will allow the State to exercise the leadership required to provide all students with the educational opportunities they are entitled to. However, it is important to note that the API does not even provide accurate, let alone useful information about student output.

- *Volatility of and Error in Test Scores:* There is substantial volatility in individual SAT-9 test scores. As an example, there is a 70% chance that a student's

[1] It should be noted that the expected change from the SAT-9 to a new NRT test in 2003 does not rectify the issue of poor alignment. Like the SAT-9, the new test will be a general test of skills and knowledge that was designed to provide normative comparisons at the national level. Moreover, like the SAT-9, the new NRT will not be developed to specifically target skills and knowledge specified by California's standards.

obtained percentile rank is five or more points away from his/her true score (at the mean). There is also a 50% chance that a student's percentile rank will change by 10 points or more over the course of a year when in fact it should have remained unchanged (Rogosa, 1999). Similarly, aggregate (or mean) test scores for schools containing fewer than 100 to 140 students fluctuate substantially from year to year. These fluctuations result largely from error in measurement and differences in the characteristics of cohorts rather than real differences in learning (Kane & Staiger, 2001; Haney, in press). As recently reported by the Orange Country Register (August 11, 12 and 13, 2002), aggregate test score error was not fully openly disclosed by the State until July of 2002 and was reported to be approximately 20 points. This 20 point error means that the API score for an "average" school could be 20 points higher or 20 points lower than the actual score reported by the State. For many schools, test score error is as large as the amount of improvement prescribed by the State. Since the current accountability system has been in place, other factors such as late delivery of tests to schools (50% of schools have reported this problem) and inaccurate reporting of results for several schools have contributed to errors in measurement (Noble, 2000).

- *Aggregating Scores at the School Level Masks the Successes and Failures At the Grade and Classroom Levels:* Although aggregation of scores across grade levels may help decrease the volatility of score changes, it blurs differences in performance and/or gains at different grade levels. Underperformance is not uniform across grade levels.

  A second problem with aggregating test scores stems from the difficulty in using

aggregates to explain what factors actually caused scores to change. As Haney and Raczek (1994, p. 17) state, "attempting to hold schools or educational programs accountable in terms of aggregate performance of students almost guarantees that accountability in terms of *explaining* patterns of individual students' learning will be largely impossible."

The API's focus on school-level performance across grade levels rather than also considering performance within grade levels or classrooms obfuscates the impact of efforts within these lower-level units to improve student learning. Additionally, the focus on school-level performance and characteristics may promote fallacious conclusions about the impacts of school-level programs and the influence other variables have on the success of these programs. While aggregation at the grade or classroom level may be a poor fix for this problem, it might promote closer examination of practices and issues within these smaller operational units.

- *National, Norm Referenced Tests Provide No Information About Student Performance in Specific Subject Areas:* Standardized tests are developed for a specific purpose. Some standardized tests are designed to diagnose learning disorders. Other standardized tests are designed to measure student skill and knowledge in a specific domain (e.g., mathematics, reading, science, etc.). Sometimes, these measures are expressed relative to desired levels of performance (i.e., criterion-referenced tests). At other times, these measures are expressed relative to the performance of other students (i.e., norm-referenced). Still other standardized tests are designed to identify misconceptions or

misunderstandings within a specific sub-domain (e.g., addition in mathematics, decoding in reading, etc.). The purpose of the test informs the type, quantity, difficulty, and, sometimes, order of items that form the test. While it is common practice to use a given test for purposes other than its intended use, this is not good practice.

The SAT-9 is a national, norm-referenced test that was designed to provide a *general* measure of student skill and knowledge relative to other students in the areas of mathematics, language arts, social studies, and science. To cover these general domains, the SAT-9 tests contain an array of items related to several sub-domains. To minimize the amount of time required for testing, a small number of items (in some cases only one or two) focus on a given sub-domain.

Being a norm-referenced test, items on the SAT-9 are selected so that student performance is distributed across a range of scores. That is, a set of items is selected to produce scores that are distributed normally. While the resulting scores provide a good indication of how a student is performing relative to other students (in the norm group), the scores do not provide useful information about the student's absolute performance within the given domain or sub-domain. Moreover, given the interaction between the need to cover several sub-domains and the need to employ items that vary widely in difficulty, useful diagnostic information about performance within sub-domains is not provided.

While some standardized tests are specifically developed to provide useful

diagnostic information or characterize student performance relative to desired levels of performance, the SAT-9 does neither. In short, the SAT-9 is a poor instrument for either identifying student weaknesses within specific sub-domains or determining whether students have achieved acceptable levels of skills or knowledge within a given domain. And because California's accountability system is heavily dependent on the SAT-9, the system has little promise for helping schools identify strengths and weaknesses in student skill and understanding in specific areas of mathematics, language arts, social studies or science.

It should be noted that the expected change to the use of CAT 6 does not correct these shortcomings. Like the SAT-9, the CAT 6 is a nationally norm-referenced test that provides poor diagnostic information at the student level.

- *Unless API Score Increases are Above Average, They Go Unnoticed:* Students must make gains to achieve school targets, not on the same test, but rather on the test they take a year later, for the next grade level. While some of the subject matter overlaps across years, additional skills and knowledge are required to perform at the same level from year to year. Although often misinterpreted as showing no growth, percentile ranks that remain the same across years actually represent substantial growth – growth that is identical to the average student nationwide. And increases in percentile ranks across years indicates even more growth than the typical student nationwide.

- *Tests that Provide Student-Level Data Provide Poor School-Level Data:* A single test administered to all students within a school, whether it be norm-referenced, like the SAT-9 and CAT 6, or criterion-referenced, like the High School Exit Exam, is inadequate for diagnosing instructional strengths and weaknesses within the school (or individual classrooms). Matrix sampling is employed by the National Assessment of Educational Progress, the Third International Mathematics and Science Study, and testing programs in states including Maryland. Matrix sampling is a far more efficient and informative approach to collecting diagnostic information that can be used by teachers and schools to improve curriculum and instruction. In essence, matrix designs randomly assign different groups of students within a classroom and/or school to perform different sets of test items. Depending upon the design employed, matrix designs can increase the amount of information collected at the school-level by two to eight times. Given the limited time available for testing, matrix designs enable a broader spectrum of the state frameworks to be tested within schools and allows the most important areas of the frameworks to be tested at a finer level. Although matrix designs do not provide comparable student-level scores, California's current accountability system only requires such student-level scores for the High School Exit Exams. For all other grade levels, a matrix design would be far more informative than is the current practice of administering the same set of test items to all students in a school and across the state.

AN IMPROVED SYSTEM

The PSAA requires that the API-based accountability system include a range of outcome variables including scores from tests aligned with the state frameworks, graduation rates, and student and teacher attendance (CDE, 2000d). The California Department of Education goes further: "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" (CDE, 1999c). Despite these proclamations, the API-based accountability system currently relies solely on test scores (several of which are poorly aligned with the state frameworks).

A truly comprehensive accountability system would ask schools to describe the programs and practices they have in place, the appropriateness of these programs and practices given specific context and background indicators, and the effect these programs have on a variety of student outcomes. Programs and practices of interest might include but should not be limited to:

- Access to quality teachers (e.g., student:teacher ratios, % of teachers with emergency credentials, % of teachers with Masters Degree or Beyond, etc.)
- Access to books, textbooks and other learning materials (e.g., ratio of library books to students, ratio of course specific textbooks to students, ratio of students:computers, ratio of students:Internet accessible computers, etc.)
- Adequacy of school facilities (e.g., overcrowding, access to sanitary facilities -- ratio of students:functioning toilets, ratio of "contaminated"

classrooms:total classrooms, etc.; availability of functional heating and cooling systems, presence of lead paint, etc.)

- Type of school calendar (e.g., multi-track year-round schools; schools operating under the concept 6 model)
- Availability of appropriate instructional materials, specially trained teachers for ELL students
- Subject area curricular materials used (e.g., math curriculum/textbooks, science curriculum/textbooks)
- Availability of Advanced Placement Courses (e.g., number of courses offered, number of sections available)
- Professional Development Opportunities (e.g., topics focused on during PD, number of hours offered, number of hours taken, percent of faculty participating)

Student outcomes might include but should not be limited to:

- Performance on tests closely aligned with the state frameworks
- Attendance rates
- Promotion/retention rates
- Graduation rates
- Drop-out rates

- Course taking patterns (higher vs. lower level mathematics, AP courses, etc.)
- Percent of students completing all courses required for UC eligibility
- Percent of students taking college entrance exams
- College entrance

Furthermore, to increase the amount of information and level of specificity of that information at the individual school level, the state should implement a longitudinal student tracking system, such as the California School Information Systems ("CSIS"), which is currently voluntary; but, according to the April 2002 minutes of the Superintendent's Advisory Committee on the PSAA, "the reality of such a system in California is likely years away." (PSAA Advisory Committee, April 2002, p.5.). Without such a system, as the State acknowledges, it has no means of accurately measuring drop-out rates, which may dramatically affect school-average performance on tests, and which may in turn be affected by perverse incentives to increase school-average scores. The State testing program should consider matrix sampling. As described above, matrix sampling provides far more information about what groups, classes, and grade levels students within a school can and cannot do. For this reason, matrix sampling provides information that will better inform the types of changes that schools might make to their curriculum and instructional practices.

By requiring schools to actively describe the impacts their inputs have on outputs, identify potential problem areas, and establish short and long term goals, educational benefits of accountability could be more fully realized. Moreover, the goals set through

this process should not be limited to changes in outcomes. At times, it is the inputs that must be altered before outcomes can change; schools must be allowed and encouraged to set goals that focus also on inputs.

As an example, there is a clear correlation between the percent of emergency credentialed teachers within a school and the school's API. It only makes sense, then, that for schools that have a high percentage of emergency credentialed teachers, interim goals should focus on decreasing the percentage of emergency credentialed teachers (ideally to 0%) rather than on increasing students' test scores. Only after significant progress towards this interim goal has been reached should attention turn to changes in test scores. A comparison of students' performance (as individuals) and the total number of years they were taught by emergency credentialed teachers might also suggest pursuing equal access to credentialed teachers for individual students.

Similarly, in several places the state frameworks state that students should be able to use a variety of tools. As an example, the grade six science content standards require students to "select and use appropriate tools and technology (including calculators, computers, balances, spring scales, microscopes, and binoculars) to perform tests, collect data and display data." (CDE, 1998.) The grade six English Language Arts standards require students to demonstrate their research skills by using "organizational features of electronic text (e.g., bulletin boards, databases, keyword searches, e-mail addresses) to locate information." (CDE, 1997.) Other standards state that students must "explain the effects of common literary devices (e.g., symbolism, imagery, metaphor) in a variety of fictional and non-fictional texts." (CDE, 1997.) To achieve these standards, students must have access and exposure to these texts and materials. In schools where these materials

are limited or non-existent, an interim goal should focus on the acquisition of these materials.

In many respects, this type of system is currently in place in Rhode Island where data is collected about a wide range of variables and schools are required to engage in active reflection, goal setting, and communication with their community (See Section 7 for more detail).

TABLE OF CONTENTS

HISTORY OF EDUCATIONAL ASSESSMENT AND ACCOUNTABILITY IN CALIFORNIA

In contrast to the steady and consistent systems in place in California during the 1970's and 1980's, large-scale student assessment in California has been tumultuous during the past decade. Over the course of ten years, teachers, students and local communities have been faced with five separate assessment systems. Some of these systems have employed a variety of test instruments that were closely linked to state frameworks and standards while others have employed off-the-shelf standardized tests. Some of the systems have given districts latitude in determining what tests to use to assess student learning while others have mandated which instruments must be used.

1.1. California Assessment Program, 1972-1990

The first large-scale and sustained attempt at assessing learning in all of California's public schools was initiated in 1972. Known as the California Assessment Program (CAP), this first iteration of educational accountability in California consisted of a series of multiple-choice tests administered in four grade levels. As Cohen and Hill describe, the CAP "gauged performance on standardized, multiple-choice tests on the 'basic skills' of writing, reading, and mathematics and in the content areas of science, history, and literature" (2001, p. 28).

Since the purpose of testing was to assess the performance of schools and districts rather than individual students, CAP employed a matrix-sampling design that required students to take only a portion of all test items. By employing a matrix design, CAP provided detailed information about a broader range of topics and skills within a given subject area while minimizing the amount of time taken away from instruction for testing. While CAP proved useful for providing information about the performance of districts

and schools, scores for individual students were not produced. As calls for individual test scores rang loudly in 1990 and content-area frameworks were developed that focused on higher-order skills, the CAP was abandoned because it was unable to "produce reliable individual student scores" (Noble, 2000) and it failed to measure many of the skills deemed important in the state frameworks. The California Assessment Program had been in place for over twenty years.

1.2. California Learning Assessment System, 1991-1993

In 1991, CAP was replaced by the California Learning Assessment System (CLAS). Three features distinguished CLAS from CAP. First, CLAS was linked closely to recently developed state frameworks. Second, since many of these frameworks focused on "higher-order thinking skills," CLAS employed a combination of multiple-choice and open-ended (or supply) test items. Third, CLAS yielded scores for individual students. It is important to note that CLAS was developed at a time when state standards and large-scale performance-based tests were in their infancy. Yet, despite the absence of a model upon which to build, California was able to produce a complex, valid and reliable testing system that employed a mix of item formats in several subject areas. This rapid development was a product of a concerted, focused and determined effort made by leaders within the Department of Education and in collaboration with several external agencies (Cohen & Hill, 2001).

Although the state did not attach any sanctions for schools that performed poorly on CLAS, the testing program was coupled with a system that supported school-level review and reflection. Schools receiving School Improvement Program funding from the

state were required to undertake a Program Quality Review (PQR) every three years. As Cohen and Hill describe, as part of the PQR,

> Teams were appointed to make site visits, and review criteria were promulgated by state officials. Criteria for school and district performance were drawn from the state's frameworks, and local educators reported on their performance. Site visitors assessed local progress on the criteria. Although PQR sought accountability, and state officials used it to encourage alignment [between instruction and the frameworks], it entailed appreciable learning, as teachers gathered to read the frameworks and review curricula and instruction within schools. (p. 28.)

1.3. Pupil Testing Incentive Program, 1995-1997

Soon after its rapid introduction, CLAS fell victim to outcry from a small but vocal group of parents who "objected to the personal nature" (Noble, 2000) of some of the questions. Concerns about the consistency of scoring on the writing tests were also raised. By 1995, CLAS was dropped. It had lasted three years. In its place, the governor signed into effect the Pupil Testing Incentive Program. Under this program, districts received five dollars for every student in grades 2-10 who took a basic skills test that was approved by the State Board of Education. Unfortunately, the Board did not select and approve tests that were specifically aligned with the state standards. In addition, the Board did not anticipate that allowing districts to select an approved test would make it impossible to compare the performance of schools and districts across the state (see Feuer et al., 1998 for a review of factors that complicate comparisons and attempts to equate scores across different standardized tests).

1.4. Standards-Based Accountability, 1997-1998

Since the approved standardized tests were not closely aligned with the state standards, the Pupil Testing Incentive Program was replaced by the Standards-Based Accountability program in 1997. This new program empowered local districts to define and implement their own standards-based assessment systems, often using multiple measures. Districts were then required to document their system and report on how well students were performing based on their district-defined standards. While this system returned much of the power and responsibility for assessing student learning to local districts, it was short-lived. In 1998, the Standardized Testing and Reporting program pushed aside district-level programs that employed multiple measures and replaced them with a single state-mandated standardized test, the SAT-9. A year later, the Public Schools Accountability Act (PSAA) folded the STAR program into California's current accountability system.

1.5. The Public School Accountability Act, 1999-Present

There are three components to the Public School Accountability Act:

- The Academic Performance Index (API), an index to measure school performance;
- The Immediate Intervention/Underperforming Schools Program (II/USP) to help underperforming schools improve academic performance; and
- The Governor's Performance Awards Program (GPA) to reward schools for improving academic performance.

The 1999 PSAA legislation, and subsequent amendments, defines the three components in a general way, and specifies what level of performance is required on the API in order for a school to be considered successful. Schools not meeting the defined standard, or demonstrating sufficient annual growth, are not eligible for monetary Governor's Performance Awards, School Site Employee Performance Bonuses, or Staff Performance Incentives. Such schools also can be identified for participation in the Immediate Intervention/ Underperforming Schools Program (II/USP). The legislation does not state specifically how the centerpiece of school accountability, the API, should be operationalized; the Department of Education, with the help of an Advisory Committee, carries out that work. The accountability index must be fully described before the system as a whole can be understood; it is detailed next.

2. CALIFORNIA'S CURRENT INDEX OF ACCOUNTABILITY – THE API

The API is a numeric index that ranges from a low of 200 to a high of 1000; the current performance target for all California schools is 800. Initially, in 1999 and 2000, the API was based entirely on student scores from the Stanford Achievement Test Version 9, Form T (SAT-9). The SAT-9 is a nationally norm-referenced achievement test that is not aligned with California standards. It is important to note that the scores provided by the SAT-9 express student performance in relation to students across the nation rather than within California. For this reason, a percentile rank of 50 does not necessarily mean that a student's performance is average in comparison to all other students in California. The SAT-9 mathematics, reading, language, and spelling subject tests are administered to students in grades 2 through 8. Students in grades 9 through 11

are administered the SAT-9 mathematics, reading, language, history/social studies, and science tests. The recent addition of a new California Standards Test to the API, the decision to switch from the SAT-9 to a different norm-referenced test after 2003, and plans for additional API components, are detailed later.

Every year, each school receives four rankings: an overall ranking, a similar school ranking[i], an overall growth ranking, and a similar school growth ranking. The overall ranking and growth ranking are used for official purposes, namely as criteria for II/USP eligibility (describe in section 4.1.1).[ii] In addition to comparing the actual changes in each school's API score to their annual growth target, schools are ranked by deciles, within elementary, middle and high schools.

Generating an API from Stanford 9 test scores requires an arcane calculation process. To calculate the API, individual student scores, in national percentile ranks (NPRs), in each subject area on the SAT-9 are combined into a single number to represent school performance . First, student NPR scores for each subject test are categorized into one of five "Performance Bands." Next, the percentage of students scoring within each of the five performance bands is weighted by a different factor.[2] These weighted proportions are combined to produce summary scores for each content area. Results for content areas are then weighted and summed to produce a single number between 200-1000, representing the school's API score.[iii]

[2] The performance bands and associated weighting factors are as follows:

- Band 1, Far Below Basic: 1-19th NPR, weighting factor 200
- Band 2, Below Basic: 20-39th NPR, weighting factor 500
- Band 3, Basic: 40-59th NPR, weighting factor 700
- Band 4, Proficient: 60-79th NPR, weighting factor 875
- Band 5, Advanced: 80-99th NPR, weighting factor 1000.

An API score is calculated for each school every year.[iv] The current target established by the state for each school is to obtain an API score of at least 800. This interim target was established by the Advisory Committee for the Public Schools Accountability Act, based on data analyses by the Committee's Technical Design Group. The Group intentionally set the target at a demanding level to represent an exemplary level of performance.[v]

For those schools that do not meet this interim target of 800, an API Growth Target is calculated. The Growth Target is determined by subtracting a school's current API score from 800 and then multiplying the difference by 5%. The Target is compared with actual change, or growth, in API the following year. In this way, schools are expected to close the gap between their current performance and the target performance level by 5% each year. For schools that are within 20 points of the target, the API index is expected to grow by at least 1 point. (CDE, 2001c.)

Beyond meeting the 5% growth target, schools whose API score is below 800 are expected "to demonstrate comparable improvement in academic achievement by all numerically significant ethnic and socio-economically disadvantaged subgroups" (CDE, 1999a, p. 17).[vi]

Because tests administered in English do not provide reliable and valid scores for students with limited English proficiency (LEP), LEP students who have been enrolled in the public school system for less that a year are exempt from taking the SAT-9. For those LEP students whose first language is Spanish, the Spanish Assessment of Basic Education, 2nd Edition (SABE/2) is administered. Similarly, for LEP students who have

been enrolled in public schools for more than 12 months but who are not yet proficient in English, each school system may opt to administer the SABE/2. However, since this decision is left to the discretion of each school system, scores on the SABE/2 are not included as part of API scores.

2.1. Addition of Criterion-Referenced Tests to the API

Recognizing that the SAT-9 "has serious limitations as an accountability instrument for California public education" because "this test is not linked to California content and performance standards" (CDE, 1999a, p. 2), California Standards Tests (CSTs) are being developed for English Language Arts[3], Mathematics, History-Social Science, Science, Writing, and Coordinated/Integrated Sciences.[vii] Unlike the SAT-9,

[3] The English-Language Arts California Standards Test was administered in 2000 and 2001 and was integrated into the API in 2001. This new indicator accounts for 36% of the API for elementary and middle schools and 24% of the API for high schools. Table 2 below illustrates both the weighting between the Stanford-9 and the California Standards Test as well as the subject area weighting within the Stanford-9.

Table 2. Weights for API Components, 1999-2001

	API Component Weights for Base API		
	1999	**2000**	**2001**
Elementary/Middle (Grades 2-8)			
Stanford 9			
Reading	30%	30%	12%
Language	15%	15%	6%
Spelling	15%	15%	6%
Mathematics	40%	40%	40%
California Standards Test			
English Language Arts	N/A	N/A	36%
	100%	100%	100%
High School (Grades 9-11)			
Stanford 9			
Reading	20%	20%	8%
Language	20%	20%	8%
Mathematics	20%	20%	20%
Science	20%	20%	20%
Social Studies	20%	20%	20%
California Standards Test			
English Language Arts	N/A	N/A	24%
	100%	100%	100%

SOURCE: http://www.cde.ca.gov/news/releases2002/rel03.asp, Attachment B
The California Standards mathematics tests were administered in 2000 and 2001, but have not yet been deemed ready for inclusion in the API, although they may be included in 2002.

which is a nationally norm-referenced test, the CST tests are criterion-referenced and specifically designed to be in sync with state standards. Rather than comparing performance of students to each other (or more accurately to a norm group), criterion-referenced tests compare each student's performance to a defined standard of performance. Over time, the state intends to incorporate all of these CST scores into the API calculation. As the CST tests are developed and administered to students across the state, the API index will be modified to incorporate scores for these tests.[viii]

Test scores are not the only components under consideration for inclusion in the accountability index. The PSAA legislation mandates that measures such as student and teacher attendance rates and high school graduation rates be incorporated into the API calculation. Additional measures, whether they be attendance and graduation rates or scores from CST or high school exit exams, "are to be included only when available, valid, and reliable [Section 520252(b)]" (CDE, 1999a, p. 4).

In April of this year, the State Board of Education released a plan outlining its vision for a "substantially complete" API that is to be in place by 2006. The plan includes a switch in 2003 from the SAT-9 to a new norm-referenced test (NRT), namely the CAT-6. The State also intends to satisfy the other requirements of the PSAA by incorporating attendance rates, the California Alternate Performance Assessment (CAPA), the California High School Exit Exam (CAHSEE—a criterion-referenced test that students must pass to graduate) and high-school graduation rates into the API. Other indicators beyond those that are legally required may also be included in the calculation.

As CST tests are added to the API, it is possible that the proportion of students classified into each Performance Band based on the SAT-9 and the CST tests will differ. As a result, API scores for schools will differ depending upon which tests are included in the calculation. In an attempt to make the transition from an API based solely on the SAT-9 to one that includes both SAT-9 and CST scores, a "scale calibration factor" was used to adjust the integrated API scores for 2001.[ix]

As additional sections of the California Standards Tests are added in 2002, the same methodology is planned to be employed. Each time a new indicator is added, the API will need to be re-calibrated and two API scores will be calculated for that year, one containing the current indicators and one containing the indicators used for the prior year. This will ensure that growth API calculations will be based on scores from the same indicators. However, the introduction of the new NRT to replace the SAT-9 in 2003 may also disrupt the comparability of that year's API with previous years'. Thus, until the measures used to calculate the API scores stabilize, multiple API scores and a new scale calibration factor will likely be calculated each year.[x]

2.2. Programs that Exist to Assist Schools with Self-Improvement

The three groups described below each use the API in some way to assess school performance. Because these groups are not coordinated at the state level, it is all too often unclear who is ultimately responsible for educational improvements.

2.2.1. Immediate Intervention/Underperforming Schools Program (II/USP)

Schools that do not meet their API targets and that are in the lowest five deciles are labeled "underperforming" and are eligible for II/USP funding.[xi] A total of 430 schools are chosen to participate in the II/ USP each year. Schools are selected in proportionate amounts of middle, elementary and high schools and proportionate amounts of urban and rural schools. The II/USP legislation also requires that no more than 86 schools per API decile are chosen.[4]

In the first year of eligibility, a school receives a $50,000 state planning grant, which is to be used to hire an external evaluator, who will help the school develop an Action Plan. At a minimum, the evaluator is required to: inform parents and guardians that the school has been selected to participate in II/USP due to its below average performance, hold a public meeting where input from parents and guardians is solicited, and notify parents and guardians that they can submit recommendations for school improvement in writing to the evaluator. The evaluator must then collaborate with a community team, elected by the school board, to develop a documentation of the school's weaknesses and make recommendations for improvement. State law requires the action plan to "focus on improving pupil academic performance, improving the involvement of parents and guardians, improving the effective and efficient allocation of resources and management of the school, and identifying and developing solutions that take into account the underlying causes for low performance by pupils." (CDE, 2000d.) After the plan is completed it is sent to the Superintendent of Public Instruction for approval, with

[4] In 1999, of the 3,144 schools that were eligible to apply for II/USP funding, over 1,400 schools did (Perry & Carlos, 1999). From this list, 350 schools received state funding and an additional 80 schools received federal funding. The total dollar amounts set aside for each of these programs was $32.4 million from

a request for additional funding to implement the plan. If the plan is approved, the school receives an implementation grant through either state funds or through the Federal Fund Comprehensive School Reform Demonstration Program (CSRD) for up to $200/pupil with a minimum of $50,000/school.

If after twelve months of receiving funding, the school does not meet its short-term growth target, a public hearing must be held to "ensure that members of the community are aware of the lack of progress." The governing board of the school district then has the option of reassigning school personnel or making other changes that are deemed appropriate. After 24 months, if the school still does not meet its target, the Superintendent of Public Instruction assumes all legal rights, duties and powers of the governing board with respect to the school. (CDE, 2000d.)

2.2.2. Western Association of Schools and Colleges (WASC)

The Western Association of Schools and Colleges (WASC) is a private, non-profit group, responsible for accrediting public schools. Schools prepare a self-study for the initial accreditation process. Thereafter, a follow-up evaluation occurs on a six-year cycle. The current accreditation process, which is called Focus on Learning, was adopted in 1995. Prior to 1995, the accreditation team reviewed inputs, for example what teachers were doing, what their credentials were, what facilities were available. Since the adoption of Focus on Learning, the accreditation team now focuses on outputs, meaning student learning, without regard to the processes. The WASC team looks at multiple assessment means to assess student learning, including the SAT-9.

---

federal Comprehensive School Reform Demonstration program funds, and $63.7 million from the state General Fund. (http://www.cde.ca.gov/bills/sb1xhigh.htm).

2.2.3. Fiscal Crisis and Management Team (FCMAT)

A separate, independent entity known as the Fiscal Crisis and Management Team (FCMAT) was created in 1992 to perform school assessments. The five operational areas that the FCMAT assesses are: personnel management, financial management, facilities management, instructional management and governance/community issues. FCMAT determines which schools to assess based on how often emergency or permit credentials are requested and on the API. Since its inception, about 85% of FCMAT work has focused on management assistance and 15% on fiscal crisis intervention (Thomas Henry Deposition, 2001). In addition to these two independent organizations, the California Department of Education has its own accountability branch. But the lines of accountability are easily blurred between the state organization and the individual school districts. Paul Warren, Deputy Superintendent of the Accountability Branch, has said that the state's role in terms of accountability is to create the incentives for schools to "do the right thing" regarding student outcomes (Paul Warren Deposition, 2001). It is then the district's responsibility to implement an action plan according to its own specific situation. In this way, the state would play a regulatory role. But school districts often disagree with this description of accountability roles. They see the state as being accountable for implementing appropriate programs to achieve intended student outcomes.

3. CALIFORNIA'S OUTCOME BASED ACCOUNTABILITY SYSTEM CANNOT HELP STUDENTS RECEIVE THE KIND OF EDUCATION THEY DESERVE

3.1. Role of Tests/Student Assessment in State Educational Accountability Systems

As of this writing, educational accountability systems have been established in all 50 states. Accountability is also a critical component of national educational policy. At both the national and state level, student testing stands at the center of educational accountability programs. This dominance is made clear in *Education Week*'s now annual attempt to rate the quality of each state's standards and accountability system. As Orlogsky and Olson (2001) describe, the factors that influence ratings for standards and accountability include:

- whether the state tests students
- whether the tests are norm-referenced or criterion-referenced
- the subject areas tested (English, mathematics, science and social studies tests are required to receive "full credit")
- the type of test items used (multiple-choice, open-ended, essay, portfolio, etc.)
- the extent to which the tests are aligned with the state standards in elementary, middle and high school
- Whether the state requires school report cards (ratings of schools)
- Whether the state rates, rewards, sanctions, and/or provides assistance to schools based on student test scores

The emphasis placed on testing by *Education Week* is reflected in President Bush's *No Child Left Behind Act of 2001* (Public Law No: 107-110). As approved by both houses of Congress in January 2002, federal education funding requires that states implement tests for all students in grades 3-8 in reading and mathematics. As stated in the White House's summary of the legislation, "These systems must be based on challenging State standards in reading and mathematics, annual testing for all students in grades 3-8, and annual statewide progress objectives ensuring that all groups of students reach proficiency within 12 years" (source: www.whitehouse.gov/news/releases/2002/01/print/20020108.html). Although the President's education policy does not stipulate how states should use test scores, the legislation itself and rhetoric surrounding the legislation exemplify the extent to which many education and political leaders equate educational accountability with student testing.

3.2. Defining Accountability in Education

When accountability systems focus primarily (or exclusively) on test scores, educational accountability becomes defined as requiring schools to improve student test scores from year to year. At the national level and within nearly all states, changes in student test scores are the sole focus of accountability systems, with no reference to school policies and practices, or educational opportunities provided to students. In this way, the operational definition of accountability in education is based on a single set of student outcome measures, namely changes in test scores. To hold a school accountable one must only count the number of points by which students' scores change over the course of a year.

This working definition of accountability, however, differs noticeably from a more formal definition of accountability. According to the Oxford English Dictionary, the word accountability means: "The quality of being accountable; liability to give account of, and answer for, discharge of duties or conduct; responsible or amenableness" (p. 65). In this more formal definition, those charged with duties are expected to provide an account -- that is, a description and an explanation -- of their duties and conduct, in order to assist in determining whether said conduct was responsible. Where this more formal definition differs most notably from the working definition in education is in the active role the leaders play in telling the story of education in their school(s) and the extent to which this education is responsible to its constituents – students, families, and the community.

Absent from this formal definition is any mention of results or outcomes. Given that a key purpose of education is to help students learn, one component in determining how adequate education is might well be the impact education has on student academic learning. A test is unquestionably one tool that can measure this student learning. But, the outcomes of education certainly extend beyond student academic learning.

This report addresses the educational accountability system in the state of California from the perspective that such systems should result in an accounting – informing consumers about what schools are doing and how well. An accountability system that is test-based alone is, by definition, a limited one. And California's further reductionist approach of developing a single accountability "index" number does not inform schools, parents or students about the quality of education. As described below,

this index fails to provide information relevant to several aspects of education which California's schools explicitly state are part of their mission.

3.3. Mission of Education in California

To develop a sense of how schools in California describe their missions to students, families, and their community, a sample of 46 mission statements were analyzed. Using the U.S. Department of Education's Common Core of Data database, 50 California schools were selected at random. An internet-based search was conducted to find each school and its associated mission statement. Of the original sample of 50 schools, 19 mission statements were found. A second set of 50 schools was then selected at random, the search process was repeated and an additional 17 school mission statements were found. The process was repeated a third time, yielding a total of 46 schools. Using this sample of 46 schools, a systematic review of the words and phrases used in the mission statements was conducted.

Within this sample of mission statements, academic learning was only one component of school goals. While 83% of these mission statements specifically referenced academic learning and/or cognitive development, several other purposes were also mentioned. Among the more common goals were:

- Developing citizenship (52%)
- Ethical/Moral development (28%)
- Helping students reach their full potential (41%)
- Providing a safe environment (35%)
- Exposing students to modern technology (24%)

- Promoting a sense of community (26%)
- Appreciation of diversity/culture (28%)
- Creating a challenging (28%) and/or nurturing (48%) environment.

These are, arguably, all important aims for public education. However, they are outcomes ignored by California's API-based accountability system.

3.4. Disjuncture Between Educational Mission and Educational Accountability

Public opinion polls indicate that the general public also focuses on more than test scores when evaluating education. Respondents to a national survey were more concerned about schools' providing safe and drug-free environments than about student performance on tests. And when asked which aspects most influence their decisions about school quality, respondents placed more emphasis on the quality of teaching staff, adequacy of financial support, discipline, and class sizes than on the test scores of the student body. In fact, in 2001, 31% of respondents believed there is too much emphasis on testing in schools, an increase from 20% in 1997 (Rose & Gallup, 2001). Among non-white respondents, this percentage has increased from 27% in 2000 to 42% in 2001.

While our reviews of school mission statements and public opinions regarding education are in no way exhaustive, their story is consistent: it is the view of the public and of the schools themselves that increasing student academic achievement is only one duty with which public schools are charged.

4. THE API IS NOT EVEN AN ADEQUATE OR USEFUL MEASURE OF STUDENT ACADEMIC ACHIEVEMENT

An API of 800 (or any value for that matter) does a poor job of characterizing the actual performance of students in a school.[xii]

The distribution of student performance on the SAT-9 in California is noticeably lower than the national norm group. As Herman, Brown and Baker (2000) report, nearly a fifth of California's students are not proficient in English as compared to less than two percent nationwide. This, and other differences in demographics, contribute to performance that is well below the national average. In grades 2 through 11, mean percentile ranks in 1999 ranged from the 32nd to the 46th percentile on the SAT-9 reading test, and from the 44th to the 52nd percentile in mathematics. Given these starting points, the lofty target of 800 establishes an admirable goal, but destines many schools to failure.[xiii]

Even if a school is successful in increasing the performance of students as they progress through the grade levels, each year two new group of students whose skills and knowledge are distributed "normally" enter a given school. One set of entering students becomes the lowest grade level in the school. In a K-5 school, this set of entering students becomes the Kindergarten class. Some of these students may come from any number of pre-school programs. Others may not have attended pre-school at all, while still others may have recently arrived in the U.S. Similarly, in grade 6-8 middle schools, the entering set of students, which becomes the sixth grade, come from elementary schools in the same district or in other districts. In addition, a fraction of these students may also be recent immigrants. The same entry pattern occurs in high schools. A second

entering set is composed of students who move into the school during the course of the school year. Some of these students may come from other schools in the same district, from other districts in California, from other states, or from other nations. (Note that a student's scores are not include in the school API unless they have been in the school for at least one year.)

Whatever the grade levels served by a school, the current API system, which compares cross-sectional performance across years, holds schools partially responsible for skills and knowledge that students may or may not have acquired before entering the school. Schools that serve large groups of disadvantaged students, students whose primary language is other than English, or students who enter from "poorly" performing schools, must have extraordinary impacts on student learning during the first year(s) of the students' educational experience in that school in order to obtain the target API.

For example, based on past and current performance of California's LEP students on SAT-9, one might expect to find that kindergarten students (or any students newly arrived in California from most other countries whose primary language is not English) would perform considerably below the national mean. If these assumptions hold, it is also reasonable to assume that even if the school is extraordinarily successful in improving the performance of these entering groups of students, by the time they are in Grade 2, and eligible for taking the tests that make up the API, a large minority of the students would still be performing below the national mean. The larger the LEP population of these entering groups, the larger the percentage of students still performing below the mean is likely to be. Thus, without some sort of miraculous effect during the two-and-a-half years of each child's formal schooling, a substantial portion of the second

grade class will perform below the national mean. In turn, the "poor" performance of this segment of the second grade must be offset by much higher performance of students in upper grade levels in order to generate a high API. [xiv]

The point here is that with the incredibly high performance target of 800, schools – especially those with a high proportion of incoming second-language speakers – are required to dramatically (perhaps, impossibly) alter the shape of the achievement distribution to one shaped quite different from that for the nation as a whole.

Not only that, but, the gains students must make are not on the same test. Rather, the gains must be made on the test for the next grade level. [xv] While some of the subject matter overlaps across years, additional skills and knowledge are required to perform at the same level from year to year. Although often misinterpreted as showing no growth, percentile ranks that remain the same across years actually represent substantial growth – growth that is identical to that of the average student nationwide. And increases in percentile ranks across years indicates even more growth than that of the average student nationwide.[xvi]

Two additional problems with the API further demonstrate that it is often irrational to use it as a diagnostic tool:

- Measurement error impacts the reliability of scores and score changes, so individual test scores will always be to some degree volatile.[xvii]

- As Haney and Raczed (1994, p. 17) state, "attempting to hold schools or educational programs accountable in terms of aggregate performance of students almost guarantees that accountability in terms of *explaining* patterns

of individual students' learning will be largely impossible."[xviii] Depending upon the level of aggregation, correlations can vary dramatically.[xix] This problem led one scholar "to warn of the ecological fallacy: the fallacious inference that statistical relationships discovered in analyzing aggregate data (such as class, school, district or state level average test scores) also pertain at the level of individuals" (Haney & Raczek, 1994, p. 20).[xx]

4.1. Score Gains are Deceptive

It is often assumed that an increase in test scores represents an increase in learning or ability. Thus, the higher a student scores, the more that student is said to have learned. Over the past decade, however, several studies suggest that this assumption becomes tenuous when schools are mandated to increase scores on a standardized test administered over several years.

4.1.1. Lessons from Kentucky

During the 1990's, Kentucky put into place a complex, multiple-measure assessment system. Between 1992 and 1996, student scores on these assessment instruments increased. In 1998, Koretz and Barron performed a series of analyses to examine the validity of these gains. Among their findings were:

- Fourth and Eighth grade teachers believed that gains in scores were more a reflection of students' becoming familiar with the tests and their formats than of changes in students' knowledge and skills.
- Score gains on KIRIS did not translate to score changes on other related tests. As an example, fourth-grade KIRIS reading scores increased by

three-fourths of a standard deviation but scores did not change on NAEP. While math scores on KIRIS and NAEP increased across the four years, the gains on KIRIS were about 3.5 times larger than the gains on NAEP. Similarly, increases in high school KIRIS scores were not mirrored by increases in ACT scores.

- Performance on items that were re-used was noticeably higher than performance on items that were used only once. This suggests that student increases may be partially due to familiarity with the items.

In addition to these findings, Koretz and Barron noted that the initial score gains for many of the tests were "very large relative to past evidence about large-scale changes in performance, and several were huge" (p. 114). The authors proceed to explain that "meaningful gains of these magnitudes would be highly unusual, but observed gains of this size are less surprising. It is common to find large gains in mean scores during the first years of administration of a new assessment, in part because of familiarization."

4.1.2. Lesson from Texas

Texas has had its accountability system (Texas Assessment of Academic Skills or TAAS) in place for over ten years now. Over this time period, the percentage of students passing TAAS has increased dramatically.

During 1999 and 2000, Haney undertook a series of analyses to investigate possible alternative explanations for these increases. Like Koretz and Barron, Haney compared performance on TAAS to performance on other indicators of student learning. Briefly, Haney found:

- Little relationship between changes in TAAS scores and high school grades
- Large gains on TAAS were not mirrored by changes in scores on the Scholastic Aptitude Test (also known as the SAT)
- Gains on NAEP were about one-third the size of gains on TAAS, and when gains on NAEP are adjusted for Texas' unusually large exclusion rate, the gap increases further.

Haney also presents considerable evidence that much of the gains in TAAS scores resulted from increases in retention and drop-out rates rather than increases in learning. These, and other findings, led Haney to conclude that the "Texas 'miracle' is more myth than real."

4.1.3. The Lessons Apply in California

SAT-9 scores in California have increased between 1999 and 2001. Across all grade levels, the largest gains occurred during the first year.[xxi]

It is interesting to note that the pattern of gains on SAT-9 in California are similar to gains in Kentucky during the early years of KIRIS—the sharpest gains occurring during the first two years, after which the gains flatten out.

Below, I present data that may begin to explain some of the causes for these early increases. Not surprisingly, these causes are similar to those associated with score gains in Kentucky and Texas, and include a focus on test-taking skills, teaching to the test, and increased retention in some schools.

As an additional indication that the score increases had less to do with learning, and more to do with factors like those just mentioned, one can look at the fact that, as was the case with the standardized tests in Kentucky and Texas, the sharp increases in California on the SAT-9 do not generalize to the NAEP (National Assessment of Educational Progress).[xxii] Whereas California's grade 4 SAT-9 Math scores saw a sharp increase, California's grade 4 NAEP Math scores increased at about the same rate as those of the nation. And, whereas California's grade 8 SAT-9 Math scores increased slightly between 1998 and 2001, California's grade 8 NAEP Math scores decreased slightly between 1996 and 2000 while the national average increased. Thus, whereas California's grade 4 SAT-9 Math scores suggest that California gained sharply on the nation, California's grade 4 NAEP Math scores suggest that the gain was negligible. And whereas California's grade 8 SAT-9 Math scores suggest that California gained on the nation, California's grade 8 NAEP Math scores suggest that the gap between the state and the nation actually increased.

Consider also the California Writing Standards Test (CST Writing test), which was administered for the first time to 4th and 7th grade students this past year (2001).[xxiii] If one believed that the increases on SAT-9 represented actual increases in students' language arts skills, one might have expected students to have performed at least moderately well on the CST Writing test. Predictably, however, both grade levels performed very poorly on these tests.[xxiv] The large difference between student performance on the SAT-9, the recent increases in SAT-9 reading scores, and the very poor performance on the 2001 writing test lead to one of two conclusions: Either the SAT-9 scores are inflated and do not represent the achievement level of 4th and 7th grade

students in California, or the performance standards for the CST Writing test were exceedingly high.

5. CALIFORNIA'S ACCOUNTABILITY SYSTEM IS A PRODUCT OF QUESTIONABLE POLICY DECISIONS MADE BY STATE OFFICIALS

The 1999 PSAA legislation, and subsequent amendments, defines in a general way what the API should be and what level of performance is required for a school to be considered successful. However, the process of defining the components of the law to a level of specificity adequate for actual implementation, involves a committee of expert advisors interpreting the intent of a policy written by legislators, and making choices about how the system should be carried out. Like any choices, the selected definition of API and the targeted performance level have consequences for California students and schools.

On the surface, the single API target of 800 seems deceptively simple – achieve at least this number, and your school is successful. Although the end-goal of the API System Index is to summarize school performance with a single, seemingly precise number, the factors and weightings used to produce that single number are based on informed, but nonetheless subjective, decisions.[xxv]

While decisions about some of these variables were informed by simulations and modeling conducted by members of a technical advisory committee, it is not clear how scientific the decision-making process was. Given the important consequences for schools based on API scores, one would hope that the decision-making process was deliberate and thoughtful. Yet, available documentation from the California Department

of Education presents the process of selecting values for this system as a murky one, carried out quickly to ensure that a law approved by the governor in April, 1999 could be implemented by that July (the State Board of Education actually adopted an API definition in November, 1999).

To get a sense of the decision making process, and to see that alternative decisions were possible, refer to Appendix A where I outline some of the key decisions made during several of the meetings that led to the current API index.

The most important thing to understand is that human judgments continue to affect the way in which API scores are calculated, and that many of the decisions that have resulted in the current system appear to have been more arbitrary than methodical. See Appendix B for a detailed analysis of how minor changes in those decisions can have major effects.

5.1. One Example of How a Questionable Policy Choice Affected the API: The Decision to Use the SAT-9

The PSAA Advisory Committee released a final report for the 1999 API in October of 1999. The report opens with four concerns of the committee, including concern about the limitations of the Stanford 9 as the sole accountability measure for California and about unintended consequences. As they stated,

> Reluctantly, the Committee has arrived at the conclusion that for 1999, the API should consist solely of norm-referenced test results from the Stanford 9 administered as part of the Standardized Testing and Reporting (STAR) Program. The Stanford 9, however, has serious limitations as an accountability instrument for California public education. The norm-referenced component of this test is not linked to California content and performance standards. As a result, the Committee advocates that as soon as possible the SBE base the API

> predominately on measures linked to California's content and performance standards….The introduction of a high-stakes accountability system may result in unintended and undesirable consequences as schools strive to obtain rewards and to avoid sanctions.…A major priority of the accountability system must be to identify, evaluate, and mitigate unintended consequences. (CDE, 1999a, p. 2.)

One might question whether a test with "serious limitations" that is unrelated to current educational standards should hold such high stakes for students and schools in California. Clearly, this decision was a matter of judgment and, without a better-aligned test in hand, was deemed the best alternative.

In another exercise of judgment, the SBE recently decided to replace the SAT-9 in 2003 with a new norm-referenced test that is equally indifferent to California standards. Not only will the new NRT have the same "serious limitations" that discredited the SAT-9, it will also hamper the State's ability to produce a consistent, comparable API.

6. THE API ENCOURAGES POOR EDUCATIONAL PRACTICES

State educational leaders establish test-based accountability systems to motivate teachers and schools, to improve student learning and to encourage schools and teachers to focus on specific types of learning. Some observers have raised concerns that this encouragement to focus on specific types of learning too often translates into "teaching to the test." As Shepard notes, however, teaching to the test means different things to different people. In many cases, state and local educational leaders, as well as classroom teachers, interpret this phrase to mean "teaching to the domain of knowledge represented by the test" (Shepard, 1990, p. 17) rather than teaching only the specific content and/or items that are anticipated to appear on the test.

6.1. Previous Findings in other States

As part of her examination of state-level testing programs, Shepard interviewed state testing directors in 40 high-stakes states. As she describes:

"When asked, 'Do you think that teachers spend more time teaching the specific objectives on the test(s) than they would if the tests were not required?' the answer from the 40 high-stakes states was nearly unanimously, 'Yes.' The majority of respondents went on to describe the positive aspects of this more focused instruction. 'Surely there is some influence on the content of the test on instruction. That's the intentional and good part of testing, probably.'…Other respondents (representing about one-third of the high-stakes tests) also said that teachers were spending more time teaching the specific objectives on the test but cast their answer in a negative way: 'Yes. There is some definite evidence to that effect. I don't know that I should even say very much about that. There are some real potential problems there…Basically the tests do drive the curriculum'" (p. 18).

6.2. Influence on Instruction

There is clear evidence that past educational leaders in California not only hoped that the state tests would influence what teachers taught, but also how they taught. As Cohen and Hill (2001) recount, CLAS (described above in section 2.2) was established in part to motivate teachers to adopt new forms of instruction, particularly for mathematics:

> Reformers who wanted to make instruction more 'accountable' argued that teachers would not move far from algorithms and rote memorization if the old tests, which called for exactly such work, remained in place. New tests, tied to

> the new framework, would remove the incentive for teachers and students to focus solely on 'the basics' and might provide an incentive for teachers to modify instruction to match the framework more closely. Those who cared about performance on the test might be motivated to investigate new forms of instruction or curriculum (p. 28).

Although a different set of educational leaders were at the helm when the STAR and PSAA were developed, the intent to influence instructional practices was implicit in the Academic Performance Index Framework developed by the State Board of Education. The third point in this framework states: "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" (CDE, 1999c). Clearly, by emphasizing that the content, skills and competencies tested must be teachable, the Board anticipated that schools and teachers would endeavor to teach them.

But beyond influencing what and, potentially, how teachers teach, state accountability programs that rely heavily (or entirely) on test results can have negative consequences. Recognizing the potential for negative consequences, Linn, Baker and Dunbar suggest that the extent to which unintended positive and negative consequences result from the standards-based assessment system should be examined (1991). Among the negative consequences, Herman, Brown and Baker (2001) list increases in retention, increases in drop-out rates, narrowing of the curriculum, and decreased attention to topics and subjects not tested. In addition, some observers have noted that the high stakes associated with some state-level testing programs lead to questionable educational practices such as focusing instruction on test-taking skills, falsely classifying poor-

performing students as SPED so that their scores are excluded from averages, altering test administration conditions, providing inappropriate instruction during testing, and, in some extreme cases, altering student response sheets.

6.3. Influence on Retention and Drop-outs

As Herman, Brown and Baker (2000, p. 9) state, "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." There is ample evidence that this unintended outcome is occurring in other states. As an example, Haney (2000) reports a clear relationship between the rise in student scores on Texas' TAAS and decreased graduation rates. In Texas, this relationship is stronger for Blacks and Hispanics than it is for White students.

Although the legislature intended to include attendance data and other indicators of school performance in the calculation of the API, as of today, drop-out rates <u>are not</u> included in the API. California currently does not have an accurate way of collecting drop-out rates for all California schools. As noted by Herman, Brown, and Baker (2000), the data "are often unreliable or inaccurate because schools across the state do not use uniform definitions or share equally careful procedures for collecting the data." Although California should be moving towards a statewide student data system that would permit more precise understanding of indicators such as drop-out and retention rates, the CSIS has yet to be uniformly implemented in California public schools, and according to the Superintendent's Advisory Committee on the PSAA, its full implementation is likely years away. Without such a system, as the State acknowledges, it has no means of accurately measuring drop-out rates, which may dramatically affect

school-average performance on tests, and which may in turn be affected by perverse incentives to increase school-average scores.

Drop-out rates are important information because they could enable the state to ensure that improvements in test scores are not coming at the cost of having more students pushed out of school. (Herman, Brown, and Baker, 2000, p.135.) California's current statistics on dropouts might be misleading, as definitions of dropouts vary from district to district (Herman, Brown, and Baker, 2000), and are probably under-inclusive, thereby hiding potential dramatic unintended consequences of high-stakes testing on drop-out rates. We used available data to calculate what imputed drop-out rates[5] might be in high schools within the Los Angeles Unified School District, and found disturbing results. In high schools with a high percentage of low income students (students receiving free or reduced price meals), imputed drop-out rates are staggering. For instance, Jefferson High School has an imputed drop-out rate of 69.23% with 81% of students receiving free /reduced price meals, and a 2001 API score of 429. (According to CDE data, Jefferson has a 4-year derived dropout rate of 24.6 % and a 1-year drop-out rate of 6.4%.) Similarly, Garfield High School has an imputed drop-out rate of 63.19%, 82% of students receiving free/reduced price meals and an API score of 487. As it was the case in Texas, there is reason for serious concerns about high level of drop-out rates in

[5] Because student level data are currently unavailable, imputed drop-out rates were calculated using the following formula: (1998 9th grade enrollment – 2001 12th grade enrolment)/1998 9th grade enrollment. The imputed drop-out rates do not account for demographic variations, such as students transience rates, migration rates, or retention rates, but provide a snapshot of what drop-out rates might be.

LAUSD high schools, and the State should conduct an in-depth analysis of the potential relationship between high-stakes testing and high levels of student drop-out rates.[6]

6.4. Patterns Emerging in California

As the Advisory Committee specifically states: "A major priority of the accountability system must be to identify, evaluate, and mitigate unintended consequences." (1999, p. 3.) In sections 6.5 through 6.9 I examine some of the consequences, both positive and negative, that are emerging in California. The analyses that follow are based on an examination of responses from a random sample of California teachers representing 433 respondants to a survey administered in the late winter of 2001 to teachers across the nation by the National Board on Educational Testing and Public Policy.[7] Specifically I use data from this survey to explore the potential impact of the testing program on five broad areas: Alignment of Instruction to the Standards and the Test, Changed emphasis on tested and non-tested subjects, Preparation for tests, Conduct of questionable educational practices, and General Beliefs About the Test.

6.5. Alignment of Instruction to State Standards and Tests

Table 16 summarizes responses from a sample of California teachers to several questions on the NBETPP survey that focus on the relationship between curriculum and instruction and the state standards. In general, the majority of teachers agree that their district's curriculum is aligned with the test. Teachers also appear to be designing tests in

---

[6] The data collected for LAUSD point to other indicators that should be of interest to the State, such as whether a school is on a year-round schedule.

[77] A stratified random sampling method was used. States were first classified into one of nine bands based on the stakes for schools and students associated with their testing program. Within each of these bands,

the classroom that have the same format as the state tests, but not necessarily the same content. Less than half the teachers also believe that the tests are not compatible with their daily instruction or their instructional materials. Moreover, nearly three-quarters of the teachers believe that the testing program is leading some teachers to teach in ways that are not consistent with what they believe is good practice.

Table 16. Percent of California Teachers Reporting Specific Effects of Tests

Please indicate the extent to which you agree with each of the following statements	**Agree**
The state-mandated test is compatible with my daily instruction	42.5%
My district's curriculum is aligned with the state-mandated testing program	61.9%
The instructional texts and materials that the district requires me to use are compatible with the state mandated tests.	40.0%
If I teach to the state standards or frameworks, students will do well on the state-mandated test.	47.8%
My tests are in the same format as the state-mandated test.	77.4%
My tests have the same content as the state-mandated test.	40.4%
The state-mandated testing program leads some teachers in my school to teach in ways that contradict their own ideas of good educational practice.	73.2%

teachers were stratified by location of school (urban or non-urban), grade level and subject area (when appropriate).

6.6. Changed Emphasis on Tested and Non-Tested Subjects

Table 17 summarizes teacher responses to survey items asking about changes in instructional emphasis due to the PSAA. Many teachers report that the amount of time spent on several activities has changed in response to the state testing program. Not surprising, the vast majority of teachers report that the amount of instructional time on subjects that are tested increased. Conversely, over half the teachers report that instruction on non-tested areas has decreased. Specifically, teachers indicate that instruction in the fine arts, physical education and foreign language have decreased; time in fine arts most markedly. Finally, just over a quarter (28%) of the teachers also reported that teachers in their school do not use computers when teaching writing because the state-mandated writing test is handwritten.

Table 17. Percent of California Teachers Reporting Changes in Instructional Emphasis

In what ways, if any, has the amount of time spent on each of the following activities changed in your school in order to prepare students for the state-mandated testing program?	**Decreased**	**same**	**increased**
Instruction in tested areas	2.5%	17.4%	80.1%
Instruction in areas not covered by the state-mandated test	58.1%	34.5%	7.4%
Instruction in fine arts	63.7%	30.2%	6.1%
Instruction in physical education	50.3%	49.0%	0.7%
Instruction in foreign language	41.5%	53.5%	5.0%

6.7. Preparation for State Tests

Table 18 indicates that teachers in California are employing a variety of methods to prepare students for the state tests. The most common practices include teaching specific test-taking skills, encouraging students to work and prepare, and teaching the standards. In addition, teachers provide students with items similar to those on the test and/or use preparation materials developed by someone outside of the school. Surprisingly, 8.5% reported providing students with released items (this is surprising because the items are not released but the items are re-used each year and can be acquired through a direct request to the publisher) and 6.8% indicated that the did not provide any special preparation for the test.

Table 18: Percent of California Teachers Reporting Preparation Activities for State Test

How do you prepare your students for your state-mandated tests?	**Yes**
I do no special test preparation	6.8%
I teach test-taking skills	86.5%
I encourage students to work hard and prepare	76.0%
I provide rewards for test completion	14.1%
I teach the standards or frameworks know to be on the test	69.1%
I provide students with items similar to those on the test	64.4%
I provide test-specific preparation materials developed commercially or by the state	54.6%
I provide students with released items from the state-mandated test	8.5%

6.8. Conduct of Practices of Questionable Educational Value

Beyond specific preparation for the test, teachers indicated that the testing program is impacting the atmosphere within schools and classrooms. Nearly two-thirds of teachers believe that students are under intense pressure to perform well on the tests, but just over half of the teachers believe that their students feel destined to do poorly on the test no matter how hard they try. Only 7% of teachers believe that the tests are motivating students who were previously unmotivated. Similarly, pressures within schools have led nearly two-thirds of teachers to focus solely on test preparation. In many cases, this preparation has led to practices that improve test scores but not learning. And a third of teachers report that retention has increased because of the tests.

Table 19. Percent of California Teachers Reporting Practices of Questionable Educational Value

Please indicate the extent to which you agree with each of the following statements	**Agree**
Many students in my class feel that, no matter how hard they try, they will still do poorly on the state-mandated test.	51.8%
There is so much pressure for high scores on the state-mandated test that teachers have little time to teach anything not on the test.	66.4%
The state-mandated test has brought much needed attention to education issues in my district.	46.4%
Students are under intense pressure to perform well on the state-mandated test.	66.4%
Teachers in my school have found ways to raise state-mandated test scores without really improving students learning.	41.7%
State-mandated testing has caused many students in my district to drop out of high school.	21.6%
State-mandated test results have led to many students being retained in grade in my district.	33.1%

The state-mandated test motivates previously unmotivated students to learn.	7.8%

6.9. General Beliefs of Teachers about the State Accountability and Assessment Practices

Table 20 also indicates that teachers are under pressure from administrators to increase scores. However, just over a quarter of teachers feel the testing program is worth the time and money. Similarly, just over a fifth of teachers feel that test scores reflect the quality of education students receive and the vast majority of teachers believe that the test scores do not provide an accurate measure of what minority or LEP students have learned. Moreover, 73% of teachers believe that changes in scores from year to year reflect changes in the characteristics of students rather than school effectiveness. Finally, less than half of the teachers believe that if they teach to the state standards, students will do well on the tests.

Table 20. California Teachers' Attitudes Toward Testing

Please indicate the extent to which you agree with each of the following statements	Agree
Overall, the benefits of the state-mandated testing program are worth the investment of time and money	27.6%
Scores on the state-mandated test accurately reflect the quality of education students have received.	21.0%
Teachers feel pressure from the district superintendent to raise scores on the state-mandated tests.	90.3%
The state-mandated test is NOT an accurate measure of what minority students know and can do.	85.4%
Score differences from year to year on the state-mandated test reflect changes in the characteristics of students rather than changes in school effectiveness.	72.6%
If I teach to the state standards or frameworks, students will do well on the state-mandated test.	47.8%
The state-mandated test is NOT an accurate measure of what students who are acquiring English as a second language know and can do.	96.1%
Differences among schools on the state-mandated test are more a reflection of students' background characteristics than school effectiveness.	82.9%
There is so much pressure for high scores on the state-mandated test that teachers have little time to teach anything not on the test.	66.4%
The state-mandated test has brought much needed attention to education issues in my district.	46.4%
State-mandated testing has caused many students in my district to drop out of high school.	21.6%
Teachers feel pressure from the building principal to raise scores on the state-mandated test.	82.6%

7. WHAT MUST BE DONE?

Although student test scores have become the predominant form of "educational accountability" in most states, it is a seriously flawed approach to helping schools improve teaching and learning. As Amrein and Berliner (2002), Koretz and Barron (1998), Haney (2000), and several other researchers have consistently shown, test-based educational accountability systems create more problems than they solve. While scores on the state tests often increase (creating the illusion of improved learning), more often than not, this "improved" learning does not translate to other tests. While these test-based systems are intended to better prepare students for college and the workplace, they do not lead to increases in SAT or ACT scores, increases in the percentage of students attending college, increases in college completion, or improvements in college readiness. To the contrary, evidence is beginning to emerge that college-going students are in fact less prepared to excel in post-secondary studies (Haney, in press).

Why is test-based accountability failing? The answers are numerous. As is explained in greater detail in Section 8, a sole focus on changes in test scores takes attention away from quality, well-rounded instruction across the disciplines and instead focuses instruction narrowly on what is tested. As an example, Amrein and Berliner (2002) recount instances of a narrowing of the curriculum in California. Here, I quote the authors at length:

> The curriculum was so narrowed to reflect the high-stakes SAT 9 exam, and the teachers under such pressure to teach just what is on the test, that

they [the teachers] felt obliged to add a half hour a day of unpaid time to the school schedule. As one teacher said:

"This year [we]…extended our day a half hour more. And this is exclusively to do science and social studies. …We think it's very important for our students to learn other subjects besides Open Court [a language arts curriculum] and math…because in upper grades, their literature, all that is based on social studies, and science and things like that. And if they don't get that base from the beginning [in] 1st [and] 2nd grade, they're going to have a very hard time understanding the literature in upper grades… There is no room for social studies, science. So that's when we decided to extend our day a half hour…But this is a time for us. With that half hour, we can teach whatever we want, and especially in social studies and science and stuff, and not have to worry about, 'OK, this is what we have to do.' It's our own time, and we pick what we want to do." (Interview, 2/19/01)

In this school the stress to teach to the test is so great that some teachers violate their contract and take an hourly cut in pay in order to teach as their professional ethics demand of them. Such action by these teachers – in the face of serious opposition by some of their colleagues – is a potent indicator of how great the pressure in California is to narrow the curriculum and relentlessly prepare students for the high-stakes test. The paradox is, that by doing these things, the teachers actually invalidate the very tests on which they work so hard to do well. It is not often pointed

out that *the harder teachers work to directly prepare students for a high-stakes test, the less likely the test will be valid for the purposes it was intended.*

In this example, several problems are identified. First, teachers recognize that the instructional practices they feel forced to employ in order to prepare students for state tests are not good practices. Second, some teachers are seeking their own remedies to providing students with a higher quality education. In the example above, teachers are effectively taking pay cuts and violating their contracts to lengthen the school day. It is unclear how long teachers will be willing to sustain these practices before they return to a normal school day or leave the profession. Third, the remedies sought by some teachers causes tension between teachers and in turn negatively impacts school morale. Fourth, the intense focus on topics included on the test invalidates the test scores as measures of increased learning. Fifth, students are denied the opportunity to acquire skills and knowledge that are not tested but are important for later course work and through out life.

This intense focus of instruction on what is tested might not be as problematic if tests tested everything that was important for students to know. But they do not, time does not permit them to do so, nor is it necessary to test everyone on everything.

Even more importantly, the single-minded focus on test scores does not provide any incentive for schools to improve their practices or to better serve students' long-term educational and social needs. Beyond taking instructional time away from non-tested areas, there is clear evidence that schools are engaging in questionable practices to improve test scores. In the worst cases, these practices include outright cheating. Mr.

Warren stated that he believes that testing irregularities are going to become an increasing problem and yet the CDE plays a minimal role when irregularities are discovered. (Minutes, Superintendent Advisory Committee PSAA, May 25, 2000.) Although there have been numerous reports of cheating, the State relies on districts alone to investigate testing irregularities, and for administering sanctions. (See Philip Spears Deposition, p.171-262).

In other cases, the practices are more subtle such as reclassifying students as special needs so that they are not tested or their scores are not included in the school average, encouraging students to be absent on the day of testing, retaining students, and/or counseling students to seek other avenues for education such as a GED or attending an alternative school. In short, the result of test-based accountability is less concern about how to improve student learning by improving the conditions that affect learning. Instead, the focus is on obtaining prescribed changes in test scores.

Reducing these problems requires an improved accountability system. To improve the current system, the types of information considered by the system must be expanded to include inputs and well as output. As an example, past research shows that several factors outside of a school's control correlate with student achievement. Yet, despite these external factors, schools still play an important role in improving student learning. One of the key variables under the control of schools that has been shown to influence student learning is the quality of teachers and the instructional practices employed by teachers (see Wenglinski, 2002, for a review of the research). To increase the quality of teachers in California's schools, the CDE requires teachers to meet specific requirements in order to be credentialed. Despite these efforts, however, a shortage of

teachers in California has forced many schools to hire teachers who have not yet met all requirements to be fully credentialed. In such cases, teachers are given Emergency Credentialing. As is shown in Table 21, there is a clear relationship between the percentage of emergency credentialed teachers within a school and API scores – as the percentage of Emergency Credentialed Teachers increases, API scores decrease. Similarly, there is a clear relationship between the socio-economic status of students within a school and the percentage of emergency credentialed teachers in a school – the lower the SES level, the higher the percentage of emergency credentialed teachers. As one might expect, there is also a clear relationship between SES levels and API scores – the lower the SES level, the lower the API scores. While several factors combine to influence the relationship between SES and API scores, teacher quality (as represented by Emergency Credentialing) is one key factor.

Table 21. Correlations of Selected Students and School Characteristics with API Scores, All Schools 2000

	API	% Emergency Credentials	% Free/ Reduced Lunch	% English Learners	% First Year in School	% Parents Not HS Graduates
API		-.46	-.81	-.68	-.21	-.73
% Teachers Emergency Credentials	-.46		.36	.36	.19	.34
% Students Free/Reduced Lunch	-.81	.36		.76	.20	.75
% Students English Learners	-.68	.36	.76		.06	.77
% Students First Year in this School	-.21	.19	.20	.06		.10
% Parents Not HS Graduates	-.73	.34	.75	.77	.10	

Given these relationships, one first step toward improving the performance of students is to replace emergency credentialed teachers with teachers that are fully credentialed. Including a measure of the percentage of Emergency Credentialed teachers in a school in the API would provide an important piece of information, and benchmarks for a desirable level of Emergency Credentialed teachers could be established (most likely, 0%). But teacher quality is only one of many inputs that may be in need of improvement. Others include adequate textbooks, curricular materials, access to current technology, classrooms and schools that are not overcrowded, sanitary conditions, an environment conducive to learning, etc.

7.1. Alternatives to the Current API-based System

As described in Section 3, California has introduced a variety of assessment and accountability systems over the past decade. While much can be learned from these various systems and components of some of these systems represent alternatives to the current API-based system, none come close to an accountability system that is likely to prevent, detect, or deter gross disparities in education or lead to meaningful improvement in the quality of education across California's public schools.

7.2. Learning from Other States

To develop a sense of reasonable alternatives, it is useful to examine systems in other states. To this end, aspects of Connecticut and Rhode Island's accountability systems and a comprehensive system proposed for Massachusetts serve as good models.

7.2.1. Accountability in Connecticut

Connecticut's state assessment system has been in place since 1986. This long term commitment to one system allows CT to track long term trend information and allows students and teachers to more easily understand the purposes and logistics of the program. The state-wide tests are the Connecticut Mastery Test (CMT), which is administered in grades 4, 6 and 8, and the Connecticut Academic Performance Test (CAPT) which is administered in grade 10. Both the CMT and CAPT are criterion reference tests and use multiple choice, grid-in, open ended, essay and performance-based items to capture student knowledge of Mathematics, Reading, Writing and Science. Connecticut recognizes and explicitly states the disadvantages of using test scores as a single measure of student achievement. When writing about CAPT results, the Commissioner of Education states: "(test results) do not provide a comprehensive picture of student accomplishments. There is a danger that overemphasizing state test scores to evaluate a student's, a school's or a district's performance can result in an inappropriate narrowing of the curriculum and inappropriate classroom instructional practices. Focused preparation for state tests should be a small fraction of a yearlong comprehensive curriculum that balances the competencies assessed on the state tests with other critical skills and objectives." (Connecticut Board of Education, 2001b.) Connecticut's five year plan (2001-2005) outlines several complimentary objectives: curriculum development through statewide frameworks, student assessment, teacher quality, accountability, equalization of school resources, targeted categorical aid for the state's neediest districts and efforts to reduce racial, ethnic and economic isolation. In order to measure success of student achievement, CT plans to focus on CMT/CAPT tests,

SAT scores, graduation rates and participation in community service programs. The test score data will be disaggregated according to economic groups, racial groups, students with disabilities, bilingual students, magnet and charter school students, Title I districts and schools and vocational-technical school students.

7.2.2. Accountability in Rhode Island

Rhode Island's current state assessment system was drafted in 1996 and signed into law in 1998. The basis for Rhode Island's system is the state frameworks. These frameworks address content standards, teaching practice, assessment and evaluation, curriculum scope and sequence, equity and access, and family and community involvement. The wide scope of the frameworks translates to multiple measures to define school success: "A wide range of data is used to make judgments about school and district needs including state assessment data (especially the disaggregations and modeling results), SALT survey data, field service team familiarity with the school and district, and local data." (Rhode Island Board of Education, 2001.) The state allows schools and districts to set three year targets for their academic growth, with input from the Department of Education. Growth is reported as a three year rolling average. Rhode Island realizes the problems with setting yearly targets: "a single year of data can woefully misrepresent the trend of a school. A particularly exemplary or challenged class can skew the results and either inflate or deflate the real achievement of the school as a whole. . . Other states have found themselves accidentally setting some schools up for failure by establishing unrealistic and undeliverable goals and setting targets that were not challenging enough for other schools." (Rhode Island Board of Education, 2001, p. 29) Rewards are not given based on achieving the self-established three year goal, but if

a school falls short of their goals "a series of support and intervention strategies" are put into effect.

Student performance on the state tests are not viewed as the only indicator of school performance. Information Works!, which is Rhode Island's annual publication of the state of schools, lists the following as diagnostic tools used for assessment:

- Basic school-level statistics: includes school enrollment, demographic makeup, eligibility for subsidized lunch, absenteeism, suspensions.

- The State Assessments: Students are tested in English Language Arts and Mathematics in 4th, 8th and 10th grade, Writing in 3rd, 7th, 10th and 11th grade and Health in 5th and 9th grade. These standards-based tests use multiple choice, short response and essay items.

- The Rhode Island Statistical Model: RI uses a "value added" model, which uses statistical methods to compare similar students according to their demographic level and educational program characteristics. Rhode Island describes this model as follows: "By using sophisticated statistical modeling, researchers can statistically weight or flavor each child according to those characteristics which research shows most adversely affect achievement. This statistical exercise levels the playing field between schools that serve children with very different levels of challenge by adjusting for those levels or characteristics." (Rhode Island Board of Education, 2001, p. 19.)

- The SALT Survey: Teachers, parents and administrators fill out surveys about classroom practice, school climate and expectations. This data is matched

with the assessment data to draw conclusions about the quality and impact of teaching and learning in that school.

- SALT Self Study: The primary mechanism for school improvement is through the School Accountability for Learning and Teaching (SALT) program. The SALT cycle begins with forming a team, which conducts self-study activities that include analyzing student test scores and parent, teacher and student questionnaire results. Based on the self-study findings, the team develops a self-improvement plan and presents this plan to the community at a school report night.

- The SALT Visit and the NEASC high school accreditation: Each school hosts a SALT visit once every five years, which provides external perspective on school practices and student learning. Every other SALT visit serves as the New England Association of School and Colleges (NEASC) visit.

- Financial Data: Tracks tax and income statistics, school-level expenditure data and district level revenue and expenditure information.

7.2.3. Proposed Comprehensive Accountability System in Massachusetts

Although school accountability is a complex process, recent advances in computer-related technologies can be combined with alternative approaches to measuring student achievement to create a more comprehensive and informative accountability system. To this end, members of the National Board on Educational Testing and Public Policy have worked with the Massachusetts Department of Education and other key

members of the educational community to propose to develop a web-based accountability system that includes the following components:

1. Student demonstration of achievements through multiple modes of assessment: Although I do not advocate testing all students on all things, information related to a fuller spectrum of student learning must be collected in order to examine adequately the impact schools have on students and their learning. To this end, the comprehensive accountability system will employ multiple methods of assessment to collect information about student achievement in traditional subject areas and in the area of essential and applied technology skills. In addition to the measures of student achievement described above, I will also explore ways to collect and integrate other data into the accountability system including demographic information, grades, attendance, behavioral referrals (e.g., suspensions and detentions) and attitudinal and survey measures.
2. Teachers actively involved in analyzing and scoring student work: Most state-level testing programs restrict teachers from viewing the work students produce during on-demand tests. In only a handful of states are even a small group of teachers involved in scoring student work. Although teachers may receive test results, there is little opportunity for teachers to actively assess the quality of their students' work on state tests. To return power for assessing student work to teachers, the comprehensive accountability system proposed here will have teachers use rubrics to score student responses to the four extended items and portfolios.
3. Teachers, school leaders, parents, community members and policy makers accessing and analyzing relevant data: A wide variety of people are interested in the results of state-level accountability systems. Parents want to know how their children and the schools they attend perform. Teachers are also interested in how their students and their school perform. School leaders and external funding agencies often use results to evaluate the impact various school improvement initiatives have on students. And policy makers want to compare the impact different programs have on achievement, especially for

different subgroups of students. To facilitate the examination of assessment results and related data by this varied group of constituents, a series of web-based tools will be developed. These tools will enable visual data exploration, disaggregation of data, and simple statistical analyses.

4. Schools accounting to the public through systematic and public explanations of results: To assist schools in sharing results and strategies for improvement with the public, the system will provide schools with a web-based report that includes a summary of school initiatives/programs implemented during the previous year(s), static summaries of assessment results, and the school's accounting of performance for that year. Rather than simply displaying results, the system will provide a vehicle for schools to explain how their programs and initiatives impacted students and their achievement. This system will also prompt schools to identify areas they believe are in need of improvement and to develop and document plans to support increased achievement in these areas during upcoming years. By displaying these accounts and plans for improvement to the public via the web, community members will be better informed about how schools are doing, why, and what changes to look for in the future. To provide the public with a concrete understanding of student achievement, random samples of student work will also be displayed in the web-based report. These samples will represent a range of performances. In addition to providing concrete examples of each level of performance (e.g., what "Excellent" or "Advanced" writing looks like), these samples can also be used to document the integrity of the scoring system (i.e., that an "Excellent" or "Advanced" writing sample should look similar across all schools).

To be clear, while this proposed accountability system has received support from a variety of educational and political leaders in Massachusetts, funding has not been provided to develop and implement the system. And unless such funding is obtained, it is likely that the system will not be fully developed and implemented. Nonetheless, several

of the principles – namely multiple-measures of student learning, teacher and community involvement in analyzing data, and schools providing accounts of their practices and impacts – hold more promise for preventing, detecting or deterring gross disparities in education and improving the quality of education across all schools.

7.3. Blueprint for California

To provide schools, constituents, funding agencies, and policy makers with a more thorough understanding of the impacts of school-based programs on student learning, a more comprehensive accountability system is needed. To better satisfy the needs of schools and their constituents and to overcome the shortcomings discussed above, comprehensive accountability systems must meet the following goals:

- Provide relevant and timely information that schools can use to examine the impact their programs have on a wide spectrum of student learning
- Focus both on inputs and on outputs
- Collect more valid and authentic measures of student achievement
- Implement a statewide coherent student level data system
- Be sensitive to local context
- Increase the responsibility of teachers and school-leaders for accounting for educational practices and their outcomes

The state explicitly states that the API-based accountability system should include a range of outcome variables including scores from tests that are aligned with the state frameworks, graduation rates, and student and teacher attendance rates (CDE, 1999a). The CDE also states: “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" (CDE, 1999c). Despite these proclamations, the API-based accountability system currently relies solely on test scores (several of which are poorly aligned with the state frameworks). As described by the CDE (1999c), a truly comprehensive accountability system would ask schools to describe the programs and practices they have in place, the appropriateness of these programs and practices given specific context and background indicators, and the impacts these programs have on a variety of student outcomes. Programs and practices might include but should not be limited to:

- Access to quality teachers (e.g., student:teacher ratios, % of teachers with emergency credentials, % with Masters Degree or Beyond, etc.)
- Access to books, textbooks and other learning materials (e.g., ratio of library books to students, ratio of course specific textbooks to students, ratio of students:computers, ratio of students:Internet accessible computers, etc.)
- Type of school calendar (e.g., multi-track year-round schools; schools operating under the concept 6 model)
- Availability of appropriate instructional materials, specially trained teachers for ELL students
- Adequacy of school facilities (e.g., overcrowding, access to sanitary facilities -- ratio of students:functioning toilets, ratio of "contaminated"

classrooms:total classrooms, etc.; availability of functional heating and cooling systems, presence of lead paint, etc.)

- Subject area curricular materials used (e.g., math curriculum/textbooks, science curriculum/textbooks)
- Availability of Advanced Placement Courses (e.g., number of courses offered, number of sections available)
- Professional Development Opportunities (e.g., topics focused on during PD, number of hours offered, number of hours taken, percent of faculty participating)

Student outcomes might include but should not be limited to:

- Performance on tests closely aligned with the state frameworks
- Attendance rates
- Promotion/retention rates
- Graduation rates
- Drop-out rates
- Course taking patterns (higher vs. lower level mathematics, AP courses, etc.)
- Percent of students completing all courses required for UC eligibility
- Percent of students taking college entrance exams

- College entrance

Furthermore, to increase the amount of information and level of specificity of that information at the school-level, the state testing program should consider matrix sampling and should move towards implementing a statewide student data gathering mechanism such as the CSIS. Because of the poor quality of the data currently collected by the CDE, and without student-level data, year-to-year comparisons of student level test scores are meaningless. As described above, matrix sampling provides far more information about what students within a school can and cannot do. For this reason, matrix sampling provides information that will better inform the types of changes that schools might make to their curriculum and instructional practices.

To be clear, simply recording data for each of these variables would be a vast improvement over the current system. Yet, without requiring schools to actively describe the impacts their inputs have on these outputs, identify potential problem areas, and establish short and long term goals, the educational benefits of accountability will not be fully realized.

Moreover, the goals set through this process should not be limited to changes in outcomes. Given that inputs affect outcomes and that at times it is the inputs that must be altered before outcomes are impacted, schools must be allowed and encouraged to set goals that focus first on the inputs. As an example, there is a clear correlation between the percent of emergency credentialed teachers within a school and the school's API. It only makes sense, then, that for schools that have a high percentage of emergency credentialed teachers, interim goals should focus on decreasing the percentage of

emergency credentialed teachers (ideally to 0%) rather than on increasing students' test scores. Only after significant progress towards this interim goal has been reached should attention then focus on changes in test scores.

As the system is reformed, educational and political leaders should also consider who is actually being held accountable and for what. As an example, we know that quality teachers, quality curricular materials, and quality facilities all result in positive impacts on student learning. But to what degree does a teacher, a school, or a district have an influence over each of these factors? Districts and the city/town and state in which they operate have control over funds that impact facilities. Is it reasonable, then to hold schools accountable for quality facilities? While local school leaders often have much say in the hiring process, they do not control the amount of funding available to pay salaries. Nor do they have control over the locations in which they operate. Is it reasonable to hold schools solely accountable for the quality of their teachers? Clearly schools should be asked to provide accounts for the conditions (physical and instructional) that they provide for students. When the school's account makes it clear that the conditions are in need of improvement, all those entities that can impact those conditions should also be expected to provide accounts of the actions they take and how those actions impact the conditions. Depending on the conditions in need of improvement, this responsibility may extend from the individual school, through the district, and up to the state. In short, an effective accountability system should not focus on a single level within the educational system. Instead, all levels from the classroom up to the state should be asked to account for their practices and the impact those practices have on students and their learning.

One question that arises as accountability systems are reconsidered relates to the role tests play in the system – that is should tests serve as a signaling effect or as an outcome measure. As a signaling effect, tests would be used to identify schools and/or districts in which poor conditions (physical and/or instructional) are being provided. In such cases, poor test performance may prompt an inquiry into the conditions that may be impacting student performance. Once identified, actions would be taken to improve these conditions. Conversely, when test scores are used as an outcome measure, schools would first identify conditions for improvement. Once these conditions were improved, the test scores would be used to examine the impact these actions had on student learning.

Currently, California's API-based system more closely resembles a signaling effect than an outcome measure. As is described in greater detail below, test scores are used to target schools that are performing poorly (e.g., have low scores and are not improving at a desired rate). Once targeted, a school becomes eligible for funding that supports an investigation into conditions that may be negatively impacting student performance. The schools are then expected to remedy these conditions, but the extent to which the conditions are actually remedied is never examined.

Two problems arise when test scores are used as a signal that conditions are in need of improvement. First, this approach assumes that if test performance is acceptable, then the conditions must be good. As described above, test scores often improve due to practices of questionable educational value. In other cases, test scores are good due to conditions outside of the school (most often in schools serving students with high socio-economic status) and in spite of poor conditions or practices within the school. Using test scores to identify schools in need of improvement overlooks many schools that appear to

perform well but have conditions that are in need of improvement. Second, since the way to avoid being targeted is to have "acceptable" test scores, schools can be attracted to remedies that improve test scores without improving conditions. As an example, schools that have invested in technology to improve student writing, research and other high-order skills may begin using the software to drill students on topics included on the test and/or may decrease the amount of student writing performed on the computer because the tests are administered on paper (Russell, 2000).

Finally, using test scores to signal potential problems puts schools, districts, and the state educational agency on a mission to discover what is already known: conditions matter. Given the large body of evidence that quality teachers, quality instruction, quality instructional materials, and quality facilities have a positive impact on student learning, it would be more efficient to ask all schools to examine their current conditions, identify those that are in need of improvement, and then hold them accountable for improving those conditions. In this context, test scores might be used to examine the impact the improved conditions have on students and their learning.

8. Ability to be deposed and testify at trial

I have agreed to testify at trial in this manner. I am also able to submit to a meaningful deposition on any opinion, and its basis, that I would give at trial.

9. Consulting Fees

I have been compensated $10,000 for consulting in this action. I have not yet been compensated for testifying in this action, but, if called upon to testify, would expect additional compensation.

Two graduate students, Stacey Raczek and Jennifer Cowan, have provided me with some assistance in preparing this report but were not compensated beyond the financial compensation they currently receive from Boston College.

_____ _____

Michael Russell Date

Appendix A

KEY DECISIONS THAT LED TO THE CURRENT API

Guiding Principles for the API

The development of the API was spurred by the Public Schools Accountability Act which was enacted in April, 1999. As noted above, the PSAA required the Superintendent of Public Instruction, with the approval of the State Board of Education (SBE), to develop an Academic Performance Index to measure public school performance by July 1999. The Act did not specify how the API should be calculated, but instead provided for an Advisory Committee and a Technical Design Group to assist in this development process. It did require that the API include student test performance data, which should count for no less than 60% of the index.

Meeting in May 1999, the Advisory Committee for the PSAA began their development process by focusing on eligibility for the II/USP. At issue was whether to establish a target that was based on a central tendency (e.g., mean or median school performance) or on the percentage of students falling into specific categories or levels of performance. The committee reasoned that the latter criterion would be less vulnerable to distributional anomalies. Examining past performance on the SAT-9, the committee also recognized that test scores tended to be higher for each successive grade level, making it difficult to equitably average across grades in a school. The committee's recommendation was that the national average (50th national percentile rank) for the STAR be used as the cut point (CDE, 1999c). This recommendation, however, was not adopted.

A month later, the Advisory Committee meeting drafted thirteen guiding principles for the new API, including:

• The API must be **technically sound** (comparable, valid, and reliable measures must be used)

• The API must **emphasize student performance**, not educational processes.

• The API must strive to the greatest extent to measure **what is actually being taught or considered important for students to know** ("validity in measurement must be a continuing interest and focus…. Adequate research and exploratory studies will need to be conducted to investigate and verify that the API accurately represents what it is intended to measure"). This was later modified to: "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 reflect changes across the distribution of scores, and it should value growth among low-achieving, average, and high-achieving students").

• 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.

(California Department of Education, 1999c; Herman, Brown, & Baker, 2000)

In addition to codifying the shift from a focus on educational processes to specific student outputs, advisors also made it clear that the composition of the API would change over time as indicators became available, especially the tests designed to be better aligned to state standards. In addition, the committee reiterated that the API should be based on the percentages of pupils scoring at or above certain score levels on an assessment (e.g., percent above cut points or PACs), stating that this methodology would best respond to the PSAA legislation intent. Like test-based accountability systems in Kentucky and Philadelphia, it was decided that the proposed methodology should calculate the percentages of pupils scoring within specified levels, and multiply those by weighting factors, summarizing all the data into a single number. The committee, however, noted that "this approach is particularly appropriate when summarizing the results of standards-based assessments…which will employ performance levels in reporting pupil results" (CDE, 1999c). Standard-based assessments were not available in 1999 or 2000, and only one component of such a test went into the 2001 Base API. Further, the Committee also made it clear that when test content areas are weighted to generate a summary statistic, "ultimately, the value of these weights is a policy question. The weight that is assigned to a content area is an expression of the relative importance that the SPI and the SBE attaches to that content area" (CDE, 1999c). In July 1999, the State Board of Education approved the "Framework for the Academic Performance Index" which included the guiding principles, design features, and uses for API. The 200 to 1000 measurement scale and cut point of 800 was not specified in that document.

API Scale

In the PSAA Advisory Committee's final report for the 1999 API in October of 1999, the 200 to 1000 API scale is first defined. The committee recommends the metric based on two criteria; acceptance by the general public and producing scores that are not susceptible to misinterpretation (CDE, 1999a).

Percentage Above Cut Points (PAC)

The weighted PAC-based scoring methodology described above is recommended for the 1999 API and for the first time, it is recommended that five performance bands, with four fixed cut-points, be used with individual student Stanford 9 test results. The five performance bands should be equal in terms of national percentile ranks, requiring each to encompass 20 NPR points (defined as: 80-99th, 60-79th, 40-59th, 20-39th, and 1-19th NPRs). There is little published documentation of why these groupings of NPRs were selected, other than the committee's comment that they "attempted to strike a balance between simplicity (keeping the number of performance bands as few as possible) and sensitivity to gains in pupil achievement and school performance" (CDE, 1999a). The report does reveal that the Technical Group performed data simulations to study different scenarios including the use of ten performance bands, instead of five, but that was not

found to offer much advantage in measuring school status or growth. It is not reported whether alternate band assignment scenarios were tested, or whether the implications of this band allocation were considered.

Standard-Setting – No Known Methodology Employed

It should be noted that there are several established approaches to standard-setting and delineating cut scores for performance levels in a standards-based assessment system (Cizek, 2001), but it is less clear how to proceed in a norm-referenced system such as the 1999 and 2000 API. Other states with test-based accountability systems use these methods to allocate students into performance bands. For example, in Connecticut, scores on the criterion-referenced CAPT are reported on a 100 to 400 scale, in four subject areas. Scale scores are summarized by four performance levels in which only the extremes are labeled, "Goal Level" at the top and "Intervention Level" at the bottom. A standard-setting process known as "item mapping" was used to define the four performance bands, with an acknowledgement that the standards are set at a high level. This process uses a panel of experts and is well suited for an assessment with multiple item types such as the CAPT (Connecticut State Board of Education, 2001b).

In Massachusetts, the state-mandated accountability test (formally known as the Massachusetts Comprehensive Assessment System or MCAS) holds high stakes for individual students. That state uses two methods for establishing score thresholds that distinguish one performance level from another, both based on expert judgment within a standardized process using test items; the Bookmark Method for a 3rd grade Reading test, and the Body of Work method for other tests (Massachusetts Department of Education, 2001).

There are not such well-defined methods for a system like the initially norm-referenced, later hybrid norm- and criterion-referenced API used in California. However, it is not clear whether any systematic approach involving expert judgment or based on test items was used. Perhaps it was just decided that five equal quintiles was adequate.

Performance Band Weighting Factors

Next, the Committee endorsed the use of a set of weighting factors in which the percentages of pupils scoring within each performance band would be multiplied by a particular factor; this would then be summed in order to arrive at a single number. Instead of an equally-weighted system, they recommended a progressive system of weighting, giving schools relatively more credit for improvement by low-achieving students than by high-achieving:

> schools would be given an incentive for focusing on the needs of low achievers…. the Committee believes that a progressive set of weighting factors is appropriate for a state with a high proportion of pupils who are in the lower part of the distribution on the Stanford 9. In the Committee's view, priority should be given to raising achievement levels in California's lowest-performing schools…. This system would provide schools with incentives to focus on the instructional needs

> of low-achieving pupils by giving schools more credit for moving pupils across lower cut points than higher cut points. (CDE, 1999a.)

This is a laudable goal, but it is important to make clear that a policy decision has been made when defining the accountability index in this way; the legislation mentions no such weighting scheme.

Content Area Weighting Factors

The next issue addressed in the report concerns the relative importance of each content area when generating the summary API score. The committee clearly acknowledges that "This is in part a policy question, not a purely technical one" (CDE, 1999a). They recommend the same weighting scheme as the May, 1999 one for identifying II/ USP schools (see above). The Advisory Committee selected these weights

> because the consensus of the Committee was that for grades 2 to 8 they reflected the curriculum priorities in California. Reading combined with language mechanics was thought to be deserving of a higher weight (a total of 60%) than mathematics (40%). For grades 9 to 11, the Committee based its recommendation on the departmentalization of curriculum in high schools where equal time is devoted to each subject, deciding that all content areas in high school should receive equal weighting.

These content area weightings were used in 1999 and 2000. When the English/ Language Arts CST became included this year, the total weight associated with Stanford 9 performance went from 100% to 64%, although within that 64% the relative contribution of the subjects was the same as before.

API Interim Target = 800

Finally, the report presents a recommended statewide API performance target, as required by the PSAA legislation. The chosen target should "define an exemplary level of performance to which all schools should aspire" (CDE, 1999a). Here the thinking gets a little unclear; the committee recognizes that there are two distinct forms of assessment involved – the current norm-referenced, standardized Stanford 9, and the not-yet-implemented, standards-based test (as well as a planned high school exit examination). They suggest waiting until the state adopts performance standards before defining an exemplary level of performance for the standards-based components of the API, but that

> It is not necessary to wait for the adoption of performance standards to define an exemplary level for these [other] components…. Since the 1999 API would consist of Stanford 9 results only, this exemplary level would become de facto the temporary, interim performance target for the 1999 API. Based on data analyses by the TDG, the Advisory Committee recommends that the SBE set this interim target at 800. These data analyses document exactly how demanding this target of 800 is. For 1999, a target of 800 represents an exemplary level of performance that was attained only by a very small percentage of California schools: an estimated eight percent of the

elementary schools in the state, six percent of the middle schools, and four percent of the high schools.

(CDE, 1999a, p 14.)

Thus, it is clear that from the beginning the API target of 800 was recognized as a lofty goal – one in which less than ten percent of elementary and middle schools, and less than five percent of high school students, met at the time. Selecting an API target of 800 as the desirable, albeit "interim," cut-off score for categorizing schools was a key decision with important consequences for schools.

Five Percent Growth

The final tasks addressed by the Committee include how to define "five percent growth" of the legislated target and a recommendation to generate a School Characteristics Index (SCI) for each school to rank API scores and growth relative to those of comparable schools. As described in the Advisory Committee report, the committee evaluated a number of ways that five percent growth could be defined, including:

1. five percent of the school's base year API
2. five percent of the schools base year API with escalating higher raters for low-performing school
3. five percent of the maximum range of the API
4. five percent of the statewide average API
5. five percent of the distance to a statewide interim performance target
6. five percent of the statewide interim performance target

After reviewing models and simulations performed by the TDG, the Committee selected the fifth method. As the report indicates, "This method is intuitively plausible, simple, and best meets the three basic criteria set forth by the Committee." No further details were provided.

Scale Calibration Factor

As a footnote, a more recent meeting of the Technical Advisory Committee in October, 2001 described the acceptance of using a Scale Calibration Factor, SCF, beginning in 2001 as an adjustment to improve comparisons of API scores that include CST information with earlier ones that do not. The SCF would be added to subsequent API scores to make them more comparable to those without the CST. As noted previously, the value of the factor was estimated to be very small for all schools (positive for elementary schools and negative for middle and high schools). There was even discussion within the Committee about whether the factor should be applied at all since it ended up being so small. However, since additional indicators will be introduced in future APIs, its use was retained (CDE, 2001d). At this meeting, it was also announced that the math section of CST will be integrated into the 2002 Base API; the High School Exit Exam (HSEE) may also be added at this point. The Technical Group is studying how to integrate what they've termed a "non-universal" indicator in the API.

Appendix B

EFFECTS OF ALTERNATIVE API DECISIONS AND PERFORMANCE EXPECTATIONS

As described above, several decisions are embedded in the API definition process. In an effort to study the effect different decisions might have on the characterizations of the performance of schools, I undertook a series of simulations. As is explained in greater detail, some of these simulations are based on actual school level data. In other cases, particularly when student level data within a school was required, the data upon which these analyses are based was simulated. While there are several different factors or decisions that impact the calculation of API base and growth scores, I focus here on three factors: use of different performance bands, use of different performance band weights, and establishment of a different target.

Alternative Performance Bands and API Scores

As stipulated by the Advisory Board, scores from each of the SAT-9 subject tests are used to classify students into one of five performance bands. The performance bands include a range of 20 national percentile points; e.g., the lowest band includes NPR scores below 20, while the highest includes NPR scores above 80. The proportion of students in each band receives differential weight when a school's API is calculated; the reasoning here is that moving low-performing students out of lower bands should count more than moving other students to a higher level. The choice of score ranges for allocating norm-referenced Stanford 9 scores into performance bands is not well described in literature from the CDE. The concern seems to have been that the bands be of equal size in the NPR metric, e.g., in quintiles. However, there are other options. The performance of large groups of students on norm-referenced tests is generally considered to fall not in a linear pattern, in which an equal number of students score at all levels along the continuum. Rather, performance more often is represented by a bell-shaped curve, with many students scoring in the middle of the range. In fact, through item selection, norm-referenced tests are specifically designed to result in this pattern of results.

Rather than allocating student scores into performance bands by equal NPR intervals, an alternative is to form bands based on standard deviation ("SD") units. NPR scores can be transformed to a standard deviation, or standardized score, metric, and Performance Band categories can be defined differently, which affects the ultimate calculation of a school's API score. Band allocation could be defined with a specific philosophy about the relative band size or worth of each level of performance instead of an equal-sized interval approach. Here, two alternative categorization definitions are compared to the current allocation method.

In "Alternative #1," the top and bottom bands include students who achieve NPRs equivalent to one and a half standard deviation units above and below the mean, and the middle band includes students who score within one-half of an SD above and below the mean. In this option, the highest and lowest categories are relatively extreme and

statistically are likely to contain a small number of students. However, some might argue that this is a way of categorizing students in a way consistent with their beliefs about student performance.

For Alternative #2 the top and bottom bands include students who achieve NPRs greater than one standard deviation unit above and below the mean, while the middle band still includes students who score within one-half of an SD from the mean. This is a "smoother" choice in which more students will be assigned the top and bottom levels.

Table 8 summarizes the current quintile-based allocation scheme used, as well as two alternatives. The standard deviation units associated with and the number of students in each band -- if the achieved distribution of scores were perfectly normal – are presented.

Table 8. Percentiles Associated with Five Performance Bands: Current Allocation and Alternative Versions 1 & 2, Assuming Normally Distributed Scores in All Subjects

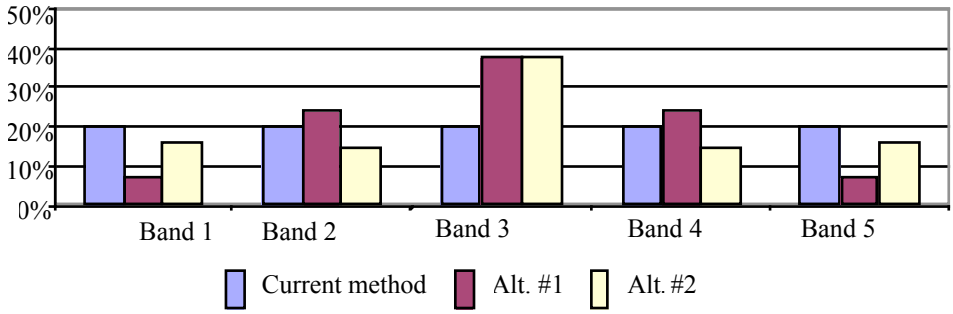
	CURRENT ALLOCATION			ALTERNATIVE #1			ALTERNATIVE #2		
PERF. BAND	Percentiles Allocated to Band	SDs Associated with Band	% in each band	Percentiles Allocated to Band	SDs Associated with Band	% in each band	Percentiles Allocated to Band	SDs Associated with Band	% in each band
1	1-19th	<-0.85	20%	1-6.7th	< -1.5	7%	1-15.9th	< -1.0	16%
2	20-39th	-0.84 to -0.25	20%	6.8-30.8th	-0.49 to –1.5	24%	16-30.8th	-0.49 to -1.0	15%
3	40-59th	-0.24 to +0.25	20%	30.9-69.1th	-0.5 to +0.5	38%	30.9-69.1th	-0.5 to +0.5	38%
4	60-79th	+0.26 to +0.84	20%	69.2-93.3th	+0.51 to +1.5	24%	69.2-84th	+0.51 to +1.0	15%
5	80-99th	>0.85	20%	93.4-99.9th	> +1.5	7%	84.1-99.9th	> +1.0	16%

In the absence of actual student-level performance data, the percent of student scores within each level must be based on a specific assumption – here, that the scores are normally distributed.[8] This assumption can of course be modified, and I test the effect of some other distributions below.

Figure 5 compares the percent of scores in a normal distribution that would be categorized into the bands based on these three different approaches. For the same set of scores, fewer or more students would be in each band depending on how the bands are defined. That is, an individual student might be placed in a different band under different allocation methods.

[8] Analyses of 2001 school-level Stanford 9 data not reported here suggest that many schools in California do present such a distribution of achievement, so this is a tenable assumption.

Figure 5. Percent in Performance Bands, Normal Score Distribution for All Tests



Given these Performance Band definitions, and the assumption of a normal distribution for all subject tests, an API can be calculated. Tables 9, 10, and 11 present API calculations using the 2001 Base definition of the Index.

Table 9. Calculating API Scores for A Sample with Normally Distributed Scores, Current Performance Band Allocation Method

			Weighted Score				
Performance Bands	**Weighting Factor**	**Proportion of Pupils in Level**	**S9 Reading**	**S9 Language**	**S9 Spelling**	**S9 Math**	**CST: Lang. Arts**
1 (FB Basic)	200	0.20	40	40	40	40	40
2	500	0.20	100	100	100	100	100
3	700	0.20	140	140	140	140	140
4	875	0.20	175	175	175	175	175
5 (Advanced)	1000	0.20	200	200	200	200	200
		Sum ("Indicator Score"):	655	655	655	655	655
		Indicator (Subject) Weight 2001	0.12	0.06	0.06	0.4	0.36
		Total Weighted Score for Indicator:	78.6	39.3	39.3	262	235.8
						2001 API **	**655**

** Actual API 2001 Base scores include a small Scale Calibration Factor. However, this has been reported to be less than 5 points.

Table 10. Calculating API Scores for A Sample with Normally Distributed Scores, Performance Band Allocation Alternative Method #1

			Weighted Score				
Performance Bands	**Weighting Factor**	**Proportion of Pupils in Level**	**S9 Reading**	**S9 Language**	**S9 Spelling**	**S9 Math**	**CST: Lang. Arts**
1 (FB Basic)	200	0.07	14	14	14	14	14
2	500	0.24	120	120	120	120	120
3	700	0.38	266	266	266	266	266
4	875	0.24	210	210	210	210	210
5 (Advanced)	1000	0.07	70	70	70	70	70
		Sum ("Indicator Score"):	680	680	680	680	680
		Indicator (Subject) Weight 2001	0.12	0.06	0.06	0.4	0.36
		Total Weighted Score for Indicator:	81.6	40.8	40.8	272	244.8
						2001 API **	**680**

Table 11. Calculating API Scores for A Sample with Normally Distributed Scores, Band Allocation Alternative Method #2

			Weighted Score				
Performance Bands	**Weighting Factor**	**Proportion of Pupils in Level**	**S9 Reading**	**S9 Language**	**S9 Spelling**	**S9 Math**	**CST: Lang. Arts**
1 (FB Basic)	200	0.16	32	32	32	32	32
2	500	0.15	75	75	75	75	75
3	700	0.38	266	266	266	266	266
4	875	0.15	131.25	131.25	131.25	131.25	131.25
5 (Advanced)	1000	0.16	160	160	160	160	160
		Sum ("Indicator Score"):	664.25	664.25	664.25	664.25	664.25
		Indicator (Subject) Weight 2001	0.12	0.06	0.06	0.4	0.36
		Total Weighted Score for Indicator:	79.71	39.855	39.855	265.7	239.13
						2001 API **	**664**

For the same set of normally-distributed scores, the current band allocation method and the two alternatives result in three different calculated API scores: 655, 680, and 664; the current method actually results in an API lower than these two alternatives. While I describe two alternatives, any number of different categorization options could be generated. And, depending upon which definition is used, the school's proximity to the interim API target (800) and the resulting growth target may change.

This procedure was carried out in the same way for two additional scenarios regarding the distribution of student scores. First, a positively-skewed distribution in which many students would score below the 50th NPR was assumed. Next, the distribution was shifted so that many students scored just above the 50th NPR, in the 60th to 80th NPR range. As presented earlier in this paper, this distributional shape is desirable because having most students above an NPR of 50 leads to an API of 800 or more. A summary of these simulations on API scores is presented in Table 12.

Table 12. Effect of Alternative Methods for Defining Performance Bands on API Scores

Distributional Shape	Method for Defining Performance Band Threshold Scores	Theoretical API*
Normally Distributed NPR Scores in All Subjects	Current	**655**
	Alternative #1: SD-based (+1.5)	680
	Alternative #2: SD-based, (±1.0)	664.25
Scenario 2: Positively-Skewed Distribution All Subjects, Many Low NPR Scores	Current	**525.25**
	Alternative #1: SD-based (+1.5)	599
	Alternative #2: SD-based, (±1.0)	539.5
Scenario 3: Negatively-Skewed Distribution All Subjects, Many High NPR Scores	Current	**674.25**
	Alternative #1: SD-based (+1.5)	792.5
	Alternative #2: SD-based, (±1.0)	680

*Weighting of subtests in the manner of 2001 Base

As Table 12 shows, these two band definitions produced API scores higher than the current method does for three different hypothesized score distributions. The difference in API was more pronounced in non-normal than normal distributions, suggesting that schools with many low- or high-scoring students would see a greater effect from modifications to the Performance Bands. These allocation alternatives demonstrate that, for a given set of scores, a school might receive a very different API if the calculation rules differed. In summary, the decision for how to define bands has important consequences.

Alternative Subject Area and Performance Band Weightings

In addition to specifying how NPR scores should be categorized in Performance Bands, the PSAA Advisory Board also made decisions regarding the weights given to each Performance Band and each subject test when calculating the API. The differential weighting factors for each of the five performance bands (e.g., 200 for Band 1, Far Below Basic; 1000 for Band 5, Advanced) were essentially selected to give more reward to moving students out of lower bands than to changes at the top of the scale. This decision, like others, also affects the calculation of API scores. In this section, analyses designed to study the effect of changing definitions for band and subject weightings are described.

Summary Stanford 9 and CST achievement data are publicly available at the school level, by grade, but not at the individual student NPR score level. Thus, it was necessary to develop a method for estimating the proportion of students in each of the five Performance Bands. For a small random sample of schools, multi-grade data for 2001 were aggregated within a school to get an estimate of a school's overall performance for each subject test. In order to approximate the actual proportion in each

band, we generated the mean, across grades, of the proportion of students scoring at or above certain reported NPR levels (e.g., the 25th, 50th, 75th), and made a few assumptions. The proportion of students with NPR scores less than 25 were considered to be in Performance Band 1 (usually NPR scores under 20); the proportion scoring 50 to 75 were labeled as being in Band 4 (usually 60 to 79), and scores over 75 were labeled Performance Band 5 (usually NPRs over 80). The remaining group, proportion of NPR scores in the range of 25 to 50, overlapped two Performance Bands, so the mean proportion was simply divided into two equal groups to form Bands 2 and 3 (usually defined as 20 to 39, and 40 to 59). This certainly is a rough estimation method, but seems to be a reasonable approach given the form in which data are available. From this, it was possible to proceed with calculating API scores with different weighting schemes than the current one.

Three alternative subject weighting schemes were developed, using only Stanford 9 subjects tests (e.g., without the English/ Language Arts CST introduced for 2001). In the first, mathematics and English/ subjects were weighted equally; the content weighting for math was 0.50, and the sum of Reading, Language Arts, and Spelling weights was also 0.50. The second version had a high English/ low Math weights, and the third was a high Math/ low English alternative. Table 13 presents the specific alternate content weights used to generate API scores.

Table 13. Three Alternative Content Weightings

	1999, 2000 Actual Subject Weightings	Alternative #1: Equal Math/English	Alternative #2: Low Math	Alternative #3: Low English
Reading	0.30	0.20	0.30	0.15
Language Arts	0.15	0.20	0.30	0.15
Spelling	0.15	0.10	0.10	0.10
Math	0.40	0.50	0.30	0.60

These content weights were used to generate four sets of API scores for fifteen schools in our random sample using the simulated Performance Bands described above (in addition to the three alternate conditions, an API with current weights was calculated for comparison purposes since we didn't have actual Performance Band data). The method described above for calculating an API score was employed.

Table 14 presents API scores calculated five different ways; the reported API for 2001, an API based on those same weights, and three from modified subject weightings. A comparison of our current-weight API with the actual reported API serves as a test of how reasonable the Performance Band allocation method was – if these two scores are quite different, it would indicate that our method was too different from reality, and this analysis would be called into question. However, as Table 13 shows, they are relatively similar. The average difference between actual and simulated 2001 API was only two points.

The three alternatives did in fact generate in API scores that varied from the current index. Weighting Math and English equally typically resulted in a higher API, while weighting Math higher than current resulted in lower API scores. Differences

between an API calculated with current weights and the equal subject method ranged from 2 to 14 points (mean across 15 schools = 7); between current and high-English ranged from 1 to 10 points (mean = 3); and between current and high-Math from 2 to 25 points (mean =12). The effect of more differential English/ Math weightings is larger for schools with greater differences between their Math and English/ Language Arts scores.

Table 14. Effect of Alternative Subject Weights on API Scores

	API01 Actual	API Current Weights, Sim. Perf. Bands	Alternative #1: Equal English/ Math	Alternative #2: High English/ Low Math	Alternative #3: Low English/ High Math
School 1	687	679.4	685.8	677.2	422.2
School 2	339	341.1	346.5	334.3	222.8
School 3	546	537.1	541.9	532.3	336.2
School 4	N/A	565.6	579.2	555.3	383.9
School 5	752	757.6	768.2	754.2	485.8
School 6	736	731.9	743.3	722.0	474.1
School 7	518	528.7	529.2	525.9	320.0
School 8	625	631.0	635.2	634.7	378.6
School 9	698	716.2	722.8	717.4	442.1
School 10	826	843.8	846.0	845.0	511.9
School 11	422	421.5	432.8	415.4	284.0
School 12	421	413.4	419.1	407.2	264.7
School 13	660	656.4	664.5	654.8	414.5
School 14	873	885.3	888.0	886.5	537.3
School 15	701	694.5	701.2	688.0	440.4

These particular weighting schemes may not be of interest for implementation. The purpose here, as for the previous analysis, was to illustrate that small modifications to the complex API definition result in changes to the outcome.

Next, the effect of modifying Performance Band weights in API calculations was studied. For the same 15 schools and simulated Performance Bands described above, we calculated an API with weights for bands 1 through 5 that are slightly more equal. As in the previous analyses, the possibilities for modifications to current practice are infinite; we simply proceeded with selected alternatives to demonstrate the effect of changes. Here, Band 1 was weighted 500, rather than 200; the Band 2 weight was changed from 500 to 700; Band 3 weight was changed from 700 to 800; Band 4 was changed from 875 to 900; and the weight for Band 5 remained 1000. Because the sum of weights was slightly larger in this alternative than actual, the resulting score was multiplied by a small factor so the results would be on the same 200 to 1000 scale. As Stecher and Arkes (2001) found, changes in the performance band weights can have substantial effects on the resulting API scores. Table 15 presents the results. Differences between an API calculated from current band weights and this alternative were greater than those produced from changes in the subject weights, although this may only be an artifact of the specific weights selected for these analyses.

Table 15. Effect of Alternative Performance Band Weights on API Scores

	API01 Actual	API Current Weights, Sim. Perf. Bands	API: More Equal Weight to Perf. Bands *	DIFF WTEQAPI/API
School 1	687	679.4	666.38	13.0
School 2	339	341.1	487.01	-145.9
School 3	546	537.1	591.54	-54.4
School 4	N/A	565.6	607.38	-41.8
School 5	752	757.6	705.37	52.2
School 6	736	731.9	694.13	37.8
School 7	518	528.7	587.71	-59.0
School 8	625	631.0	641.76	-10.7
School 9	698	716.2	684.87	31.3
School 10	826	843.8	748.85	94.9
School 11	422	421.5	530.60	-109.1
School 12	421	413.4	526.10	-112.7
School 13	660	656.4	653.94	2.5
School 14	873	885.3	768.46	116.8
School 15	701	694.5	673.33	21.2
				-10.9

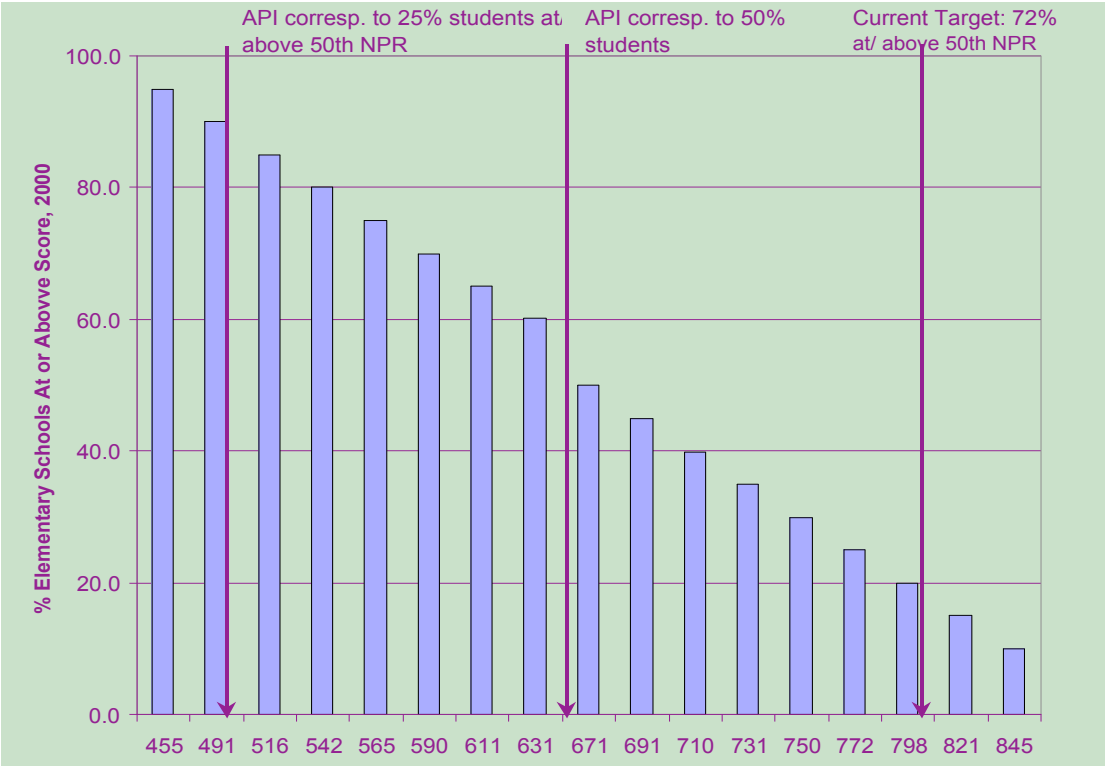
* NOTE: A weighting factor was applied to the generated API to adjust for differences in the sums of weights; this serves to place it on the same metric as actual API.

Alternative Performance Targets and School Success

As noted above, the current interim API target of 800 represents a very high (and for many schools unrealistic) expectation. It is instructive to calculate how many schools in California might be considered to meet standards of accountability should the standards be defined differently. In 2000, 20% of elementary schools, 14% of middle schools, and 5% of high schools scored at least an 800 for the API. Those figures rose by 1% in 2001 for elementary and middle schools, but stayed stable for high schools. Figure 6 presents the percent of elementary schools achieving 2000 Base API scores at or above particular points. Ninety-five percent of elementary schools attained at least a score of 455; at the other end of the scale, only ten percent scored at or over 845. The figure effectively represents what proportion of schools would be considered to have met the API target if it were not 800, but some other figure. At an API of 565, 75% of schools meet the target, at an API of 671, 50% meet it, and 25 % of schools meet an API target of 772. About 52% of elementary schools met an API target corresponding to the point at which about 50% of students score at or above the national 50th percentile on the API component tests (about 664; see section 6.1 above for calculation of API scores given a normal distribution of achievement on the Stanford 9; Rogosa, 2000).

Given the wide variation in API scores and the fact that approximately half of the schools are performing below the national average, a more reasonable interim goal for California might be to move all schools in the state above the API score corresponding to the national mean. For those schools already above this point, alternative targets might also be set.

Figure 6: Percent of Elementary Schools Meeting Different API Cut-Scores, 2000 Base API



APPENDIX C

Los Angeles Unified School District High School Imputed Drop-out Rates – 2001*

SNAME	API01	% MEALS	% EL	YR_RND	FULL CRED	TOTAL ENROLL 01	9th GRADE ENROL. / 98	12th GRADE ENROL. 01	IMPUTED DROP OUT RATE
Huntington Park Senior High	482	91	29	Yes	83	3399	1446	619	57.19%
Franklin (Benjamin) Senior High	521	90	25	Yes	76	2483	1165	466	60.00%
Bell Senior High	490	88	30	Yes	79	3581	1607	693	56.88%
San Fernando Senior High	492	87	31	Yes	79	3329	1593	662	58.44%
Manual Arts Senior High	415	87	35	Yes	78	3042	1349	610	54.78%
South Gate Senior High	522	83	19	Yes	78	3630	1418	746	47.39%
Garfield (James A.) Senior High	487	82	31	Yes	88	3279	1600	589	63.19%
Roosevelt (Theodore) Senior High	477	82	29	Yes	73	3816	1677	751	55.22%
Bravo (Francisco) Medical Magnet High	733	82	13	No	93	1320	509	386	24.17%
Jefferson (Thomas) Senior High	429	81	38	Yes	64	2692	1352	416	69.23%
Cleveland (Grover) High	614	80	24	No	80	2188	783	499	36.27%
Lincoln (Abraham) Senior High	508	79	39	No	74	2379	1001	384	61.64%
Banning (Phineas) Senior High	494	78	22	No	73	2590	1177	433	63.21%

SNAME	API01	% MEALS	% EL	YR_RND	FULL CRED	TOTAL ENROLL 01	9th GRADE ENROL. / 98	12th GRADE ENROL. 01	IMPUTED DROP OUT RATE
Francis (John H.) Polytechnic	493	78	36	Yes	77	3035	1128	543	51.86%
Locke (Alain Leroy) Senior High	385	76	35	No	63	1806	738	281	61.92%
Jordan (David Starr) Senior High	417	76	36	No	67	1783	601	247	58.90%
Hollywood Senior High	477	74	36	Yes	76	2375	1120	435	61.16%
Canoga Park Senior High	522	74	33	No	79	1695	653	263	59.72%
Marshall (John) Senior High	545	74	27	Yes	84	3163	1012	706	30.24%
Downtown Business High	601	74	19	No	74	771	253	181	28.46%
Los Angeles Senior High	473	73	37	Yes	81	3354	1484	623	58.02%
Wilson (Woodrow) Senior High	507	73	23	No	65	2011	896	396	55.80%
Monroe (James) High	525	73	40	Yes	71	3457	1308	672	48.62%
Verdugo Hills Senior High	579	69	19	No	82	1644	785	298	62.04%
Van Nuys Senior High	613	68	28	No	75	2812	1313	646	50.80%
Fairfax Senior High	549	68	25	No	84	1820	675	431	36.15%
Crenshaw Senior High	455	67	9	No	63	2026	817	424	48.10%
Reseda Senior High	574	67	24	No	77	1698	694	420	39.48%

SNAME	API01	% MEALS	% EL	YR_RND	FULL CRED	TOTAL ENROLL 01	9th GRADE ENROL. / 98	12th GRADE ENROL. 01	IMPUTED DROP OUT RATE
Sylmar Senior High	513	65	22	No	76	2617	849	502	40.87%
North Hollywood Senior High	571	64	29	No	76	3107	1319	531	59.74%
Fremont (John C.) Senior High	431	63	42	Yes	62	3627	1573	438	72.16%
Grant (Ulysses S.) Senior High	571	62	28	No	80	2462	1142	601	47.37%
Washington (George) Preparatory High	433	59	15	Yes	63	2872	992	421	57.56%
Kennedy (John F.) High	568	58	18	No	89	2037	705	505	28.37%
Narbonne (Nathaniel) Senior High	588	50	15	No	81	2401	930	445	52.15%
Dorsey (Susan Miller) Senior High	442	48	18	No	63	1633	609	325	46.63%
Gardena Senior High	484	44	16	No	70	2386	1024	474	53.71%
Birmingham Senior High	585	42	20	No	73	2477	1032	502	51.36%
Eagle Rock Junior-Senior High	629	42	14	No	74	2290	631	376	40.41%
Carson Senior High	541	41	6	No	74	2505	914	562	38.51%
Venice Senior High	570	35	24	No	84	2169	822	455	44.65%
King/Drew Medical Magnet High	619	34	4	No	59	1177	395	274	30.63%
San Pedro Senior High	640	30	9	No	83	2327	866	480	44.57%

SNAME	API01	% MEALS	% EL	YR_RND	FULL CRED	TOTAL ENROLL 01	9th GRADE ENROL. / 98	12th GRADE ENROL. 01	IMPUTED DROP OUT RATE
Palisades Charter High	714	29	7	No	84	1904	766	465	39.30%
Chatsworth Senior High	599	28	18	No	72	2484	967	603	37.64%
Hamilton (Alexander) Senior High	597	27	11	No	80	2159	827	500	39.54%
University Senior High	564	27	23	No	82	1933	708	418	40.96%
Taft (William Howard) Senior High	649	19	12	No	69	2264	940	568	39.57%
Granada Hills Senior High	739	16	8	No	88	2863	1022	689	32.58%
Westchester Senior High	577	15	6	No	77	1589	703	372	47.08%
El Camino Real Senior High	725	12	10	No	79	2506	967	694	28.23%

*Source: California Department of Education website

REFERENCES

(1971). *Oxford English Dictionary, Compact Edition.* Oxford: Oxford University Press.

Amrein, A.L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning. Education Policy Analysis Archives, 10(18). Retrieved May 15, 2002, from http://epaa.asu.edu/epaa/v10n18/.

California Department of Education. (1997). *English Language Arts Content Standards for California Public Schools Kindergarten through Grade 12.* Available: http://www.cde.ca.gov/board/pdf/reading.pdf

California Department of Education. (1998). *Science Content Standards for California Public Schools Kindergarten through Grade 12.* Available: http://www.cde.ca.gov/board/pdf/science.pdf

California Department of Education. (1999a). *The 1999 Base Year Academic Performance Index (API): The Report of the Advisory Committee for the Public Schools Accountability Act of 1999.*

California Department of Education. (1999b). *Recommendations of the Advisory Committee for the Public Schools Accountability Act of 1999 on the selection of schools for the Immediate Intervention/ Underperforming Schools Program, May 1999.* Available: http://www.cde.ca.gov/psaa/board/may.

California Department of Education. (1999c). *Framework for the Academic Performance Index.* Available: http://www.cde.ca.gov/psaa/board/june.

California Department of Education. (1999d). *The 1999 Base Year Academic Performance Index (API): The Report of the Advisory Committee for the Public Schools Accountability Act of 1999.*

California Department of Education. (2000a). *Explanatory notes for the 1999 Academic Performance Index.* Available: http://www.cde.ca.gov/psaa/api/api99/explan.pdf

California Department of Education. (2000b). *Explanatory notes for the 2000 Academic Performance Index Base Report. Available:* http://www.cde.ca.gov/psaa/api/yeartwo/base/explan2kb.pdf

California Department of Education. (2000c). *Explanatory notes for the 2000-2001 Academic Performance Index (API) Growth Report. Available:* http://www.cde.ca.gov/psaa/api/yeartwo/growth/expnotes.pdf

California Department of Education. (2000d). *Key elements of Senate Bill IX (Chapter 3 of 1999), Public Schools Accountability Act of 1999.* Available: http://www.cde.ca.gov/psaa/keyelements.pdf

California Department of Education. (2001a). *The 2001 Base Academic Performance Index (API): Integrating the California Standards Test for English-Language Arts into the API. Available:* http://www.cde.ca.gov/psaa/api/yeartwo/growth/integrate.pdf

California Department of Education. (2001b). *Alternative Schools Accountability Model 2001-2002 Indicator Selection and Reporting Guide. Available:* http://www.cde.ca.gov/psaa/ASAMfnl.pdf

California Department of Education. (2001c). *Explanatory notes for the 2001 Academic Performance Index Base Report. Available:* http://www.cde.ca.gov/psaa/api/api0102/base/expn01b.pdf

California Department of Education. (2001d). *Public School Accountability Act (PSAA) Advisory Committee Meeting Minutes, October 25, 2001*. Available: http://www.cde.ca.gov/psaa/minutes.

California Department of Education. (2001e). *Explanatory notes for the 2000-2001 Academic Performance Index (API) growth report*. Available: http://www.cde.ca.gov/psaa/api/yeartwo/growth/expnotes.pdf

California Department of Education. (2002a). *Fact Book 2002 Handbook of Education Information. Available:* http://www.cde.ca.gov/resrc/factbook/pubaccount.htm

California Department of Education. (2002b). *Minutes. Superintendent's Advisory Committee, Public Schools Accountability Act (PSAA) of 1999, January 17, 2002*. Available: http://www.cde.ca.gov/psaa/minutes/PSAAjan1702.pdf

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Clark-Thomas, Eleanor Deposition from *Williams v. State of California* (No 312236), April 5, 2001.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. New Haven: Yale University Press.

Connecticut State Board of Education. (2001a). *CAPT Connecticut Academic Performance Test second generation. Interpretive guide*. Available: http://www.csde.state.ct.us/public/der/s-t/testing/capt/interpretive_guide_2001_147.pdf

Connecticut State Board of Education. (2001b). *CAPT Connecticut Academic Performance Test second generation. 2001 program overview*. Available: http://www.csde.state.ct.us/public/der/s-t/testing/capt/proovr.pdf

Connecticut State Board of Education. (2001c). *Greater Expectations Connecticut's Comprehensive Plan for Education 2001-2005*. Available: http://www.state.ct.us/sde/whatsnew/greater_expectations.pdf

Feuer M., Holland P., Green B., Bertenthal M. & Hemphill F. (1998). Uncommon Measures: Equivalence and Linkage Among Educational Tests. Washington DC: National Academy Press.

Guth, G. J. A., Holtzman, D. J., Schneider, S. A., Carlos, L., Smith, J. R., Hayward, G. C., & Calvo, N. (1999). *Evaluation of California's standards-based accountability system. Final report*. San Francisco, CA: WestEd.

Haney, Walt & Raczek, Anastasia. (1994). Surmounting outcomes accountability in education. Paper prepared for the US Congress Office of Technology Assessment.

Haney, Walt. (2000). The myth of the Texas miracle in education. Education Policy Analysis Archives, 8(41). Retrieved on February 7, 2002 from : http://epaa.asu.edu/epaa/v8n41/.

Henry, Thomas. Deposition from *Williams v. State of California* (No 312236), September 26, 2001.

Herman, J. L., Brown, R. S., Baker, E. L. (2000). *Student assessment and student achievement in the California public school system. CSE Technical Report 519*. Los Angeles, CA: Center for the Study of Evaluation, Center for Research on Evaluation, Standards, and Student Testing.

Herman, Joan L., Brown, Richard S., & Baker, Eva L. (2000). Student Assessment and Student Achievement in the California Public School System. CSE Technical Report 519. Center for the Study of Evaluation, University of California, Los Angeles.

Kane, Thomas J. & Staiger, Douglas O. (2001). Improving school accountability measures. Cambridge, MA: National Bureau of Economic Research.

Kohn, A. (2000). The Case Against Standardized Testing: Raising the Scores, Ruining the Schools. Portsmouth, NH: Heinemann.

Koretz, Daniel M. & Barron, Sheila I. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Massachusetts Department of Education. (2001). *Spring 2001 MCAS tests: Summary of state results*. Available: www.doe.mass.edu/mcas/2001.

Meier, D. (2002). In Schools We Trust: Creating Communities of Learning in an Era of Testing and Standardization. Boston, MA: Beacon Press.

Moriarty, J. (2001). School safety 1st, tests last, parents say. *Union-News. Sunday Republican.* Retrieved October 29, 2001 from: http://masslive.com/newsindex/holyoke/index.ssf?/news/pstories/ae517edt.html

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. The Future of Children, 5(2), pp. 113-127.

Noble, M. (2000). STAR Program: Ongoing Conflicts Between the State Board of Education and the Superintendent of Public Instruction as Well as Continued Errors Impede the Program's Success. California State Auditor. http://www.bsa.ca.gov/bsa/

Orlofsky, Greg F. & Olson, Lynn. (2001). The state of the states. Education Week, 20(17), pp. 86-88.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.) (1999). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. *Chapter 5: Setting reasonable and useful performance standards.* Washington, DC: National Academy Press.

Perry, M. & Carlos, L. (1999). What to Expect from California's New School Accountability Law. EdSource Report.

Rhode Island Board of Education, (2001). Information Works! State Analysis Looking Through Rhode Island's School Accountability Lenses 2001. Retrieved on February 26, 2002, from: http://www.infoworks.ride.uri.edu/2001/default.asp

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15, 351-357.

Rogosa, David. (1999). How Accurate Are the STAR National Percentile Rank Scores for Individual Students? An Interpretive Guide. Retrieved February 27, 2002, from www.cse.ucla.edu/CRESST/Reports/drrguide.html

Rogosa, David. (2000). Interpretive Notes for the Academic Performance Index. Retrieved February 6, 2002 from http://www.cde.ca.gov/psaa/api/fallapi/apnotes.pdf

Rose, L. C. & Gallup, A. M. (2001). The 33rd Annual Phi Delta Kappa/Gallup Poll Of the Public's Attitudes Toward the Public Schools. *Phi Delta Kappan*, 83, 41-58. http://www.pdkintl.org/kappan/k0109gal.htm, downloaded on February 4th 2002.

Russell, Michael (2000). Using expected growth size estimates to summarize test score changes. Practical Assessment, Research & Evaluation, 7(6). Accessed online on 2/7/02 at: http://ericae.net/pare/getvn.asp?v=7&n=6.

Russell, M. & Haney, W. (2000). Bridging the gap between testing and technology in schools. Education Policy Analysis Archives, 8(19). Retrieved March 20, 2002, from http://epaa.asu.edu/epaa/v8n19.html.

Shepard, L. (1990). Inflating Test Score Gains: Is the Problem Old Norms or Teaching the Test. *Educational Measurement: Issues and Practice,* Fall 1990, 15-22.

Spears, Philip. Deposition from *Williams v. State of California* (No 312236), October 31, 2001.

Stecher, B. & Arkes, J. (2001). Rewarding schools based on gains: It's all in how you calculate the index and set the target. RAND Report DRU-2532.

Warren, Paul. Deposition from *Williams v. State of California* (No 312236), May 23, 2001.

Wenglinsky, H (2002). How Schools Matter: The Link Between Teacher Classroom Practices and Student Academic Performance. Education Policy Analysis Archives, 10(12). Retrieved on February 7, 2002 from : http://epaa.asu.edu/epaa/v10n12/.

The White House. Fact Sheet: No Child Left Behind Act. Retrieved on January 9, 2002 from: http://www.whitehouse.gov/news/releases/2002/01/print/20020208.html

[i] Schools are compared to other schools with similar characteristics. These characteristics include pupil mobility, pupil ethnicity, pupil socioeconomic status, percentage of teachers who are fully credentialed, percentage of teachers who hold emergency credentials, percentage of students who are English language learners, average class size per grade level, and whether the school operates multi-track year-round educational programs. A statistical model is used to create an index for similar schools (CDE, 2001c).

[ii] Financial awards are also tied to the API under the Governor's Performance Award (GPA) program, Certificated Staff Performance Incentives, and School Site Employee Performance Bonus. The GPA program will distribute $157 million in the 2001-2002 school year to schools that meet three criteria. First, the schools must meet or exceed their API growth target or have an API increase of five points, whichever is greater. In addition, all subgroups must meet or exceed 80% of the school target or have an API increase of four points, whichever is greater. Finally, the schools must have 95% SAT-9 participation rate for elementary and middle schools and 90% participation rate for high schools.

The Certificated Staff Performance Incentive will distribute a proposed $100 million of bonuses in 2001-02 to certified staff in schools who rank in the top half of the API and who show at least two times the annual growth target set for their school. Similar restrictions for subgroups and participation rate are in effect. Finally the School Site Employee Performance Bonus will distribute a proposed $350 million in 2001-02 to all site staff, based on the same eligibility requirements as the GPA. An amount of money equal to what the staff receives will be given to the school for school-wide use. (http://www.cde.ca.gov/psaa/awards/#AWARDS) It should be noted that, as this paper was being produced, the Certified Staff Performance Incentive program was eliminated due to budget cuts.

[iii] In grades 2-8, in 2000, the weights given to each content area were: Mathematics, 40%; Reading 30%; Language Arts, 15%; and Spelling, 15%. In grades 9-11, the weights were Mathematics: 20%; Reading: 20%; Language: 20%; History/Social Science: 20%; and Science: 20%. (CDE, 2000d.)

[iv] Below I use an example, based on 2000 rules, to demonstrate how scores from the Stanford 9 were transformed via weightings for each test and performance level into a school's total API Index. As Table 1 indicates, ten percent of the students in this school received national percentile ranks between 80 and 99 on the SAT-9 Reading test. The weighting factor for this performance band (80-99) is 1000. Thus, the weighting factor (1000) is multiplied by the proportion of students in this performance band (.10) to

produce a weighted score of 100. This process is repeated for each performance band on each test. The weighted scores for each subject area are then summed to produce a weighted total score. In the case of the Reading test here, the weighted total score is 710.

To produce an API score, the weighted total score for each test is multiplied by the Content Weight. For Reading, the Content Weight is 30%. Thus, the weighted total score of 710 for Reading is multiplied by .30 to yield a Content score of 213 for Reading. The same process is repeated for each subject area. The API scores for each subject area are then summed to yield a total weighted score, the API.

Table 1. Calculating a School API with 1999, 2000 Weights

Stanford 9		**Reading**		**Mathematics**		**Language**		**Spelling**	
Performance Bands (NPR)	Weighting Factors	% Pupils in Band	Weighted Score	% Pupils in Band	Weighted Score	% Pupils in Band	Weighted Score	% Pupils in Band	Weighted Score
80-99th	1000	10%	100	20%	200	10%	100	20%	200
60-79th	875	20%	175	10%	87.5	10%	87.5	30%	262.5
40-59th	700	50%	350	40%	280	60%	420	40%	280
20-39th	500	15%	75	20%	100	10%	50	10%	50
1-19th	200	5%	10	10%	20	10%	20	10%	20
		Total	710	Total	687.5	Total	677.5	Total	812.5
Content Weighting, %:			30%		40%		15%		15%
Content Score			213		275		102		122
							API: Weighted total (213+ 275+ 102 +122)		**712**

[v] In the Advisory Committee's 1999 Report, the Committee emphasizes that this target is demanding: These data analyses document exactly how demanding this target of 800 is. For 1999, a target represents an exemplary level of performance that was attained only by a very small percentage of California schools: an estimated eight percent of the elementary schools in the state, six percent of the middle schools, and four percent of the high schools (p. 14).

[vi] Operationally, numerically significant sub-groups are defined as sub-groups that comprise at least 15% of the total school enrollment and consist of at least 30 people OR a sub-group that contains a minimum of 100 students. For these sub-groups, comparable improvement is defined as 80% of the school-wide Growth Target. Under these rules, a school is said to have made its target if the API based on all students in the school increases by at least 5% AND the API for each numerically significant sub-group increases by at least 4% (CDE, 2001e.)

[vii] (source: www.cde.ca.gov/statetests/star/s2blueprt.html)

[viii] To facilitate this integration, student performance on the CST will be classified into five Performance Band categories, as for the Stanford 9 subject tests. In this way, performance band weighting factors can be applied for both the SAT-9 and the CST tests. As of this writing, it is unclear which of several methods will be used to establish performance standards (see Cizek, 2001 for a review of several methods of establishing performance standards). The publisher does provide a table for the English/ Language Arts CST specifying, by grade, which Performance Band is associated with ranges of score points correct (star.cde.ca.gov/star2001/help/ScoreTypes.html), but does not describe what process was used to assign scores to the five bands. Typically, students are placed in the bottom "Far Below Basic" band if they achieve a score corresponding to about a third or fewer of items correct, while students in the top band, "Advanced," have scores equivalent to more than 80% correct (e.g. second-graders need to achieve 66 points to be "advanced," which corresponds to 88% correct answers).

[ix] The scale calibration factor was determined by calculating the difference between the mean API score across the state based solely on SAT-9 and the mean score based on the SAT-9 and the CST. This mean difference becomes the scale calibration factor and is added to each school's integrated API. For 2001, the

scale calibration factor increased each elementary school's integrated API by approximately 5 points, and decreased middle and high schools' API by approximately 5 points (CDE, 2002b). In other words, for each elementary school in 2001, about five points were added to the API score that was calculated using the procedure summarized in Table 1. For middle and high schools, about five points were subtracted.

[x] With the addition of CST English Language Arts test scores to the index, three separate API scores were produced in 2001:

- 2001 API Growth using just the SAT-9 (in order to calculate 2000-2001 growth)
- 2001 API Base using both the SAT-9 and the English Language Arts section of the CST (will be used to calculate 2001-2002 growth)
- 2000-2001 API Growth (based on the SAT-9 2000 and SAT-9 2001 results)

[xi] Traditional schools with 100 valid STAR scores or more are ranked according to the API formula and are eligible for rewards and interventions through the Public Schools Accountability Act. Schools with 11-99 students with valid STAR scores receive an API with an asterisk, which is not used in calculating the cut points for the decile ranks. But small schools do receive a report indicating what their decile rank would be if their score were included in the ranking (CDE, 2000c.) "Small schools" are eligible for II/USP funding and other monetary awards. Schools with fewer than 11 valid STAR scores and alternative schools, serving primarily high-risk students, use the Alternative Schools Accountability Model (ASAM). These schools are accountable through valid STAR results and two state approved indicators such as number of suspensions, student punctuality, attendance and course completion (CDE, 2001b.) More than 1000 schools were registered to participate in the ASAM in 2000 (http://www.cde.ca.gov/psaa/api/2000-2001Cycle/letters/alted42501.pdf) Schools participating in the ASAM are not eligible for rewards and interventions.

[xii] As the 1999 Advisory Committee report states, the interim API target of 800 is very demanding. By way of example, an elementary or middle school in which student SAT-9 scores are distributed identically to the national norm group would receive an API of 655. To obtain an API of 800, Rogosa (2000, p. 1) estimates that "a little less than three-quarters of the students" in the school must exceed the national 50th percentile on each SAT-9 test. Table 3 presents several extreme combinations of percentile ranks that produce an API of 800. The first example shows that a school in which 75% of students performed at the 99th percentile and 25% performed at the 1st percentile would earn an API of 800. Similarly, a school in which 57.5% of students performed at the 60th percentile and 42.5% performed at the 40th percentile would also obtain an API of 800. An API of 800 would also be obtained for a school in which 57.5% of students performed at the 79th percentile and 42.5% performed at the 59th percentile.

Table 3: Binary Combinations of Percentile Ranks that Produce an API of 800

	Percentile Rank	% of Students	API	Mean PR
Example A	99	75		
	1	25	800	88
Example B	80	75		

	1	25	800	52
Example C	99	75		
	19	25	800	94
Example D	80	75		
	19	25	800	66
Example E	60	57.5		
	40	42.5	800	51
Example F	79	57.5		
	40	42.5	800	64
Example G	79	57.5		
	59	42.5	800	71

While all of these examples are extreme and highly improbable, they illustrate three points. First, an API of 800 (or any other number for that matter) can be obtained by many different combinations of scores that may be relatively uniform or may differ dramatically within a given school. Second, in all cases the mean national percentile rank for a school that obtains an API of 800 will always be greater than 50, and in most cases substantially higher than 50. Third, beyond indicating that on average students are performing above the 50th percentile, an API of 800 (or any value for that matter) does a poor job characterizing the actual performance of students in a school.

[xiii] The effect size required to move students in California, on average, from where they are now to above the 60th percentile[xiii] (the level at which a school would meet the 800 target) range from .20 to .73. An effect size above .2 is considered to be of practical significance, while an effect size of .73 represents an extraordinary change. At first brush, it does not seem all that unreasonable that, on average, students within a school perform above the national average. To some, it also may not seem unreasonable to expect students to perform above the 60th percentile, on average. Recognize, however, that the 60th percentile is just over a quarter of a standard deviation above the mean. Thus, if student performance on the SAT-9 within a California school were distributed identically to the national norm group, this accomplishment would represent an effect size of approximately .25. In education, an effect size of .25 is considered moderate and is often viewed as having important practical significance. As Mosteller (1995, p. 120) states, "Although effect sizes of the magnitude of 0.1, 0.2, or 0.3 may not seem to be impressive gains for a single individual, for a population they can be quite substantial."

[xiv] To illustrate this situation, Table 4 models the relationship between the percentage of LEP students, the performance of second graders (assumed to be "normally" distributed for all non-LEP students and distributed evenly between the two lowest performance bands for all LEP students), and the performance required of grade 3-5 students in order to obtain an API of 800. In a school that contains no LEP students, the performance of grade two students is assumed to reflect that of students across the nation. Thus the API score based only on grade 2 students would be about 655.[xiv] Given this starting point for grade 2 students, 15% of students in grades 3-5 would need to perform between the 40th and 59th percentile rank and 85% of students must perform between the 60th and 79th percentile rank.

For a school that contains 20% LEP students, the performance of second grade students would be lower (on average), resulting in an API of 594 in this example. To offset the second grade API, 96% of students in grades 3-5 must perform above the 60th percentile. Clearly, this level of performance is unrealistic. As shown in table 5, however, this unrealistic expectation applies to nearly 50% of schools in California.

Table 4. Proportion of Grade 3-5 Students in Top Performance Bands Required to Offset Second Graders with Limited English Proficiency

% LEP	Grade 2 API	K-5 API	%3-5 40-59	%3-5 60-79	%3-5 80-99
0%	655	848	.15	.85	.00
10%	625	859	.09	.91	.00
20%	594	869	.04	.96	.00
30%	564	879	.00	.97	.03

40%	533	889	.00	.89	.11
50%	503	899	.00	.81	.19
60%	472	909	.00	.73	.27
70%	442	920	.00	.64	.36
80%	411	930	.00	.56	.44
90%	381	940	.00	.48	.52
100%	350	950	.00	.40	.60

Table 5. Percentage of Elementary Schools Serving LEP Students in California, 2001

% LEP	% Elementary Schools
0	4.1
1-10	30.7
11-20	16.2
21-30	12.3
31-40	10.8
41-50	8.2
51-60	6.5
61-70	5.5
71-80	3.8
81-90	1.5
91-100	0

Similarly, in middle and high schools, each year a large group of students enters whose skills and knowledge were shaped in different schools. Depending upon the success of the schools these students attended previously, this entering class could perform above, below, or at the national mean. Even if the school is successful in increasing the skills and knowledge of these entering students at a rate faster than is typical nationwide, these students may inflate or deflate the school API, depending upon their initial level and distribution of skills and knowledge. It is important to note that in the case of grade 6-8 middle schools and 9-12 high schools, this entering class represents roughly one-third of the total group of students tested each year (the SAT-9 is currently not administered to students in grade 12).

[xv] California's PSAA/API accountability system requires schools to close the gap between their API and the interim target of 800 by at least 5% each year. Since a school's API is based solely on student test scores, meeting this expectation requires an increase in students' scores. Rogosa (2000) estimates that a school's API score would increase by roughly 8 points if the performance of all students in the school increased by one percentile point. Thus, for schools whose current API is at or above 640, a universal increase of one percentile rank on all tests would produce satisfactory growth of at least 5% toward the target. Similarly, a universal increase of two percentile points on all tests would produce satisfactory growth for schools whose API is at 480. And below this, student scores must increase roughly 3 points. At first brush, Rogosa's estimation creates the impression that a one to three percentile point increase is reasonable. On most of the SAT-9 tests, this growth would be obtained if all students answered one more question correctly. No doubt, it seems reasonable that with additional instruction and another opportunity to take a given test, a student's score should increase by a few or more points.

[xvi] A school that moves the minimally required 5% a year toward the interim target is actually making relatively slow progress toward that goal. For example, a school whose students, on average, perform at the mean would obtain an API of 655. As Table 6 indicates, it would take 48 years for a school whose API began at 655 to reach the interim target of 800 if that school closed the gap by the minimum 5% each year. Similarly, for the median California high school, whose API is currently 636, it would take 52 years to reach the interim target. And for a low performing school, whose current API is 354, 71 years are required. In other words, if all schools met their growth expectation each year, four to seven generations of students would pass through California's schools before all schools met the interim benchmark.

Table 6. Years to Meet Interim API Target of 800, at 5% Annual Growth

	Base API	Years to Target
API Corresponding to All Students at Mean	655	48
Median Elementary School 2001 API	690	43
Median Middle School 2001 API	669	47
Median High School 2001 API	636	52
Low Observed 2001 API	354	71
Lowest Possible API	200	77

[xvii] A key requirement established by the Advisory Committee for the use of any measure as part of the API is that the measure be reliable. In 1999, Rogosa applied classical test theory to perform a series of analyses that explore the reliability of SAT-9 percentile rank scores. In his presentation, Rogosa poses a series of questions with accompanying answers. As an example, Rogosa asks, "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 points away from the 50th percentile?" His answer – 70%. Only 30% of the time will the student's actual score on the test be between the 45th and 55th percentile. As a second example, Rogosa asks, "What are the chances that a ninth-grade math student who is actually at the 60th percentile in both years experiences a 10 point change (up or down) in his percentile rank?" Rogosa's answer – 50%. That is, 50% of the time, a students' percentile rank may change by ten points when in fact the student is performing at the same level both years.

Rogosa's analyses identify a major problem in using percentile ranks and changes in percentile ranks to make decisions about the performance and improvement of individual students. These problems stem from the error in measurement that occurs for all tests, the SAT-9 included.

[xviii] It is important to note that California currently does not use test scores to make decisions about individual students. Instead, the scores of students within a school are aggregated. Because error in measurement is assumed to be random, the mean error across sufficiently large numbers of students approaches zero. Thus, when the sample of students is large, the use of aggregate scores is less vulnerable to measurement error.

To illustrate the volatility of test scores aggregated to the school in another way, Kane and Staiger (2001) ranked schools in North Carolina by their average test score levels and average score gains between 1994 and 1999. The proportion of times each school ranked in the top 10 percent over the six-year period was then counted. As the authors describe,

> If there were 'good' and 'bad' schools which could be observed with certainty, I might expect to see 90 percent of schools never ranking in the top 10 percent and 10 percent of schools always ranking at the top. At the opposite extreme, where schools were equal and the top 10 percent were chosen by lottery each year, I would expect 47 percent [of] schools ranking in the top 10 at least once over 6 years and only 1 in a million ranking in the top 10 percent all 6 years.

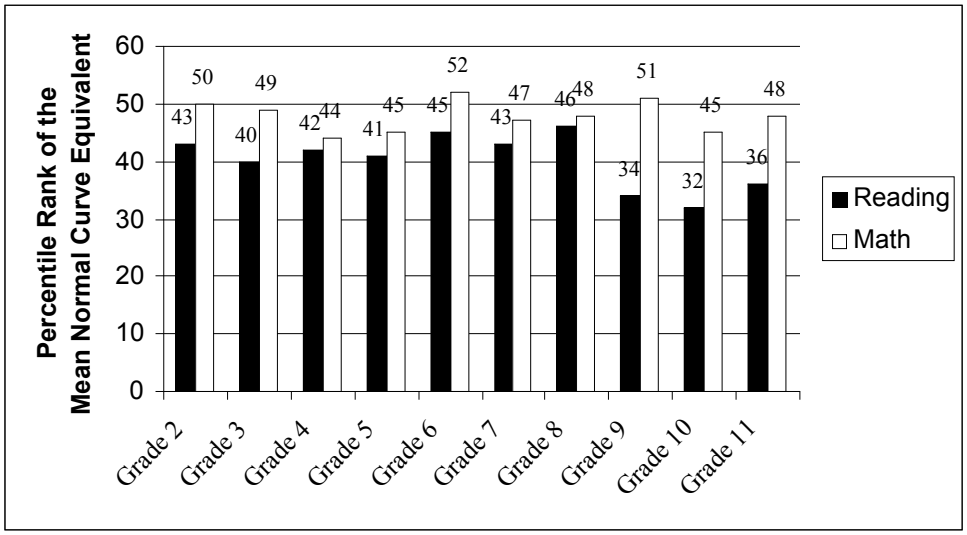
For the math scores, Kane and Staiger found that between 31 and 36 percent of schools ranked in the top 10 percent at least once over the six year period based on mean test score or on mean gain score. In addition, less than one percent of schools consistently ranked in the top 10 percent all six years. For reading scores, they found that no school ranked in the top 10 percent across all six years. The authors concluded "the rankings generally resemble a lottery, particularly in gain scores" (p. 9-10).

California addresses the problem of volatility in two ways. First, scores are aggregated across all grades within a school rather than within each grade level. As a result, even in schools that have relatively small numbers of students in each grade, the total number of students the school API is calculated from is usually

larger than 100. Second, for those schools that contain fewer than 100 students, PSAA specifies that an alternative API system will be established.

Although aggregation of scores across grade levels may help decrease the volatility of score changes, it presents at least two additional challenges. First, aggregation across grade levels masks differences in performance and/or gains at different grade levels. As noted above, students in California perform worse on average than students across the nation on the SAT-9. But this underperformance is not uniform across grade levels. Figure 1 indicates that grades 9-11 perform noticeably worse than all other grades on the SAT-9 Reading test. For grades 2-8, mean SAT-9 scores differ between grade levels by as much as 6 points on the reading test and 8 points on the math test. And, whereas grade 3 has the lowest mean NPR for grades 2 to 8 on the reading test, it is one of the top three scorers in math for those grades; grade 4 is at the bottom.

Figure 1: California 1999 SAT-9 Mean Performance by Grade Level



[xix] By way of example, Haney and Raczek recount the work of Robinson (1950) who performed a series of analyses to explore the relationship between race and illiteracy using 1930 U.S. Census data. When regions of the country were the unit of analysis, the correlation was 0.95. When state averages were calculated and then correlated, the correlation dropped to 0.77. And when individuals were the unit of analysis, the correlation was only 0.20. Thus, depending upon the unit of analysis, correlations can vary widely.

[xx] The ecological fallacy associated with using aggregates to summarize student performance is relevant to the API and PSAA in at least two ways. First, the focus on school-level performance across grade levels rather than within grade levels or classrooms obfuscates the impact of efforts within these lower-level units to improve student learning. Second, although the Similar School Index is not used to inform formal decisions about the success or shortcomings of schools, the focus on school-level performance and characteristics may promote fallacious conclusions about the impacts of school-level programs and the influence other variables have on the success of these programs. While aggregation at the grade or

classroom level may be a poor fix for this second problem, it might promote closer examination of practices and issues within these smaller operational units.

[xxi] Figure 2. SAT-9 Reading Mean Percentile Ranks for 1998-2001

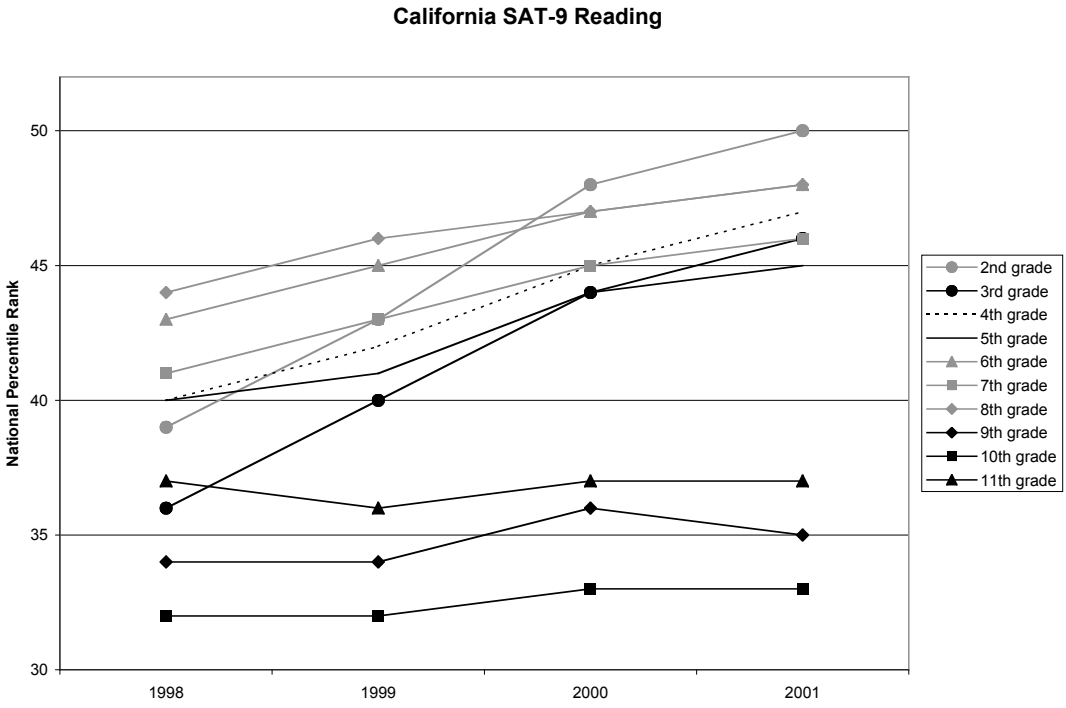
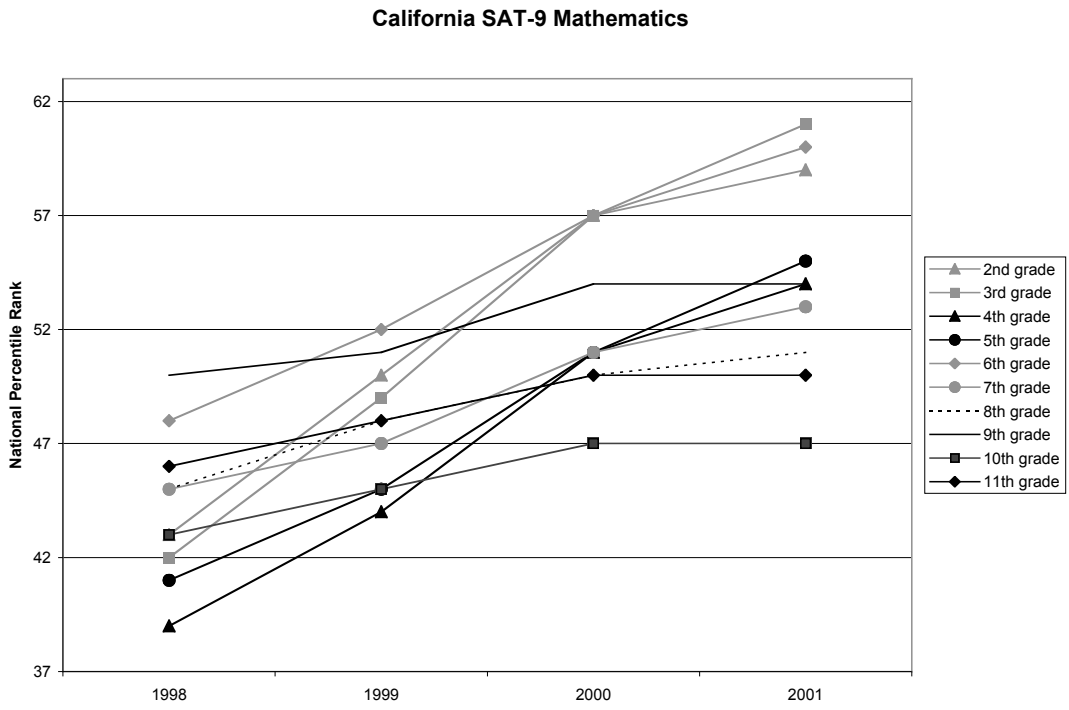
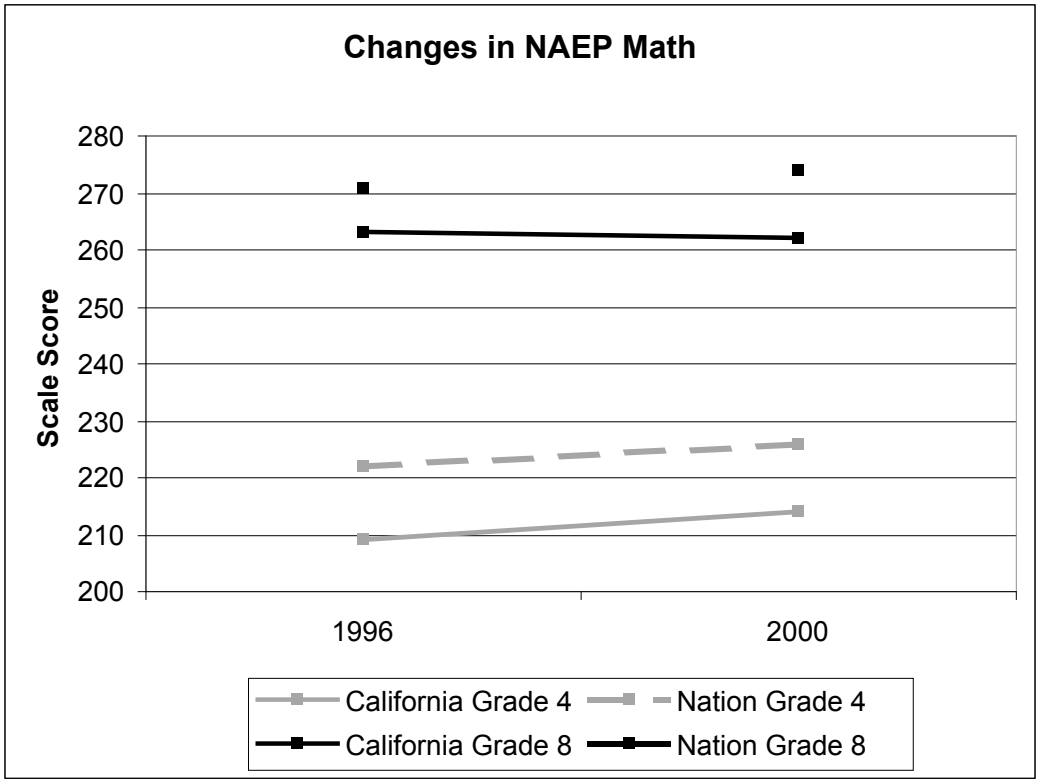


Figure 3. SAT-9 Mathematics Mean Percentile Ranks for 1998-2001



[xxii] The PSAA/API system has not been in place long enough to compare gains on the SAT-9 Reading test with changes on NAEP (note that the most recent NAEP Reading administration was 1998). However, gains in the SAT-9 Mathematics test can be compared to gains on the NAEP Mathematics test. As figure 4 displays, between 1996 and 2000, California's NAEP Math mean scale score increased from 209 to 214 in grade 4 while the national average increased from 222 to 226. In eighth grade, the national average also increased from 271 to 274 while California's average decreased slightly from 263 to 262. In both grade levels, California performed below the national average. This low performance, however, likely results in part due to the high percentage of ESL students in California. The more important comparison, however, is the comparative gain/loss made at the national level and in California.

Figure 4. Changes in NAEP Mathematics Scores



[xxiii] On the 4th grade test, students were asked to summarize information after reading a short article. In 7th grade, students were asked to write a response to literature after reading a short story. Responses were scored on a four point scale by two raters (htpp://www.cde.ca.gov/statetests/star/cst2001writing.pdf). These scores were then added together to produce a score ranging from 2 – 8.

[xxiv] Table 7. Summary of 2001 CST Writing Test Scores

	4th Grade (451,492 students)	**7th Grade (429,973 students)**
% scoring 8	0%	0%
% scoring 6-7	14%	6%
% scoring 4-5	62%	33%
% scoring 2-3	23%	60%

Source: http://star.cde.ca.gov/star2001

[xxv] Developing the API required decisions about operational definitions for five distinct variables embedded in the accountability system. Four of these decisions focused on the calculation of an API:
• the selection of indicators of which the API is composed,
• the relative weight of each chosen indicator,
• the selection of cut-scores for "performance band" allocation for indicators, and
• the relative weight for each performance band.

The fifth decision focused on the choice of an API target score.