

PUTTING EDUCATION TO THE TEST: A VALUE-ADDED MODEL FOR CALIFORNIA

By Harold C. Doran and Lance T. Izumi
June 2004



755 Sansome Street
Suite 450
San Francisco, CA 94111
Phone: 415-989-0833 / 800-276-7600
www.pacificresearch.org

Putting Education to the Test: A Value-Added Model for California

**By Harold C. Doran and Lance T. Izumi
June 2004**

755 Sansome Street, Suite 450
San Francisco, California 94111
Phone: 415-989-0833 / 800-276-7600
www.pacificresearch.org

Putting Education to the Test: A Value-Added Model for California

**By Harold C. Doran and Lance T. Izumi
June 2004**

ISBN 0-936488-90-5

Pacific Research Institute
755 Sansome Street, Suite 450
San Francisco, CA 94111
Tel: 415-989-0833 / 800-276-7600
Fax: 415-989-2411
Email: info@pacificresearch.org
www.pacificresearch.org

Additional print copies of this study may be purchased by contacting us at the address above,
or download the PDF version at www.pacificresearch.org.

Nothing contained in this briefing is to be construed as necessarily reflecting the views of the Pacific Research Institute or as an attempt to thwart or aid the passage of any legislation.

©2004 PACIFIC RESEARCH INSTITUTE. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopy, recording, or otherwise, without prior written consent of the publisher.

TABLE OF CONTENTS

Executive Summary	1
I. Introduction: A California Overview	8
II. California’s Testing and Accountability System	10
III. The Evolution of Test-Based Accountability	13
IV. Accountability Systems and the Inadequacy of Simple Design	14
Figure 1: Targeting Instruction for “Bubble” Students	15
V. Value-Added Analysis	17
VI. VAM: A Brief History	18
VII. A Statistical Model for Value-Added Analysis	20
VIII. A Contextual Growth Model	22
Figure 2: An Illustration of REACH for Two Students	23
Table 1: Sample REACH Computations	24
IX. The School-Level Contextual Model	25
Figure 3: A Multidimensional View of School Effectiveness	25
Figure 4: An Illustration of a Multidimensional School Plot	26
X. Preparing for Value-Added Analysis	27
XI. Common Criticisms	28
XII. Using the Model for Accountability System Design	28
XIII. The REACH VAM and NCLB	30
XIV. Legislative Action in Other States	32
XV. Conclusion	33
Endnotes	36
References	37
About the Authors	39
About PRI	41
Donation Form	42

ACKNOWLEDGEMENTS

The Pacific Research Institute is grateful to the many generous donors who support its education program. In particular, we thank the Arthur N. Rupe Foundation and several anonymous donors for providing funding for this publication.

EXECUTIVE SUMMARY

This paper examines the following:

- the concept of a value-added model,
- the merits, limitations, and common criticisms of this model,
- value-added models used in two jurisdictions,
- and a value-added model that could be adopted by California to improve the state's accountability system.

California's Testing and Accountability System

California's school accountability system was enacted in 1999. The testing devices, the California Standards Test and the California Achievement Test, are aligned with the state's academic content standards. Student test scores are used to calculate a school's score and ranking on the state's Academic Performance Index (API). An annual API growth target is also calculated for each school.

Overview

- Determinations of school performance on the API are based on the average performance of students at the school on the tests in a given year. The state uses these schoolwide averages to calculate test-score growth targets for each individual school.
- Using annual schoolwide averages causes the "snapshot" reporting problem. For example, student populations at schools change from year to year, especially in schools with large immigrant populations. Yet, the school's growth targets are based on student performance from the year before, even though many of those students are no longer at the school and different students with different achievement levels have taken their place.
- Also, because the API is based on average schoolwide student test scores, focus on growth or decline in achievement of individual students is de-emphasized. Rather than focusing attention on the individual needs of students, the current California testing system, and API, causes schools to focus on raising the performance of a large enough group of students in order to meet the state's growth targets.

Current Issues

- To avoid the “snapshot” problem, the performance of individual students on California’s tests should be tracked over a period of time. Such a tracking system is currently being assembled.
- In 2002, a law was enacted that required the state Department of Education to contract for the development of proposals for a longitudinal pupil achievement database consisting of student achievement data on the state’s various testing devices.
- In 2003, California adopted a law that provided for a system of unique student identification numbers that would allow the tracking of individual student performance on state assessments through the student’s school career in elementary, middle, and high schools.
- According to the current timeline, all students in California would receive a unique identification number by the 2005–06 school year. The data system would be available for analytical use by 2007.
- The 2003 law also set up a state-appointed advisory committee that would make recommendations to the state superintendent of public instruction “on the appropriateness and feasibility of a methodology for generating a measurement of academic performance by utilizing unique student identifiers for pupils in kindergarten and any of grades 1 to 12, inclusive, and annual academic achievement growth to provide a more accurate measure of a school’s growth over time.”
- It would be up to the state superintendent, with the approval of the state Board of Education, to implement this recommended value-added measurement of student performance.

Inadequacy of Current Accountability Systems

- Most accountability models wrongly focus on current status. Current status measures are concerned with how groups of students have performed at a single point in time, ignoring how much they have improved over a specific period of time.

- Accountability models based on current status encourage schools to target instruction to middle-performing students rather than encouraging growth for all students within a school. It is easier to get middle-performing students past the proficiency benchmark score.
- Because students are not randomly assigned to schools, accountability models based on current status provide misleading and invalid results. Differences among schools may not be due to differences in instructional quality, but may be due to other factors such as students' economic status.
- Simply comparing percentages of students from year to year that hit the proficiency benchmark score overlooks the progress that a school may be making with students who are still some way off from hitting the proficiency target.
- Current status scores are cumulative. For example, a test score for grade 8 is invalid for evaluating grade 8 instruction since it reflects the cumulative impact of schooling from all previous years. It represents the sum of all good and bad teachers that students have during their school careers, not the isolated impact of the grade 8 teacher.
- Badly conceived and constructed accountability systems result in unreliable and uninformative data, thus failing to paint an accurate picture of school quality. This means that high-stakes decisions regarding school effectiveness are likely to be flawed and result in incorrect decisions.

Value-Added Analysis

- In general, value-added models are a class of statistical procedures that use longitudinal test score data, i.e. data collected over a period of time, to measure the change in a student's performance during a specific period of time.
- Value-added models can measure how a student's performance is growing toward a targeted outcome, such as the "proficient" standard on a state test.
- Tennessee was the first state to implement a value-added model as the basis of a school accountability program.

- The Tennessee model analyzes individual student performance on state tests and also includes the cumulative effect of teachers on student performance.
- The Tennessee model allows for individual student data to be linked to specific teachers, meaning that the effectiveness of individual teachers could be estimated.

A Model for Reform

The REACH (Rate of Expected Academic Change) value-added model focuses on the achievement growth of individual students and measures that growth not in comparison to other students but instead against the goal of subject-matter proficiency. Under the REACH value-added model (REACH VAM), the growth rate toward proficiency can serve as a tool for targeting remedial assistance to students.

The model can also form the basis for important state education reforms:

- **Better evaluation of policies and programs.** Because the REACH model measures and projects how each individual student is progressing toward proficiency, it can be used to evaluate whether a student's exposure to a particular education program or reform helped or hurt that progress.
- **Promotion of better instruction.** By measuring student achievement gains under individual teachers who may be using similar or different teaching methodologies, the REACH VAM can inform lawmakers, education officials, teachers, and the public about which instructional practices are best able to move students toward subject-matter proficiency.
- **Better measurement of teacher effectiveness.** Since the REACH VAM focuses on student achievement growth toward subject-matter proficiency, it can help identify schools that raise student achievement and ineffective schools that do not. Based on this identification, incentives can be given to effective teachers to teach in classrooms with low-performing students and compensation systems can be crafted based on teacher effectiveness.
- **Improve teacher professional development.** By showing if students are not growing toward subject-matter proficiency, the REACH VAM can individualize professional development to address teacher weaknesses.

The Proposed Value-Added Model for California

- The proposed model suggests that all students should reach the proficient benchmark on state tests within a specified timetable, such as by the time they are in the highest grade in a school. This encourages students to grow towards an outcome of value—proficiency—rather than making comparisons among students.
- By comparing a student’s estimated growth rate with the growth rate the student will need to attain in order to hit the proficiency benchmark, a ratio called the REACH ratio is obtained.
- The REACH score will tell how much each student’s scores have to grow in order for that student to become proficient. The yearly growth varies by student, but the end goal, proficiency, is the same for all.
- The REACH score answers the question, “given a student’s current location on the ability scale, how much does he need to grow each year in order to be proficient by the time he leaves this school?”
- A REACH ratio of “1” or greater means that a student is likely to be proficient by the time he leaves the highest grade in the school, while a ratio of less than “1” means that he is unlikely to hit the proficiency benchmark.
- The REACH score can also be used to compute the percentage of students within a class, grade, school, or district that are on track towards meeting the proficiency benchmark.
- It is the linkage of the REACH score to the state’s proficiency benchmark that differentiates the REACH model from other value-added models.
- The REACH model can also enable policymakers to differentiate between two schools that have the same percentage of students at or above the proficiency benchmark, but are making different gains. Schools making higher yearly gains are more effective.

The REACH Model and NCLB

The REACH VAM can also help the state meet the federal No Child Left Behind Act (NCLB) proficiency requirements. At present, the federal and state accountability systems are not in sync. Also, the way the state has structured its targets for meeting the NCLB proficiency goals almost guarantees that schools will come up short.

- The state's NCLB growth targets for the percentage of students in a school hitting the proficient mark encourage schools to focus year-by-year on those students closest to hitting the proficiency benchmark in order to meet the target requirements.
- The REACH model will help ensure that all students meet NCLB achievement goals. By focusing on achievement progress among all students, the REACH VAM will help prevent schools from concentrating only on those students just below the proficiency bar and will help ensure that the lowest-performing students are not left behind.
- The REACH model will also reduce pressure to redefine "proficiency" downward. Since the REACH VAM provides the information needed to bring all students up to the current state definition of proficiency, reducing the stringency of the definition may not be necessary if the information provided by the model is used to target instruction and assistance to individual student needs, and if it is used to support effective education programs and reforms.

Legislative Action in other States

As California takes initial steps to create a value-added testing system, policymakers should be aware of similar efforts in other states. In Colorado, legislation has been introduced that would establish a value-added system that would use students' state assessment scores over time to measure academic growth.

- Under the Colorado proposal, school participation in the value-added system would be voluntary for the first year, but then be made mandatory for all school districts.
- The state Department of Education would be instructed to use collected longitudinal data on test scores to determine the levels of increase that constitute a full year of academic growth in reading, math, and writing for each grade level tested.

- The department would also be required to provide school districts with an academic growth information report for each student. The report would include the student's test scores and growth amounts in reading, math, and writing over the period between administrations of the test.
- The department would contract with a private or public entity to calculate annually the amount of each student's academic growth, based on test scores, in all three core areas.
- Each student's academic growth profile would be given to the principal at the student's school. The profile would also be shared with the student and his or her parents in discussing the student's academic strengths and weaknesses, as well as strategies to increase the student's academic growth and achievement.
- The Colorado proposal lays out a structure for a value-added testing system, but does not specify the exact type of value-added model to be used.

Conclusion

The REACH VAM discussed in this study is consistent with the intent of No Child Left Behind and would provide useful information related to individual students and their progress towards proficiency.

Value-added analysis is a critical component for all states in their reform of the education system. And it is particularly essential for California's current effort to improve accountability and student achievement.

I. INTRODUCTION: A CALIFORNIA OVERVIEW

Over the past few years, California has made important strides in improving its K–12 public education system. The state has adopted rigorous academic content standards, implemented a standards-aligned state testing system, and crafted a school accountability system. These reforms, however, can be advanced even further through the adoption of a value-added analysis.

Under a value-added model, individual student testing data are analyzed longitudinally, that is, over a period of time. Using this longitudinal data, policymakers can better evaluate education policies and programs, promote better instruction, better measure school effectiveness, and improve teacher professional development.

Increased accountability for student learning as measured by educational tests has become the *sine qua non* of school reform. This notion became most clearly institutionalized via the strong bipartisan support for the federal No Child Left Behind Act of 2001 (NCLB). Its strong emphasis on test-based accountability has resulted in states developing methods for measuring the effectiveness of all public schools.

The capstone of accountability within the legislation—Adequate Yearly Progress (AYP)—quantitatively defines the expected increase in the percentage of students scoring at or above “proficient”¹ on the state test each year. Making AYP in reading and math on a yearly basis permits a school to avoid a series of sanctions associated with NCLB that may ultimately lead to significant changes to the governance and organization of the school.

While the goals of increased accountability are laudable, the means by which schools are held accountable has been the source of great debate (Hess, 2003). For example, some argue for gentler means such as providing educators with more resources and training. Others prefer tougher approaches involving incentives and sanctions to ensure improved student achievement. However, those responsible for developing and implementing accountability plans are likely to share two common goals irrespective of the manner in which the accountability system is fashioned.

First, accountability systems should provide useful data to educational practitioners working within a school system. As such, practitioners should be provided with meaningful and relevant information that can be used to improve the quality of instructional delivery. Second, the accountability system must also provide accurate and reliable information reflecting the quality of the educational program.

In other words, the data should serve an internal function to support appropriate *classroom action*, i.e., instructional consequences that originate from within a school system. At the same

time the data should serve an external function to support appropriate *public action*, or social consequences that originate external from the school.

O'Day (2002) succinctly characterized the manner by which accountability information can support the goal of improved classroom action:

In particular, I argue that accountability systems will foster improvement to the extent that they generate and focus attention on information relevant to teaching and learning, motivate individuals and schools to use that information and expend effort to improve practice, build the knowledge base necessary for interpreting and applying the new information to improve practice, and allocate resources for all of the above. (p. 294).

Public action, on the other hand, will be more likely to be spurred when high-quality data sources are provided that represent a fair and accurate portrait of school effectiveness. When classroom and public action are maximized via reliable accountability models, they should both converge upon the same end result—improved levels of student achievement.

Achieving this goal necessitates the development of accountability systems that use information from multiple sources that more reasonably relate to the notion of individual student learning. In particular, one must ask, “What methods of accountability system design are likely to provide information that will maximize appropriate classroom and public action?”

This study describes in detail one method likely to support test-based accountability systems that appropriately address these issues, value-added models (VAMs) of student achievement. From the outset, we acknowledge that VAMs alone do not represent an accountability system. Instead, multiple measures of educational quality should be included within the design of high-stakes accountability systems. However, for clarity of presentation, we intentionally limit our discussion to the analysis of student achievement test score data using longitudinal methods of measuring student progress.

This report provides the historical basis of test-based accountability systems and presents a general modeling strategy, its potential role to support the design of accountability systems, and its limitations. Our aim is to provide adequate information such that educational leaders and policymakers may fully understand VAMs, their potential as a method for measuring AYP, and the prerequisites for implementation.

This study consists of two parts. The first section examines California's current testing and accountability system and critiques the system's flaws which could be remedied by a

changeover to a value-added model. The second section contains the specific discussion of VAMs and the suggested model, plus recommendations for reform.

For brevity and illustration, we limit our discussion to a general latent growth model; however, more detailed descriptions of additional value-added statistical models may be found in Doran and Lockwood (2004), Lockwood, Doran, and McCaffrey (2003), McCaffrey et al (2003a), McCaffrey et al (2003b), and Thum (2003). Conceptual, and less statistical, descriptions of value-added analysis may also be found in Doran (2003) and Drury and Doran (2003).

II. CALIFORNIA'S TESTING AND ACCOUNTABILITY SYSTEM

In April 1999, California passed the Public Schools Accountability Act (PSAA), the brainchild of then newly elected Governor Gray Davis. The PSAA has three major components.

First, the Academic Performance Index (API) provides individual schools with a numerical score based originally on multiple measures of performance, e.g., test scores, dropout rates, and attendance rates. For the time being, however, the API is based exclusively on student test scores.

Second, the rewards program called the High Performing-Improving Schools Program (HP/ISP) awards schools and staff monetary bonuses if they meet or surpass API growth targets. The third component is the Immediate Intervention-Underperforming Schools Program (II/USP) that allows the state to intervene in schools that fail to meet targets for improving test scores. The intervention-program portion of the legislation also included sanctions, such as state takeover of individual schools and monetary grants to pay for the interventions. California's accountability system, thus, is a high-stakes program based on standardized tests as the measurement device of student learning.

From 1999–00 to 2000–01, the state calculated the API using only scores from the Stanford-9 standardized test, a norm-referenced test that compares student performance within California to a national sample of students who have taken the same exam. The state Board of Education selected the Stanford-9 to be the state's assessment device in 1997, several years before passage of the new accountability law.

The Board chose this test, in part, because as an "off-the-shelf" exam, it was readily available. As a nationally norm-referenced test, however, the SAT-9 was not well aligned with the state's academic content standards, meaning that it did not directly measure student performance relative to the California State Standards.

The state used the Stanford-9 to test students in grades 2–11, with students in grades 2–8 tested in reading, mathematics, written expression, and spelling. Students in grades 9–11 are

tested in reading, writing, mathematics, science, and history and social science. The state requires districts to provide individual student scores to parents. The state Department of Education's website publicly listed aggregate Stanford-9 scores by grade level for schools, districts, counties, and the state.

Because the Stanford-9 was not directly aligned to the state's rigorous academic content standards, teachers could teach to the standards, but the Stanford-9 may not be sensitive to what students were learning in the classroom. As a remedy, the state picked out parts of the Stanford-9 that tested the topics listed in the content standards and developed a set of standards-aligned questions that were added onto the Stanford-9 exam.

Later the state expanded this approach into separate standards-based tests called the California Standards Tests (CST). The CST is not a nationally norm-referenced test, but is a standards-referenced test, i.e., an exam aligned with the state's academic content standards that measures the performance of California students relative to those standards as opposed to a national sample of students.

The state gives CSTs in English and mathematics in grades 2–11. Students in grades 4 and 7 are tested on writing, and those in grades 9–11 are tested on the state content standards in science, history, and social science.² Scores on the CSTs are categorized using performance levels: advanced, proficient, basic, below basic, and far below basic. The goal is for all students to score at the proficient level or above.

In 2001–02, results from the English Language Arts CST were added to the Academic Performance Index calculation, counting for 36 percent of elementary and middle school scores and 24 percent of high school scores.³ In 2002–03, the CSTs took center stage. In grades 2–8, the English Language Arts and Math CSTs counted for 80 percent of the API calculation, with 20 percent from the norm-referenced test. For grades 9–11, the English, Math, and History/Social Science CSTs counted for 73 percent of the API calculation, 15 percent from the High School Exit Exam, and 12 percent from the norm-referenced exam.⁴

In 2003, California replaced the Stanford-9 exam with another norm-referenced exam, the California Achievement Test-6 (CAT-6) which is aligned to California's academic content standards. California, therefore, continues to use two tests, the standards-referenced CST and the norm-referenced CAT-6. The reasons for retaining a nationally normed test like the CAT-6 include the necessity of having a "basic skills" test that acts as a check on how California's self-assessed progress measures up against national norms of performance.⁵

The initial complaint that California's state tests were not aligned with the state standards has now been addressed. However, the state testing system continues to be plagued by the equally

important “snapshot” reporting problem. In other words, determinations of school performance are based on the average performance of students at the school on the tests at one point in time. The state uses these schoolwide averages to calculate test-score growth targets for each individual school.

When using only Stanford-9 results, the state Department of Education calculated a score ranging from a low of 200 to a high of 1,000 for each school. The interim statewide API target for all schools is 800. The state Department of Education also ranked schools on a 1-to-10 decile ranking scale with 10 being the best. The department uses a separate “similar schools” ranking to compare schools with other schools having similar demographic characteristics. These various features continue under the new standards-aligned assessment exams.⁶

Schools scoring below 800 must close the gap between their current score and the state performance target by at least five percent to meet their growth target for the year. For example, if a school’s 1999 Academic Performance Index score was 500, the school’s growth target would be $(800-500) * 5 \text{ percent} = 15 \text{ points}$.⁷

Each numerically significant ethnic or socio-economically disadvantaged subgroup at a school (that constitutes at least 15 percent of the school’s total pupil population and consists of at least 30 students) must have a growth target of 80 percent of the school’s growth target. Thus, if a school’s growth target was 15 points, each numerically significant subgroup at the school must improve by at least 80 percent of 15 points, i.e., by 12 points.

Again, however, all these calculations are based on average schoolwide student performance. This approach fails to consider a number of critical factors. For example, student populations at schools change from year to year, especially in schools with highly mobile populations. Yet, the school’s growth targets are based on student performance from the year before, even though some of those students may no longer be at the school and new students with different achievement levels may have entered the school and taken their place.

Also, because the API is based on average schoolwide student test scores, focus on growth or decline in achievement of individual students is de-emphasized. Rather than focusing attention on the individual needs of students, the current California testing system and API cause schools to focus on raising the performance of a large enough group of students in order to meet the state’s growth targets. A more detailed explanation of all these negative phenomena is contained later in this paper.

To avoid the “snapshot” problem, the performance of *individual students*, not successive groups of student cohorts, on California’s tests should be tracked over a period of time. Such a tracking system is currently being assembled.

In 2002, Governor Davis signed Senate Bill 1453, authored by State Senator Dede Alpert, that required the state Department of Education to contract for the development of proposals for a longitudinal pupil achievement database consisting of student achievement data on the state's various testing devices.

In 2003, Governor Davis signed Senate Bill 257, also by Senator Alpert, that provided for a system of unique student identification numbers that would allow the tracking of individual student performance on state assessments through the student's school career in elementary, middle, and high schools.

According to the current timeline, all students in California would receive a unique identification number by the 2005–06 school year. The data system would be available for analytical use by 2007.

The 2003 law also set up a state-appointed advisory committee that would make recommendations to the state superintendent of public instruction “on the appropriateness and feasibility of a methodology for generating a measurement of academic performance by utilizing unique student identifiers for pupils in kindergarten and any of grades 1 to 12, inclusive, and annual academic achievement growth to provide a more accurate measure of a school's growth over time.” It would be up to the state superintendent, with the approval of the state Board of Education, to implement this recommended value-added measurement of student performance.

California is, therefore, well on its way to setting up the foundation for a value-added model. The question, then, is what value-added model will eventually be adopted and to what uses will it be put? The following section of this paper provides potential answers to these questions.

III. THE EVOLUTION OF TEST-BASED ACCOUNTABILITY

Prior to the 1970s, inputs were the primary indicator of school quality. These inputs traditionally included per pupil funding allocations, student-teacher ratios, and teacher qualifications (Drury and Doran, 2003). However, *A Nation at Risk* (1983) sparked increased dissatisfaction with public education, resulting in public demand for improved levels of student achievement. Subsequently, standards-based reform initiatives led to a greater demand for student achievement outcomes to be used as an indicator of school quality.

Three intimately connected components of this reform movement, *standards*, *assessment*, and *accountability*, act in tandem encouraging educators to focus on a defined set of skills, measure student progress, and hold educators and students accountable for the achievement

results. Consequently, the role of tests as a primary agent of education reform became central to educational policymaking.

In particular, Linn (1993) observed that assessments became more prominent within the policy context for two reasons. First, the data from large-scale assessment document school quality and serve as evidence that change may be needed. Second, the test itself can be used as the agent of change by attaching high-stakes, such as graduation, merit pay, or school takeover, to test results.

For example, data from a large-scale state test can be used to document that the quality of an educational program is poor if very few students score at or above a specified cutpoint, such as the 50th percentile or the “proficient” category. The same test can now be used as the agent of change by holding the school accountable for increased student achievement as measured by the test.

While the policy motivation for test-based accountability is clear, its appropriateness as an instrument of accountability has not been without legal challenge. In 1981, the precedent setting case, *Debra P. v. Turlington*, was argued in the Fifth Circuit Court of Appeals. The opinion of the court supported the use of the Florida competency test as a high school accountability requirement if it could be demonstrated that students had been given adequate opportunity to learn (OTL) the tested material (Jaeger, 1989; Wang, 1998). Consequently, the legality of tests as instruments of accountability is substantiated when it can be demonstrated that students have been provided with adequate learning opportunities.

IV. ACCOUNTABILITY SYSTEMS AND THE INADEQUACY OF SIMPLE DESIGN

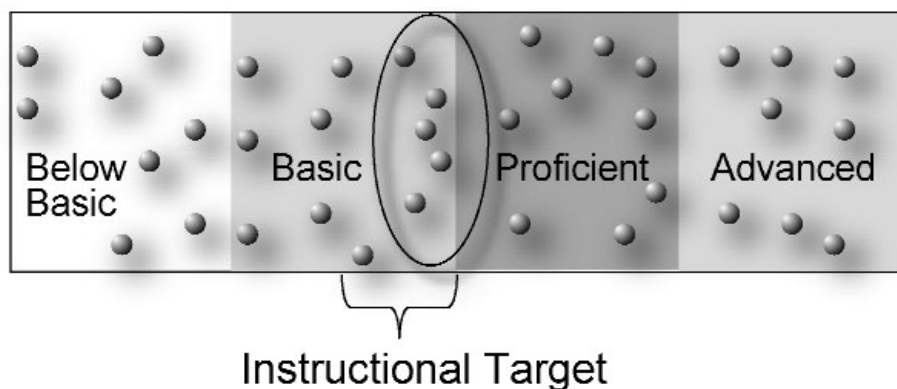
As tests became more formally established as mechanisms for change, their role within the design of accountability systems blossomed. Most often, test results were analyzed using simple statistical procedures, such as computing group averages or the percentage of students at or above a specific cutpoint on the test. For example, it is common to find the scores of fourth grade students in 1999 compared with the performance of fourth grade students in 2000.

In other words, completely different cohorts of students are being compared at a similar point in time. Despite the obvious lack of a meaningful comparison, these so-called cross-sectional comparisons have been at the heart of accountability system design. This has certainly been the case in California. When state test results are released, news stories and comments by education officials focus on cross-sectional comparisons.

Accountability models designed in this fashion have the wrong central focus—current status. Current status measures are concerned with how groups of students have performed at a single point in time with no regard to how much they have improved over a specific period of time. As such, status-based accountability systems are likely to encourage schools to target instruction for middle performing students rather than encourage growth for all students within a school. Why is this?

The reason has to do with the ease of getting middle-level students past the scoring benchmark, or cutpoint, that has been denoted as signifying proficiency. As illustrated in Figure 1, status-based systems persuade schools to identify and target instruction for students nearest the proficiency cutpoint, while ignoring the instructional needs of high achieving and low achieving students.

Figure 1: Targeting Instruction for “Bubble” Students



When schools focus on the “bubble” students, they are likely to show an increase in the percentage of students at proficient, at the expense of leaving many students—both at the high and low ends—without targeted instruction to meet individual needs. As a result, very high and very low performing students are set aside, while schools focus on those middle performing students in the high end of the “basic” category.

In addition, status-based accountability systems built upon simple approaches provide misleading and invalid results for at least three reasons (Doran, 2003). First, students are not randomly assigned to schools or classrooms, resulting in preexisting differences across school populations to begin with. As such, differences among schools may not be due to instructional quality, but instead due to other characteristics such as a student’s economic status.

To illustrate the problem of differences between student populations, consider a hypothetical school with 50 percent of its students in 2003 coming from a low-income background. In 2004,

because of students moving into and out of the school, 70 percent of its students now come from a low-income background. Test score differences from one year to the next, therefore, may be due to the change in its student population that is not detected by cross-sectional comparisons under a status-based model.

Second, comparing the percentage of students at or above a cutpoint inherently dismisses the fact that a school may be making tremendous progress with its students, but is some distance away from having all students at or above the cutpoint. Because cutscore categories are gross measures of academic performance, it is possible for students to make significant progress within a category, but remain within the same category on subsequent measurements.

It is also possible that schools that start with high-performing students will decline in performance, but remain at the cutpoint. Clearly, this type of school is doing worse each year, but remains above the AYP threshold, thereby avoiding the NCLB series of consequences.

For example, the total reading scale scores corresponding to the basic category on the Intermediate 1 version of the Stanford 9⁸, which was used as the main California assessment until 2002, ranges from a lower bound score of 607 to a higher bound score of 646. Translated into percentile ranks, this means a student can grow from the 26th percentile to the 56th percentile, yet still remain within the same performance category (Harcourt Brace, 1997). It is practically and statistically inefficient not to capture this information.

Third, current-status measures are cumulative in nature (Meyer, 1994). For example, a test score for a Grade 8 student is invalid for evaluating Grade 8 instruction as it reflects the cumulative impact of schooling over all previous school years. In other words, the Grade 8 score is the sum of all good and bad teachers a student has had over his entire schooling career to date—not the isolated impact of the Grade 8 teacher.

The end result of these poorly conceived and constructed accountability plans is twofold. First, the data are both irrelevant and uninformative, thereby failing to meet the criteria described for maximizing classroom action. For example, cross-sectional methods often provide teachers with information regarding students who are no longer enrolled in their classrooms. Second, they fail to paint an accurate picture of school quality, a necessary component of public action.

For example, large fluctuations in test scores may be completely due to changes in the school's population, not instructional quality. Therefore, high-stakes decisions regarding school effectiveness, when based on faulty information, are likely to be highly flawed and will result in many incorrect decisions regarding the quality of the educational program. California's accountability system, unfortunately, is characterized by both these defects.

Beyond these issues, the primary issue of neglect in status-based accountability systems is that they fail to appropriately measure student learning. By definition, learning must be measured by the extent to which an individual student has changed in knowledge and skills over time from one measurement occasion to the next. In other words, *learning* is synonymous with *change*. Consequently, measuring individual student growth is both reasonable and appropriate. Clearly, knowing where a student *is* describes nothing unless we know where the student *was*!

However, status-based systems simply describe where a group is—i.e., the current mean score—not how individuals within that group have increased in knowledge and skills during their tenure at any given school.

Thus, in California, the API ranking provides an aggregate snapshot of student performance at a school rather than showing how much growth in learning there has been among individual students.

V. VALUE-ADDED ANALYSIS

In contrast to the cross-sectional methods commonly found in accountability systems, value-added models more reasonably align with the notion of learning and measure the extent to which an individual student has increased in knowledge and skills over a specific time period. That is, gains in test scores serve as an indicator of increased academic knowledge. By using a student's entire test score record (e.g., Grades 2, 3, 4, and 5) it is possible to evaluate trends in student/group growth over time to assess how much learning has occurred while attending a particular school.

In general, value-added models are a class of statistical procedures that use longitudinal test score data to measure the extent to which a student has changed during a specific period of time. Motivation for VAMs has increased due to a belief that they can adequately determine how a student is growing over time and statistically attribute the gain to a school or teacher (Sanders, Saxton, Horn, 1997). Essentially, VAMs are an attempt to determine “how much *value* has a school *added* to a student's learning?”

There are some inherent problems in this question. It assumes that one can isolate the “effect” of the school aside from other non-school factors, such as family background, that may have affected a student's test score. Because students are not randomly assigned to schools, all school evaluations are quasi-experiments rather than true scientific experiments, and suffer from difficulty in determining the causal effect of a school on student learning and achievement. Second, the

effect is commonly defined in most value-added models as the deviation of a student or a school from a mean growth trajectory—hardly a useful measure in a standards-based environment.

Any failure to reconcile the influence of non-school-related factors in quasi-experiments of school performance confounds the causal inferences sought and produces estimates of school performance plus that of the exogenous variables. Controlling for exogenous variables may take the form of research design or statistical controls, or both. While the statistical methods can be applied to the analysis of quasi-experimental data (i.e., non-random data sets), they can never fully isolate the impact of the school from other factors believed to affect a student’s test score.

Therefore, a second question, “how is a student growing towards an outcome of *value*?” is posed to examine how a student is growing in relationship to an expected outcome, such as the “proficient” standard on the state test. In this regard, each student’s growth rate is compared to an expected growth trajectory, resulting in criterion-based decisions, that is, decisions based on the performance standards/benchmarks determined for the test. This is discussed in detail in a later section.

All this having been said, the benefits of a value-added approach, when performed correctly, even with the limitations of quasi-experimental data, are far superior to the conventional approaches that have often been used to evaluate school performance. As Thum and Bryk noted (1997):

As Meyer (1993) has forcefully demonstrated, significant school improvements can be occurring but are masked under status-based accountability systems, especially under conditions of high student mobility and other factors that depress average scores. From a purely technical perspective, the arguments seem very clear: Anything other than a value-added-based approach is simply not defensible. (p. 102).

VI. VAM: A BRIEF HISTORY

A variety of analyses falling under the banner of value-added analysis have been in place for years. But it is likely that the Tennessee Value-Added Accountability System (TVAAS), developed by Dr. William Sanders, is the most well-recognized and first system to have been employed for an entire state (Ceperly and Reel, 1997). TVAAS is an aggressive, linear, mixed-effects statistical model that analyzes student achievement data collected from individual students over time.

The TVAAS system simultaneously analyzes individual student performance on state tests and “layers” teacher effects. The layering, or cumulative effect of teachers on student performance, permits for the student’s level score to be the sum of the current teacher in addition to all previous teachers experienced by this student.

The longitudinal nature of responses presents the opportunity to use each student as his or her own control, a statistical process otherwise known as a “blocking.” No other covariates (e.g., economic status, gender, ethnicity) are used to control for non-random assignment in the TVAAS model.

Because a student’s performance is compared to his own past performance, this “blocking” technique attempts to control for differences among students, such as race, socioeconomic status, and other factors. There is strong evidence, however, that blocking alone may not fully eliminate systematic bias (McCaffrey et al, 2003b).

The fact that longitudinal student responses are used also presents the threat of missing data. However, mixed-effects statistical models, such as those used by Sanders, do not necessarily require imputation or “filling-in” techniques if they meet a statistical assumption referred to as missing at random (MAR). Instead, the data that are available, regardless of the extent to which data are missing, can be used rather than deleting students from the analysis when they have a missing data point.

A second well known value-added example can be found in the Dallas Independent School District (DISD). The methodology developed by Webster and Mendro (1997) also uses longitudinal test score information to estimate the contributions of schools to each student’s learning. However, their approach differs from the Sanders model in a few ways. While the Tennessee system compares each student’s scores to his or her previous scores in order to control for biasing factors, the Dallas system uses demographic characteristics to control for biasing factors in addition to the blocking design.

Specifically, the methodology employed makes use of two statistical procedures. In the first stage, student results from a standardized test are regressed using Ordinary Least Squares (OLS) on “fairness” variables. These fairness variables are factors, or more technically “covariates,” thought to influence a student’s learning, including gender, ethnicity, language proficiency, and free and reduced lunch status.

The residuals from the OLS model are then used as the outcome in a less sophisticated mixed model, where covariates are again entered as statistical controls. Holland (2001) notes that the Dallas system:

attempts to ensure fair assessments of gain by comparing scores of individual members of a given [socioeconomic] group (such as black, urban, low-income) not to their own prior performance but to a mark adjusted by the average performance for their ‘group.’

Both the TVAAS and DISD approaches have been used to identify effective teachers and included as components in state and district accountability plans. In Tennessee, for example, individual student data has been linked to specific teachers. As a result, the effectiveness of individual teachers could be estimated.

Additionally, both approaches include the use of longitudinal standardized test scores and a complex statistical approach to estimate school effects. Yet, they differ from a number of perspectives. Namely, TVAAS does not use any covariates to adjust for differences among students. Second, the DISD model uses OLS regression in the first stage of analysis, a statistical method that makes assumptions concerning the data that may be untenable given the correlation among students within the same school.

VII. A STATISTICAL MODEL FOR VALUE-ADDED ANALYSIS

A mixed-effects statistical model taking the following general form $Y_i = X_i\beta + Z_i\theta_i + \epsilon_i$ (Pinheiro & Bates, 2000) is the basis for the modeling approach described in the remainder of this report where Y_i is an n_i -dimensional response vector of test score data for a single subject (e.g., reading or math), X_i is an $n_i \times p$ design matrix, β is the p -dimensional vector of fixed effects, Z_i is the $n_i \times q$ design matrix for the q -dimensional vector of θ random effects, and ϵ is the n_i -dimensional within-group error term.

Using a multivariate mixed model, it is possible to simultaneously account for the correlation among outcomes within students, the correlation among outcomes by students within the same school, use the entire vector of student observations regardless of missing observations, and “shrink” parameter estimates to adjust for unreliability. Though a full description of mixed-effects linear models is beyond the scope of this paper, the interested reader may consult Pinheiro and Bates (2000), Searle (1971), and Raudenbush and Bryk (2002) for a review of the statistical techniques used.

The general structure of the data considered in this report consists of repeated observations nested within students, with students nested within schools. We assume that each test score can

be linked to students over time using unique student IDs and that these students can be linked to schools over time.

Furthermore, we assume that the test has been vertically equated (Peterson, Kolen and Hoover, 1989) such that test scores represent a continuous development scale. Though not a mathematical assumption, we assume that the test instrument is appropriately aligned with classroom curricular goals and objectives such that the results obtained from the analysis provide information relevant to the instructional process.

As previously described, an array of value-added models exist that can be applied to the analysis of test score data. Contingent upon the availability of the data and the specific inferences sought, VAMs may range from simple gain score models to more computationally intensive models with crossed random effects (Lockwood, Doran and McCaffrey, 2003).

For simplicity and generality we present an unconditional VAM that can be used given many of the common data elements likely to be maintained by a state department of education (SEA). A full description of this model, its properties, and methods for fitting it in a statistical software package can be found in Doran and Lockwood (2004).

We currently ignore models with crossed-random effects for two reasons. First, these models require unique teacher identification numbers linked to student achievement test scores over time. It is uncommon to find that these numbers are maintained with a high degree of integrity.

Second, software programs are not fully optimized to adequately handle the very large data matrix created for this type of analysis. Though these models present significant advantages, they are currently not computationally feasible, a limitation soon to be overcome (Bates and Debroy, 2003).

The first goal of a value-added model is to formulate and fit a statistical growth model that accurately reflects the observed data. The following unconditional growth model is posed to assess the degree to which students and schools are growing with respect to time.

$$Y_{tij} = [\mu + \beta_0(\text{time})] + [\theta_{0j} + \theta_{1j}(\text{time}) + \delta_{0ij} + \delta_{1ij}(\text{time}) + \varepsilon_{tij}] \quad (1)$$

where t indexes time, i indexes student, and j indexes schools.

The specification of Equation (1) indicates that \mathbf{Y} is a linear combination of the structural portion of the model given by the grand mean (μ) and the main effect for time (β_0), and the entire stochastic portion of the model including the school and student level random effects and the within-group residual.

Extensions to this model can easily accommodate additional fixed effects, such as gender and ethnicity, be reformulated as doubly-multivariate (i.e., more than one response variable),

as well as incorporate conditional standard errors of measurement estimated from an IRT-based test.

Given the parameterization of (1), computing the estimated true growth rate of student i in school j is simply:

$$ETGR_{ij} = \beta_0 + \theta_{1j} + \delta_{1ij} \quad (2)$$

which produces the average estimated gain of each student in the scale score metric.

VIII. A CONTEXTUAL GROWTH MODEL

Using the proposed modeling framework above, one obtains the estimated true growth rate for each student relative to the scale used in the analysis from Equation (2). For example, one may find that student i has grown by 20 scale score units while another has grown by 50. Because scale score units are often arbitrary units designed to operationalize a latent trait, such as reading or math, gauging the “adequacy” of this growth is difficult to interpret without placing each student’s growth within the context of an expectation. In other words, knowing whether a gain of 50 is “adequate” must be considered.

However, determining whether growth is adequate is highly value-laden and must fit within an espoused system of educational and social values. In particular, all inferences regarding student achievement should be made within a standards-based context. Parallel with the idea of NCLB, but operationalized differently, we suggest that all students should (at least) reach the proficient cutscore within a specified timetable, such as by the highest grade in a school.

In this regard, students are growing towards an outcome of *value*—proficiency—rather than making normative comparisons among the varying growth rate of students. For this expectation to be realistic, we must assume that the cutscore for proficiency represents an ambitious, yet attainable, standard for all students to achieve.

Consequently, we define the Rate of Expected Academic Change, or REACH Score, for each student. In particular we ask, “given this student’s current location on the ability scale, how much does he need to grow each year in order to be proficient by the time he leaves this school?” This question is operationalized in the following manner:

$$REACH_{ij} = \frac{\lambda_{pE} - y_{ij}}{T - \alpha_i} \quad (3)$$

where λ represents the lower bound cutscore for proficiency in subject p at the highest grade in a school, g , T is the highest grade in the school (or end of a specified timeline), and α represents the current grade level of student i .

The student’s estimated true growth rate obtained from Equation (2) is then compared to the value produced in Equation (3). Thus, the extent to which a student is making “adequate” progress is judged by the following ratio:

$$REACH\ Ratio = \frac{ETGR_{ij}}{REACH_{ij}} \tag{4}$$

From this perspective, a REACH Ratio of “1” or greater indicates that, given the estimated rate of change, this student is likely to be proficient by the time he leaves the highest grade in the school. A REACH Ratio less than “1” indicates that, unless instruction is modified for this student, he is unlikely to reach the Proficient cutpoint by the time he leaves the highest grade in the school.

The left panel of Figure 2 illustrates the idea of a REACH Ratio less than “1” and the right panel illustrates the idea of a REACH Ratio greater than “1.” In the figure below, the heavy bold line represents the historic growth rate for two students from grades 1 to 3. The dashed line is an extrapolation, or “best guess,” of their most likely growth trajectory given their historical record of growth. The solid line is the REACH score computed for this student from Equation (3).

As evident from the left panel of the figure, this student is unlikely to reach the proficient cutpoint by Grade 5 if he continues to grow at the same rate. In other words, instruction needs to be modified for this student for him to reach proficiency.

Figure 2: An Illustration of REACH for Two Students

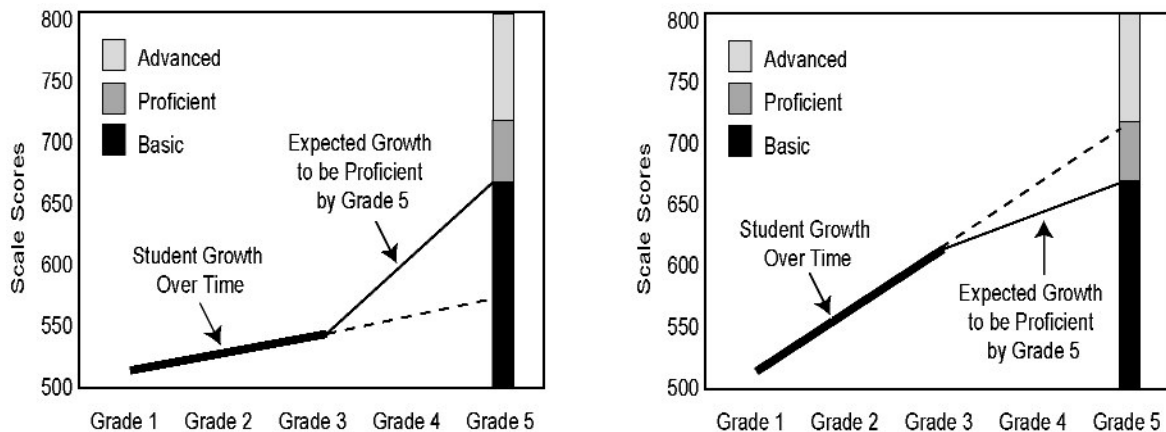


Table 1 illustrates the computations used to calculate the REACH Ratio for each student in the analysis. One may note that the yearly expected growth varies by student, but the end goal—that all students be (at least) proficient—is the same for all students. Also note that the REACH Ratio is not calculated for students who have already reached the proficient cutpoint.

Table 1: Sample REACH Computations

Student	y	ETGR	α	λ	$(\lambda-y)$	REACH	REACH Ratio
A	400	60	3	500	100	50	1.2
B	460	45	4	500	40	40	1.1
C	440	25	4	500	60	60	.42
D	510	40	5	500	Proficient	*	*

Essentially, the REACH score may serve an instructional diagnostic function as it projects which students, given their record of growth, are likely to reach the proficient cut score. In addition, by couching the estimated growth rate for each student within the context of an expected growth rate, in this case the REACH score, one can easily compute the percentage of students within a class, grade, school, or district that are on track towards meeting an outcome of value.

The ultimate benefit is that the amount of growth observed is not judged by its statistical significance. Instead, the adequacy of the observed growth is judged on its practical significance. A value-added model constructed as such not only expects students to grow, but to reach a goal. Furthermore, a simple metric, such as REACH, has more instructional relevance to classroom teachers and school administrators as they can identify which students need instructional remediation or enrichment.

It is the linkage of the REACH score to the state’s proficiency benchmark that differentiates the REACH model from other VAMs. It is often the case that changes in student achievement are reported in terms of “gains in scale score points and in the form of comparisons to local, state and national averages.” (Stone, 1999). The REACH scores, in contrast, focus not on comparisons to average scores, but on the progress toward the proficiency cutpoint. Since the state’s goal is to have all students hit proficiency, the REACH VAM may be seen as preferable. Indeed, Stone (1999) says:

An alternative to Tennessee's reporting system is one in which the annual gains produced by a given teacher, school, or system are compared to the annual learning gain necessary to bring students to an externally referenced benchmark. Although not currently in use by any state, such a report would make it possible to consider both indicators simultaneously.

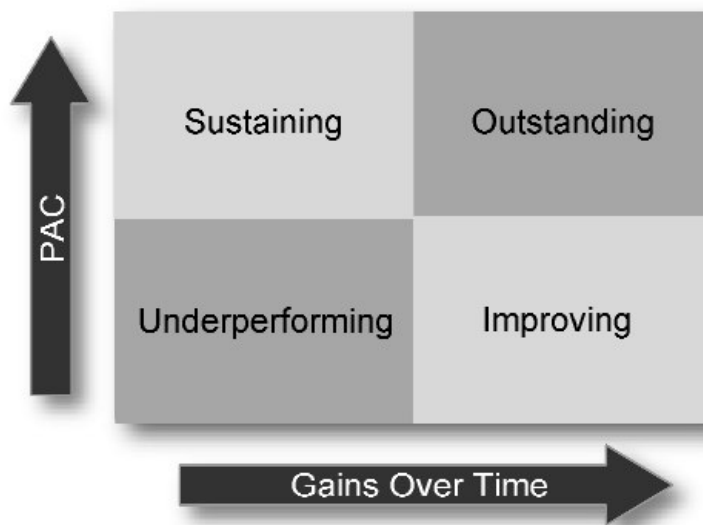
IX. THE SCHOOL-LEVEL CONTEXTUAL MODEL

The REACH method is the contextual model used to judge the adequacy of a student's rate of change. At the school level, we pose a slightly different method for evaluating performance. At this level, we calculate the percentage of students at or above the Proficient cutpoint (PAC) across all tested grade levels in the school and use the estimated growth rate of the school, which, given the parameterization of Equation (1) is

$$ETGR_j = \beta_0 + \theta_{1j} \quad (5)$$

Combining these data, we can organize all schools on a two-dimensional coordinate plane ($ETGR_j$, PAC_j). As such, we can compare the performance of all schools in the multidimensional manner detailed in Figure 3.

Figure 3: A Multidimensional View of School Effectiveness



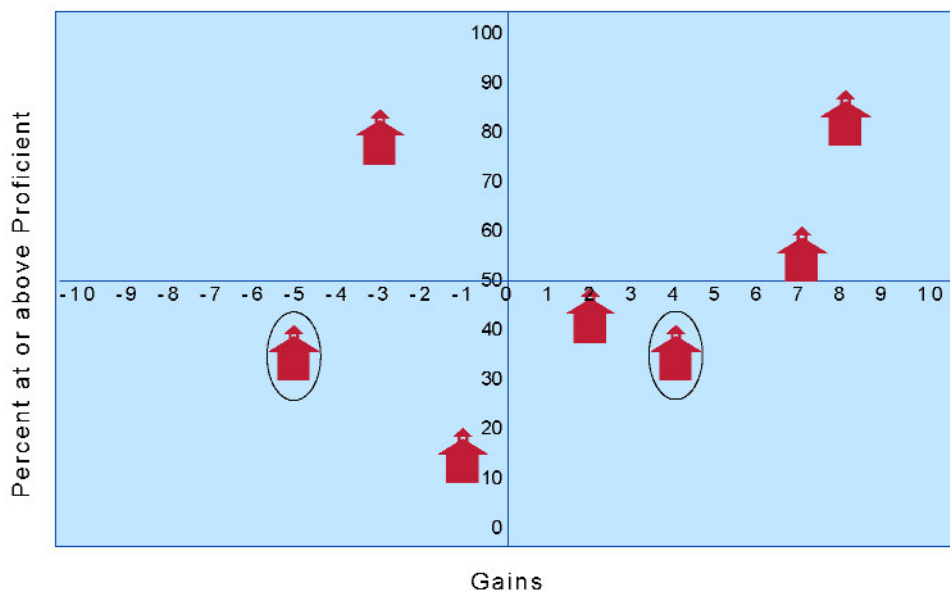
The y-axis represents the percentage of students within the school at or above proficient and the gains are presented along the x-axis as determined by Equation (2). The horizontal line within the matrix corresponds to the fixed effect for time (β_0) and the vertical line corresponds to 50 percent of students at or above PAC.

In other words, the origin represents a school with 50 percent of its students at or above PAC and has a growth rate equal to β_0 . When data are organized as such, schools in the *Underperforming* quadrant have fewer than 50 percent of their students at or above proficient and are making year-to-year gains lower than average. Schools in the *Improving* quadrant also have fewer than 50 percent of their students at or above proficient, but are making yearly gains larger than average. Schools in the *Sustaining* quadrant have more than 50 percent of their students at or above proficient, but are making yearly gains less than average. Last, schools in the *Outstanding* quadrant have more than 50 percent of their students at or above proficient and are making yearly gains greater than average.

Schools in the *Sustaining* quadrant have more than 50 percent of their students at or above proficient, but are making yearly gains less than average. Last, schools in the *Outstanding* quadrant have more than 50 percent of their students at or above proficient and are making yearly gains greater than average.

Organizing schools into the four quadrants allows policymakers to differentiate between two schools that have the same percentage of students at or above proficient, but are making different gains. Consider the example in Figure 4. The two highlighted schools have approximately 30 percent of their students at or above proficient. However, the school in the lower right quadrant is making gains much larger than the other. Viewed from a single dimension—PAC—one would not observe the differences between these two schools.

Figure 4: An Illustration of a Multidimensional School Plot



X. PREPARING FOR VALUE-ADDED ANALYSIS

While VAMs are likely to provide more meaningful information than conventional accountability strategies, most states and districts are ill prepared to implement value-added models. California does, however, have many of these requirements in place.

Specifically, VAMs require:

- an annual testing system
- longitudinal test score data organized in an electronic database
- unique student and teacher identification numbers
- a vertically equated test
- the computational capacity to implement the model

The assessment provisions of NCLB require that an annual testing system in reading/language and math be implemented in grades 3 through 8 by the year 2005. This will, to some extent, fulfill the requirement of an annual testing system.

The organization of the electronic database can take many forms. For example, relational databases, SPSS files, or simple ASCII text files are suitable.

Students' names often change or are misspelled. For this reason, a unique identification number that remains consistent over time regardless of which school the student has attended is required. Given the high probability of intra-state mobility, it is more appropriate for assignment and maintenance of the IDs to be a function of the State Education Agency rather than individual districts or charter schools.

Measuring progress requires a test metric or scoring system designed to measure growth over an extended period of time. Testing companies have for some time reported developmental score scales that are designed specifically for this purpose. When tests have been vertically equated, all forms and levels of the test have been placed on a single, continuous scale, and the same ruler is used to measure student progress over time.

The computational capacity required to analyze such large data sets is substantial. However, software programs such as the `n1me` library in the R software package are capable of fitting value-added models on common desktop computers (Doran and Lockwood, 2004; Lockwood, Doran, and McCaffery, 2003).

XI. COMMON CRITICISMS

Though VAMs present advantages over conventional strategies for measuring student progress, there are a number of criticisms that are common. For example, Ballou (2002) argues that, given their complexity and particular uncertainties, policymakers should avoid their use in accountability system design.

However, avoiding VAMs in lieu of models with significant methodological flaws, such as cross-sectional “snapshot” approaches, in light of the high-stakes consequences attached to the results is problematic. While VAMs do not represent perfect strategies for measuring school effects, they more reasonably align with the notion of student learning, do not encourage schools to target instruction for middle-performing students, and set expectations for growth rates for individual students towards an expected learning outcome. Even with their imperfections, they are more likely to provide meaningful information than conventional methods of analysis.

Second, it has been claimed that VAMs set lower expectations for learning for certain populations of students. This is untrue for at least two reasons. VAMs are a measurement method designed to measure how much students have increased in knowledge and skills as measured by a given test. The VAM itself does not set any expectations for what students are expected to know and be able to do. In no way are teachers or parents precluded from setting high expectations for learning. Second, using the proposed REACH methodology above, students who start lower are required to meet the same end goal—proficiency.

Third, some have argued that NCLB does not permit for longitudinal analyses. The legislation, as it is currently written, is primarily focused on the increase in the percentage of students at or above a cutpoint each school year, where 100 percent of the students in a school will be proficient by the year 2013–2014. A letter from the Secretary of Education, Rod Paige (2002), encourages schools to develop accountability systems that measure improvements over time.

XII. USING THE MODEL FOR ACCOUNTABILITY SYSTEM DESIGN

The REACH value-added model described in this paper can be an important tool to support test-based accountability systems in California. Not only does REACH use longitudinal student test-score data to map out how much individual students must improve to hit the state’s proficiency goal, it could also serve as an evaluative device for determining the effectiveness of education programs, teaching methodologies, school personnel, and teacher training programs.

Because the REACH VAM measures and projects how each individual student is progressing toward proficiency, it can be used to evaluate whether a student's exposure to a particular education program or reform helped or hurt that progress. For example, the REACH VAM can be used to determine whether reducing class size has helped individual students hit their proficiency growth goals.

Also, since the REACH VAM looks at student achievement over time and relates that achievement to the state's proficiency benchmark, there is a cumulative aspect to the model that is a great advantage over unsophisticated "snapshots" or current-status indicators.

By measuring student achievement gains under individual teachers who may be using similar or different teaching methodologies, the REACH VAM can inform lawmakers, education officials, teachers, and the public about which instructional practices are best able to move students toward subject-matter proficiency.

While the use of VAMs to identify effective teachers holds significant promise, the complexity of VAMs for evaluating teacher effects, or more reasonably termed "residual classroom effects," should be fully explored and understood before implementation of a high-stakes system (McCaffrey et al, 2003b). Furthermore, building a VAM that evaluates teacher effects requires that individual scores be linked to classroom teachers over time through the maintenance of unique teacher IDs, a metric unlikely to be maintained by most state departments of education.

Some value-added proponents have argued that teacher effectiveness is the single greatest predictor of school-related student learning. Consequently, it may be reasonable to construct teacher evaluation systems that include, among other indicators of instructional effectiveness, a value-added component. In this regard, the value-added indicator can serve as a weighted indicator within the broader context of the accountability system design similar to the model used in Dallas (Webster and Mendro, 1997).

There have been recent attempts in different parts of the country to craft compensation systems based on teacher effectiveness. However, most of these efforts have relied upon subjective evaluations of teachers by administrators or fellow teachers. The complaint that pay increases could end up in the pockets of the principal's "favorites" has some legitimacy.

Also, the complaint that effectiveness-based pay systems will simply reward teachers at affluent high-performing suburban schools has merit, especially if the compensation system uses only cross-sectional comparisons. A value-added-based compensation system, however, would address these concerns.

Under a value-added-based compensation system, rewards would be based, in part, not on subjective evaluations, but on gains in test scores. Further, since the REACH VAM focuses on

growth toward proficiency, it would reward those teachers or schools whose students demonstrate the required growth.

Teachers of disadvantaged students in an inner city could earn rewards if their students showed growth, while teachers of advantaged students may not earn rewards if their students do not show growth, i.e., simply coast. A compensation system based on the REACH VAM would reward teachers who are successful in helping students progress toward proficiency, which is, after all, what the state expects and what it is supposedly paying teachers to accomplish.

In addition to linking compensation to effectiveness, a value-added system could also help target professional development assistance to individual teachers. In California and other states, professional development has too often been formulated as a one-size-fits-all program that ignores the differences among teachers in terms of their strengths and weaknesses. In contrast, by showing whether students are growing toward subject-matter proficiency, value-added analysis can individualize professional development to address teacher weaknesses.

Under the REACH VAM, class “profiles” could be produced that identify skill areas where students require help in order to reach the goal of proficiency. Similar profiles are produced under the Dallas school system’s VAM.

Using the profiles, along with teacher portfolios and structured interviews, administrators could create individualized blueprints for staff development at the beginning of the school year. Teachers would then be expected to take professional development coursework geared toward strengthening their areas of weakness.

XIII. THE REACH VAM AND NCLB

Under the federal No Child Left Behind Act (NCLB), all students must reach proficiency in English/language arts and mathematics by 2013–14. While the federal and California governments both have student proficiency as their goals, the California accountability system currently focuses on growth in overall school achievement from year to year.

Thus, the state’s accountability system, as presently structured, does not provide the means to ensure that the goal of student proficiency is reached. The REACH VAM, through its measurement of student growth toward proficiency, would allow the state to have the means to meet the end goal of student proficiency.

Under the state’s current plan to meet the requirements of NCLB, targets referred to as Adequate Yearly Progress (AYP) are set for increases in percentages of students in a school

meeting the proficiency benchmark. Thus, in 2003–04 the state set a target of 13.6 percent of students proficient in English/language arts and 16 percent proficient in mathematics.

In 2004–05, the targets will increase by 10.8 percent for English/language arts and 10.5 percent for mathematics. In 2007–08, the targets will increase by another 10.8 percent for English/language arts and 10.5 percent for mathematics, with the targets increasing by similar percentages per year until 2013–14.

The problem for California, however, is its focus on school-level growth.

Under the state accountability system, schools rather than students are given growth targets and those targets require minimal annual growth. This means that even if schools hit their yearly growth targets, many students may not be proficient by the NCLB deadline. Also, the state's NCLB growth targets for the percentage of students hitting the proficient mark encourage schools to focus year-by-year on those students closest to reaching the proficiency benchmark in order to meet the target requirements.

As Drury and Doran (2003) point out:

It makes little sense to continue to define AYP solely in terms of the percentage of students crossing an arbitrary bar of “proficiency,” while ignoring the growth that occurs within broad performance categories. This is tantamount to measuring a child's height with a yardstick but acknowledging growth only when his or her height exceeds 36 inches. Relying exclusively on average test scores and proficiency cut-offs to measure AYP can lead to unintended consequences, such as leading teachers and administrators to target those students most likely to cross the proficiency cut point, while leaving others without focused instruction.

The REACH VAM provides necessary annual growth information for each individual student to meet the eventual goal of subject-matter proficiency. With this information, schools will have to ensure that all students are hitting their growth targets rather than concentrating on groups of students just under the proficiency bar. By changing the focus of schools to achievement progress among all students, the model will help prevent the lowest-performing students from being left behind due to short-term expediency.

It should be noted that the prospect of low-performing students being left behind increases pressure to redefine “proficiency” downward. For example, the Legislative Analyst's Office (LAO), the research arm of the California legislature, has suggested reducing the stringency of

the proficiency definition (LAO, 2002). By focusing on the growth of all students, the REACH VAM could lessen that pressure.

The REACH VAM provides the diagnostic information needed by schools and provides more accurate information regarding growth than conventional cross-sectional methods of analysis. Consequently, it is consistent with the intention of NCLB—to be diagnostic—and aligned with the principles of NCLB, to measure student progress towards proficiency by the end of a specific timeline.

XIV. LEGISLATIVE ACTION IN OTHER STATES

As California takes initial steps to create a value-added system, state policymakers should be aware of similar efforts in other states.

In Colorado, State Rep. Keith King, speaker of the Colorado House of Representatives, has introduced legislation that would establish a value-added system that would use students' state assessment scores over time to measure academic growth.

Under the Colorado proposal, school participation in the value-added system would be voluntary for the first year, but then be made mandatory for all school districts. The state Department of Education would be instructed to use longitudinal test score data to determine the levels of increase that constitute a full year of academic growth in reading, math, and writing for each grade level tested.

The department would also be required to provide school districts with an academic growth information report for each student. The report would include the student's test score and the growth amounts that indicate the student's level of growth in reading, math, and writing over the period between administrations of the test. The department would contract with a private or public entity to calculate annually the amount of each student's academic growth, based on test scores in all three core areas.

Each student's academic growth profile would be given to the principal at the student's school. The profile would also be shared with the student and his or her parents in discussing the student's academic strengths and weaknesses, as well as strategies to increase the student's academic growth and achievement.

The Colorado proposal, therefore, lays out a structure for a value-added system, but does not specify the exact type of value-added model to be used. It is likely that any future California legislation on value-added would similarly leave the decision regarding a model to the state Board of Education. By presenting a specific model, this paper seeks to stimulate discussion

about the most appropriate and effective methodology to maximize the impact and usefulness of a value-added system.

Although it does not specify a model, the Colorado legislation cogently sums up the advantages of a value-added system:

Establishing a system for measuring actual academic growth will increase parents' understanding of their children's actual academic progress, assist teachers in meeting each student's academic needs and raising each student's rate of academic growth, and increase each public school's and school district's level of accountability for the educational services it provides.

XV. CONCLUSION

There are many commendable aspects to California's school accountability system. However, there are serious flaws in the system, especially in the way testing data are used.

Policymakers and educational leaders must consider programs and models most likely to result in consequences consistent with the intended effect of accountability—the accurate identification of schools that meet or exceed the system's espoused values of quality.

Conventional cross-sectional or “snapshot” methods, while easy to compute and simple to explain, are inadequate to measure school effectiveness within California's high-stakes accountability system, with its rewards and sanctions. More logical alternatives, such as value-added models, should play a more significant role as the measurement component of the state's test-based accountability system.

The REACH VAM suggested in this paper has the primary benefit of focusing on the achievement growth of individual students. Second, it measures that growth not in comparison to other students but against the state's goal of subject-matter proficiency. Under the REACH model, the growth rate toward proficiency can serve as a tool for targeting remedial assistance to students.

The model can form the basis for the following important state education reforms.

- **Better evaluation of policies and programs.** Because the REACH VAM measures and projects how each individual student is progressing toward proficiency, it can be used to

evaluate whether a student's exposure to a particular education program or reform helped or hurt that progress.

- **Promotion of better instruction.** By measuring student achievement gains under individual teachers who may be using similar or different teaching methodologies, the REACH VAM can inform lawmakers, education officials, teachers, and the public about which instructional practices are best able to move students toward subject-matter proficiency.
- **Better measurement of school effectiveness.** Since the REACH VAM focuses on student achievement growth toward subject-matter proficiency, it can help identify schools that raise student achievement and ineffective schools that do not. Based on this identification, incentives can be given to effective teachers to teach in classrooms with low-performing students and compensation systems can be crafted based on teacher effectiveness.
- **Improve teacher professional development.** By showing if students are not growing toward subject-matter proficiency, the REACH VAM can individualize professional development to address school/teacher weaknesses.

The REACH VAM can also help the state meet the federal NCLB proficiency requirements. At present, the federal and state accountability systems are not in sync. Also, the way the state has structured its targets for meeting the NCLB proficiency goals almost guarantees that schools will come up short. The REACH VAM will:

- **Help ensure that all students meet NCLB achievement goals.** By focusing on achievement progress among all students, the REACH VAM will help prevent schools from concentrating only on those students just below the proficiency bar and will help ensure that the lowest-performing students are not left behind.
- **Reduce pressure to redefine “proficiency” downward.** Since the REACH VAM provides the information needed to bring all students up to the current state definition of proficiency, reducing the stringency of the definition may not be necessary if the information provided by the model is used to target instruction and assistance to individual student needs and if it is used to support effective education programs and reforms.

Value-added analysis holds great potential for improving the way in which education services are delivered to students, making state accountability systems more effective, and improving the achievement of all students. When appropriately implemented within a state accountability system, the desired consequences—improved classroom and public action—may be maximized.

As California moves forward in its efforts to improve its accountability system in order to meet its own goals for student achievement and those of the federal government, policymakers should consider the various value-added models, including the REACH VAM detailed in this paper. In combination with other valid indicators of educational quality, these models provide a more accurate portrait of school effectiveness and provide better opportunities for improvement.

ENDNOTES

¹ We use the names of the four common performance categories throughout this report: below basic, basic, proficient, and advanced.

² In late 2001, the State Board of Education announced plans for a history test in grade 8 (and elimination of the current history test in grade 9) as well as a science test in grade 5.

³ “Lowest Performing Schools,” (Palo Alto, CA: EdSource, February 2003), p. 4.

⁴ *Ibid.*

⁵ Suzanne Tachney in David Fleishhacker and Tacheny, “Arguments: Do Tests Add Up?” *San Francisco Chronicle*, September 2, 2001; Larry Crabbe, quoted in Erika Chavez, “School test results out this week,” *Sacramento Bee*, August 13, 2001.

⁶ According to EdSource, “Whenever new elements are added to the [Academic Performance Index], Base scores [from 1999] are adjusted so that they are comparable to the Growth scores from the previous cycles.” See “Lowest Performing Schools,” *op. cit.*: p. 4.

⁷ The state Department of Education uses this example.

⁸ Harcourt Brace refers to this as Level 2.

REFERENCES

- D. Ballou, "Sizing up test scores," *Education Next*, 2, (2002), pp. 10-15.
- P. Ceperley and K. Reel, "The impetus for the Tennessee value-added accountability system," in J. Millman (ed.), *Grading Teachers, Grading Schools: Is student achievement a valid education measure?*, (Thousand Oaks, CA: Corwin Press, 1997).
- H.C. Doran, "Adding value to accountability," *Educational Leadership*, 61(3), (2003), pp. 55-59.
- H.C. Doran and J.R. Lockwood, "Fitting Value-Added Models in R," submitted for publication, 2004.
- D. Drury and H.C. Doran, "The value of value-added analysis," *National School Boards Association*, 3(1), (January 2003).
- H. Goldstein, *Multilevel Statistical Models*, (London, England: Oxford University Press, 1995).
- Harcourt Brace, *Spring Norms Book*, (San Antonio, TX: Harcourt Brace, 1997).
- F. Hess, "The case for being mean," *Educational Leadership*, 61(3), (2003), pp. 22-26.
- R. Holland, "Indispensable tests: How a value-added approach to school testing could identify and bolster exceptional teaching," *Lexington Institute*, (2001), <http://www.lexingtoninstitute.org/education/schooltesting.htm>.
- R.M. Jaeger, "Certification of Student Competence," in R.L. Linn (ed.), *Educational Measurement*, (Macmillan Publishing: New York, NY), 1989.
- Legislative Analyst's Office, *Analysis of the 2002-03 Budget Bill*, (Sacramento, CA: Legislative Analyst's Office, 2003).
- R.L. Linn, "Educational assessment: Expanded expectations and challenges," *Educational Evaluation and Policy Analysis*, 15(1), (1993), pp. 1-16.
- J. Lockwood, H.C. Doran, and D.F. McCaffrey, "Using R for estimating longitudinal student achievement models," (2003), <http://cran.r-project.org/doc/Rnews/Rnews2003-3.pdf>.
- D.F. McCaffrey, J. Lockwood, D. Koretz, T. Louis, L. Hamilton, and S. Kirby, "Models for value-added modeling of teacher effects," submitted for publication, (2003).

D.F. McCaffrey, J.R. Lockwood, D.M. Koretz, and L.S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*, (RAND Education: Santa Monica, CA), 2004b.

A Nation at Risk: The Imperative for Educational Reform, A Report to the Nation and the Secretary of Education, United States Department of Education, by the National Commission on Excellence in Education, April 1983.

J. O'Day, "Complexity, accountability, and school improvement," *Harvard Educational Review*, 72, (2002), pp. 293-321.

J. Pinheiro and D. Bates, *Mixed-Effects Models in S and Splus*, (New York, NY: Springer, 2000).

S.W. Raudenbush, "Toward a coherent framework for comparing trajectories of individual change," in L. Collins and A. Sayer (eds.), *New Methods for Analysis of Change* (second ed.), (Washington, D.C.: American Psychological Association, 2001).

S.W. Raudenbush and A.S. Byrk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, (second ed.), (Newbury Park, CA: Sage, 2002).

W.L. Sanders, A. Saxton, and S. Horn, "The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment," in J. Millman (ed.), *Grading Teachers, Grading Schools: Is student achievement a valid measure?*, (Thousand Oaks, CA: Corwin Press, 1997).

S. Searle, *Linear Models*, (New York, NY: John Wiley and Sons, 1971).

J.E. Stone, "Value-added assessment: An accountability revolution," in M. Kanstoroom and C.E. Finn (eds.), *Better Teachers, Better Schools*, (Washington, D.C.: Thomas B. Fordham Foundation, 1999).

Y.M. Thum, "Measuring Progress Towards a Goal: Estimating Teacher Productivity Using a Multivariate Multilevel Model for Value-Added Analysis," *Sociological Methods & Research*, 32 (2), 2003, pp. 153-207.

U.S. Department of Education, *A Nation at Risk*, 1983.

U.S. Secretary of Education, *Dear colleague*, (2002),
<http://www.ed.gov/policy/elsec/gui/secletters/020724.html>.

J. Wang, "Opportunity to learn: the impacts and policy implications," *Educational Evaluation and Policy Analysis*, 20(3), (1998), pp. 137-156.

W.J. Webster and R.L. Mendro, "The Dallas value-added accountability system," in J. Millman (ed.), *Grading Teachers, Grading Schools: Is student achievement a valid measure?*, (Thousand Oaks, CA: Corwin Press, 1997).

ABOUT THE AUTHORS

Harold C. Doran

Harold C. Doran is the director of assessment for the Council of Chief State School Officers in Washington, D.C. In this role, he oversees the State Collaborative on Assessment and School Standards (SCASS) for the Council and all assessment-related activities.

Dr. Doran is a member of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

His interests include methods for measuring student achievement, assessment and accountability policy, consequential validity, psychometric theory, and statistical computing in the R programming environment.

Dr. Doran was formerly the director of research and evaluation at New American School, a national educational consulting firm, an elementary school principal, and a classroom teacher.

He earned his Bachelor and Masters degrees from Northern Arizona University and his doctorate from the University of Arizona.

Lance T. Izumi

Lance T. Izumi is senior fellow in California studies and director of education studies at the Pacific Research Institute. He is the author of many PRI studies, including the *California Education Report Card: Index of Leading Education Indicators* (1997, 2000, and 2003 editions), *Developing and Implementing Academic Standards* (1999), *Facing the Classroom Challenge: Teacher Quality and Teacher Training in California's Schools of Education* (2001), and *They Have Overcome: High-Poverty, High-Performing Schools in California* (2002).

Mr. Izumi is the co-editor of two books: *School Reform: The Critical Issues* (Hoover Institution Press and Pacific Research Institute, 2001) and *Teacher Quality* (Hoover Institution Press and Pacific Research Institute, 2002). He is also the co-author of "State Accountability Systems" in *School Accountability* (Hoover Institution Press, 2002).

In 2004, Governor Arnold Schwarzenegger appointed Mr. Izumi as a member of the Board of Governors of the California Community Colleges. In 2003, U.S. Secretary of Education Rod Paige appointed Mr. Izumi to the Teacher Assistance Corps, a task force of experts assigned to review state teacher quality plans as they relate to the federal No Child Left Behind Act. Mr.

Izumi is also a former member of the California Postsecondary Education Commission and the Professional Development Working Group of the California Legislature's Joint Committee to Develop a Master Plan for Education.

Mr. Izumi is a regular contributor to KQED-FM, the National Public Radio affiliate in San Francisco. His articles have been published in the *Notre Dame Journal of Law*, *Harvard Asian American Policy Review*, *National Review*, *Wall Street Journal Europe*, *Sunday Times* (of London), *Los Angeles Times*, *Investor's Business Daily*, *San Francisco Chronicle*, *California Journal*, *Orange County Register*, *Sacramento Bee*, and many other publications.

Mr. Izumi was formerly chief speechwriter and director of writing and research for California Governor George Deukmejian, and speechwriter to U.S. Attorney General Edwin Meese III in the Reagan administration.

Mr. Izumi received his master of arts in political science from the University of California at Davis and his juris doctorate from the University of Southern California School of Law. He received his bachelor of arts in economics and history from the University of California at Los Angeles.

ABOUT PRI

The Pacific Research Institute champions freedom, opportunity, and personal responsibility for all individuals by advancing free-market policy solutions. It provides practical solutions for the policy issues that impact the daily lives of all Americans. And it demonstrates why the free market is more effective than the government at providing the important results we all seek—good schools, quality health care, a clean environment, and economic growth.

Founded in 1979 and based in San Francisco, PRI is a non-profit, non-partisan organization supported by private contributions. Its activities include publications, public events, media commentary, community leadership, legislative testimony, and academic outreach.

Education Studies

PRI works to restore to all parents the basic right to choose the best educational opportunities for their children. Through research and grassroots outreach, PRI promotes parental choice in education, high academic standards, teacher quality, charter schools, and school finance reform.

Business and Economic Studies

PRI shows how the entrepreneurial spirit—the engine of economic growth and opportunity—is stifled by onerous taxes and regulations. It advances policy reforms that promote a robust economy, consumer choice, and innovation.

Health Care Studies

PRI demonstrates why a single-payer, Canadian model would be detrimental to the health care of all Americans. It proposes market-based reforms that would improve affordability, access, quality, and consumer choice.

Technology Studies

PRI advances policies to defend individual liberty, foster high-tech growth and innovation, and limit regulation.

Environmental Studies

PRI reveals the dramatic and long-term trend towards a cleaner, healthier environment. It also examines and promotes the essential ingredients for abundant resources and environmental quality property rights, markets, local action, and private initiative.

DONATION FORM

JOIN PACIFIC RESEARCH INSTITUTE IN “PUTTING IDEAS INTO ACTION.”

Pacific Research Institute promotes the principles of individual freedom and personal responsibility. The Institute believes these principles are best encouraged through policies that emphasize a free economy, private initiative, and limited government. By focusing on public policy issues such as education, the environment, law, economics, and social welfare, the Institute strives to foster a better understanding of the principles of a free society among leaders in government, academia, the media, and the business community.

“This Institute has done so much to further the idea of a law-governed liberty.”
—FORMER BRITISH PRIME MINISTER MARGARET THATCHER

“PRI is one of the more innovative and effective think tanks in the world.”
—NOBEL LAUREATE MILTON FRIEDMAN

NAME	TITLE
ORGANIZATION	PHONE
ADDRESS	FAX
CITY/STATE/ZIP	EMAIL

Your contribution:		You will receive:
COLLEGIATE SPONSOR	\$75	<ul style="list-style-type: none"> • 25% discount on books and studies • Discounted admission to select PRI events
INDIVIDUAL SPONSOR	\$500	All the above plus: <ul style="list-style-type: none"> • Complimentary copy of Steven Hayward’s book <i>The Age of Reagan</i>
POLICY LEADER	\$1,000	All the above plus: <ul style="list-style-type: none"> • Invitations to all VIP events • Gift certificate toward the purchases of any items from the PRI bookstore • 2 complimentary tickets to be used for a PRI breakfast, lunch, or reception
BENEFACTOR	\$2,500	All the above plus: <ul style="list-style-type: none"> • 4 complimentary tickets to be used for PRI breakfasts, lunches, or receptions
CORPORATE SPONSOR	\$5,000	All the above plus: <ul style="list-style-type: none"> • Direct access to PRI research scholars for timely policy analysis
CHAIRMAN’S CIRCLE	\$10,000	All the above plus: <ul style="list-style-type: none"> • Invitation to Annual Benefactors’ Summit
PRESIDENT’S CLUB	\$20,000+	All the above plus: <ul style="list-style-type: none"> • Annual personal briefing/presentation to you and/or your company from PRI Research Directors

I would like to make a donation in the amount of \$_____.

I would like to make a contribution of stocks/securities; please call me for details.

Please send me information on PRI’s current programs.

Please make check payable to:

PACIFIC RESEARCH INSTITUTE

DAVID NAKAYAMA, VICE PRESIDENT OF DEVELOPMENT
EMAIL: DNAKAYAMA@PACIFICRESEARCH.ORG

A 501(C)(3) ORGANIZATION

“In light of No Child Left Behind, states are currently looking at how best to design testing and student accountability systems that fulfill both the letter and spirit of the law. This important study lays out a blueprint that states can follow to ensure that student performance is accurately measured over time, and that parents, teachers, and administrators have the right information to help them in their critical jobs.”

—**Bill Hansen**, Former U.S. Deputy Secretary of Education

“Doran and Izumi’s model does a superior job of tracking the one outcome that parents, policymakers, and taxpayers consider indispensable—student achievement. And it does so in a way that is fair even for schools with large numbers of disadvantaged students.”

—**J. E. Stone**, Education Consumers ClearingHouse

“Although specifically designed for California, the model set forth in this report is readily applicable to other states. Just as practical data management tools during the 1970s ushered in a new era of data-based decision support for U.S. corporations, the system proposed by PRI would herald a similar new era for U.S. public schools.”

—**George A. Clowes**, Managing Editor, *School Reform News*