Improving School Accountability in California

April 2011

S. Eric Larsen, Stephen Lipscomb, and Karina Jaquet

Summary

Federal education policy will soon undergo a major revision, with significant consequences for the state's own policy and practices. This report seeks to help federal and state policymakers consider this restructuring and one of its core questions: How should schools and school districts be held accountable for the academic progress of their students?

At present, California schools and districts are held accountable based on rules set forth in the federal No Child Left Behind (NCLB) Act of 2002. NCLB requires schools and districts to show that increasing shares of their students are attaining specified levels of proficiency in English and math. These levels must be attained by NCLB-set deadlines until the final 2014 deadline, when all students in the state are to be proficient. Schools and districts can face sanctions if they fail to meet proficiency deadlines—up to and including being closed down and their students dispersed to other schools.

However, it is now generally acknowledged in the education community that few California schools will actually meet the 2014 NCLB goal of 100 percent proficiency. Theoretically, that means a majority of California schools could be facing federal sanctions.

In light of these realities and of the upcoming policy restructuring, policymakers may want to consider an alternative method for measuring student proficiency. A value-added model makes it possible to identify schools where students' test scores are higher or lower, on average, than one would expect given their prior achievement histories and other background characteristics. A value-added model measures each school's contribution to student learning rather than the share of students at each school who have attained set proficiency *levels*. (Value-added models can also be used to evaluate individual teachers, but that is not the object of this research.)

A value-added model may provide a more accurate measure of how schools are actually doing. Here's why: The current accountability system may be judging schools partly on factors that schools cannot control. Our analysis of schools that are least likely to meet NCLB's 2014 goals finds that they tend to have more economically disadvantaged students and English learners. Conversely, those that consistently meet their yearly NCLB goals have fewer such students, as well as smaller overall enrollments.

This finding suggests that the attainment of proficiency levels is not purely a measure of school quality, but also a measure of the type of students a school happens to serve, along with other salient school characteristics such as size. A value-added model, which would measure the school's contribution to student learning, would diminish the impact the composition of a particular student body has on a school's accountability rating. Using a value-added approach would therefore be a more accurate and fair means of assessing school effectiveness.

Contents

Technical appendices to this paper are available on the PPIC website:
http://www.ppic.org/content/pubs/other/411ELR_appendix.pdf

Tables

Figures

Introduction

With the passage of NCLB in 2002, school accountability became the principal policy model by which the federal government tries to improve the academic outcomes of the nation's K–12 students. This was a significant change from earlier policy efforts, which had focused on providing more money to stimulate improvements. Like many of the state accountability programs that preceded it—including California's—NCLB sets annual student proficiency levels, and schools and school districts are required to make sure that increasing shares of their students meet those levels. It also requires schools and districts to report publicly which goals have and have not been met, and sanctions those that repeatedly fail to meet these goals.

The primary tool for measuring accountability now in use in California is standardized test performance in English language arts and in math. High-school graduation rates and scores on the Academic Performance Index (API) tests are also used. Each year, NCLB requires schools and school districts to meet the four performance goals shown in Table 1. Schools and school districts that do so are designated as having made Adequate Yearly Progress (AYP). Missing one of these four targets, however, means schools and districts have failed to demonstrate AYP, and if they fail two years in a row, sanctions can follow.[1] A school or district reverts back to normal status after it makes AYP two years in a row.[2]

TABLE 1
Adequate Yearly Progress requirements

Requirement	Description
Annual Measurable Objectives (AMO), English language arts	The share of students who achieve the level of *proficient* or above must meet or exceed that year's target, and 95 percent of students must be tested. These requirements must be met both overall and within each numerically significant subgroup.[3]
Annual Measurable Objectives, math	The share of students who achieve the level of *proficient* or above must meet or exceed that year's target, and 95 percent of students must be tested. These requirements must be met both overall and within each numerically significant subgroup.
Academic Performance Index (API) score	The API score must either increase by one point or meet the annual API target.
High school graduation rate	The graduation rate must increase by 0.1 percentage points in one year, increase by 0.2 percentage points in two years, or reach the annual graduation rate target.

NOTE: The graduation rate requirement only applies to schools and districts serving high school students. Elementary and middle school grades use the California Standards Tests (CSTs) and high school grades use the grade 10 grade administration of the California High School Exit Exam (CAHSEE) to measure student proficiency.

Unmistakably, English and math proficiency rates have been improving in California. Steady gains are clear not only for all students on average, but also within demographic subgroups. Figure 1 displays average proficiency rates from the California Standards Test (CST) in English and math across California students between 2003

---

[1] Only schools and districts that receive federal funds under NCLB face sanctions.

[2] Schools or districts that fail to make AYP for two consecutive years are designated for Program Improvement (PI). PI's sanctions increase in severity each year, beginning with restrictions on the way a district can spend funds and concluding with "alternative governance" of schools and "corrective action" against districts.

[3] A subgroup is numerically significant in California if it represents at least 100 students at a school or district, or 50 students making up at least 15 percent of the school's or district's enrollment.

and 2010. Student proficiency in English improved 19 percentage points over that period. This translates into growth in excess of 50 percent above the 2002 level. In most cases math trends mirror the English trends.

FIGURE 1
Math and English proficiency have risen steadily this decade

Math

English
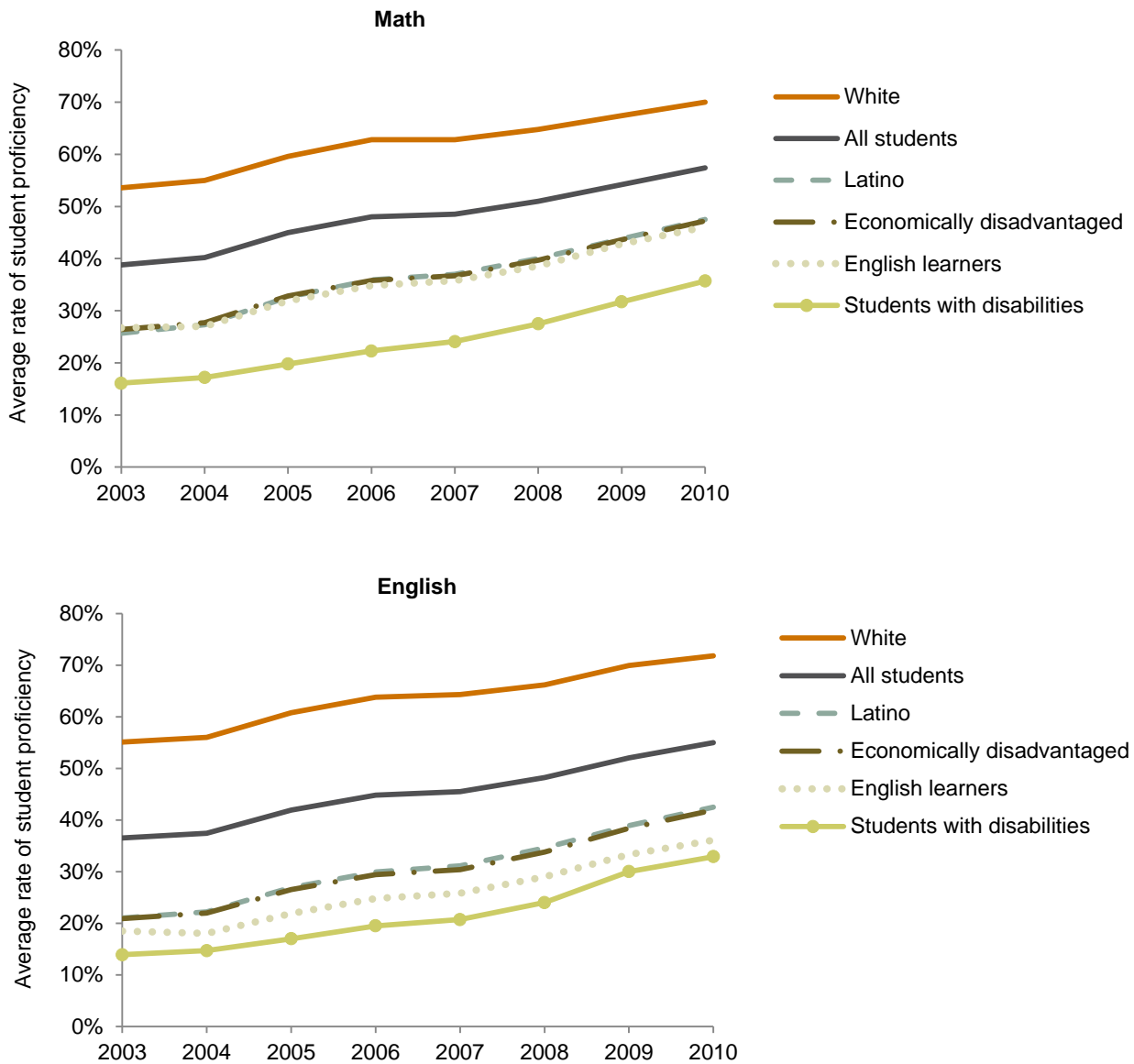
Moreover, California has rigorous standards for proficiency. The Thomas B. Fordham Foundation reviewed state content standards in 2006 and gave California's the highest marks; it was the only state to receive a grade of A in each of the five subject areas examined (Finn, Julian, and Petrilli 2006). A similar analysis by Peterson and Hess (2008) ranks California fifth. A report by the National Center for Education Statistics (Bandeira de Mello et al. 2009) does not rank California as high, but does find that California's reading standards are among the most difficult and its math standards are above the median.

Although these improvements have encompassed all subgroups, not all achieved at the same level: proficiency by white and Asian students exceeded the average, while economically disadvantaged students, English

learners, and students with disabilities had below-average rates of proficiency.[4] Nonetheless, subgroup trends do show consistent patterns of growth—proficiency improved at least 16 percentage points in each subgroup over the seven-year period.

It is not clear that these CST score improvements are a direct result of NCLB. Scores on the Stanford 9 achievement test, which is no longer in use, were rising before NCLB was instituted, and scores on the National Assessment of Educational Progress were rising before 2002 and continued to do so thereafter. The lack of any available control group—because NCLB applies to all students in the nation—is the chief obstacle to gauging NCLB's effectiveness.

Despite these improvements, it is unlikely that the majority of California schools will meet the NCLB target of 100 percent student proficiency by the 2014 deadline. Thus, policymakers could conceivably be faced with a situation in which sanctions are triggered and imposed against a majority of California schools, even though proficiency rates have increased significantly.

---

[4] Students can belong to more than one subgroup.

Accountability and NCLB

Now more than eight years old, NCLB was a reauthorization of the federal Elementary and Secondary Education Act (ESEA) of 1965—which is now due for another reauthorization by Congress. Much debate about what the new legislation might or should look like has already occurred. It is clear that the accountability model based on standardized tests will continue, as will performance targets for schools and student subgroups. A fundamental question is how to most fairly assess the effectiveness of a school.

As policymakers undertake ESEA renewal, they may wish to consider ways to refine accountability in order to encompass factors such as the rising CST scores noted above. One concern is that schools that succeed in meeting NCLB requirements do so not because they are more effective schools, but because of the students they inherit. There are clear differences between the composition of schools and districts that do meet their Annual Measurable Objectives (AMO) and those that do not. In our analysis (Table 2), the former have smaller total enrollments, higher math and English proficiency rates, and lower percentages of economically disadvantaged and English learner students. They also have slightly fewer subgroups that are subject to NCLB requirements.[5]

Conversely, schools that fail to meet their AMOs generally have larger enrollments overall and more economically disadvantaged students. This suggests that the current system evaluates schools and districts at least in part based on the student population in the geographic region they serve. That is, the current school accountability system may be judging schools in part on factors that schools cannot actually do anything about.

Moreover, our projections[6] of California school AMO attainment levels out to the 2014 deadline indicate that the discrepancies between these two groups are unlikely to diminish (Table 3). This is true whether proficiency growth continues at the same rate or slows down, as might be expected in the later stages of NCLB, when student proficiency rate requirements begin to approach 100 percent. (Comparable projections for school districts can be found in the Technical Appendices.)

---

[5] Novak and Fuller (2003) find that schools in California were more likely to make AYP if they had fewer subgroups because they are responsible for meeting fewer AMOs.

[6] Descriptions of the data and methodology we use to project proficiency rates are described in the Technical Appendices.

TABLE 2
Characteristics of schools and districts that met their AMOs in 2009

	Schools		Districts	
2009 Averages	Made AMOs	Did not make AMOs	Made AMOs	Did not make AMOs
Proficiency (%)				
English language arts	**63.6**	**43.3**	**65.9**	**49.4**
Mathematics	**68.4**	**47.2**	**65.8**	**51.3**
Subgroup populations (%)				
African American	**6.0**	**8.5**	**2.3**	**4.9**
American Indian	1.0	1.0	2.0	2.1
Asian American	**11.1**	**6.0**	6.1	5.8
Filipino	2.9	2.7	**1.5**	**2.2**
Latino	**38.3**	**60.2**	**27.0**	**49.7**
Pacific Islander	**0.7**	**0.8**	**0.4**	**0.6**
White	**38.9**	**19.9**	**59.1**	**33.5**
Economically disadvantaged	**44.7**	**70.4**	**35.7**	**59.2**
English learners	**26.8**	**42.5**	**16.8**	**34.0**
Students with disabilities	**11.3**	**11.6**	10.9	10.8
Other school and district data				
Valid subgroups	**5**	**4**	7	7
Tested students	**425**	**556**	**2103**	**5908**
Number of schools/districts	3,507	3,044	275	505

NOTE: Students are tested in grades 2–8 and 10. Values for which we reject a t-test of the null hypothesis of zero difference between the two averages at $p < 0.05$ are in bold type.

TABLE 3
Characteristics of schools by projected AMO passage in 2014

2009 Averages	Constant growth		Slowing growth	
	Projected to make AMOs	Not projected to make AMOs	Projected to make AMOs	Not projected to make AMOs
Proficiency (%)				
English language arts	**62.3**	**43.9**	**75.3**	**47.9**
Mathematics	**67.8**	**47.0**	**78.7**	**52.6**
Subgroup populations (%)				
African American	**6.0**	**8.5**	**4.6**	**7.9**
American Indian	1.0	1.0	0.9	1.0
Asian American	**10.7**	**6.4**	**16.0**	**6.6**
Filipino	2.8	2.9	2.9	2.8
Latino	**40.8**	**58.1**	**22.4**	**56.1**
Pacific Islander	**0.6**	**0.8**	**0.6**	**0.7**
White	**36.9**	**21.5**	**50.8**	**23.9**
Economically disadvantaged	**47.7**	**67.9**	**26.3**	**65.6**
English learners	**28.5**	**41.2**	**16.9**	**39.2**
Students with disabilities	**11.3**	**11.6**	**11.1**	**11.5**
Other school characteristics				
Valid subgroups	**5**	**4**	**4**	**5**
Tested students	**409**	**581**	**354**	**525**
Number of schools	3,648	2,907	1,496	5,059

NOTE: Values for which we reject a t-test of the null hypothesis of zero difference between the two averages at $p < 0.05$ are in bold type.

Accounting for Student Improvement

The alternative accountability model we suggest[7] is based on measuring student improvement and has become known in education circles as a value-added model. In its simplest form, a value-added model of school effectiveness estimates the relationship between a student's score on a year-end test and three other factors: the school attended by the student, a baseline test taken at or before the beginning of the year, and typically a set of student background characteristics. Such a model makes it possible to identify schools where students' test scores are higher, on average, than one would expect given their baseline scores, and schools where students score lower, on average, than one would expect given their baseline scores.

Little research has been done on district-level value-added models, and it is not clear if it is possible to identify the direct effect of districts on student learning. One possibility for holding districts accountable is to evaluate them based on the value-added scores of schools in the district. We do not assess here the value-added-model as a basis for evaluating districts, nor do we assess the value-added model as a basis for evaluating individual teachers. The capability of value-added models to evaluate teachers has attracted much media attention, but it is not any part of our focus here. We examine only the potential of value-added as a school accountability measurement tool, and ultimately as a way to improve schools.

Value-added models have the potential to improve student outcomes in California, but can only do so if the accountability system and the value-added model on which the accountability system is based are designed well. The validity of the inferences that can be drawn from a value-added model depends in part on the specification of the model, and the effectiveness of the accountability system depends on the specific details of its implementation. A number of states and districts, including Colorado, Louisiana, North Carolina, Ohio, Pennsylvania, Tennessee, Utah, Chicago, Dallas, and Washington, D.C., are in the process of developing or have already developed value-added models, although many of these areas plan to use value-added for program evaluation, rather than for accountability purposes. Before deciding on the specifics of California's accountability system, policymakers should study and learn from the experience of states and districts that have implemented value-added models. Value-added models have served different purposes in different areas; when studying different areas' implementation experiences, policymakers should keep in mind the goals behind these different value-added models.

The Safe Harbor Alternative

Policymakers may decide that switching to an accountability system based on value-added may be too complicated, and may search for a simpler alternative that still incorporates growth in student achievement. Over the past few years, increasing numbers of schools have been making AYP through a number of alternative methods, including one called Safe Harbor (Table 4). Safe Harbor is based on growth in proficiency rates, is far

---

[7]Education researchers have been using value-added models for at least four decades (e.g. Hanushek 1971), but research on value-added models has flourished in the past decade, and most of the research has focused on teacher-level value-added modeling. A complete review of the literature on value-added models is outside the scope of this paper. Todd and Wolpin (2003) provide a formal overview of the model and the assumptions on which different specifications of the model are based. Harris (2009), Reardon and Raudenbush (2009), and McCaffrey et al. (2003) provide comprehensive overviews of the strengths and weaknesses of value-added models. The findings from these reviews indicate that there is measurable variation in teacher effectiveness, but that analysts developing these models should take care to ensure that their results are robust to several methodological challenges. For example, Rothstein (2010) provides evidence that non-random assignment of students into classrooms biases value-added estimates for teachers, while Koedel and Betts (2011) suggest that this bias is mitigated if one includes in the estimation model students taught over several recent school years.

less complicated than a value-added model, and is already used in California. Under Safe Harbor, schools and districts can make AYP if the share of students who are proficient increases enough to meet a target, which is based on the proficiency rate at that school or district during the previous year. Although Safe Harbor is based on growth in proficiency, it is different from value-added models, which are based on the growth in achievement of individual students. Instead, Safe Harbor compares the proficiency rates of consecutive cohorts of students.

TABLE 4
Alternative methods for making Adequate Yearly Progress

Requirement	Description
Safe Harbor	Schools and districts that met their API and Graduation requirements but did not make one of their AMOs can meet AYP if the share of students in the district, school, or subgroup performing below the proficient level in either ELA or math decreased by at least 10 percent from the preceding school year.
Adjustment for students with disabilities	Schools and districts that did not make AYP solely due to their students with disabilities subgroup not making AMOs were allowed to add 20 percentage points to their percent proficient in mathematics for this subgroup during the years 2005–2007 and 20 percentage points for their percent proficient in ELA for this subgroup during the years 2005–2006.
Pass using a two-year average	Schools, districts, or subgroups that do not meet an AMO can make AYP if the share of students who were proficient in that AMO over the past two years meets the AMO target for the current year.
Pass using a three-year average	Schools, districts, or subgroups that do not meet an AMO can make AYP if the share of students who were proficient in that AMO over the past three years meets the AMO target for the current year.

Figure 2 shows the rising rates of schools' reliance on Safe Harbor in recent years in relation to the falling rates of using standard criteria to meet AMO. (The comparable figure for school districts can be found in the Technical Appendices.) Before 2005, few schools or districts relied on alternative methods to meet their AMOs.[8] Safe Harbor usage began to rise in 2005, the year in which the proficiency target climbed by 11 percentage points. Usage then diminished through 2007 as the proficiency target remained at its new (higher) plateau. When the proficiency targets increased again, so did reliance on Safe Harbor.

---

[8] The main exception was district use of a temporary adjustment for the students with disabilities subgroup. From 2005 to 2007, schools and districts that only missed their AMOs for the students with disabilities subgroup could add 20 percentage points to their proficiency rate for that subgroup. In June 2005 the United States Department of Education allowed the California Department of Education (CDE) to grant this exception temporarily while the CDE developed an alternative assessment designed to measure the achievement of students with moderate cognitive disabilities. The exception expired once the California Modified Assessment (CMA) was developed. This additional flexibility boosted passage rates for districts by over 10 percentage points in 2005 and 2006, indicating that the students with disabilities subgroup jeopardized AMO passage for a sizable number of districts. The adjustment for students with disabilities had a smaller effect on AMO passage for schools because fewer schools have numerically significant subgroups of students with disabilities. This adjustment factor applied only to math in 2007, which practically eliminated its effect on passage rates because districts fell short in the English content area. The adjustment phased out entirely in 2008.

FIGURE 2
Percentage of schools meeting Annual Measurable Objectives, 2002–2009
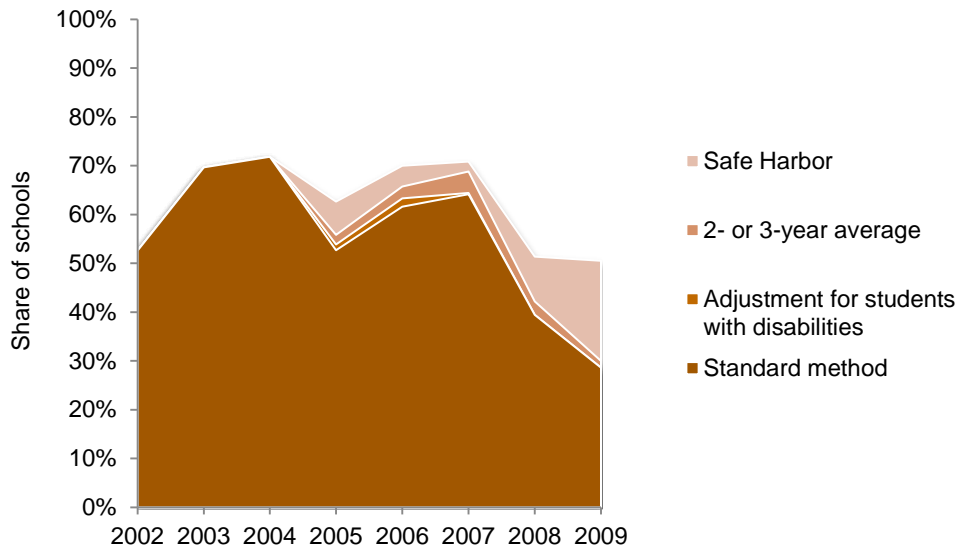
Safe Harbor says that schools and districts are responsible for reducing their rate of non-proficiency by 10 percent annually. As proficiency rates grow, the percentage point gain in proficiency needed to meet Safe Harbor requirements shrinks. For example, a school or district needs to raise proficiency by 8 percentage points if it has 20 percent of its students proficient, but only 4 percentage points once it has 60 percent proficient (Table 5).

TABLE 5
Safe Harbor requirements for high- and low- proficiency schools

	High-proficiency school	Low-proficiency school
Percent proficient	60	20
Percent not proficient	40	80
10% of share not proficient	4	8
Proficiency target	64	28

We project[9] that under NCLB, reliance on the Safe Harbor alternative is likely to increase until 2014. Our projections show that between one tenth and one third of school districts are likely to make their AMOs in 2014, and all of them will rely on Safe Harbor to do so. Similarly, 23 percent to 56 percent of schools are likely to make their AMOs in 2014 once alternative methods—and Safe Harbor in particular—are considered. That there will be greater use of Safe Harbor is not surprising because the proficiency gains needed to meet Safe Harbor requirements are now always less than those that are needed under the standard method. The share of schools and districts that rely on Safe Harbor to meet their AMOs is likely to increase every year after 2010 (Figure 3).

---

[9] Descriptions of the data and methodology we use to project proficiency rates are described in the Technical Appendices.

FIGURE 3

**Actual and projected rates of schools meeting their AMOs
(assuming constant growth)**
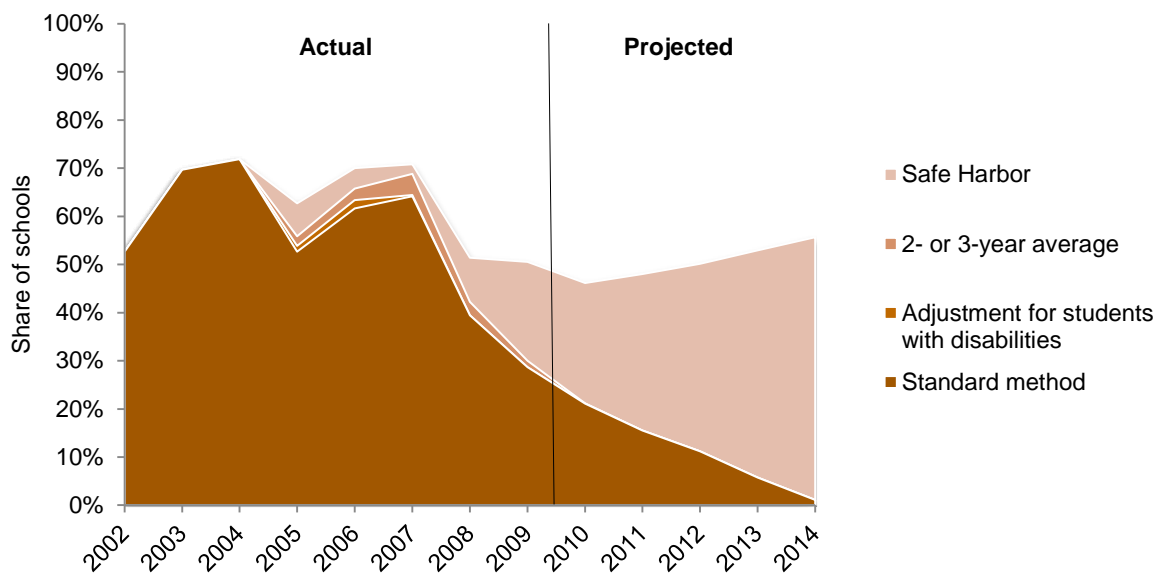
NOTE: Comparable figures for schools' slowing-growth scenario and for school district constant- and slowing- growth scenarios can be found in the Technical Appendices.

An accountability system that incorporated Safe Harbor would not solve the basic accountability problem, because it would still evaluate schools and districts based on factors outside their control. This finding derives from Safe Harbor rules that, as mentioned above, hold lower-performing schools and districts responsible for bigger gains in student proficiency.

Value-Added Models and Implementation Issues

A number of choices and decisions would face policymakers in any move to a value-added model. The experience of other states that have already developed a value-added model could certainly inform the decision-making process. Key questions that policymakers would need to consider in an accountability system based on value-added include the following:

- Should a student's baseline score be derived from a test taken at the beginning of the school year or from a test the student took in the previous year and grade?

- Other than student baseline scores, which student or other characteristics should be controlled for in a value-added model?

- How will schools and districts be held accountable for improving dropout rates?

- How should rewards and sanctions be determined?

The choice of baseline test used in the value-added model can affect the value-added model's validity. Two candidates for the baseline test are a test the student took in the previous year and grade and a test taken at the beginning of the current year. The benefit of using the student's score on test taken during the previous year as the baseline is that it requires students to be tested only once a year, reducing the amount of time diverted from

teaching and learning and toward testing. Another argument against using the student's score on a beginning-of-the-year test as the baseline is that it may present an opportunity for some schools to bend the rules, perhaps by discouraging students from doing their best on the test, or in some way creating an environment that results in artificially low baseline scores. This potential for abuse would tend to argue for the alternative of using the student's score on a test taken in the previous year as the baseline.

However, students experience considerable learning loss over summer vacation: on average, a student tested at the beginning of the summer and the end of the summer will perform worse on the end-of-summer test.[10] If students who experience greater learning loss over the summer are not distributed evenly across schools, some schools will have to work harder to help students relearn what they lost over the summer. As a result, using scores from a test taken the previous year as a baseline may be unfair to schools where students experience a relatively high degree of summer learning loss.

Students differ in ways other than their baseline test scores, and it is possible to control for these differences in a value-added model. A value-added model's validity can be improved by including controls for differences across students that are correlated with factors that influence the rate at which students learn but over which schools have little control. Policymakers will need to choose from a number of potential value-added models, including models that control for observable student characteristics such as ethnicity, English fluency, and special education status. Different value-added models may yield different school effectiveness ratings based on the factors that are controlled for. Policymakers and analysts will need to select a model specification that most effectively assesses the performance of schools, rather than simply reflecting the characteristics of the students that schools inherit.

An accountability system based on value-added may unintentionally inflate the scores of schools where large shares of students drop out, because more students whose growth potential is believed to be low would not be included in the evaluation of those schools. To reduce these unintended consequences, the new accountability system should continue to hold schools and districts accountable for dropout rates. Unfortunately, a value-added model cannot be used to hold schools accountable for dropout rates: there is no baseline test for dropping out. Statistical modeling can be used to estimate the association between the likelihood that a student drops out and a number of other factors, including the school the student attends, the student's age and grade, the student's eligibility for free- or reduced-price lunches, the student's ethnicity, and other factors. Estimates that control for student characteristics are a better measure of a school's effectiveness at reducing dropout rates than the unadjusted dropout rates on which NCLB's sanctions are based.

NCLB holds schools and districts accountable for reaching specific proficiency rate and graduation rate targets, and in theory all schools could meet those targets. In contrast, a value-added model compares a school's performance with the performance of other schools. Thus, a value-added model is best used to identify the most and least effective schools. Policymakers may choose to reward or sanction a set share of schools; for example, policymakers may choose to reward the top 10 percent of schools and sanction the bottom 10 percent, or those schools that are persistently near the top or bottom of the distribution.

Finally, although the accountability system can be based on just one year of value-added data, rewards and sanctions could instead be based on two or three years of value-added data, as a system based on two or three years of data provides a more valid and reliable estimate of school effectiveness.

---

[10] See Cooper et al. (1996) for a review of research on summer learning loss.

Conclusion

Policymakers will soon consider how to renew and restructure both state and federal education policy. Some examination of ways that NCLB could be improved would not be out of order as part of this process. It is clear that since NCLB's 2002 implementation, California student performance on math and English achievement tests has been rising. Yet it is generally acknowledged that a majority of California schools will not meet the NCLB 2014 deadline for 100 percent proficiency. As a result, a majority of California schools that have seen rising achievement test scores would face sanctions.

Schools where student learning occurs at a fast pace should be recognized and emulated. At the same time, the state should sanction and intervene at schools where student learning occurs at a slow pace. To improve state and federal accountability programs, policymakers should do the following:

- Use a value-added model to evaluate school effectiveness.
- Increase the state's efforts to ready the California Longitudinal Pupil Achievement Data System (CALPADS).
- Postpone sanctions until a value-added system is implemented.

Below, we discuss each of these recommendations.

Use Value-Added Models to Evaluate School Effectiveness

Federal policymakers should reauthorize ESEA as an accountability system that uses a value-added model to measure school effectiveness. Because value-added models control for students' initial levels of achievement, measures of school effectiveness obtained from value-added models are likely to be more accurate than California's current measures of school effectiveness, which do not control for students' initial levels of achievement. An accountability system based on value-added has the potential to identify the most and least effective schools more effectively than the current system does. By doing a better job of identifying exemplary schools and schools where changes are needed, an accountability system based on a value-added model has the potential to improve student outcomes in California.

Although an accountability system based on value-added would be an improvement over the current system, which is based on levels of proficiency and comparisons of different cohorts of students, even the best value-added model is not perfect. The validity of the model depends on how it is specified. Although we would like to have a highly accurate measure of a school's effectiveness at improving math and English proficiency, value-added can provide only one measure of effectiveness in these areas. And it should not be forgotten that schools and districts do much more than teach math and English. Any judgment of a school's success or failure should be based on unusually high or low proficiency growth over a number of years, and policymakers should include other performance measures as well.

Increase the State's Effort to Ready CALPADS

The goal of CALPADS is to merge student-level data that had previously been stored in separate data systems into a single, centralized database, and to make it possible to track an individual student across all of the years the student is enrolled in California schools. CALPADS has the potential to play an important role in an

accountability system based on value-added. Schools should be held accountable for the achievement growth for all students, regardless of whether the student attended the same district or a different district during the previous year. If the value-added model's pre-test scores are obtained from a test taken at the beginning of each school year, districts do not need to share student test-score information with one another. But if the value-added model's pre-test scores or obtained from the previous year's post-test, districts must identify the district from which each new student has transferred and request student test-score information from those districts. A centralized longitudinal student-level data system such as CALPADS can decrease districts' costs associated with sharing data and may be essential to estimating value-added measures of school effectiveness in districts where transfer rates are high.

There are other significant benefits of a longitudinal data system such as CALPADS. A statewide longitudinal student-level data system can be essential to holding schools and districts accountable for dropout rates, because it can make it much easier to determine whether a student who has left a district has re-enrolled at another. A statewide longitudinal data system will increase the likelihood that value-added will be used for purposes other than estimating school effectiveness. Value-added is an important tool for evaluating school policies and interventions, which can direct educators and policies toward successful programs and away from those found to be ineffective.

The development CALPADS is behind schedule. Given the potential of a longitudinal data system to improve school accountability and program evaluation, it is important that the state not only continue to focus on, but also increase, its efforts to make ready a statewide longitudinal student-level data system.

Postpone Sanctions

This report presents evidence that under NCLB, schools and districts are evaluated in part on the basis of the students they inherit, rather than on their effectiveness. But schools with low levels of achievement are not necessarily schools with ineffective teachers and administrators. In fact, the rate of student learning at some of these schools may be relatively high, but because students enter school with very low ability levels, the success of these teachers and administrators goes unnoticed. Or the teachers and administrators at these schools may simply be ineffective. Until California evaluates schools on the basis of individual student achievement gains, it will not be possible to distinguish between schools where teachers and administrators are effective and where teachers and administrators are not. When ESEA is renewed, the sanctions it requires should be based on the gains in achievement made by individual students at a school, not by levels of student achievement at a school.

Some may argue that schools with low levels of student achievement should be sanctioned regardless, because "anything is better" than the current school. This argument overlooks the possibility that some—perhaps many—of the schools with the lowest levels of achievement may prove themselves to be very effective schools once a valid measure of school effectiveness is put in place.

At the same time, most research that has focused on the effects of sanctions has found no significant positive effects. Most studies of charter schools (e.g. Buddin and Zimmer 2005, Edwards et al. 2008) find that they are no more effective on average than non-charters; closing schools with low levels of student achievement does not typically lead to improved outcomes for the students who transfer from the closed schools (de la Torre and Gwnne 2009); the effects of additional teacher training are limited (Jacob and Lefgren 2004); extending the school year appears to have few positive effects (Card and Krueger 1992, Pischke 2007), nor does allowing an outside entity, including the state, to take over the school (Gill et al. 2007). At some schools, the effect of

sanctions may be to disrupt an effective school in order to implement expensive changes for which there is little track record of success.

The ultimate goal of accountability programs is to improve student achievement. Certainly one positive outcome of accountability programs thus far is that they have forced states and schools to adopt content standards and measure student performance relative to those standards. However, accountability programs have not used the right metrics to identify successful and failing schools. Moving to a value-added model is a step in the right direction and would be a more accurate way to measure school performance. Ultimately, however, we must find better strategies for intervening in schools where the rate of student learning is too slow.

References

Bandeira de Mello, Victor, Charles Blankenship, and Don McLaughlin. 2009. *Mapping State Proficiency Standards onto NAEP Scales: 2005–2007*. Washington DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Buddin, Richard, and Ron Zimmer. 2005. "Student Achievement in Charter Schools: A Complex Picture." *Journal of Policy Analysis and Management* 24 (2): 351–71.

Edwards, Brian, Eric Crane, Heather Barondess, and Mary Perry. 2008. *California's Charter Schools: 2008 Performance Update*. Mountain View, CA: EdSource.

Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100 (1): 1–40.

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research* 66 (3): 227–68.

de la Torre, Marisa, and Julia Gwynne. 2006. *When Schools Close: Effects on Displaced Students in Chicago Public Schools*, Chicago: Consortium on Chicago School Research.

Finn, Chester E., Jr., Liam Julian, and Michael J. Petrilli. 2006. *The State of State Standards: 2006*. Washington DC: Thomas B. Fordham Institute.

Gill, Brian, Ron Zimmer, Jolley Christman, and Suzanne Blanc. 2007. *State Takeover, School Restructuring, Private Management, and Student Achievement in Philadelphia.* Santa Monica, CA: RAND Corporation.

Hanushek, Eric. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review* 61 (2): 280–88.

Harris, Douglas N. 2009. "Would Accountability Based on Teacher Value-Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives." *Education Finance and Policy* 4 (4): 319–50.

Jacob, Brian A., and Lars Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-experimental Evidence from School Reform Efforts in Chicago." *Journal of Human Resources* 39 (1): 50–79.

Koedel, Cory, and Julian R. Betts. 2011. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy* 6 (1): 18–42.

Legislative Analyst's Office. 2009. "Improving Academic Success for Economically Disadvantaged Students." Available at www.lao.ca.gov/2009/edu/academic_success/academic_success_0109.aspx.

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability*, Santa Monica, CA: RAND Coporation.

Novak, John R., and Bruce Fuller. 2003. "Penalizing Diverse Schools? Similar Test Scores, But Different Students, Bring Federal Sanctions." PACE Policy Brief 03-4.

Peterson, Paul E., and Frederick M. Hess. 2008. "Few States Set World-Class Standards." *Education Next* 8 (3): 70–73.

Pischke, Jörn-Steffen. 2007. "The Impact of the School Year on Student Performance and Earnings: Evidence from the German Short School Years." *Economic Journal* 117 (523): 1216–42.

Reardon, Sean F., and Stephen W. Raudenbush. 2009. "Assumptions of Value-Added Models for Estimating School Effects." *Education Finance and Policy* 4 (4): 492–519.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175–214.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113 (485): 3–33.

About the Authors

Eric Larsen is a research fellow at PPIC, where he focuses on the economics of public education, particularly school accountability programs, the relationship between school resources and student achievement, and labor markets for teachers. He holds a M.Ed. from the University of California, Los Angeles, and a Ph.D. in economics from the University of California, Davis. He was an English teacher in California public middle schools for eight years.

Stephen Lipscomb is a researcher at Mathematica Policy Research and an adjunct fellow at PPIC. His current work focuses on special education and measures of teacher and school effectiveness. Before joining Mathematica in 2009, he was a research fellow at PPIC. He holds a Ph.D. in economics from the University of California, Santa Barbara.

Karina Jaquet is a research associate at PPIC, where her work focuses on K–12 education policy, including early grade retention, special education, special education finance, and accountability issues under the No Child Left Behind Act. Before joining PPIC, Karina worked as a seasonal proxy research analyst at Glass, Lewis & Co., as a field researcher in Guatemala with a BASIS grant from the U.S. Agency for International Development, and as an outreach coordinator for FIRST 5, Santa Clara County. She holds a B.A. in economics from the University of California, Davis, and a M.A. in international and development economics from the University of San Francisco.

Acknowledgments