Large Mandates and Limited Resources:
State Response to the *No Child Left Behind Act* and Implications for Accountability

By

Jimmy Kim and Gail L. Sunderman

February 2004

This publication should be cited as:

Kim, J., & Sunderman, G. L. (2004). *Large Mandates and Limited Resources: State Response to the No Child Left Behind Act and Implications for Accountability*. Cambridge, MA: The Civil Rights Project at Harvard University.

Additional copies of this report may be obtained from our website at:
http://www.civilrightsproject.harvard.edu

Produced with generous support from The National Education Association.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The authors of this report are grateful to numerous individuals for their assistance during the course of the study. Christopher Edley Jr. and Gary Orfield provided leadership, counsel, and feedback on successive drafts of this report. We also thank many of our colleagues at the Civil Rights Project at Harvard University for their invaluable assistance in the production of the final report. Special thanks go to Marilyn Byrne, Laurent Heller, Cathy Horn, Al Kauffman, Michal Kurlaender, Lori Kelley, Chungmei Lee, Dan Losen, Patricia Marin, Jerry Monde, and Christina Safiya Tobias-Nahi. Several members of our NCLB Academic Advisory Board provided enormously useful comments, including Jaek-Young Lee at SUNY-Buffalo and Lori Shepard at the University of Colorado-Boulder. Many graduate students and volunteers provided able research assistance, including Tim Bazzle, Maya Harris, Mei Mei Peng, and Chris Tracey. We are especially grateful to state education officials and state board of education members in Arizona, California, Illinois, New York, Virginia, and Georgia for answering questions about their accountability systems. A previous version of this paper was presented at the 25th Annual Research Conference of the Association for Public Policy Analysis and Management in Washington, D.C., November 8, 2003. The support of the National Education Association (NEA), the Carnegie Corporation of New York and the Charles Stewart Foundation are gratefully acknowledged. However, the views and opinions expressed in this report are solely those of the authors.

EXECUTIVE SUMMARY

Title I of the *No Child Left Behind Act* of 2001 (NCLB) represents the largest source of federal funding for K-12 schools and commits our nation's public schools to the unfinished work of narrowing achievement disparities among different groups of students. Central to the federal legislation are the accountability provisions that require all students become proficient on reading and math assessments within 12 years. NCLB's primary mechanisms for bringing all students to proficiency are test-based accountability policies that include sanctions and subgroup rules. These policies require all students, including minority and low-income students, students with disabilities, and students with limited English proficiency, to meet the same achievement goals.

This report examines how state policymakers designed their accountability systems to meet the NCLB Title I requirements and the implications of its provisions for schools with large numbers of low-income and minority students. We conducted our study in six states—Arizona, California, Illinois, New York, Virginia, and Georgia—which are geographically, politically, and demographically diverse. First, we examine how these six states designed their accountability systems to meet the Title I accountability requirements, including the interaction of the federal requirements and state accountability systems. Second, we examine the effect of the Title I adequate yearly progress (AYP) requirements on high-poverty and high-minority schools in these six states.[1] Third, we explore the impact of subgroup accountability rules in California's public schools. We focus our subgroup analysis on California, since it is the state with the most ethnically and socially diverse public schools. Our analyses reveal three broad findings:

- **The federal accountability rules complicated state efforts to build a coherent accountability system.** Even though NCLB requires policymakers to create a "single statewide accountability system" for all public schools, **s**tates layered the federal accountability requirements on top of pre-existing plans. They retained their own state developed accountability systems in addition to adopting plans that complied with NCLB. This created mixed messages about school performance. Moreover, the federal rules imposed unrealistic expectations for annual test score gains and held schools in different states to very different expectations. The federal NCLB requirements had the following consequences:

    - The dual accountability systems meant that schools received conflicting signals about their performance. For example, in Arizona, 289 schools were identified as needing improvement under NCLB, but these same schools met the state's performance targets and earned either a "performing" or "highly performing" label. In Virginia, 723 (40% of all schools) failed to make federal AYP goals while only 402 (22%) failed to meet state accreditation standards.

[1] To make adequate yearly progress (AYP), all public schools must meet both achievement targets and participation requirements. Subgroups of students must meet statewide proficiency targets in reading and math, which are known as annual measurable objectives (AMO), and 95% of all subgroups must take the reading and math test. For these rules to apply, the number of students in a given subgroup must meet or exceed the state's minimum group size criterion.

- The federal law created an uneven playing field among states toward the requirement that individual states have all students proficient within 12 years. Although this common goal applies to all states, it ignores the fact that states started at different points in their 12-year climb toward the 100% proficiency target. As a result, some states will have to make small annual improvements in student achievement while others will have to make extremely large gains over 12 years. This creates incentives for states to lower their proficiency standards and to promise the largest achievement gains toward the end of the 12-year timeline.
- The term "proficiency" is used in NLCB as if it describes a stable and meaningful concept, but it has no consistent definition across states. Instead, the percentage of students meeting the proficiency level varies depending on the state where the school is located. In Georgia, schools needing improvement had on average 75% of their students meeting proficiency in reading. This is substantially higher than in California, where schools meeting AYP had on average only 34% of their students meeting proficiency in reading. This means that low-performing schools in Georgia had higher average proficiency rates than high-performing schools in California.

- **The federal law identified schools as needing improvement on the basis of their demographic characteristics rather than their contribution to student learning.** This was the case in the first year of implementing NCLB and reflects the overlap between student background characteristics and test score levels. That is, low-income and minority students and students with limited English proficiency tend to score lower on standardized achievement tests than their white, middle class peers.

  - In all six states, schools identified as needing improvement enrolled a disproportionately large number of minority and low-income students and students with limited English proficiency. The demographic differences were especially large in Illinois and New York, which are two of the most segregated states for Black and Latino students. In these two states, schools needing improvement enrolled over twice as many minority and low-income students, on average, than schools meeting AYP.
  - NCLB relies on average test scores, which usually reflect differences in student background characteristics more than differences in school quality. When other measures of achievement are used that show trends in achievement over time, average improvements in reading and math proficiency scores were similar in both schools needing improvement and schools meeting AYP. This was the case in Arizona, Illinois, New York, Virginia, and Georgia.

- **Subgroup accountability rules put disadvantaged schools segregated by race and poverty and multiracial schools at a higher risk of failing AYP than white and middle-class schools.** This is because subgroups are not mutually exclusive, that is, a student can be counted as a member of several subgroups—a low income, Asian student with limited English proficiency counts for three subgroups. Title I schools can be identified as needing improvement if a single subgroup does not meet a reading or math cut score or if any subgroup does not meet the 95% participation requirement. Since

schools serving disadvantaged students and integrated schools are more likely to contain several subgroups, they are also more likely to be identified for improvement. In California, our analyses revealed:

- Schools needing improvement were more likely to contain a Black, Latino, socio-economically disadvantaged, and limited English proficient subgroup than schools meeting AYP.
- Schools needing improvement were also held accountable for more achievement targets than schools meeting AYP. In California, 93% of schools needing improvement were held accountable for three or more targets compared to 69% of schools meeting AYP. Conversely, only 7% of schools needing improvement had to meet two or less targets as compared to 31% of schools meeting AYP.

This report shows that the Title I accountability policies have unintended consequences and ambiguous benefits for minority and low-income students and the schools they attend. In light of these findings, the federal government should reject a one-size-fits-all approach to test-based accountability. Instead:

- The federal government should encourage states to experiment with different approaches to accountability, including the use of multiple criteria for evaluating school performance and longitudinal analyses of student learning.

- It should revise the subgroup rules so that a student's scores count for only one subgroup while retaining the requirements that schools disaggregate test scores.

- It should require further evidence about how testing accommodations and modifications influence scores for students with limited English proficiency and students with disabilities before scores for these two subgroups are used for school accountability.

- In undertaking future revisions of Title I, Congress should require evaluations that assess both the intended and unintended consequences of accountability policies on minority and low-income students and the schools they attend.

To bridge the achievement gap, state and federal policymakers must work together to design and build more effective accountability policies that support the work of school leaders and classroom teachers.

(1) INTRODUCTION

The Promise of The No Child Left Behind Act of 2001 (NCLB)

Title I of the *No Child Left Behind Act* of 2001 (NCLB) represents the largest source of federal funding for K-12 schools and commits our nation's public schools to the unfinished work of narrowing achievement disparities among different groups of students. Broadly stated, the purpose of Title I is to secure equal educational opportunities and equal achievement outcomes for historically underserved and low-performing students, including racial and ethnic minorities, low-income and disabled students, and children with limited English proficiency. A number of recent developments underscore the educational imperative of improving achievement outcomes for minority students in particular, including the widening of racial test score disparities in the 1990s (Grissmer, Flanagan, & Williamson, 1998; Lee, 2002), and the hope among our nation's leaders that race-sensitive admissions in selective colleges and universities will no longer be needed in 25 years (Grutter v. Bollinger, 123 S. Ct. 2325 (2003)). By making educational equity a central goal of Title I policy, NCLB promises to hold all public schools responsible for eliminating achievement disparities between White and minority students.

Central to the federal legislation are the accountability provisions that require all students become proficient on reading and math assessments within 12 years. To meet this goal, state education agencies must develop reading and math assessments in grades 3 to 8, set short- and long-term achievement goals, hold schools accountable for the achievement of all students, and provide technical assistance to low-performing schools. These requirements place enormous demands on state education agencies, which are being required to assume major responsibilities with fewer staff and smaller budgets. NCLB is based on the assumption that these policies will create performance pressures for schools to improve student proficiency in reading and math and narrow achievement disparities based on student background characteristics. Furthermore, NCLB assumes that states and districts will mobilize resources to help struggling schools and that a series of sanctions, ranging from school choice to restructuring, will motivate low-performing schools to improve student achievement scores.

NCLB's two mechanisms for bringing all students to the proficiency level in reading and math are test-based accountability policies and subgroup rules. To make adequate yearly progress (AYP), schools must improve the proficiency rates of different subgroups of students. Since NCLB establishes a single performance standard for all students, schools with lower-scoring subgroups will face especially strong pressures to raise test scores in order to avoid federal sanctions. Since schools must meet both achievement targets and participation requirements for different subgroups of students, schools serving disadvantaged minority students will have to meet multiple targets to make AYP.[2] As a result, federal sanctions may fall disproportionately

[2] To make adequate yearly progress, all public schools must meet both achievement targets and participation requirements. Subgroups of students must meet statewide proficiency targets in reading and math, which are known as annual measurable objectives (AMO), and 95% of all subgroups must take the reading and math test. For these rules to apply, each subgroup must meet or exceed the state's minimum group size criterion. For details on minimum group size criterion in all 50 states, see Table 1 in Erpenbach, Forte-Fast, and Potts (2003). Adequate yearly progress rules also include a "safe harbor" provision, which allows a subgroup to make AYP if the number of students falling below proficiency is reduced by 10%. Finally, all states are required to include an additional indicator in determining whether schools made adequate yearly progress.

on schools with disadvantaged minority students, whose scores often count for multiple subgroups.

With accountability elevated to a central position in educational reform by NCLB, it is important to consider the functions that a good accountability system should provide. Ideally, a high quality accountability system would provide reliable and valid information about school and student performance. This information should be related to what students learn, be able to show the school's contribution to that learning, and be useful to teachers and principals. This includes providing instructionally useful information so that teachers can revise their instructional strategies to promote greater student learning. Moreover, if accountability systems are to assist schools in narrowing the achievement gap among different groups of students, they must provide information about the performance of these students. Finally, an accountability system should be coherent and flexible, that is, provide one system that is easy to understand and, at the same time, is responsive to the local context.

Keeping in mind these functions, this paper examines how states designed their accountability systems to meet the NCLB requirements and the implications of the accountability provisions on schools that enroll large numbers of low-income and minority students. In the remainder of this section, we summarize recent research on test-based accountability, the cornerstone of NCLB accountability. This review includes the principles it is based on, the unintended consequences of test-based accountability for minority students, and the implications of subgroup accountability rules for schools that enroll large numbers of minority students. This research supports two general findings: first, test-based accountability has inconsistent benefits and several unintended consequences for minority students, and, second, subgroup accountability rules may sanction a disproportionately large number of predominantly Black and Latino schools.

NCLB and Test-Based Accountability

What are the principles and policies underlying test-based accountability?

NCLB codifies several policies concerning test-based accountability into federal law. According to Hamilton and Koretz (2002), test-based accountability systems incorporate a "set of policies and procedures that provide rewards and/or sanctions as a consequence of scores on large-scale achievement tests" (p. 3). Four components form the backbone of most systems—goals, measures, targets, and incentives. Goals are usually articulated by performance standards, which describe how well students have mastered state curriculum (Hambleton, 1998). The use of performance standards has prompted state policymakers to replace norm-referenced test results with standards-based measures that report student achievement in terms of different levels of performance (e.g., basic, proficient, and advanced). By moving away from norm-referenced data and relying more on achievement in terms of performance levels, policymakers assume that educators will understand how well students have mastered learning standards and will focus on helping all students become proficient in reading and math. Test-based accountability systems also rely on targets, such as adequate yearly progress, to measure progress toward a long-term goal, such as NCLB's 100% proficiency target. It should be emphasized that NCLB's approach to setting targets is based on "status measures," which compare each school's performance to a

fixed state standard. This means that initially low-performing schools will have to make larger improvements than higher-performing schools since all schools are required to meet the same goals (Linn, 2003b). Incentives, which usually involve a combination of rewards and sanctions, are intended to motivate students, teachers, schools, and districts to meet the achievement targets. Under NCLB, Title I schools failing to make AYP for two or more consecutive years are subject to a series of federally mandated sanctions, ranging from school choice to school restructuring and reconstitution.[3]

What are the intended and unintended consequences of test-based accountability for minority students?

Although some researchers find that test-based accountability is positively related to aggregate measures of student learning on both state tests and the National Assessment of Educational Progress (NAEP) (Carnoy & Loeb, 2002; Raymond & Hanushek, 2003), there is inconsistent evidence that these relationships exist for different racial subgroups and across different regions within the same state (Grissmer & Flanagan, 2001).[4] For example, NAEP trends (1992-1996) suggest that White, Black, and Latino fourth-graders made statistically significant math gains only in Texas (Reese, Miller, Mazzeo, & Dossey, 1997), and significant reading gains only in Connecticut (Donahue, Voelkl, Campbell, & Mazzeo, 1999). Since Texas adopted a high-stakes test while Connecticut did not, it is difficult to attribute these achievement gains to test-based accountability policies. There is also ample evidence that achievement gains in both states were tied to comprehensive policies intended to upgrade curriculum, instruction, and assessments, as well as to remedies to narrow funding inequities among low- and high-income districts. Texas, for instance, enacted funding equalization legislation that significantly reduced funding disparities between high- and low-wealth districts, increased per-pupil expenditures in poor schools, and improved educational opportunities and outcomes for minority students (Treisman & Fuller, 2001). Similarly, reading gains among White and minority students in Connecticut have been linked to a number of policies, including efforts to increase funding and support for the state's poorest districts, the use of diagnostic reading tests to guide and improve classroom instruction, and the development of a strong infrastructure for recruiting, retaining, and training teachers (Baron, 1999; Wilson, Darling-Hammond, & Berry, 2001).

[3] Title I schools failing to make adequate yearly progress for two consecutive years are identified for their first year of school improvement and must offer transfer options to all students, develop a two-year improvement plan, and receive technical assistance from the state. If a school fails adequate yearly progress for three consecutive years, the district must also provide supplemental educational services to students in improvement schools. If a school fails adequate yearly progress for four consecutive years, the district must implement corrective actions to improve the school, such as the replacement of staff members or the implementation of a new curriculum. Finally, if a school fails adequate yearly progress for five consecutive years, it may be taken over by the state, a private management contractor, converted to a charter school, or reconstituted with a new staff. For a complete list of possible interventions resulting from corrective action and restructuring, see P. L. 107-110, § 1116 (b)(7)(8).

[4] There is a large body of evidence that examines the unintended consequences of high-stakes testing for minority students. In particular, while we do not focus on the impact of test-based accountability on high school graduation rates, a number of recent studies have explored the link between graduation tests and high school completion rates for minority students (Carnoy & Loeb, 2002; Carnoy, Loeb, & Smith, 2001; Dee, 2003; Haney, 2000; Orfield & Kornhaber, 2001). More recently, the Urban Institute (Swanson, 2003) has undertaken work that focuses on the implementation of NCLB's high school graduation accountability requirements.

In addition to the intentional aims of test-based accountability, studies have also documented several unintended and negative consequences of test-based accountability in predominantly minority schools and districts. In one study of the Texas accountability system, a team of researchers at RAND (Klein, Hamilton, McCaffrey, & Stecher, 2000) explored the validity of inferences about minority student achievement using test scores from the Texas Assessment of Academic Skills (TAAS). The results indicated that four-year gains on TAAS were much larger than four-year gains on the National Assessment of Educational Progress (NAEP). The RAND researchers explained the divergence in score trends by speculating that "schools are devoting a great deal of class time to highly specific TAAS preparation [and that] schools with relatively large percentages of minority and poor students may be doing this more than other schools" (pp. 13-14). Similarly, a study of high-stakes testing in a high-poverty district with large Black and Latino enrollments found that performance on high-stakes tests did not generalize to other assessments that covered similar content (D. M. Koretz, Linn, Dunbar, & Shepard, 1991). As a result, the researchers concluded that teachers were "focusing on content that is specific to the particular test used for accountability, rather than trying to improve achievement in the broader sense that we would all desire" (p. 21). More recently, a national survey found that teachers in high-stakes testing situations reported feeling more pressure to align their instruction with the test, to engage in intensive test preparation activities, and to devote less time to untested subjects (Pedulla et al., 2003). Such practices may produce large score gains on high-stakes tests that do not reflect meaningful improvements in student achievement. Other research on high-stakes testing also indicates that high-minority and high-poverty schools are often subjected to the strongest performance pressures (Madaus & Clarke, 2001; Reardon, 1996). If these schools focus substantially, if not exclusively, on tested content, disadvantaged minority students may be further restricted from a rigorous academic curriculum, and rapid test score gains may not reflect broad learning gains. Under NCLB, the problem of inflated score gains may worsen as schools with low-scoring students focus on making AYP.

How do subgroup accountability rules affect minority students and their schools?

To fulfill the promise of leaving no child behind, the federal legislation requires all schools to improve the performance of historically under-performing and under-served subgroups—in particular, racial and ethnic minorities. In principle, NCLB's subgroup rules underscore the need for schools to focus explicitly on race in monitoring and eliminating achievement disparities between White and minority students. In practice, however, subgroup rules may adversely affect minority students for two reasons. First, given the large test score disparities between White students, on one hand, and Black and Latino students, on the other, schools with a minority subgroup will have to make large achievement gains to make AYP. Under NCLB, all schools are required to meet an absolute achievement standard, which will be more challenging and difficult to reach for Title I schools serving low-scoring students than other schools in the state. Moreover, the federal law does not require or assure that districts have equal access to key educational resources, which are associated with test score gains and are less available in high poverty schools.

Second, the subgroup rules are likely to put predominantly minority schools and racially integrated schools at a statistical disadvantage. Schools with large minority enrollments will have to meet more achievement targets than predominantly White schools. Given the strong

correlation between minority status and poverty status and language ability (Miller, 1995; Puma et al., 1997), Black and Latino students are more likely than White students to be counted in multiple subgroup categories, including race, ethnicity, economic disadvantage, and limited English proficiency. Moreover, racially integrated schools may have a difficult time meeting the AYP goals since they will be required to meet more achievement targets than racially homogenous schools. As Kane and Staiger (2002b) point out, requiring schools to meet several subgroup targets "is analogous to correctly calling three or four coin tosses in a row, instead of a single toss" (p. 258), thereby increasing the chances of failing to make AYP. Furthermore, subgroup scores based on small samples of students are likely to yield more unreliable estimates than schools means, which are based on a larger sample of students (Kane & Staiger, 2002b; Linn & Haug, 2002). The imprecise nature of average scores based on a limited number of students suggests that some schools will be incorrectly identified as failing AYP while others will be incorrectly classified as making AYP.[5]

[5] As Kain and Staiger (2002b) point out, both sampling error and one-time factors contribute to the imprecision in test score measures. They estimate that between 50 and 80 percent of the variation in annual changes in test score measures is the result of one of these two factors.

Key Questions and Study Summary

This research review raises three questions about NCLB's accountability provisions:

1. How did the federal accountability requirements affect the design of state accountability systems and the public schools?
2. How did the transitional adequate yearly progress definitions affect public schools in six states with racially and ethnically diverse enrollments?
3. How did the subgroup accountability rules affect California public schools?

Our analysis of state accountability systems reveals the following findings:

- States added the NCLB requirements on top of exiting accountability policies, resulting in a fragmented approach to accountability within states. Contributing to this fragmented approach, states often retained two methods for measuring student achievement, one for NCLB and another for the state. These different systems produced different performance labels for schools, creating confusion among educators, policymakers, and the public.

- The development of dual accountability systems by states sent mixed messages about school performance, imposed inconsistent expectations for annual test score gains, and yielded divergent proficiency definitions in reading and math.

Our analysis of transitional adequate yearly progress rules highlights the following results:

- NCLB identified schools as needing improvement largely on the basis of a school's demographic characteristics rather than a school's contribution to student learning. That is, schools that enrolled minority, low-income, and limited English proficiency students were more likely to be identified as needing improvement than schools without these subgroups.

- The federal adequate yearly progress rules do not consider how students performed before they came to a school, how long they were there, or what difference the school makes in improving student achievement.

Our analysis of subgroup accountability rules in California suggest the following results:

- Schools in California were identified as needing improvement based on the scores of two subgroups: students with limited English proficiency and students with disabilities.

- Schools with multiple subgroups have a more difficult time making adequate yearly progress because they have to hit multiple performance targets and participation requirements. Schools identified as needing improvement were more likely than schools making AYP to contain three or more subgroups and to contain a Black, Latino, socio-economically disadvantaged, and limited English proficient subgroup. As a result, there were more opportunities for schools with multiple subgroups to miss a single target due to random fluctuations in the scores of a small number of students.

In the following analysis, we examine how state accountability plans have been redesigned to meet the NCLB requirements and the effects of the accountability rules on schools with diverse enrollments. First, we describe the state selection criteria and the data collection strategies used to address our research questions. In the second section, we examine how states developed their accountability plans to comply with the NCLB requirements. Specifically, we examine how states constructed their 12-year timeline for improving student performance, how proficiency rates differ across states, and the implications of these decisions for Title I schools. The third section analyzes the implications of AYP rules on schools that enrolled large numbers of minority, low-income, and limited English proficient students. We use data from the six states in our study to show the disparate impact of these rules on certain groups of students. To provide a different perspective on school performance, we also examine improvements in proficiency scores for both schools identified as needing improvement and schools meeting AYP. In the fourth section, we focus on the impact of the subgroup rules in California schools. California, which enrolls large numbers of minority students, identified 10 subgroup categories for accountability purposes under NCLB. We conclude with a discussion of the implications of these findings for meeting the goals set forth in NCLB and offer recommendations on how the accountability rules can be improved.

State Selection Criteria

We purposefully chose six states—Arizona, California, Illinois, Georgia, New York, and Virginia—to study the implementation of NCLB. Each state offers a unique opportunity for understanding how the federal law affects schools with large minority enrollments. Four criteria guided the selection process. First, the six states are geographically and politically diverse. At least one state is located in the West, Central, Northeast, and Southeast regions, including Arizona and California (West), Illinois (Central), New York (Northeast), and Georgia and Virginia (Southeast). Second, each state has a large proportion of minority students. Minority students in California are the numerical majority, since Asian, Black, and Latino students comprise over half of the K-12 enrollment. Arizona has large Native American and Latino enrollments; New York and Illinois have large Black and Latino enrollments; and, Georgia and Virginia have large Black enrollments (Table 1).

Table 1: Racial/Ethnic Breakdown of K-12 Enrollment (2001-02) in Arizona, California, Illinois, New York, Georgia, and Virginia.

State	Native American	Asian	Black	Latino	White
Arizona	7%	2%	5%	35%	51%
California	1%	11%	8%	44%	35%
Illinois	<1%	3%	21%	16%	59%
New York	<1%	6%	20%	19%	55%
Georgia	<1%	2%	38%	5%	54%
Virginia	<1%	4%	27%	5%	63%

Source: Common Core of Data (2001-2002)

Third, the degree of state control over local education policy varies across the six states (Wirt, 1977). Some state governance systems are highly centralized (Virginia), some are highly

decentralized (Arizona), and others are in between (California, Georgia, Illinois, and New York). In addition to each state's unique governance structure, there are important differences in each state's approach to improving student achievement, underscoring the different state policy contexts in which federal policies are being implemented. For example, Virginia has been cited as a leader in adopting state-mandated standards and testing requirements (Ravitch, 2002). Arizona, on the other hand, has relied more heavily on local districts to improve achievement through choice mechanisms and charter schools (Keegan, 1999).

Fourth, we selected states based on where they were in the reform process as it relates to the new federal requirements. To compare states with different starting points, we included states where some elements of the state policy align with NCLB's accountability requirements and other states where few policies meet the requirements. We used state compliance with the 1994 Improving America's Schools Act (IASA) as a measure of the status of state accountability policy. Two states in our sample—Virginia and New York—fully complied with the 1994 IASA mandate that assessments be aligned with content standards. The other four states received waivers from the federal government, allowing them extra time to comply with the 1994 requirements.[6]

Data Sources

We relied on both qualitative and quantitative data sources to conduct this study. First, we conducted interviews with (1) state superintendents; (2) state administrators responsible for assessment, accountability, and information technology; (3) directors of federal programs, research and evaluation, and teacher staffing; and (4) members of the state boards of education. In addition to interview data, we reviewed the state consolidated applications for federal funding under NCLB, state accountability workbooks submitted to the U.S. Department of Education, and the final regulations governing NCLB implementation. We supplemented information from these documents with local and national newspaper articles. The triangulation of qualitative data sources enabled us to verify information from individuals and institutions with different perspectives. Second, we collected quantitative data for all public schools in each state on (1) Title I program status and number of years in school improvement; (2) student background characteristics; and (3) achievement outcomes.[7] We constructed six state databases that included information on these variables for all public schools in the state.

[6] The waiver expired on December 31, 2002 for Illinois, June 30, 2003 for Georgia, August 31, 2003 for Arizona, and November 30, 2003 for California.

[7] These files were obtained from different divisions from each state department of education and merged with the three other data sources. See Appendix 1 for a complete description of the state data files.

(2) FRAGMENTED ACCOUNTABILITY SYSTEMS

Background

To comply with NCLB, states were required to develop and implement an accountability plan that would have all students proficient in reading and math in 12 years. Since provisions of the law went into effect immediately, states had less than a year to develop their accountability plan. Preliminary accountability plans were due to the U.S. Department of Education on January 31, 2003, and states had to submit final plans on May 1, 2003. If the implementation timeline was tough, states were also challenged with meshing the new requirements with existing systems. Since Congress did not consider how state context could affect a state's ability to interpret and implement the new requirements, the constraints imposed by previous educational decisions were ignored. In this section, we describe how states redesigned their accountability systems to meet the federal requirements, examine the obstacles to implementing a coherent accountability system, and discuss the political and educational implications of a fragmented accountability system.

The Interaction of State and Federal Accountability Policies

In building accountability systems to comply with NCLB, states layered the new federal requirements on top of existing systems. All six states retained their own state developed accountability systems in addition to adopting plans that complied with NCLB. This allowed states to retain educational goals they felt were important, maintain broad based support for the state system, and, at the same, satisfy the federal goals. Nonetheless, it created fragmented systems where schools often had to meet two sets of performance measures—one for the state and one to meet the federal law's AYP goals. The dual accountability systems also raised questions about the legitimacy of the federal requirements since many state officials objected to the imposition of the federal system onto the state system. In particular, state policymakers rejected NCLB's emphasis on setting a single performance expectation for all schools and subgroups of students and, instead, they embraced their own accountability systems, which gave schools credit for making progress toward achievement goals. Given the considerable variability in state accountability systems to begin with, states grafted NCLB's requirements onto their pre-existing test-based accountability systems in several ways. The summaries below describe how states incorporated the federal requirements into their state accountability systems.

- Arizona embraced one system for NCLB accountability and another for state accountability—Arizona Learns. The NCLB accountability system gives schools a pass/fail label based on the performance of different subgroups while the state system, which retained four performance levels, gives schools credit for an overall improvement in achievement. To satisfy the U.S. Department of Education requirements for a single accountability system, Arizona incorporated the AYP requirements into the state accountability plan. In practice, the two systems operate separately since schools receive both an AYP calculation and an Arizona Learns calculation, which is a composite measure of school performance based largely on absolute levels of performance on the

Arizona Instrument to Measure Schools (AIMS) and gains on the Stanford 9.[8] The retention of AZ Learns suggests that many policymakers and educators believe that gains are a fairer and more credible method for evaluating schools than NCLB's model of accountability. The Arizona superintendent of instruction underscored the importance of using Stanford 9 gains scores (i.e., measure of academic progress) in evaluating schools "…[W]e have something called the measure of academic progress, which measures how much progress you make given where the students were when you got them. So a good school in a poor neighborhood can shine by showing they made a lot of progress even though on the absolute percentage of students reaching proficiency they don't. . . that (i.e., measure of academic progress) used to be a marginal element in the Arizona system, [but] I've raised that to 40% and in 2005, the plan is to raise it to 50% of the measure (T. Horne, personal communication, June 18, 2003).

- California incorporated its Academic Performance Index (API), the cornerstone of its state system, into the accountability system as the additional indicator required by NCLB. However, it also retained the accountability requirements developed prior to NCLB. For example, California relies on a uniform growth target for all students and subgroups in a given school, which differs from NCLB's absolute standard for measuring student achievement. California policymakers were reluctant to rely solely on AYP, since this clashed with the API model. According to one state official, "there was a lot of long and hard debate before the state decided that was the most reasonable way to go, and it [API] takes a more holistic look at what students do, and how schools perform. We're going now into a model (NCLB) where a relatively small number of students in a school can affect a school's status, and even to the fact, just by a certain number of children not taking the test, and that slaps a label on a school as low performing. That's vastly different than what we. . . [have] under our API system, [and] how we looked at schools (D. Rury, personal communication, August 18, 2003).

- Illinois will continue the previously implemented Academic Watch List and Academic Early Warning List in addition to the policies adopted in response to NCLB. Schools that had fewer than 50% of students meeting or exceeding standards for two consecutive years were placed in Academic Early Warning Status (AEWS), and if they failed to show improvement after two more years, were placed in the Academic Watch list. Both lists target failing schools for additional assistance to help them meet the state standards. NCLB will produce a third list that will likely identify a different set of schools for improvement.

- New York deviated from a strict focus on annual increases in the percent of students who are proficient. Although NCLB requires states to establish three performance levels and credits schools only if they increase the percentage of students meeting the proficiency standard, state officials retained four performance levels and assigned each school a

[8] AZ Learns relies on a complicated weighting system based on multiple indicators of performance in different grades and subjects, and it incorporates both levels and gains in student achievement scores. However, unlike NCLB, it does not set separate subgroup accountability requirements. AZ Learns places very little emphasis on AYP scores. Under the state system, AYP counts for only 1 point out of a possible 19-37 possible points, which are used to determine one of four school labels: excelling, highly performing, performing, underperforming.

performance index. The New York school performance index credits schools for moving students from Level 1 (basic) to Level 2 (basic proficiency), which are both below the Level 3-proficiency standard required by NCLB. According to state officials, "the prime motivation of the index is to ensure that schools . . . focus effort and resources on students who start the school year so far below state standards that it unlikely that they will be able to demonstrate proficiency by the time of the administration of the state test" (New York State Department of Education, 2003, January 6). The Deputy Commissioner of Elementary and Secondary Education elaborated on the purpose of the state accountability system: "New York's system is not intended to rate or rank schools or hold them up to public derision. Nor is it a system that seeks to impose intrusive or draconian interventions upon a school or district. Instead, the New York system is intended to help policymakers determine how well schools and districts are performing in relation to preparing students to meet standards in key subject areas and then to provide assistance and support to those with the greatest need. The focus is always on helping schools and districts to help themselves, with the recognition that the continued failure to provide adequate educational opportunities to students is unacceptable and must be remedied" (Kadamus, 2001, p. 14).

- Georgia has an accountability system that assigns Absolute and Progress grades (A, B, C, D, F) to each school. Georgia law requires that the state's accountability system provide rewards and interventions for all public schools based on both absolute student achievement and progress on improving student achievement (O.C.G.A. § 20-14-33, 20-14-37, and 20-14-38). The state's accountability application for NCLB differentiated state interventions, which will be based on the state grading system, from the federal interventions required by NCLB (Georgia Board of Education, 2003). The state grading system is under development and will be merged with the federal AYP determinations. According to one state policymaker, "AYP is just a portion of Georgia's accountability system" (D. Nelson, personal communication, February 13, 2003).

- Virginia will continue to issue its own accreditation ratings in addition to the accountability requirements under NCLB. Virginia officials strongly objected to how they were required to use assessment data for limited English proficient students to determine AYP. In a letter to Under Secretary Hickok, the president of the Virginia State Board of Education claimed, "the formula for determining AYP is irrational and lacks common sense, certainly as applied to Virginia. As a consequence, the AYP results for Virginia will be seriously flawed. . ." (Virginia Department of Education, 2003). Prior to the enactment of NCLB, state policymakers and lawmakers invested enormous political capital and fiscal resources to build its test-based accountability program based on the Standards of Learning (SOL) assessment, which was first administered in 1998. By 2006-07, schools are required to have 70% pass-proficient rates, which obviously differsdifferent from NCLB's 100% proficiency rate. In many ways, the SOL has become so deeply ingrained in Virginia's education policy landscape that many educators view it as more legitimate than NCLB's model of school accountability. For example, one principal in a northern Virginia suburb district, in which five of 18 schools made AYP, observed that NCLB is "not a meaningful way to judge schools. The state has a far

better plan in terms of looking at accreditation—that's still our focus" (Helderman, 2003, September 12).

These six state summaries converge on a clear finding: states preferred their own accountability system over that imposed by NCLB. Since state policymakers had invested considerable political capital and education planning in developing their accountability systems prior to NCLB, they were reluctant to abandon them to meet the federal requirements. The development of these systems required the political buy-in of various constituencies, political actors, and local educators that took place over a considerable length of time. For example, New York adopted a plan to revise the state's learning standards and to upgrade the rigor of its assessments that evolved throughout the 1990s. By the end of the decade, New York had developed an accountability system that incorporated many of the requirements of the federal law, but was also structured to meet state goals and to show growth in student achievement. In Virginia, the state began developing its accountability system in 1995 when it adopted the K-12 SOLs. The state tied high-stakes to performance on the SOLs, but phased in sanctions over time in order to obtain support among the public and local educators. Under pressure from low-performing schools and districts, they modified the system to ease the effects of sanctions and to maintain support for the overall system. Both states were reluctant to relinquish their system, not only because of the considerable time they had invested in developing them, but also because they had gained widespread support among the political leadership and various constituencies with an interest in education. In general, state officials believe their systems will work and are not convinced that NCLB is good education policy.

Other states did not want to jeopardize the fragile political support that they were just beginning to build for their own accountability system. NCLB added considerable uncertainty to the accountability system in California, a state that had gone through many contentious political battles among policymakers over the adoption of a statewide testing and accountability system that dates to the 1970s (Kirst, 2002; Citizens' Commission on Civil Rights, 2002). Throughout this period, changes in political leadership, shifts in the political winds, or the collapse of political coalitions led the state to dismantle earlier reforms and assessments or layer on new ones (Kirst, 2002). The most recent reform, adopted in 1999, has the backing of state policymakers and business but only lukewarm support from local educators and parents. NCLB is likely to undermine this support if large numbers of schools fail to meet the federal proficiency standards or other requirements prove unworkable.

The Unintended Consequences of Fragmented Accountability Systems

Although NCLB requires policymakers to create a single statewide accountability system for all public schools, the federal accountability rules made it exceedingly difficult for states to do that. Two policies in particular—the adequate yearly progress rules and proficiency standards—created mixed messages about school performance, imposed inconsistent and unrealistic goals for annual test score gains, and yielded divergent and meaningless proficiency definitions in reading and math. While none of these consequences were intentional aims of federal policy, they challenged the capacity of states to develop coherent accountability systems and altered the rules governing school accountability.

Sending mixed messages about school performance

The transition from a state system to one that that had to meet both state and federal goals made it difficult for schools to understand the achievement expectations and performance labels associated with different accountability systems. The most difficult task for state policymakers was in the implementation of transitional AYP rules. There are key differences in the definition of AYP in the first year (2002-03) and second year (2003-04) of NCLB implementation. In the first year, states had to apply AYP rules developed under the 1994 Title I law to identify schools in need of improvement.[9] Under the transitional definition, AYP was based on the percentage of students in a school that met the state defined proficient level. States were not required to disaggregate test scores by subgroup. Since states retained their own definition of proficiency, the transitional definition often differed from the state definition. For instance, Arizona, Illinois, New York, Georgia, and Virginia required schools to reduce the percentage of students in the lowest performing category (e.g., "basic") whereas NCLB defined adequate yearly progress by the percentage of students scoring at or above the proficiency threshold.[10]

Because of these differences in state and federal definitions of proficiency, results from the two accountability systems often conflicted. In Arizona, for example, some schools failed to make adequate yearly progress and were labeled as needing improvement under federal law, but were able to meet or exceed the state criterion for annual progress on state assessments. In 2002-03, there were 289 Title I schools in Arizona that were identified for school improvement under NCLB, but were labeled "performing" or "highly performing" under the state system. In short, these Title I schools were failing to make sufficient progress under the federal system but were doing well according to the state system. These differences in school ratings confused district administrators, who were responsible for interpreting and explaining the results to principals, teachers, and parents (J. Sullivan, personal communication, June 16, 2003).

For the 2003-04 school year, states were required to incorporate the subgroup requirements into their accountability systems. Under these rules, each subgroup must reach the state defined proficiency level in reading and math. In addition, 95% of the students in each subgroup must take the state assessment. Schools would be identified for improvement if any one subgroup did not make the performance target in reading or math or if any subgroup did not meet the 95% participation requirement. That created four indicators based on state assessments—percent proficient in reading, participation in reading exam, percent proficient in math, participation in math exam—that were required for each subgroup. These new requirements had two immediate consequences for states. First, many states were unprepared to disaggregate data for all the subgroups required by NCLB. For example, Education Week's 2003 summary of state education policy (Education Week, 2003) indicated that only 20 states actually disaggregated performance by race and ethnicity in 2002-03. In our six state sample, only New York disaggregated results for all the subgroups required under NCLB and the five other states were

[9] In this report, we refer to schools as "needing improvement" or "meeting AYP."
[10] See Appendix 3 for a description of each state's transitional adequate yearly progress.

still working to report the performance of at least one subgroup.[11] Thus, many states had to construct an infrastructure for disaggregated reporting of performance data in order to hold schools accountable for meeting separate racial subgroup targets. Second, the subgroup rules made it more difficult for schools to make AYP than to meet state accountability goals. In Virginia, 723 (40% of all schools) failed to make federal AYP goals while only 402 (22%) failed to meet state accreditation standards. Indeed, in Arlington County, a suburban district in northern Virginia, over half of all schools failed to make AYP, even though most were meeting state accreditation standards.

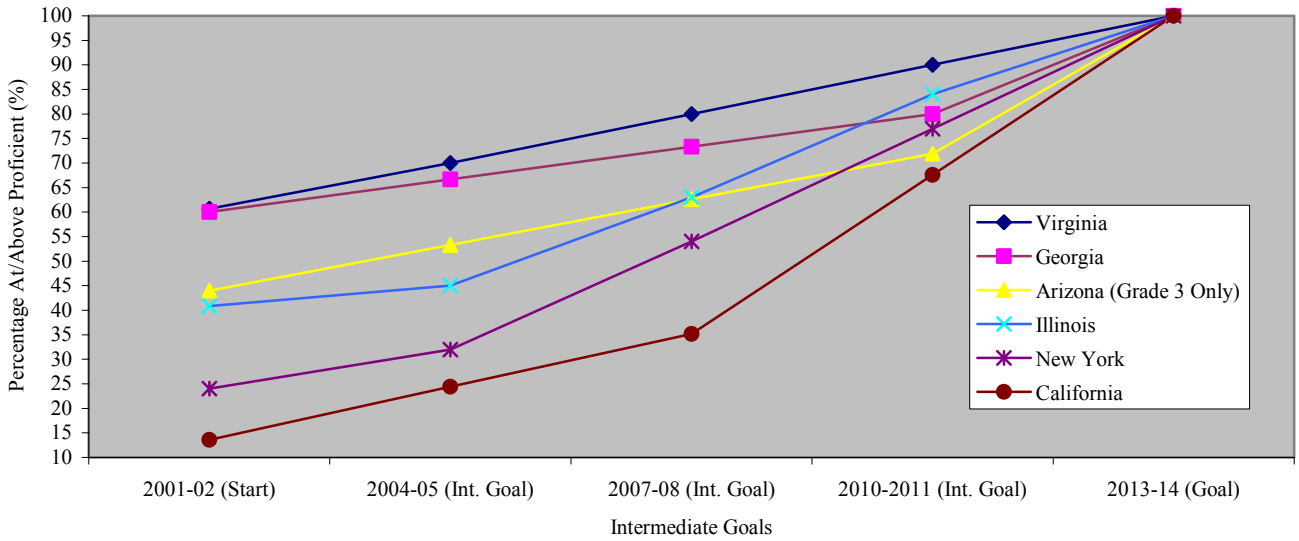
Imposing inconsistent expectations for annual test score gains

The NCLB rules for establishing proficiency standards imposed inconsistent achievement goals on schools and held schools in different states to highly uneven expectations for annual test score gains. NCLB allowed states to establish their own definitions of what it means to be proficient and, at the same time, required states to bring all students up to 100% proficiency within 12 years. To meet the requirement, states must establish a "starting point," that is the percentage of students scoring at the proficient level on state tests. States must also establish intermediate targets along the 12-year timeline to 100% proficient. Even though states must follow a very prescriptive formula for establishing the starting point, intermediate goals, and the 12-year timeline, the formula ignores differences in state definitions of proficiency. As a result, states with a smaller percentage of students at proficiency (e.g., California) will have lower initial starting points than states with a higher percentage of proficient students (e.g., Virginia) (see figure 1).

Although the federal legislation established a common achievement goal for all public schools, it ignored large differences in state academic standards, standardized tests, and proficiency levels among the states. Figure 1 shows the different challenges schools face in meeting the 100% target within 12 years—the uniform expectation applied to all public schools and students. The slope of the 12-year timeline in Virginia and Georgia is not as steep as that for New York and California. Since New York and California started well behind the four other states, as well as behind most other states in the nation, theses states will have a much steeper climb to 100% proficiency. For example, California needs to improve reading proficiency rates by 86 percentage points over the next 12 years compared to 40 percentage points in Georgia and Virginia. These slopes vary because states established different starting points (2001-02). These range from a high of nearly 60% in Virginia and Georgia, to approximately 40% in Arizona and Illinois, and below 25% in New York and California.

[11] According to the Education Commission of States' website on NCLB implementation in 2002-03, Arizona disaggregated only by race/ethnicity, California needed to include the performance of students with disabilities, Illinois needed to include limited English proficient and migrant students, Georgia needed to include migrant students, and Virginia needed to include the performance of economically disadvantaged and migrant students. See http://nclb.ecs.org/nclb/

Figure 1: Differences in Reading Proficiency Rates from the Starting Point (2001-02) to the End (2013-14) of the 12-Year Timeline in Six Selected States.



Source: "Consolidated State Application Accountability Workbook" for Virginia, Georgia, Arizona, Illinois, New York, and California.

The emphasis placed on annual improvements in proficiency rates means that states have a strong incentive to backload improvements at the end of the 12-year timeline. Table 2 compares the amount of improvement in reading proficiency during the first and second half of the 12-year timeline. Only Virginia established equal 20-percentage point improvement targets in the first and second half of the 12-year timeline while the other states set more ambitious achievement gains for the later years. As shown in column 7 of Table 2, California targeted about 75% of the improvement for the last six years, and Georgia, Arizona, Illinois, and New York targeted over 60% of the improvement for the last six years. By setting lower performance targets in the first half of the 12-year timeline, state officials realized that schools would have an easier time meeting the reading and math targets and making AYP. Consequently, potentially fewer schools would be subjected to federally mandated sanctions, at least initially.

Table 2: Differences in the Amount of Improvement in Reading Proficiency Rates During the First and Second Half of the 12-Year Timeline in Six Selected States

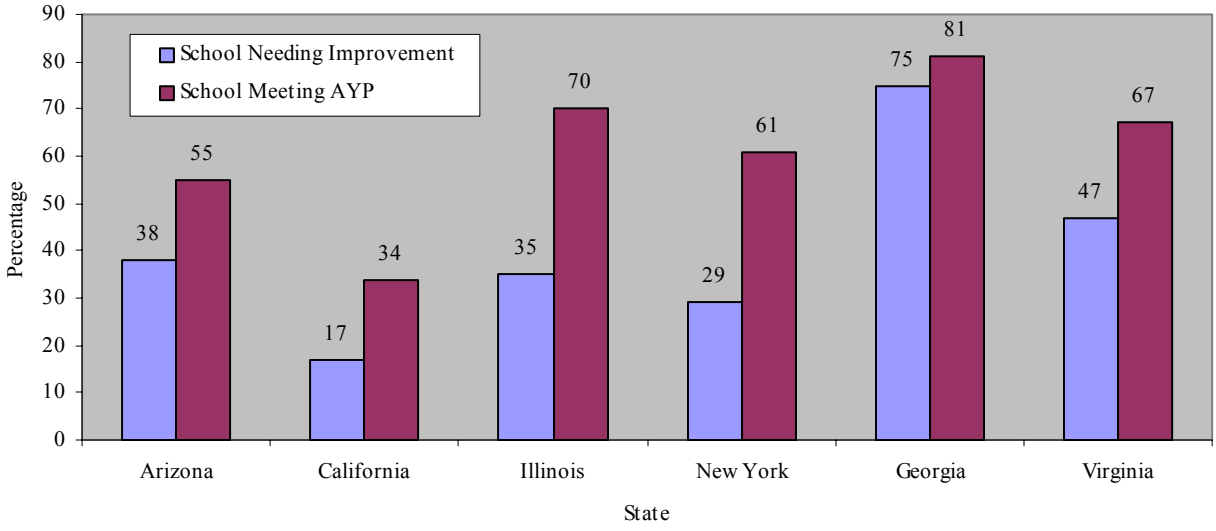
(1) State	(2) Starting point	(3) Total Improvement over 12 years	(4) Improvement over first half	(5) % of Improvement in first half	(6) Improvement over second half	(7) % of Improvement in second half
Virginia	60	40	20	50.0	20	50.0
Georgia	60	40	13	32.5	27	67.5
Arizona	44	56	19	33.9	37	66.1
Illinois	41	59	22	37.3	37	62.7
New York	24	76	30	39.4	46	60.5
California	14	86	21	24.4	65	75.6

Source: "Consolidated State Application Accountability Workbook" for Virginia, Georgia, Arizona, Illinois, New York, and California.

Yielding divergent and meaningless proficiency definitions

Because of differences between states in their performance standards and cut scores, it is impossible to compare proficiency across states. In comparing mean proficiency rates, we found that proficiency rates were vastly different in similar schools, that is, among "schools needing improvement," on one hand, and "schools meeting AYP," on the other.[12] In some instances, the mean proficiency rates were higher in failing schools in one state than in schools meeting AYP in another state. Two comparisons in Figure 2 below illustrate these points.[13] First, among schools needing improvement, mean reading proficiency rates are nearly four times higher in Georgia (75%) than in California (17%), and there is wide variability in proficiency rates in reading among schools meeting AYP. Second, average proficiency rates in schools needing improvement in Arizona (38%), Illinois (35%), Georgia (75%), and Virginia (47%) exceed the average proficiency rate of schools meeting AYP in California (34%). Put another way, average scores for the lowest-performing schools in these four states are higher than the average score for schools meeting AYP in California.

Figure 2: Mean Reading Proficiency in Schools Needing Improvement and Schools Meeting AYP in Six Selected States, 2002-03.



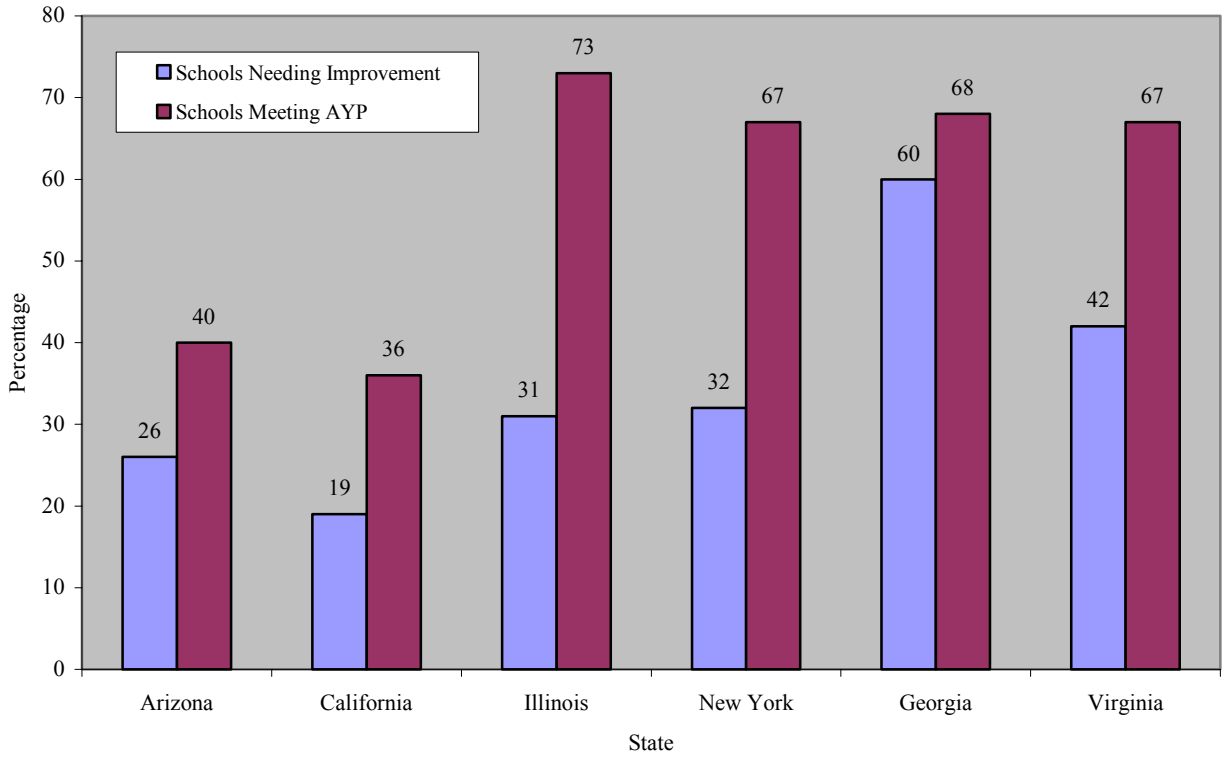
Note: Sample sizes for schools needing improvement and schools meeting AYP are as follows: Arizona (n=399, n=1404), California (n=814, n=7837), Illinois (n=527, n=3378), New York (n=434, n=6030), Georgia (n=430, n=1681), Virginia (n=34, n=1926).
Source: See Appendix 1, 2, 4.

[12] Appendix 4 lists the number of schools needing improvement in each of the six states as of June 2003. Many of these lists changed considerably over the course of the 2002-03 school year. For instance, the number of schools needing improvement in California decreased from slightly over 1,000 to 815 schools due to changes in rules. In Arizona, the initial list of 399 schools needing improvement was changed to 244 in October 2003. For our analysis, we used the 399 count, but we also conducted analyses using the 244 count, which are reported in subsequent footnotes.

[13] We created a school average score in both reading and math. Since each state administers tests in different grades, the average is based on the percentage of students at or above proficiency across the tested grades.

The math results also underscore how radically different proficiency standards are across states. Figure 3 compares math proficiency rates in schools needing improvement and schools meeting AYP. Among schools needing improvement, nearly two-thirds of all students in Illinois, New York, Georgia, and Virginia were at or above proficiency in math whereas fewer than half of the students in Arizona and California met or exceeded proficiency. When comparing schools with different labels across states, in some cases schools needing improvement had higher average test scores than schools meeting AYP. Average math proficiency rates for failing schools in Georgia (60%) and Virginia (42%) were higher than proficiency rates for schools meeting AYP in Arizona (40%) and California (36%).[14] Put differently, students attending schools that failed to make AYP in Georgia and Virginia had higher math proficiency rates than students in adequately performing schools in Arizona and California. These results underscore how radically different proficiency standards are across states.

Figure 3: Mean Math Proficiency in Schools Needing Improvement and Schools Meeting AYP in Six Selected States, 2002-03.



Note: Sample sizes for schools needing improvement and schools meeting AYP are as follows: Arizona (n=399, n=1404), California (n=814, n=7837), Illinois (n=527, n=3378), New York (n=434, n=6030), Georgia (n=430, n=1681), Virginia (n=34, n=1926).
Source: See Appendix 1, 2, 4.

[14] When using the 244 count for Arizona in 2002-03, average reading scores were 32% for schools needing improvement and 54% in schools meeting AYP. In math, the average was 19% in schools needing improvement and 39% in schools meeting AYP.

Implications of Fragmented Accountability Systems

During the first year of NCLB implementation, states continued to operate an accountability system that emphasized state and local goals over federal objectives. It should not be surprising to find that state goals take precedence over federal priorities since state policymakers have invested considerable political capital to develop their own systems and states have historically resisted federal coercion. Since many believe that the Constitution defers education authority to state governments rather than the federal government, the new requirements raise interesting questions about who controls education.[15] In this case, competition between the federal and state goals created a fragmented system and caused confusion among educators throughout state education systems.

Since NCLB requires all states to move 100% of students to the proficient level on a state test, strong incentives exist for states to create a dual system of accountability—that is, proficiency standards for state accountability and another to comply with the federal law. The law also creates incentives for states to lower the definition of proficiency for federal accountability purposes. Colorado, for example, will count students reaching the state's "partially proficient" standard on the Colorado State Assessment Program (CSAP) as meeting the federal definition of "proficient." Connecticut recently established a new proficiency level to comply with federal law, but it will be lower than the state's definition of proficiency (Education Week, 2002). More states may adjust their performance standards to establish more realistic targets for having all students meeting proficiency within 12 years. And states with lower proficiency definitions may have little incentive to raise the bar. Moreover, the differences in state accountability systems have especially troubling implications for schools, which are subject to federal sanctions for failing AYP. Current proficiency standards are so wildly divergent that one measurement expert (Linn, 2003, September 1) has argued that the term "proficiency becomes meaningless" (p. 12). Linn adds that many schools will be "subject to substantial sanctions based on the progress that is made against arbitrary performance standards that lack any semblance of comparability from state to state" (p. 12). Proficiency is used in the law as if it is describing a stable and meaningful concept but it actually has no consistent meaning and can be manipulated by policymakers.

Politically, it is also unclear how the federal government and the U.S. Department of Education will accommodate differences in state accountability policies and the variability in proficiency standards. Since NCLB requires all states and districts to participate in biennial administrations of the National Assessment of Educational Progress (NAEP), this federally funded assessment may become an increasingly important educational tool for measuring student achievement and a political tool for spotlighting states with low proficiency definitions.

[15] Our federal report discusses in greater detail and depth the relationship between federal and state government during the first-year implementation of NCLB (Sunderman & Kim, 2004).

(3) THE DISPARATE IMPACT OF TRANSITIONAL AYP RULES

Background

During the first year of NCLB, states used the transitional definition of adequate yearly progress to identify schools needing improvement. Under the transitional definition, AYP was based on the percentage of students in a school that met the state defined proficient level. States were not required to disaggregate test scores by subgroup. Since states retained their own definition of proficiency, the transitional definition often differed from the state definition. For instance, Arizona, Illinois, New York, Georgia, and Virginia required schools to reduce the percentage of students in the lowest performing category (e.g., "basic") whereas NCLB defined adequate yearly progress by the percentage of students scoring at or above the proficiency threshold. Schools that were identified as needing improvement were required to offer their students the option to transfer to another public school and to develop a plan for improving teaching and learning. In future years, schools that fail to make AYP for three or more consecutive years also will have to implement more intrusive sanctions, ranging from corrective action policies to reconstitution and restructuring. Although it is too early to determine how sanctions affect student achievement, it is clear these sanctions will be implemented disproportionately in schools with a large number of minority, low-income, and limited English proficient students. In this section, we examine the impact of the AYP rules on schools serving minority and low-income students and whether these rules identify schools needing improvement on the basis of a school's demographic characteristics or a school's contribution to learning.

The Disparate Impact of Adequate Yearly Progress Rules

Using the transitional adequate yearly progress rules, states identified schools needing improvement based on student performance in reading and math. Given the overlap between student background characteristics and test score levels, schools needing improvement enrolled a disproportionately large number of minority and low-income students and limited English proficient students. Figure 4 compares the average percentage of minority, low-income, and limited English proficient students in schools needing improvement and schools meeting AYP in Arizona, California, Illinois and New York for the 2002-03 school year. The bar graphs show that schools needing improvement enroll a larger percentage of minority and low-income students, on average, than schools meeting AYP in all six states. The magnitude of these demographic differences is especially large in Illinois and New York, which are two of the most segregated states for Black and Latino students (Frankenberg, Lee, & Orfield, 2003).[16] In both states, sanctioned schools enrolled two to three times more minority, low-income, and limited English proficient students, on average, than schools meeting AYP. Moreover, schools needing improvement are concentrated in the largest urban districts: sanctioned schools in New York City make up 69% of all sanctioned schools in New York state, and sanctioned schools in Chicago make up 66% of all sanctioned schools in Illinois.[17]

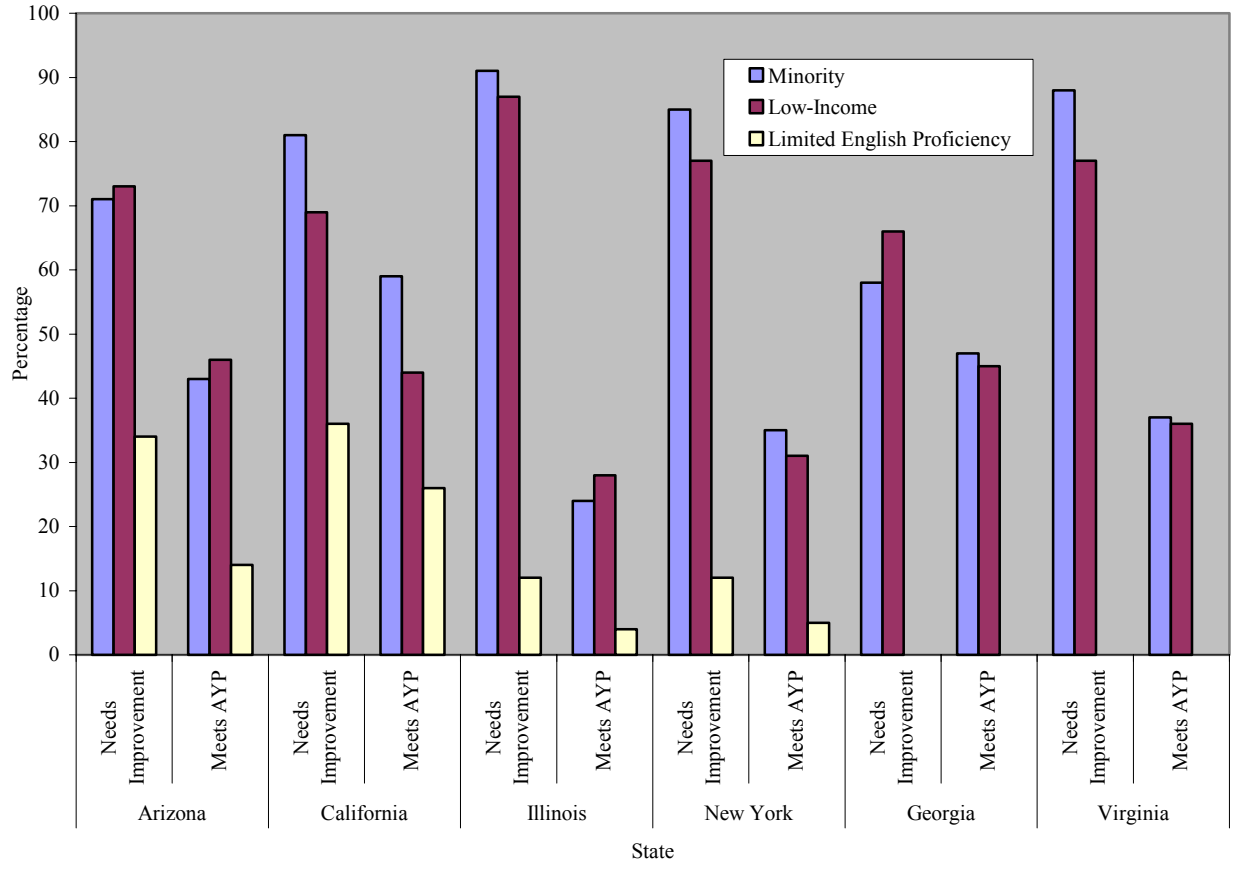
---

[16] See Table 16 and 18 in this report, which is available from http://www.civilrightsproject.harvard.edu/research/reseg03/resegregation03.php.

[17] Although 347 Chicago public schools were identified as needing improvement under NCLB, only 179 were required to offer school transfers during 2002-03. Our analysis of the transfer option provides additional details about school choice implementation in Chicago (Kim & Sunderman, 2004).

Since Georgia and Virginia enroll comparatively fewer students with limited English proficiency than schools in the other four states in our study, we limited our comparison to the percentage of minority and low-income students in schools needing improvement and schools meeting AYP.[18] In Georgia, schools needing improvement had a somewhat higher proportion of minority and low-income students than schools meeting AYP, while in Virginia the difference between the two types of schools was larger. This is because half of all sanctioned schools in Virginia (n=34) are located in Richmond City Public Schools, where minority students make up over 90% of the district enrollment.

Figure 4: Mean Demographic Characteristics (Percentage of Minority, Low-Income, Limited English Proficient Students) of Schools Needing Improvement and Schools Meeting AYP in Six Selected States, 2002-03.



Note: Sample sizes for schools needing improvement and schools meeting AYP are as follows: Arizona (n=399, n=1404), California (n=814, n=7837), Illinois (n=527, n=3378), New York (n=434, n=6030), Georgia (n=430, n=1681), Virginia (n=34, n=1926).
Source: See Appendix 1, 2, 4.

[18] Georgia identified 436 schools in need of improvement for 2002-03. However, since 6 schools had missing demographic and achievement data, our analytic sample for Georgia includes 430 schools needing improvement.

Using Proficiency Gains to Compare Performance Trends

NCLB relies on average test scores to identify schools that need improvement. Average test scores, however, usually reflect differences in student background characteristics more than differences in school quality (Ladd, 1996). Value-added measures of achievement, which are based on student achievement gains, are intended to isolate a school's contribution to student learning and provide a more valid estimate of school performance (Meyer, 1996). However, since we did not have access to student-level data, we were unable to compute value-added scores that adjust for different background characteristics (Bryk, Thum, Easton, & Luppescu, 1998; Sanders & Horn, 1998). Instead, we examined two-year changes in proficiency rates both in schools needing improvement and schools meeting AYP. These comparisons are intended to provide an alternative perspective on school quality and to assess performance trends over time in ostensibly low- and high-performing schools.

Table 3 compares two-year changes in the percentage of students at or above proficiency in reading and math in schools needing improvement and schools meeting AYP. The two-year proficiency gains in reading and math are reported for schools in five states.[19] The first row of Table 3 compares proficiency gains for schools in New York. On average, fourth-grade reading scores increased by 3.8 percentage points in schools needing improvement and 2.6 percentage points in schools meeting AYP. These two-year gains are quite similar in both types of schools, even though average proficiency levels are substantially higher in schools meeting AYP than in schools needing improvement (see Figures 2 and 3). Overall, in all five states, there are small differences in two-year gains between schools needing improvement and schools meeting AYP. In sum, schools appear to make similar progress even though their initial starting points are quite different.

[19] We were unable to calculate two-year trends in California, since it administered a new state test in 2002.

Table 3: Two-Year Changes (2000-02) in the Percentage of Students At or Above Proficiency in Reading and Math by School Category (Needs Improvement vs. Meets AYP) in Five States.

	Reading Proficiency Gain		Math Proficiency Gain	
	Needs Improvement	**Meets AYP**	**Needs Improvement**	**Meets AYP**
New York (2000-02)				
Grade 4	3.8	2.6	5.3	2.0
Grade 8	-1.5	-0.11	6.4	8.0
Virginia (2000-02)				
Grade 3	16.5	10.4	20.1	9.1
Grade 5	14.4	9.5	19.2	8.0
Grade 8	12.2	0.2	15.9	9.5
Georgia (2000-02)				
Grade 4	14.7	14.6	2.1	5.6
Grade 6	12.1	8.7	3.1	3.0
Grade 8	6.9	5.6	10.8	12.6
Illinois (2000-02)				
*ISAT (Grade 3, 5, 8)	1.4	-0.01	1.4	-0.01
Arizona (2000-02)				
Grade 3	3.0	3.6	9.3	9.9
Grade 5	-9.6	-6.5	6.5	12.1
Grade 8	2.5	3.4	1.2	2.0

*The ISAT measure is an average of reading and math, grades 3, 5, 8. Samples sizes for sanctioned schools and schools meeting AYP are as follows: AZ (399, 1404), IL (527, 3378), NY (434, 6030), GA (430, 1681), VA (34, 1926)
Source: See Appendix 1, 2, 4.

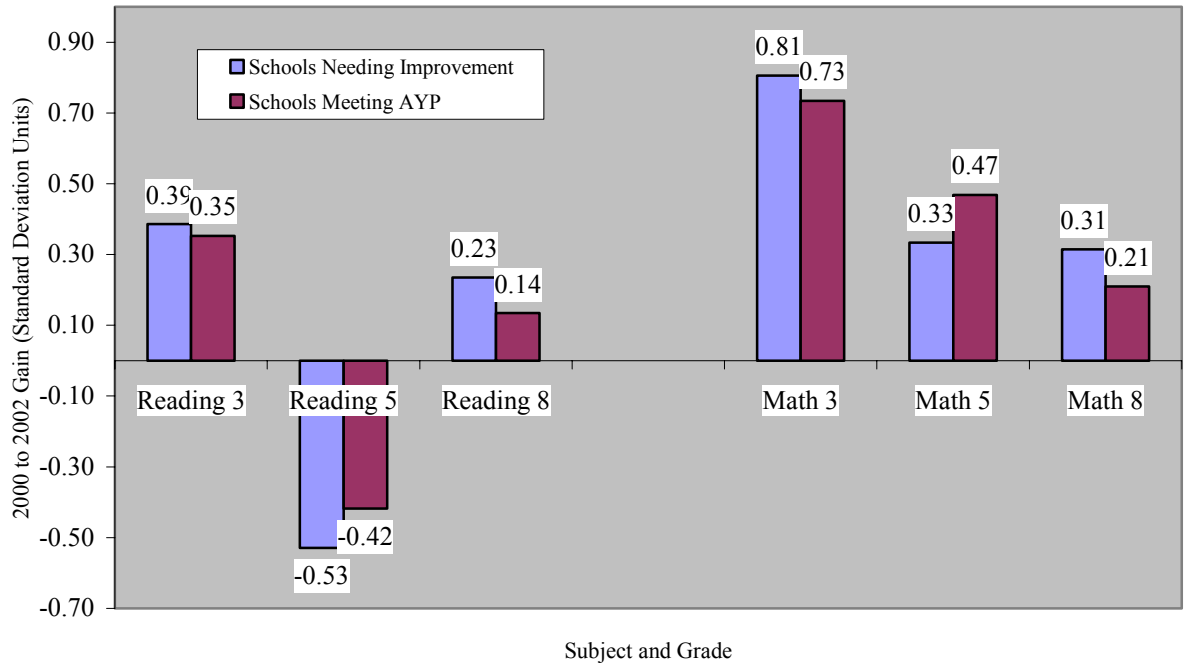
Although it appears that both types of schools made similar two-year gains, these results should be interpreted in light of several caveats. In particular, since schools needing improvement have lower average test scores than schools meeting AYP, they also have more room for increasing achievement scores over time.[20] Schools meeting AYP, however, may have less room for growth since most of their students are already meeting proficiency standards.[21] Nonetheless, the results in Table 3 underscore the importance of using multiple performance criteria, including gain scores, in evaluating school quality.

Because proficiency scores only credit improvements in the number of students who cross over the proficiency threshold, we undertook another analysis using effect size indicators.[22] These measures credit schools for improving achievement scores across different performance levels (e.g., basic, proficient, advanced) rather than simply calculating the percentage of students who are at or above the proficiency cut score. Using data from the Arizona AIMS, we examined reading and math gains from 2000 to 2002 in grades 3, 5, and 8 in schools needing improvement and schools meeting AYP. Since average scaled scores were available for each school, by subject and grade, we computed effect sizes that express two-year gains in terms of standard

[20] The large two-year gain for schools needing improvement may stem from "regression to the mean." Since schools needing improvement had proficiency rates below 50% in reading and math for two or more consecutive years, the average score for these low-performing schools will tend to move toward the mean on subsequent tests.
[21] This "ceiling effect" may suppress the amount of growth that schools meeting AYP are able to make over time.
[22] To compute an effect size that captured improvements over two-years, we took the difference in mean scaled scores between 2000 and 2002 and divided this number by the pooled standard deviation from both years.

deviation units. The bar graphs in Figure 5 below suggest that the magnitude of the two-year reading and math gain was quite similar in both types of schools, with slightly larger gains for schools needing improvement in grades 3 and 8 in both reading and math.[23] These effect sizes suggest very little difference in two-year gains for schools needing improvement and schools meeting AYP.[24]

Figure 5: Reading and Math Gains (2000-02), Grades 3, 5, 8, for Schools Needing Improvement and Schools Meeting AYP in Arizona.



Note: Average scaled scores by subject and grade were used to compute effect sizes.
Source: See Appendix 1, 2, 4.

[23] Assuming scores are normally distributed, a .31 standard deviation gain suggests that the average school needing improvement increased grade 8 math scores by 12 percentile ranks from 2000 to 2002 (i.e., from the 50th percentile to the 62nd percentile). Similarly, the .21 standard deviation gain implies that grade 8 math scores increased by 8 percentile ranks, on average, (i.e., from the 50th percentile to the 58th percentile) in schools meeting AYP. Although the effect size varies across grades and subjects, the magnitude of the two-year improvement is quite similar for both types of schools. It should also be noted that the effect sizes for grade 3 math are particularly large, since it is rare for achievement scores to increase by over three-quarters of a standard deviation in two years. For example, reading and math gains on the National Assessment of Educational Progress rarely exceed one-quarter of a standard deviation over four to six years (Grissmer & Flanagan, 2001). Some analysts (Linn, Baker, & Betebenner, 2002) have suggested that an effect size of .05 might be a reasonable adequate yearly progress target.

[24] Nonetheless, we urge caution in interpreting these gains since they are based on school-level information for only two-years. Analyses involving student-level data for several years would permit stronger inferences about learning gains over time in both schools needing improvement and schools meeting AYP.

Implications of Transitional Adequate Yearly Progress Rules

Requiring all students to be proficient in reading and math is a deceptively simple education goal fraught with political dangers if too many schools and students fail to meet unrealistically high performance standards for multiple subgroups. Since different methods are often used to determine performance levels, determining the cut score denoting proficiency is an inherently political decision. If decisions about proficiency cut scores are to survive political scrutiny, they will require enormous buy-in from stakeholders. In developing their state accountability systems, many states sought to do this by slowly phasing in requirements over time and using gains in school performance that capture improvement over time. Rather than pursuing this strategy, however, NCLB immediately failed hundreds of schools and concentrated the sanctions in schools with large numbers of minority, low-income, and limited English proficient students. As more educators question the legitimacy and fairness of Title I accountability policies, the federal law may lose the political support it needs to sustain long-term improvements in minority student achievement.

Being labeled as "needing improvement" does not imply that a school is failing to improve the performance of the students served. Rather, schools identified as needing improvement usually serve more disadvantaged students, whose test scores start lower, on average, than their peers in schools meeting AYP. In fact, we find smaller differences in two-year gains than in mean test score levels between schools needing improvement and schools meeting AYP. Thus, schools are identified as needing improvement based on their demographic characteristics rather than their contribution to student learning. If NCLB continues to concentrate failure in schools serving the most disadvantaged students with lower test scores, it may further undercut administrative and political support for public schools that are doing a good job educating the neediest students. The use of achievement gain scores and trend data would provide much-needed evidence to provide a more meaningful and comprehensive picture of school quality.

(4) THE DISPARATE IMPACT OF SUBGROUP RULES IN CALIFORNIA

Background

Beginning with the 2003-04 school year, NCLB requires states to incorporate the subgroup requirements into their state accountability systems. Under these rules, each subgroup must reach the state defined proficiency level in reading and math. In addition, 95% of the students in each subgroup must take the state assessment.[25] Title I schools can be identified as needing improvement if any one subgroup does not meet a reading or math cut score or if any subgroup does not meet the 95% participation requirement. In this section, we describe how the new subgroup rules were applied in California public schools, analyze the impact of these policies on schools, and explain why subgroup rules put schools segregated by race and poverty and integrated schools at greater risk of being identified as needing improvement.

NCLB Subgroup Rules Applied to California

Under the adequate yearly progress rules, California identified 10 major subgroup categories. Since eight categories (Black, Latino, Asian, Pacific Islander, Filipino, Native American, White, Socio-economically Disadvantaged) were already included in the state accountability system, California needed to add two categories (students with limited English proficiency and students with disabilities) for determining AYP.[26] If any subgroup meets California's minimum group size criterion, it must be counted for AYP calculations.[27] In California and elsewhere, subgroup accountability is most likely to affect schools enrolling disadvantaged minority students and schools with racially diverse enrollments. For example, a school that enrolls a large number of low-income Latino students can potentially be held accountable for meeting separate targets by race/ethnicity, socio-economic disadvantage, and limited English proficiency. A school that is ethnically diverse, that is a school that enrolls a fairly equal number of White, Black, Latino, and Asian students, may have to meet four separate performance targets. This illustrates the potential impact of subgroup rules on disadvantaged schools and diverse schools. However, the number of students in a subgroup must meet or exceed California's minimum group size criterion to be used for school accountability.

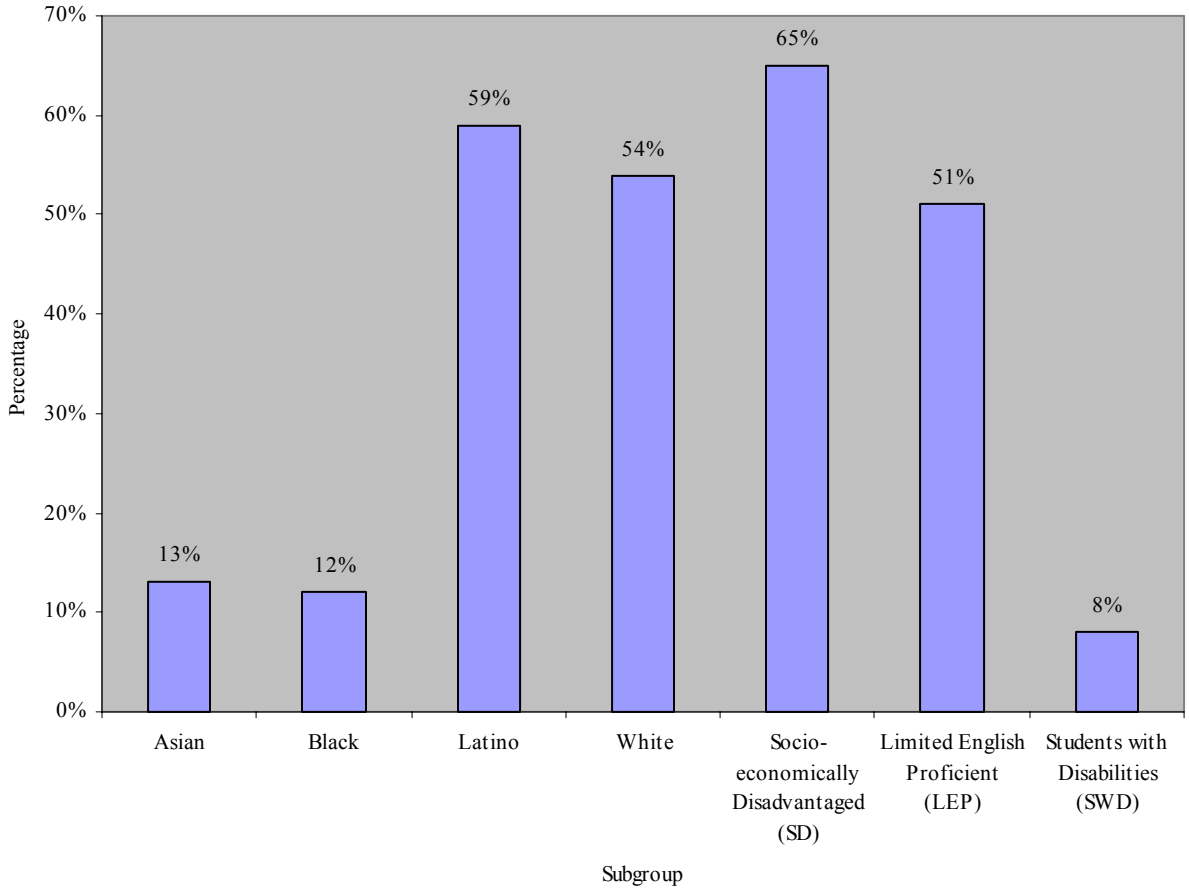
[25] As stated on page 9, there are additional requirements for making adequate yearly progress, but we focus on the proficiency targets and participation requirements for our analysis.

[26] According to the California consolidated accountability application, "California defines a socio-economically disadvantaged group rather than an economically disadvantaged group. A student is included in the socio-economically disadvantaged group if they participate in the Free or Reduced Price Lunch program or if the highest level of education of either of the student's parents is less than a high school diploma" (California Department of Education, 2003b) (p. 33).

[27] Under the definition of adequate yearly progress adopted in California, a subgroup is counted for school accountability if it represents 100 students *or* 50 students and 15% of the students with valid test scores in a given school.

Figure 6 shows the percentage of California schools, which were held accountable for meeting separate achievement targets for one of seven subgroup categories. The bar graph shows that over 50% of all California schools contained subgroups for Latinos, Whites, socio-economically disadvantaged, and limited English proficient students whereas a much smaller percentage of schools contained an Asian, Black, or special education subgroup. In most schools, then, the performance of Latino, White, socio-economically disadvantaged, and limited English proficient students will determine whether a school makes AYP. A much smaller number of schools will be held accountable for meeting separate targets for the other three groups of students.

Figure 6: Percentage of California Schools (n = 8,665) Required to Meet Separate Targets in Reading (Spring 2003 Administration) for Selected Subgroups.

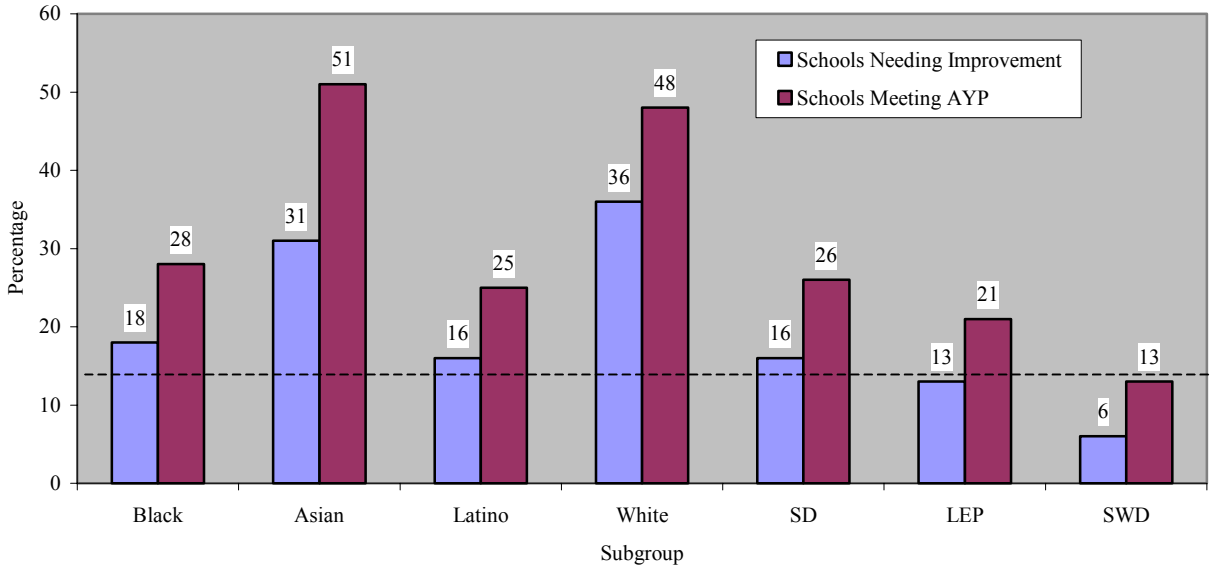


Note: We excluded Native American, Filipino, and Pacific Islanders in our subgroup analyses, because very few schools contained these subgroups: Only 14 schools (<1%) contained a Native American subgroup, 176 schools (2%) contained a Filipino subgroup, and 2 schools (<1%) contained a Pacific Islander subgroup.
Source: See Appendix 1, 2, 4.

The Disparate Impact of Subgroup Rules on Minority Students and Their Schools

Schools needing improvement and schools meeting AYP differ in their performance levels, demographic characteristics, and the number of subgroup targets used to determine AYP. First, Figure 7 suggests that schools needing improvement were likely to fail to make AYP because of the performance of two subgroups—students with limited English proficiency and students with disabilities.[28] To make AYP, 13.6% of students in each subgroup must score at or above the state defined proficiency level. In schools needing improvement, the mean reading scores for these two groups of students fall below the 13.6-percentage point cutoff in reading for 2002-03. However, in schools meeting AYP, the average scores of these two subgroups exceed the 13.6-point target. This is the case for the other subgroups as well. For example, the 26-point mean proficiency rate for socio-economically disadvantaged (SD) students in schools meeting AYP exceeds the mean of 16-points in schools needing improvement. In other words, all subgroups of students perform better in schools meeting AYP, on average, than their counterparts in schools needing improvement.

Figure 7: Average Reading Proficiency Rates by Subgroups in Schools Needing Improvement and Schools Meeting AYP, California, Spring 2003 Administration.

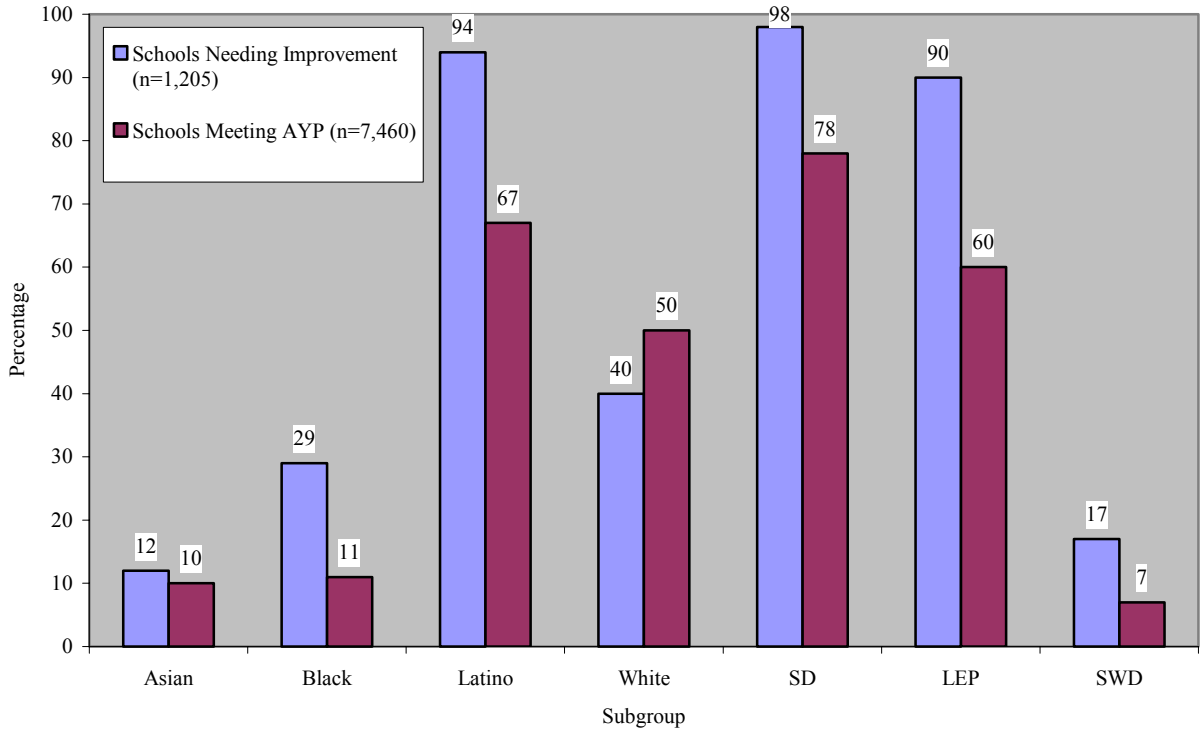


Note: These scores refer to subgroups, which meet the minimum population requirements in order to be considered valid. Sample sizes for schools needing improvement are as follows: Total, 1205; Black, 793; Asian, 526; Latino, 1169; White, 912; Socio-economically Disadvantaged, 1180; Limited English Proficient, 1156; Special Education, 1095. Sample sizes for schools meeting AYP are as follows: Total, 7441; Black, 1881; Asian, 1704; Latino, 3532; White, 3147; Socio-economically Disadvantaged, 3730; Limited English Proficient, 3241; Special Education, 3230. Source: See Appendix 1, 2, 4.

[28] See Appendix 5 for proficiency rates by subgroup for mathematics.

Second, schools identified as needing improvement are more likely than schools meeting AYP to be required to meet separate performance targets for disadvantaged subgroups. Figure 8 highlights demographic differences based on whether schools were identified as needing improvement. Close to 90% of schools needing improvement contain a Latino, socio-economically disadvantaged, or limited English proficiency subgroup. Note, however, that a smaller percentage of schools meeting AYP contain these three subgroups. Moreover, schools needing improvement were more likely to have a Black and special education subgroup and less likely to have a White subgroup than schools meeting AYP.

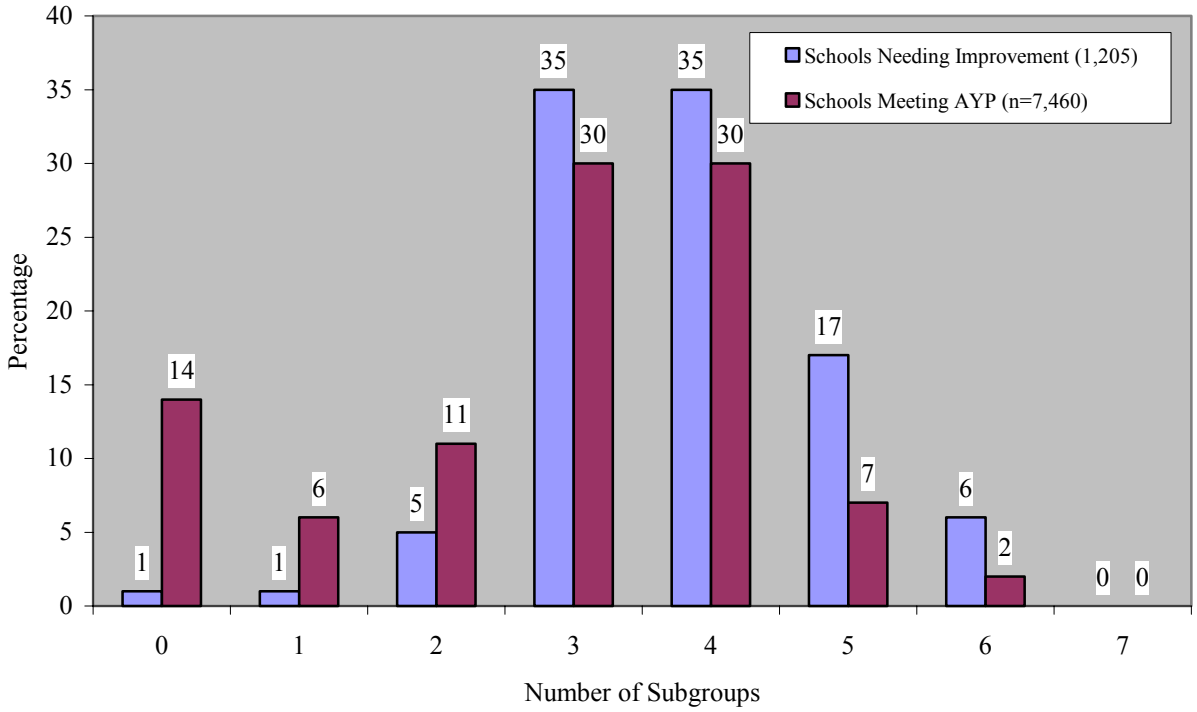
Figure 8: Percentage of Schools Needing Improvement and Schools Meeting AYP with Different Subgroups in Reading, California, Spring 2003 Administration.



Source: See Appendix 1, 2, 4.

Third, schools needing improvement had to meet more subgroup targets than schools meeting AYP. Figure 9 indicates that 93% of schools needing improvement contained three or more subgroups compared to 69% of schools meeting AYP. Conversely, 31% of schools meeting AYP had two or fewer subgroup targets compared to 7% of schools needing improvement. These figures clearly demonstrate that schools needing improvement had to meet more performance targets for different subgroups, thereby increasing their chances of missing a single target and failing to make AYP.

Figure 9: Percentage of Schools Needing Improvement and Schools Meeting AYP with a Particular Subgroup in Reading, California, Spring 2003 Administration.



Note: Fewer than 1% of schools had 7 subgroups.
Source: See Appendix 1, 2, 4.

Why, then, do schools needing improvement have more subgroups than schools meeting AYP? This difference arises because schools needing improvement enroll a large proportion of disadvantaged students, who fall into several subgroup categories, including minority, low-income, and limited English proficiency status. Since these subgroup categories are not mutually exclusive, a single student can count for multiple subgroups—for instance, a low-income, limited English proficient, Asian student counts in three subgroup calculations. Schools meeting AYP, however, enroll more White students, who are less likely than minority students to be counted for two or more subgroup categories. That is, White students are less likely than Latino and Black students to be socio-economically disadvantaged or lack proficiency in English.

We conducted additional analyses to explore the relationship among different subgroups and the likelihood that a particular subgroup attends a school with other subgroups. Table 4 shows that

Latino and Black students, and to a lesser extent Asian students, are more likely to attend a school with multiple subgroups than White students. The correlations between the White subgroup and the other subgroup categories are generally small, that is, near or below .10. This means that schools with a White subgroup are unlikely to have multiple subgroup targets. The situation for Latinos, however, contrasts sharply with that of Whites. The Latino subgroup is strongly correlated with the socio-economically disadvantaged subgroup (r = .73) and limited English proficient subgroup (r = .70), moderately correlated with a Black subgroup (r = .20), and very weakly correlated with a White subgroup (r = .08) and Asian subgroup (r = .02). Schools with a Black subgroup are somewhat likely to have a Latino (r = .20), socio-economically disadvantaged (r = .24), and limited English proficient subgroup (r = .19). Finally, schools with an Asian subgroup are also likely to have a limited English proficient subgroup (r = .22).

Table 4: Correlations Among Different Subgroups in California Schools, 2002-03.

	(1) Black	(2) Asian	(3) Latino	(4) White	(5) SD	(6) LEP	(7) SWD
(1) Black		0.05	0.20	-0.04	0.24	0.19	0.12
(2) Asian	0.05		0.02	0.10	0.04	0.22	0.05
(3) Latino	0.20	0.02		0.08	0.73	0.70	0.17
(4) White	-0.04	0.10	0.08		0.12	-0.07	0.11
(5) SD	0.24	0.04	0.73	0.12		0.64	0.15
(6) LEP	0.19	0.22	0.70	-0.07	0.64		0.15
(7) SWD	0.12	0.05	0.17	0.11	0.15	0.15	

Note: n=8646, most p-values are statistically significant at p<.01 level. Given the sample size, however, most all correlations are likely to be significant. As a result, it is more important to examine the size of the correlation coefficients.
Source: See Appendix 1, 2, 4.

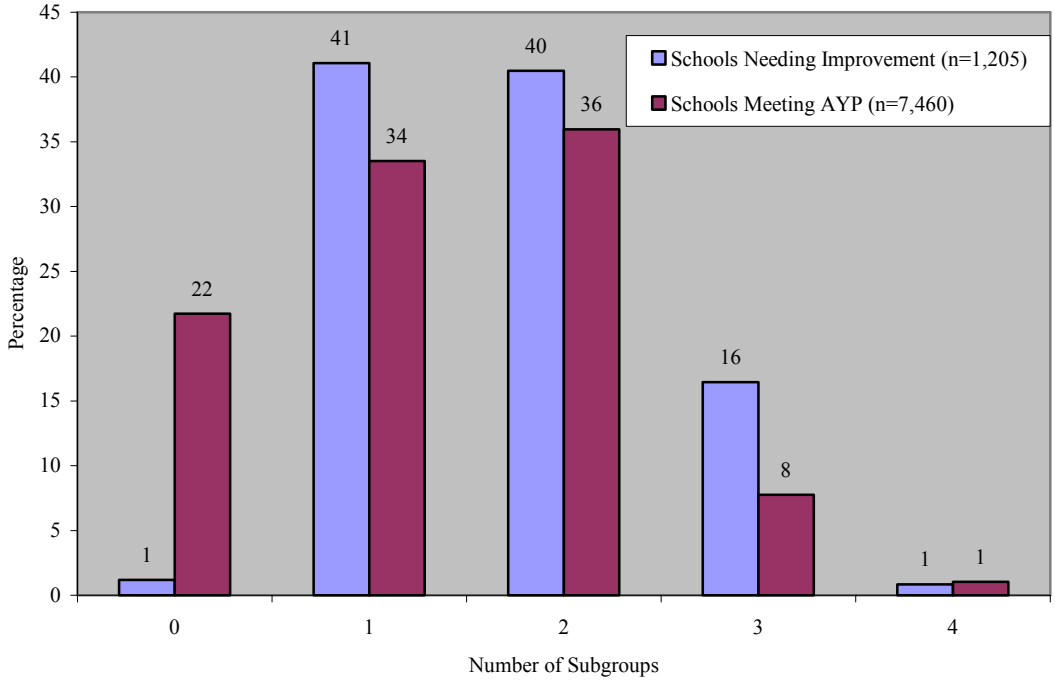
Focusing Subgroup Accountability on the Four Major Racial/Ethnic Subgroups

For school accountability, subgroups should be mutually exclusive categories so that a student's score counts only once in making AYP determinations.[29] One straightforward method for leveling the playing field between predominantly minority and White schools is to reduce the number of subgroups for determining AYP. For instance, if schools were required to meet subgroup targets only for racial and ethnic subgroups, such a policy change would mainly benefit minority students and their schools. Since there is a strong correlation between the minority subgroup and socio-economically disadvantaged and limited English proficient subgroups, excluding one of these categories for school accountability would reduce the number of subgroup requirements in schools with Latino, Black, and Asian students. Thus, schools with minority students would have to meet fewer performance targets, thereby reducing the chance of failure due to the imprecise nature of average test scores based on a small number of students.

[29] For diagnostic purposes, it is entirely reasonable and justifiable to disaggregate test scores for all of these different categories.

Figure 10 below shows that a similar percentage of both types of schools would have to meet one to two subgroup targets if California schools had to meet performance targets only by race and ethnicity (Asian, Black, Latino, White). For example, nearly 40% of schools needing improvement and about 35% of schools meeting AYP would have to meet either one or two subgroup targets in reading. Although an emphasis on the four major racial and ethnic subgroups levels the playing field for all schools, there are still nearly twice as many schools needing improvement (16%) than schools meeting AYP (8%) that would have to meet the AYP targets for three subgroups. Finally, nearly one-fifth of schools that meet AYP would have no subgroup targets since, on average, they have smaller enrollments than schools that do not meet AYP and since subgroups in small schools are unlikely to meet California's minimum group size criterion.[30]

Figure 10: Percentage of Schools Needing Improvement and Schools Meeting AYP with 0 to 4 Subgroups, Based on Separate Targets for the Major Racial and Ethnic Subgroups (Asian, Black, Latino, White) in Reading, California, Spring 2003 Administration.



Source: See Appendix 1, 2, 4.

Implications of Subgroup Accountability Rules

Although well-intentioned, NCLB's subgroup accountability policies have the unintended effect of unfairly and disproportionately sanctioning schools serving the most disadvantaged minority students. Our analysis of California's subgroup rules suggest that schools needing improvement enroll students who belong to multiple subgroups defined by minority, poverty, and limited

[30] In 2002-03, schools needing improvement had an average enrollment of 995 students compared to 654 students in schools meeting AYP.

English proficiency status. Since scores for disadvantaged minorities usually count for more than one subgroup, schools with large Latino and Black enrollments have to meet several achievement targets and testing participation requirements to make adequately yearly progress. Moreover, our results have policy implications not only for California, but also for states with multiracial schools and states with smaller minimum group size requirements than California. Subgroup accountability is likely to concentrate failure rates in states with multiracial schools such as Delaware (Frankenberg et al., 2003).[31] For example, over half of all public schools in Delaware failed to make AYP as compared to 15% of public schools in Wyoming, a state with racially homogenous schools (Robelen, 2003, September 3). It should also be noted that California's minimum group size requirement of 50 or more students is larger than the criterion set by most states. Among the five states in our report, only Virginia set a minimum group size of 50 while Arizona, Georgia, Illinois, and New York had smaller minimum group sizes (Erpenbach, Forte-Fast, & Potts, 2003).[32] Thus, in most states, fewer than 50 students will be needed to hold schools accountable for meeting separate subgroup performance targets. This means that a smaller minimum group size requirement will force a larger percentage of schools to meet separate subgroup targets. Therefore, the problems with subgroup accountability in California will be especially relevant in states with smaller minimum group size requirements.

Since disadvantaged and racially integrated schools are more likely to be identified for improvement under NCLB, those schools will also be disproportionately impacted by the mandated sanctions. It remains to be seen how these federal sanctions will affect schools and student achievement outcomes. Some analysts (Kane & Staiger, 2002a) speculate that the impact of sanctions on schools will be minimal, since "the consequences for failing schools (at least in the short term) —such as creating a school improvement plan or providing students with school choice—may not be very serious in practice" (p. 22). Evidence on the effect of different types of federal sanctions, such as school choice, on minority children's educational opportunities is mixed. On one hand, if children in low-performing schools have meaningful options, then NCLB's transfer policies may open up access to schools with academically strong peer groups and skilled teachers. On the other hand, if children are merely transferred from one low-performing school to another weak school, choice will do little to expand schooling options for minority students. In addition, the potential risks and benefits of more intrusive federal sanctions, ranging from corrective action to school restructuring, are equally unclear (Brady, 2003; Hamman & Schenck, 2002, October 19). Limited experience with these interventions suggests that results are highly uneven across districts and that success depends on an unusual mix of charismatic leadership from principals and teachers, sustained support from district administrators, and a long-term commitment to upgrade curriculum and instruction. In other words, these policies are extremely difficult to scale up across multiple schools, districts, and states, yet this is likely to occur in future years as more schools fail to meet federally mandated proficiency targets. As more and more schools are required to implement these interventions, local and state government may simply lack the personnel and expert knowledge to assist schools subject to federal sanctions. If "scientifically-based" research is to inform and shape educational

[31] Among the 50 states, Delaware ranked second nationally on a measure on racial integration. See Table 13, p. 47 (Frankenburg, Lee, Orfield, 2003).

[32] The minimum group size requirements for determining the percentage of proficient students is as follows: Arizona (n>= 30 & confidence interval), Georgia (n>=40), Illinois (n>=40, +/- 3%), New York (n>=40), and Virginia (n>=50).

policies, there is a clear need to scale back the implementation of many federally mandated sanctions until there is solid evidence about their effectiveness in improving school performance.

The need to speed up federal research is nowhere more apparent than in questions surrounding the assessment of students with disabilities and English-language learners. In the context of NCLB, schools in California and elsewhere are likely to fail to make AYP because of lower scores among students with disabilities and English-language learners. Yet, there is a serious mismatch between the ambitious goals of federal policy and the knowledge base on assessing students with disabilities and English-language learners. For example, the National Research Council (1997b) recently concluded that the "meaningful participation of students with disabilities in large-scale assessments and compliance with the legal rights of individuals with disabilities in some instances require steps that are beyond current knowledge and technology" (p. 193). In other words, the current demands of federal policy far exceed the capacity of research to provide guidance on using scores of disabled students for high-stakes accountability purposes. With very few exceptions, there is a serious shortage of research that explores the appropriateness of using different accommodations and modifications for assessing students with disabilities (Koretz & Hamilton, 2000; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). Despite recent federal regulations permitting districts and schools to count the scores from alternate assessments administered to severely disabled children (*Federal Register*, Vol. 68, No. 236, Section 200.13), this decision still ignores the broader need for the research community to supply guidance on policies for assessing students with disabilities.

In addition to addressing the research gap on testing special education students, the National Research Council (1997a) has underscored the need "to develop psychometrically sound and practical assessments and assessment procedures that incorporate English-language learners into district and state assessment systems" (p. 52). Given the urgent need to establish a more solid research base for assessing the academic skills of English-language learners, serious research needs to be undertaken before scores from these students' scores are included for high-stakes, accountability purposes.

Since the historic mission of Title I is to improve educational opportunities and outcomes for low-income and minority students, there has always been a national commitment to improving achievement outcomes for the major racial and ethnic minorities and socio-economically disadvantaged students. Therefore, these subgroup categories should be the focus of Title I's accountability policies. In addition, there are several technical reasons for reducing the number of subgroup categories used for school accountability. For one, classifications of race/ethnicity and poverty are also more reliable across the 50 states than classifications of NCLB's other subgroup categories. Whereas states often use different rules and policies for identifying students with disabilities and limited English skills, all schools must conform to federal regulations governing children's racial classifications and eligibility for free- and reduced lunch. If schools are required to meet separate subgroup scores based only on race/ethnicity and poverty status, such a policy change would also benefit disadvantaged schools and multiracial schools. It would do so by reducing the disparate impact on disadvantaged schools and racially integrated schools and by decreasing the number of schools required to implement federal sanctions. Such changes would give districts and states the opportunity to target their resources on a fewer number of schools and offer the federal government a chance to examine for whom and under

which conditions sanctions are likely to work. In many respects, NCLB's testing requirements and accountability policies should be scaled back until they are supported by research that meets the federal standards governing "scientifically-based" evidence.

(5) CONCLUSION AND RECOMMENDATIONS

Although NCLB requires all states to adopt a "single statewide accountability system," states layered the federal requirements onto pre-existing state systems. Because states kept their own accountability systems, schools had to meet state mandated goals as well as the AYP requirements for NCLB. The two systems often produced different performance labels for schools, creating confusion among educators, policymakers, and the public. Since proficiency rates vary dramatically across states, the climb toward 100% proficiency will be easier for some states than others. Over the next 12 years, California's schools, on average, will have to make larger test score gains than schools in Georgia and Virginia. Moreover, it is virtually impossible to interpret the meaning of proficiency across states since proficiency rates in schools identified as needing improvement vary by as much as 50-percentage points across the six states in our study.

Our analysis of the impact of the transitional AYP rules on Title I schools showed that minority, low-income, and limited English proficient students were disproportionately concentrated in schools needing improvement. The major difference, then, between schools needing improvement and all other schools was in the demographic composition of the student body. There were very small differences, however, in two-year gains in reading and math proficiency rates in schools needing improvement and in schools meeting AYP. In other words, both types of schools appear to contribute equally to student learning, even though schools needing improvement educate a larger percentage of disadvantaged students with lower test scores than schools meeting AYP.

The final analysis focused on the impact of subgroup accountability rules in California public schools. Our results suggest that subgroup rules unfairly disadvantaged schools with large minority enrollments by requiring them to hit multiple performance targets. Virtually all schools in California that were identified for improvement had to meet three or more subgroup targets. As currently constructed, subgroup policies concentrate failure in schools with disadvantaged minority students. By failing hundreds of schools, NCLB is likely to undercut the political consensus needed to reduce and eliminate the achievement gap.

We encourage federal policymakers to build on the consensus that led to the passage of the *No Child Left Behind Act* and revise portions of the law that have a disparate impact on minority students and the schools they attend.

- Instead of imposing a single model of test-based accountability on all public schools, states should be allowed to experiment with different models for measuring student learning and school performance. Specifically, achievement gains permit accurate assessments of how schools contribute to student achievement. Schools should also receive credit for improving student achievement along the entire distribution of test scores. That is, schools should receive credit for improving student achievement within different performance levels, not only improvements in the number of students who cross over the proficiency threshold.

- As suggested by one of the nation's leading testing experts (Linn, 2003a), accountability systems need to broaden their definitions of what counts as evidence of success and set goals that are realistically grounded in past experience. This might mean, for example, examining the largest average improvement in Title I schools and using that as a benchmark for developing performance targets. If policymakers set goals in light of actual performance trends, such expectations for improvement are more likely to be attainable with sufficient resources, effort, and time than current AYP expectations.

- NCLB should reinforce the historic commitment of Title I to improving the educational opportunities and outcomes for low-income and minority students. As such, school accountability should focus on narrowing the achievement gap between low-income and middle-income students, on one hand, and minority and White students on the other. The focus of this accountability should be broad and include students, educators, policymakers, administrators, researchers, and parents.

- Given the dearth of research on appropriate testing practices for students with disabilities and English-language learners, more information is needed before high-stakes decisions about school performance are based on the test scores of these two subgroups of students. Instead, we strongly support federal laws guaranteeing special education students' right to participate in state and federal assessment programs, such as the National Assessment of Educational Progress (NAEP). We also wholeheartedly support efforts to collect diagnostic information on the performance of students with limited English proficiency through NAEP and other state and local assessments. However, more systematic and careful study is needed before scores for these two subgroups are used for high-stakes accountability.

- The notion of "scientifically-based research" plays a prominent role in NCLB and is mentioned over 100 times in the federal statute. These standards should be applied to the accountability policies that lie at the heart of NCLB. They should also support ongoing research that examines the central assumptions of the law. For example, how do low-performing schools respond to the threat of increasingly intrusive federal sanctions, and under what conditions might these policies work best? How do we know if scores from students with disabilities and English-language learners provide reliable information and permit valid inferences about their academic achievement? Without better answers to these and other questions, it will be exceeding difficult for legislators to understand how best to minimize the disparate and adverse impact that accountability policies have on minority students and the schools they attend.

(6) APPENDIX

Appendix 1: Description of School-Level Data (Title I Information and School Demographics).

Table A.1 includes general descriptions of school-level data. Column 1 includes the unique school identifier assigned to each public school in the state. These school identifiers were used to merge data across different files. We concatenated the division and school code to create the Virginia school-identification number, since the state does not assign a unique ID to each school. Columns 2 and 3 list the relevant divisions and websites where we obtained Title I school identifiers (improvement status, years in improvement, schoolwide program, targeted assistance program) and school demographic characteristics (race/ethnicity, free lunch status, English learners, students with disabilities). Since this information is maintained in different types of electronic files (e.g., Access, Excel, Text Files), we converted this data into SPSS files using DBMS/COPY. We obtained missing information through personal contacts in each of the six departments of education.

Table A.1: Description of Title I Information and School Demographics in Six State Sample.

State/ School ID	**Title I Information**	**School Demographics**
Arizona ENTITY_ID	Academic Achievement Division (personal communication)	Academic Achievement Division (personal communication)
California CDS_CODE	Policy and Evaluation Division http://api.cde.ca.gov/datafiles.html	Educational Demographics Office http://data1.cde.ca.gov/dataquest/
Illinois RCDS	Data Analysis and Progress Reporting http://www.isbe.net/research/reports.htm#Statistics	Data Analysis and Progress Reporting http://www.isbe.net/research/reports.htm#Statistics
New York BEDS_CD	Information and Reporting Services (personal communication)	Information and Reporting Services http://www.emsc.nysed.gov/repcrd2003/database/guide.html
Georgia KEY	Policy Division-Title I Programs http://www.doe.k12.ga.us/support/plan/nclb.asp	Administrative Technology http://techservices.doe.k12.ga.us/reportcard/default.htm
Virginia DIV_SCH	Office of Information Technology (personal communication)	Office of Information Technology http://www.pen.k12.va.us/VDOE/Publications/rep_page.htm

Appendix 2: Description of School-Level Data (Achievement Outcomes).

We obtained achievement outcomes for all public schools in each of the six states through contacts in the department of education. Arizona maintains both AIMS and Stanford 9 scores for both regular public and charter schools in downloadable Excel files. Illinois, New York, Georgia, and Virginia maintain zip files that include comprehensive school-level test score trends since the late 1990s. Illinois and Virginia provide school-level averages in reading and math for all tested grades. Illinois and New York also disaggregate scores by race/ethnicity, economic disadvantage, special education status, and English learner status.

California keeps STAR results in Access files. We merged the STAR results with the 2002 and 2003 AYP Phase I Datafile (DBF format), which included information on enrollment counts, participation rates, and proficiency rates by 10 subgroup categories (i.e., African American, American Indian, Asian, Filipino, Hispanic, Pacific Islander, White, Socio-economically Disadvantaged, English Learner, Students with Disabilities). Several state administrators in the California Department of Education answered questions regarding subgroup analyses (A. Just, R. Bernstein, personal communication with J. Kim, October 6, 2003). We used two key main variables to conduct the subgroup analyses. First, in determining whether a subgroup counts for the purpose of the 95% testing requirement, the enrollment variable had to meet one of the group size requirements (100 or more or 50 and at least 15% of total enrollment). Each of these variables was coded "ee"—for instance, "ee_aa" corresponded to enrollment counts for African-American students. Second, in determining whether a subgroup counted for the performance requirement, the valid scores for subgroup had to meet the minimum group size requirement. Each of these variables was coded "ev"—for instance, "ev_aa" corresponded to the number of African-American students with valid test scores. Finally, we conducted two separate re-analyses to verify the results in section two and three.

Table A.2: Description of Achievement Outcomes in Six State Sample.

State/School ID	Achievement Outcomes
Arizona	Research and Policy
ENTITY_ID	http://www.ade.az.gov/standards/aims/Results/Default.asp
California	Standards and Assessment Division
CDS_CODE	http://star.cde.ca.gov/star2002/help/ResearchMDB.asp
Illinois	Data Analysis and Progress Reporting
RCDS	http://www.isbe.net/research/reports.htm#Statistics
New York	Information and Reporting Services
BEDS_CD	http://www.emsc.nysed.gov/repcrd2003/database/guide.html
Georgia	Administrative Technology
KEY	http://techservices.doe.k12.ga.us/reportcard/default.htm
Virginia	Virginia Report Card
DIV_SCH	http://www.pen.k12.va.us/VDOE/Assessment/2002SOLpassrates.html

Appendix 3: Transitional Adequate Yearly Progress Rules Used to Identify Title I Schools Needing Improvement in Six States for the 2002-03 School Year.

Arizona

Arizona used a subset of items on the Stanford 9 to determine AYP as required by the 1994 IASA. Although the state administered the Stanford 9 in consecutive grades (3-12), only items from the grade 3, 8 and 12 assessment in reading comprehension and math were used for AYP purposes. Moreover, to ensure that the Stanford 9 items were aligned with Arizona Academic Standards, the state conducted a correlation study and "deleted items from the Stanford 9 that did not correlate with Arizona Academic Standards" (Arizona Department of Education, 1999) (p. 8). Theses items formed the AYP transitional assessment until 2000-01. Cut scores were then applied to the tests to create four performance levels: advanced, proficient, basic, and below basic (Arizona Department of Education, 1999).

Under the transitional plan, the state also set two achievement goals for 2005, and defined a procedure for defining AYP and identifying local districts and schools that failed to meet these targets. Specifically, the state used a gap reduction model (GRM) to assess the amount of yearly improvement Title I schools had to make in reaching two goals: (1) 90% of students will score at proficient or above in reading comprehension and math on the SAT9 extracted items; and, (2) No students will score below basic on either reading comprehension or math from the SAT9 extracted items.

A school meets AYP if either goal 1 or 2 is met. The annual targets are set using the GRM, which describes the progress rate from a baseline (current school performance) to a goal value. Moreover, districts are responsible for identifying Title I schools that fail to make AYP, using the procedure outlined below: AYP determinations will use the spring 1997 scores from the Stanford 9 in reading comprehension and mathematics as a baseline. The achievement of 90% of students scoring proficient or better by year 2005 is goal 1. The difference between the two, divided by the time span of 8 years, will determine how much a particular school must improve per year; the greater the gap between the current performance and the goal, the greater the expected increase per year. The required annual improvement rate is graphed as a straight line from current performance 1997 to Goal (2005) (Arizona Department of Education, 1999) (p. 3).

California

Prior to NCLB, California used the Academic Performance Index (API) score to determine AYP. According to the California Department of Education (2003a), "California's previous definition of AYP used student achievement data from the Standardized Testing and Reporting (STAR) program to calculate API to (1) determine the AYP of all schools funded with federal Title I funds, and (2) exit existing eligible schools or identify new schools for Program Improvement (PI)" (p. 3).

Illinois

According to the Illinois State Board of Education, AYP before NCLB was defined as the following:

> Schools that had composite Meeting/Exceeding standards assessment scores of less than 50 percent for two consecutive years were placed in Academic Early Warning Status (AEWS). Title I schools were also placed in School Improvement Status under the 1994 ESEA federal legislation. In order for a school to make AYP the school had to have a composite Meeting/Exceeding standards assessment score of 50 percent within five years. Therefore, each year 20 percent of the gap had to be narrowed, between the 50 percent composite score and the school composite when originally placed on the AEWS or in School Improvement. For example, school A was placed on the AEWS in 1998 with aggregate test scores of 30 percent Meeting/Exceeding standards. The school must achieve 50 percent within five years, closing the gap (50% minus 30% = 20% gap). Thus the AYP target for school A is to achieve 4 percent gains per year. For 1999, the target would be 34 percent Meeting/Exceeding; for 2000, 38 percent; for 2001, 42 percent, etc. Under this method, each school had its own AYP targets (http://www.isbe.state.il.us/ayp/faq.htm).

New York

New York administers the Regents examination in grades 4 and 8 and includes four performance levels, including basic (Level 1), basic proficiency (Level 2), proficiency (Level 3), and advanced (Level 4). Based on these performance standards, New York developed a school performance index (PI) that credits schools for moving students from Level 1 to Level 2, although Level 3 is defined as the proficient level of performance that all schools must meet within 12 years. Since 1999, the Commissioner of Education has also used the performance index to set AYP goals. For the grade 4 and 8 tests, these targets are set in three-year intervals during which time a school must close the gap between its performance and the state standard by 15% each year.

Georgia

With respect to school accountability, the Title I AYP measure required schools to reduce the proportion of students at the "does not meet standard" level on the CRCT by 5-percentage points. This AYP definition, however, was not based on the school grading system, and it did not focus on increasing the percentage of students at the proficient level, as required by NCLB. In choosing a 5% move out rate from the basic level, Superintendent Schrenko and staff in the testing and evaluation division expected teachers to move at least one student out of basic level each year. Thus, schools with a large fraction of students meeting state standards could be identified for failing to make AYP if they did not reduce the proportion of students in the "did not meet standard" category by at least 5-percentage points. In 2002, when Georgia had to identify Title I schools that failed to make AYP, the CRCT-based accountability system had been in place for only two years. Using this accountability system, 437 Georgia schools were identified for failing to make AYP and were required to offer choice and/or supplemental services, forcing the SDE to provide support in nearly 20% of all Georgia public schools.

Virginia

By creating four performance levels to categorize schools, the Board could focus resources on schools in the lowest performance category, "accredited with warning"—that is, schools with pass rates generally below 50% on all four SOL subjects. In complying with the 1994 IASA's requirement for states to define "adequate yearly progress" for Title I schools, Virginia defined schools as failing to make AYP if they were labeled "accredited with warning." Title I schools failed to meet AYP if they remained in the "accredited with warning" category in English and/or math for two consecutive years (Consortium for Policy Research in Education, 2000a). By 2002, when the stricter NCLB requirements took effect, Virginia had identified a relatively small number of schools (34) for improvement. Most schools had shown improvement in SOL scores since 1998, leading to large increases in the number of schools meeting "provisional accreditation" benchmarks and large reductions in the number of schools "accredited with warning" (Virginia Board of Education, 2000).

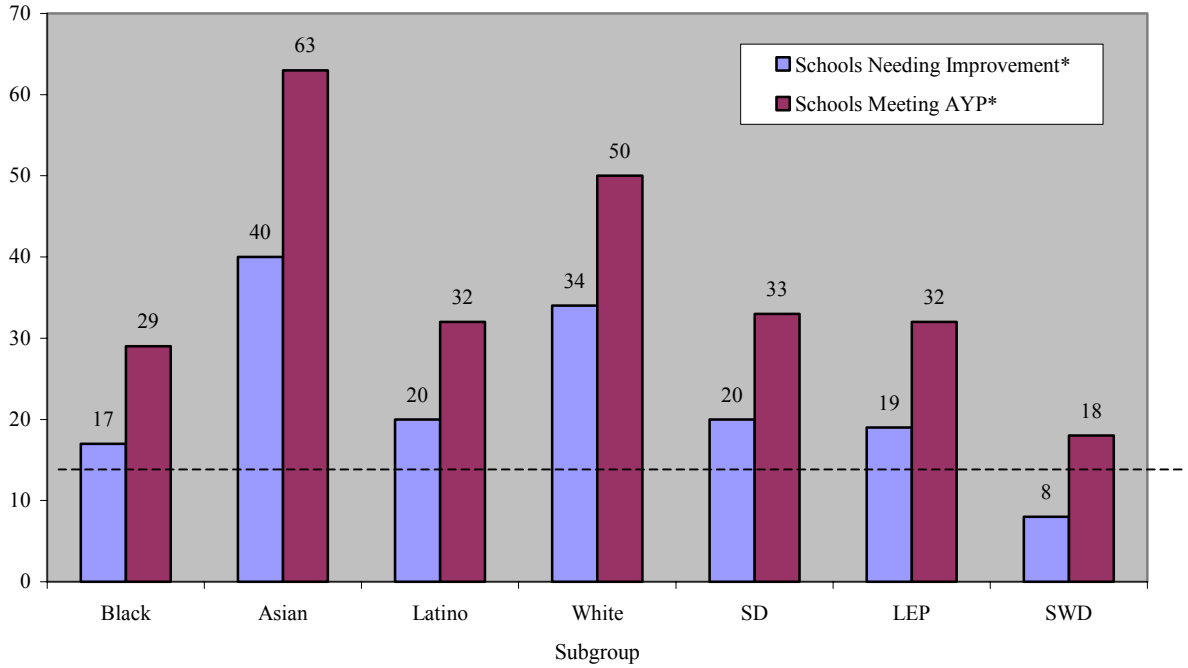
Appendix 4: Lists of Schools Identified as Needing Improvement for 2002-03 and 2003-04.

Table A.4: Number of Title I Schools Needing Improvement for 2002-03 and 2003-04.

State	2002-03 (Last Updated, June 2003)	2003-04 (Last Updated, December 9, 2003)
Arizona	399	244
(Source)	Personal Communication-Title I	http://www.ade.az.gov/profile/publicview/aypschoollist.asp.
California	815	1,205
(Source)	http://www.cde.ca.gov/iasa/titleone/pi/	http://www.cde.ca.gov/ayp/2003/titleone/titleI_layout.htm
Illinois	527	581
(Source)	http://www.isbe.net/research/reports.htm#Statistics	http://www.isbe.net/research/pdfs/2003_StateReport_E.pdf
New York	434	517
(Source)	http://www.emsc.nysed.gov/deputy/nclb/nclbhome.htm	http://www.emsc.nysed.gov/deputy/nclb/nclbhome.htm
Georgia	436	846
(Source)	http://www.doe.k12.ga.us/support/plan/nclb.asp	http://www.doe.k12.ga.us/support/plan/nclb.asp
Virginia	34	43
(Source)	http://www.pen.k12.va.us/VDOE/src/vasrc-title1.pdf	http://www.pen.k12.va.us/VDOE/src/vasrc-title1.pdf

Appendix 5: Average Math Proficiency Rates by Subgroups in Schools Needing Improvement and Schools Meeting AYP, California, Spring 2003 Administration.

Figure A.5: Average Math Proficiency Rates by Subgroups in Schools Needing Improvement and Schools Meeting AYP, California, Spring 2003 Administration.



**Note*: These scores refer to subgroups, which meet the minimum population requirements in order to be considered valid. Sample sizes for schools needing improvement are as follows: Total, 1205; Black, 793; Asian, 526; Latino, 1169; White, 912; Socio-economically Disadvantaged, 1180; Limited English Proficient, 1156; Special Education, 1095. Sample sizes for schools meeting AYP are as follows: Total, 7441; Black, 1881; Asian, 1704; Latino, 3532; White, 3147; Socio-economically Disadvantaged, 3730; Limited English Proficient, 3241; Special Education, 3230.
Source: See Appendix 1, 2, 4.

REFERENCES

Arizona Department of Education. (1999). *Arizona's process and progress toward meeting Arizona's high content and performance standards*. Phoenix, AZ: Arizona Department of Education.

Baron, J. B. (1999). *Exploring high and improving reading achievement in Connecticut*. Washington, D.C.: National Education Goals Panel.

Brady, R. C. (2003). *Can failing schools be fixed?* Washington, D.C.: Thomas B. Fordham Foundation.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998). *Academic productivity of Chicago public elementary schools*. Chicago, IL: The Consortium on Chicago School Research.

California Department of Education. (2003a). *Adequate yearly progress for 2002 base*. Sacramento, CA: California Department of Education, Policy and Evaluation Division.

California Department of Education. (2003b). *State of California consolidated state application accountability workbook*. Sacramento, CA: Author.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? *Education Evaluation and Policy Analysis, 24*, 305-332.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher state test scores in Texas make for better high school outcomes? CPRE Research Report No. RR-047*. Philadelphia, PA: Consortium for Policy Research in Education.

Consortium for Policy Research in Education. (2000a). *Assessment and accountability in the fifty states: 1999-2000 (Virginia)*. Philadelphia, PA: CPRE.

Dee, T. S. (2003). Learning to earn. *Education Next, 3*, 64-70.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo, J. (1999). *NAEP 1998 reading report card for the nation and the states*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Education Week. (2002). *The changing definition of proficient.* Retrieved [October 22, 2002] from http:www.edweek.org/ew/ew_printstory.cfm?slug=06tests-s2.h22.

Education Week. (2003). *Quality Counts*. Bethesda, MD: Education Week.

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Accountability systems and reporting*. Washington, D.C.: Council of Chief State School Officers.

Frankenberg, E., Lee, C., & Orfield, G. (2003). *A multiracial society with segregated schools: Are we losing the dream?* Cambridge, MA: The Civil Rights Project, Harvard University.

Georgia Board of Education. (2003). *Consolidated state application accountability workbook*. Atlanta, GA: Author.

Grissmer, D., & Flanagan, A. (2001). Searching for indirect evidence for the effects of statewide reforms. In D. Ravitch (Ed.), *Brookings Papers on Education Policy 2001* (pp. 181-230). Washington, D.C.: Brookings Institution Press.

Grissmer, D., Flanagan, A., & Williamson, S. (1998). Why did the black-white score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 182-228). Washington, D.C.: Brookings Institution Press.

Hambleton, R. K. (1998). Setting performance standards on achievement tests. In L. N. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I, Chapter 10*. Washington, D.C.: Council of Chief State School Officers and U. S. Department of Education.

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In S. P. Klein (Ed.), *Making sense of test-based accountability in education* (pp. 13-49). Santa Monica, CA: RAND.

Hamman, D., & Schenck, E. A. (2002, October 19). Corrective action and school choice in NYC: An analysis of district funding applications. *Education Policy Analysis Archives, 10*(45). Retrieved [October 9, 2003] from http:epaa.asu.edu/epaa/v10n45.html).

Haney, W. (2000). *Report for testimony in GI forum v. Texas Education Agency*. Boston, MA: Boston College, School of Education.

Helderman, R. S. (2003, September 12). Schools in VA fail federal standards, no area district satisfies new law. *Washington Post,* p. B1.

Kane, T. J., & Staiger, D. O. (2002a, June 10-11). *Racial subgroup rules in school accountability systems.* Paper presented at the Taking account of accountability: Assessing politics and policy, Kennedy School of Government, Harvard University.

Kane, T. J., & Staiger, D. O. (2002b). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on educational policy, 2002* (pp. 235-283). Washington, D.C.: Brookings Institution Press.

Keegan, L. G. (1999). The empowerment of market-based school reform. In A. Gresham (Ed.), *School choice in the real word, lessons from Arizona charter schools* (pp. 189-197). Boulder, CO: Westview Press.

Kim, J., & Sunderman, G. (2004). *The implementation of school choice under the No Child Left Behind Act: Educational options under the Title I transfer policy*. Cambridge, MA: The Civil Rights Project at Harvard University.

Kirst, M. W. (2002). Swing state, the downs and ups of accountability in California. *Education Next, 2*, 44-49.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND (Issue Paper IP-202).

Koretz, D., & Hamilton, L. S. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Education Evaluation and Policy Analysis, 22*, 255-272.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalizations across tests.* Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Ladd, H. F. (Ed.). (1996). *Holding schools accountable*. Washington, D.C.: Brookings Institution Press.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*, 3-12.

Linn, R. L. (2003a). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*, 3-13.

Linn, R. L. (2003b). *Requirements for measuring adequate yearly progress*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L. (2003, September 1). *Performance standards: Utility for different uses of assessments*, Education Policy Analysis Archives, 11, Retrieved [September 12, 2003] from http://epaa.asu.edu/epaa/v11no31/.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*, 3-16.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Education Evaluation and Policy Analysis, 24*, 29-36.

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools, the role of incentives* (pp. 197-224). Washington, D.C.: National Academy Press.

Miller, L. S. (1995). *An American dilemma*. New Haven, CT: Yale University Press.

National Research Council. (1997a). *Educating language-minority children*. Washington, D.C.: National Academy Press.

National Research Council. (1997b). *Educating one and all: Students with disabilities and standards-based reform*. Washington, D.C.: National Academy Press.

New York State Department of Education. (2003, January 6). *Accountability peer review, New York state*. Albany, NY: New York State Department of Education, The University of the State of New York.

Orfield, G., & Kornhaber, M. L. (Eds.). (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston College: The National Board on Educational Testing and Public Policy.

Puma, M. J., Karweit, N., Price, C., Ricciuti, A. E., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes*. Bethesda, MD: Abt Associates.

Ravitch, D. (2002). Introduction. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 1-12). Washington, D.C.: Brookings Institution Press.

Raymond, M., & Hanushek, E. A. (2003). High-stakes research. *Education Next, Summer 2003*, 48-55.

Reardon, S. (1996). *Eighth-grade minimum competency testing and early high school dropout patterns.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states.* Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

Robelen, E. W. (2003, September 3). *State reports on progress vary widely*. Retrieved [November 11, 2003, 2003] from Education Week, http://www.edweek.org/ew/ewstory.cfm?slug=01ayp.h23&keywords=delaware%20and%20race.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247-256.

Sunderman, G., & Kim, J. (2004). *Expansion of federal power in American education: Federal-state relationships under the No Child Left Behind Act, year one.* Cambridge, MA: The Civil Rights Project at Harvard University.

Swanson, C. B. (2003). *NCLB Implementation Report: State approaches for calculating high school graduation rates*. Washington, D.C.: Urban Institute, Education Policy Center.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating studnets with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64*, 439-450.

Treisman, P. U., & Fuller, E. J. (2001). Comments by Philip Uri Treisman and Edward J. Fuller. In D. Ravitch (Ed.), *Brookings papers on education policy 2001* (pp. 208-229). Washington, D.C.: Brookings Institution Press.

Virginia Board of Education. (2000). *Regulations establishing standards for accrediting public schools in Virginia*. Richmond, VA: Virginia Board of Education.

Virginia Department of Education. (2003). *State of Virginia consolidated state application accountability workbook.* Richmond: Author.

Wilson, S. M., Darling-Hammond, L., & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning*. Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.