Measuring Academic Proficiency Under the No Child Left Behind Act: Implications for Educational Equity

by James S. Kim and Gail L. Sunderman

The accountability requirements of the No Child Left Behind Act of 2001 place high-poverty schools and racially diverse schools at a disadvantage because they rely on mean proficiency scores and require all subgroups to meet the same goals for accountability. In this article, student achievement data from six states are used to highlight differences in the demographic characteristics of schools identified as needing improvement and schools meeting the federal adequate yearly progress requirements. School-level data from Virginia and California are used to illustrate that these differences arise both from the selection bias inherent in using mean proficiency scores and from rules that require students in racially diverse schools to meet multiple performance targets. The authors suggest alternatives for the design of accountability systems that include using multiple measures of student achievement, factoring in student improvement on achievement tests in reading and mathematics, and incorporating state accountability ratings of school performance.

The achievement gap on standardized tests increasingly is viewed as the most significant educational challenge facing American society in the 21st century. Currently, the only national policy that aims specifically to narrow racial disparities in academic performance is the Elementary and Secondary Education Act (ESEA) of 1965, reauthorized in 2001 as the No Child Left Behind Act (NCLB). The accountability provisions of this federal law now govern Title I schools and non–Title I schools in the 50 states, Puerto Rico, and the District of Columbia. These provisions are intended to close "the achievement gap between high- and low-performing children, especially the achievement gaps between minority and non-minority students, and between disadvantaged children and their more advantaged peers" (NCLB, 2001, Sec. 1001 [3]). By embracing the goals of NCLB, both Congress and the president also agreed, at least implicitly, that the adequate yearly progress (AYP) requirements would be the central mechanism for improving school performance and the academic achievement of different subgroups of students.

Educational Researcher, Vol. 34, No. 8, pp. 3–13

The purpose of AYP is to ensure that "all schools" and "all students" meet the same academic standards in reading and mathematics by the 2013–2014 academic year. The NCLB legislation requires all schools to meet an absolute level of performance in reading and mathematics that is uniformly applied to all subgroups of students within a school. The law describes performance in terms of "annual measurable objectives" (AMO), indicating the minimum percentage of students who must meet the proficiency level of performance on reading and mathematics assessments, and defines subgroups as economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency. Moreover, NCLB requires 95% of students overall and 95% of each subgroup of students within a school to take the standardized reading and mathematics tests. Each state establishes its own AMO targets and a minimum group size—that is, the minimum percentage of students in each subgroup who are required to meet or exceed the AMO targets (NCLB, 2001 Sec. 1111 [b][2][G][iii]). Both the AMO targets and the minimum group size criterion vary widely across states (Erpenbach, Forte-Fast, & Potts, 2003). A school can fail to make AYP if a single subgroup of students does not meet the performance target in reading or mathematics or fails to meet the 95% participation requirement. Schools that fail to make AYP for two or more consecutive years are identified as "in need of improvement" and are subject to a series of sanctions that increase in number and severity the longer a school remains in improvement status.[1] The AYP requirements apply to both Title I and non-Title I schools, although only schools that receive federal funds are subject to the mandatory sanctions. To reduce failure rates in schools with multiple subgroup targets, NCLB contains a "safe harbor" provision that allows a school to make AYP if the percentage of students below proficiency is reduced by 10% and if the school demonstrates improvement on an additional indicator of performance, such as graduation or attendance rates. However, some research suggests that the safe harbor provision does little to reduce the number of schools identified as needing improvement (Lee, 2004).

The AYP requirements under the No Child Left Behind Act of 2001 are more stringent than those under the Improving America's Schools Act of 1994 (IASA), the predecessor to NCLB (Abedi, 2004). IASA allowed states to use a compensatory model of accountability where high scores in one subject area compensated for low scores in another subject. The low performance of a

single subgroup of students did not cause an entire school to fail making AYP. Nor did IASA impose a firm deadline for ensuring that 100% of students met the proficiency level in reading and mathematics. In contrast, NCLB requires policymakers to employ a "conjunctive accountability model," which requires each subgroup of students to meet the same minimum proficiency levels on both reading and mathematics assessments regardless of their previous proficiency levels. Ultimately, all students must reach 100% proficiency within 12 years.

Meeting AYP: Challenges Facing High-Poverty and Racially Diverse Schools

The requirements for meeting AYP pose the greatest challenges to high-poverty schools, which enroll a large percentage of students who have traditionally scored poorly on standardized achievement tests. Given the strong correlation between race and poverty (Orfield, 1996; Orfield & Lee, 2005), many high-poverty schools also enroll large concentrations of Black and/or Latino students whose average test scores are likely to fall below the minimum proficiency level required to meet AYP. Using school-level data from six geographically diverse states with different school accountability policies, we examine how the AYP requirements affect high-poverty schools with significant Black and Latino enrollments. Our state sample includes Arizona, California, Georgia, Illinois, New York, and Virginia. We focus our analysis on race/ethnicity and low-income status because the definitions for these two subgroups are well established by the federal government and are generally more consistent across states than definitions for classifying students with limited English proficiency and students with disabilities. For example, low-income status is usually determined by whether a student is eligible for free or reduced-price lunch under federal guidelines (U.S. Department of Education, 2002). Definitions for the other NCLB subgroups, including students with limited English proficiency and students with disabilities, vary widely across states because each state determines its own criteria for classifying students in these two categories (Abedi, 2004; Koretz & Hamilton, 2000; National Research Council, 1997a, 1997b). Our analyses also focus prominently on school poverty levels and student achievement levels because of the historic objective of Title I to support educational opportunities for low-income students (Borman, Stringfield, & Slavin, 2001; Murphy, 1971).

This article is organized into three sections. First, we provide a review of recent studies exploring the effect of using mean proficiency levels and subgroup accountability policies on schools serving predominantly minority and low-income students. In the context of this discussion, we compare the racial composition of schools meeting AYP and schools identified as needing improvement in each of our six states. These analyses show how high-poverty schools fare under an accountability system based largely on mean proficiency levels.

Second, we compare improvements in reading and mathematics proficiency levels by school AYP status (i.e., schools identified for improvement as opposed to schools meeting AYP) and school poverty levels. To conduct this analysis, we draw on evidence from Virginia, which has administered the same reading and mathematics tests since 1998. Performance data from the same standardized achievement tests can show how much rates of proficiency in reading and mathematics improved over 5 years, from 1998 to 2003, in both high- and low-poverty schools. These analyses reveal how much performance levels improve over time among schools that start off at different proficiency levels.

Third, we examine how the NCLB subgroup accountability policy affects the likelihood of a school's failing to meet AYP and subsequently being identified as needing improvement. Our analysis of subgroup accountability focuses on California because it has the largest number of schools in the nation with multiple subgroup targets and incorporates subgroup accountability into the state formula for determining each school's Academic Performance Index (API) score. Unlike the NCLB subgroup accountability policy, California's policy does not employ a conjunctive accountability model where the failure of a single subgroup causes an entire school to fail. Because there are key differences in the design of federal and state subgroup accountability policies, we examine the performance of high-poverty schools under NCLB and California's school accountability policy. While some of our findings were predicted by simulations involving national data sets and by analyses of state-level data before NCLB (Kane & Staiger, 2003; Linn & Haug, 2002; Raudenbush, 2004), our results provide an assessment of how AYP policies affect the very schools and students that are the intended beneficiaries of Title I.

Using Mean Proficiency and Subgroup Performance for School Accountability

The NCLB accountability system requires all schools and students to meet a single mean proficiency level in reading and mathematics. Proponents of NCLB assert that high expectations for achievement are needed to address the learning needs of public school students who are "segregated by low expectations" (U.S. Department of Education, 2001). Accordingly, by applying uniform annual measurable objectives in reading and mathematics to all students, the adequate yearly progress requirements are intended to create strong incentives for schools to improve the achievement of underperforming students. Although the broad goals of the federal law command widespread support among state and local policymakers, federal legislators, educators, and parents (Rose & Gallup, 2003; Rudalevige, 2003), the disparate impact of AYP on high-poverty schools has generated controversy over the NCLB accountability requirements (Chubb, Linn, Haycock, & Wiener, 2005; Finn & Hess, 2004; Sunderman, Kim, & Orfield, 2005).

The use of mean proficiency levels and subgroup rules in federal accountability policy has prompted many researchers to challenge the validity of AYP as a measure of school effectiveness. In particular, mean differences in test score outcomes are often inaccurate indicators of school effectiveness because they fail to account for selection biases. A substantial body of education research (Bryk & Raudenbush, 1988; Linn, 2000; Mosteller & Moynihan, 1972; Raudenbush, 2004) has consistently shown that differences in mean test score levels usually reflect differences in children's cognitive skills and background characteristics before they enter school. Thus inferences about school quality must account for the initial differences in student background. An alternative to mean proficiency is a value-added indicator (Meyer, 1996) that attempts to address differences in student background characteristics by isolating the contribution of schools to student learning.

To show how high-poverty schools perform under a system based on mean proficiency and value-added measures, Raudenbush (2004) analyzed student achievement scores from three national data sets and the Washington, DC, public schools. Spanning the elementary, middle, and high school grades, these analyses highlighted large differences in the cognitive skills and background characteristics of students enrolled in high-poverty and low-poverty schools. In Raudenbush's simulations, accountability systems based on mean proficiency levels were systematically biased against high-poverty schools, and the "tendency of mean proficiency to disproportionately target high-poverty schools as failing appear[ed] to result primarily from selection bias" (Raudenbush, 2004, p. 26). Despite these initial differences in mean test score levels, high-poverty schools generated annual learning gains that were similar, on average, to those of low-poverty schools. Nonetheless, an accountability system based on mean proficiency measures would treat many high-poverty schools as failing, even though their students' achievement test scores improve at a rate equal to that in low-poverty schools.

Under NCLB, mean proficiency levels are also used to determine the academic progress of different subgroups of students within a given school. Research by Kane and Staiger (2002, 2003) suggests that subgroup accountability policies have ambiguous benefits and clear costs for minority students and their schools. They found little evidence that subgroup accountability rules in California and Texas improved minority student achievement. Moreover, subgroup policies in both states produced high failure rates in schools with a large percentage of Black and Latino students. Kane and Staiger's research has two direct implications for AYP policy. First, because Black and Latino students have lower average test scores than White and Asian students, schools with either a Black or a Latino subgroup have a higher probability of failing to meet the AYP requirements. Second, because Black and Latino students often belong to other subgroup categories defined by NCLB, including subgroups for economically disadvantaged students and limited English proficient students, schools with a minority subgroup will have to meet multiple targets, which further increases the chances of failing to make AYP.

Prior research predicts higher AYP failure rates in high-poverty schools, which have low mean proficiency levels, and in schools with large minority enrollments, which are often responsible for meeting multiple subgroup targets. In our sample of six states, there is a clear demographic divide between schools identified as needing improvement and schools meeting AYP. Table 1 compares the racial and ethnic composition of schools based on their AYP status during the 2003–2004 school year. The final column in Table 1 shows the total number of students in schools classified as "in need of improvement" and schools "meeting AYP" in Arizona, California, Georgia, Illinois, New York, and Virginia. In each of the six states, schools needing improvement have majority Black and Latino student enrollments, while schools meeting AYP have predominantly White and Asian student enrollments. More specifically, Black and Latino students constitute more than 75% of all students in schools needing improvement in five states (Arizona, California, Illinois, New York, and Virginia) and 61% of all students in schools needing improvement in Georgia.

When we examine the racial and ethnic composition of schools in different years of program improvement, we find that a disproportionate number of minority students attend Title I schools that are required to implement federal sanctions. Our analysis focuses on the two states in our study—California and New York—that have the largest total number of minority students enrolled in schools identified by NCLB as needing improvement. In particular, Table 2 displays the percentages of Black and Latino students in all schools identified for improvement in California and

Table 1
Racial and Ethnic Composition of SINI and AYP Schools in Six States, 2003–2004

State	Status	Black (%)	Latino (%)	Asian American (%)	White (%)	Total (*n*)
Arizona	SINI	7	70	1	22	123,853
	AYP	5	32	2	61	737,535
California	SINI	12	68	5	15	1,154,633
	AYP	8	42	9	41	4,616,208
Georgia	SINI	54	7	2	38	316,754
	AYP	34	5	3	58	1,096,415
Illinois	SINI	59	37	1	3	271,408
	AYP	12	13	4	71	1,651,377
New York	SINI	39	42	5	14	486,770
	AYP	16	14	6	64	2,329,569
Virginia	SINI	87	2	0	11	20,570
	AYP	26	6	5	63	1,143,097

Note. SINI = schools in need of improvement, AYP = schools meeting adequate yearly progress requirements. Data on school performance levels are from websites of the departments of education of Arizona, California, Georgia, Illinois, New York, and Virginia and are for public use. Calculations are our own.

Table 2
Percentages of Black and Latino Students Enrolled in Schools by Their Year of Program Improvement in California and New York, 2003–2004

Year of Program Improvement (2003–2004)	California		New York	
	Total (*n*)	Black and Latino (%)	Total (*n*)	Black and Latino (%)
Year 1 (Choice)	394,141	74	126,768	73
Year 2 (Supplemental Educational Services)	150,209	67	57,343	78
Year 3 (Corrective Action) and Year 4 (Restructuring)	380,802	87	188,179	87

Note. Data on program improvement status are from websites of the departments of education of California and New York and are for public use. Calculations are our own.

New York. In each state, Black and Latino students represent 87% of all students attending schools that are in the 3rd or 4th year of improvement, which are subject to corrective action or school restructuring. The figures from Tables 1 and 2 indicate that federally mandated sanctions will be disproportionately applied to minority students and their schools. Because there is little evidence to date on the effects of federal sanctions on school performance, it is unclear how these interventions will affect minority student achievement or whether they will lead to school improvement.

Our analyses of school-level ethnicity and free and reduced-price lunch data from the 2003–2004 school year indicate that many schools with large Black and Latino enrollments also have high poverty rates. This fact is substantively important for AYP policy because it suggests that schools with a Black or Latino subgroup will also contain a subgroup for economically disadvantaged students. In contrast, predominantly White schools, which usually have low poverty rates, are less likely to have a separate accountability target for low-income students. As shown in Table 3, there is a strong correlation between school poverty and race in each of the six states. In Georgia, Illinois, New York, and Virginia the correlation between the percentage of Black students and the percentage of low-income students is between .63 and .69. In Arizona and California, schools with a high percentage of Latino students have a high percentage of low-income students (r = .65 in Arizona, r = .66 in California). However, in most states, there is a strong and negative correlation between the percentage of low-income students and the percentage of Asian and White students. These negative correlations imply that the percentage of low-income students is smaller, on average, in schools with a large percentage of White and Asian students. These findings parallel results from previous studies, which suggest that segregated Black and Latino schools are more likely to have highly concentrated poverty than segregated White schools (Orfield &

Table 3
Pearson Correlations (r) Between School Poverty and Race/Ethnicity in Six States, 2003–2004

State	Black and Low-Income	Latino and Low-Income	Asian American and Low-Income	White and Low-Income
Arizona	0.28**	0.65**	–0.02	–0.18**
California	0.16**	0.66**	–0.19**	–0.62**
Georgia	0.69**	0.11**	–0.30**	–0.70**
Illinois	0.68**	0.39**	–0.19**	–0.79**
New York	0.65**	0.68**	0.11**	–0.84**
Virginia	0.63**	0.06*	–0.27*	–0.56**

Note. School poverty is the percentage of students eligible for free and reduced-price lunch. The number of schools in each state is as follows: Arizona (n = 1,732), California (n = 9,087), Georgia (n = 1,965), Illinois (n = 3,804), New York (n = 4,276), Virginia (n = 1,846). Data are from websites of the departments of education of Arizona, California, Georgia, Virginia, Illinois, and New York and are for public use.
* $p < .05$. ** $p < .01$.

Lee, 2005). Given the large racial and socioeconomic gaps in test scores, the mean proficiency standard used to determine AYP poses the greatest challenge to schools with a Black or Latino subgroup or a subgroup of low-income students. Because AYP is based largely on mean proficiency levels that are more apt to highlight differences in school demographics than school performance, the next section uses measures of improvement in reading and mathematics proficiency levels to assess school performance.

Using Reading and Mathematics Proficiency to Assess School Performance in Virginia

In the following analysis, we use data from Virginia to first compare improvements in reading and mathematics proficiency levels for different categories of schools and then examine school performance ratings under NCLB's conjunctive accountability system and Virginia's compensatory accountability system. Because Virginia has administered the same state Standards of Learning (SOL) assessments in reading and mathematics since 1998, it provides useful data for making long-term comparisons of school improvement. Before examining these performance trends, however, it is important to consider some key differences between the federal AYP requirements and the Virginia standards of accreditation, the cornerstone of the state accountability system. Virginia has implemented a compensatory accountability system in which SOL scores from all students are averaged to compute a school mean in each of the four academic subjects. To earn full accreditation, a school must ensure that 70% of all students score at or above "proficient" in reading, mathematics, history/social science, and science (Virginia Department of Education, 2003). Moreover, low scores from a single subgroup do not result in a school's failing to meet the Virginia standards of accreditation.

The amount of improvement in proficiency levels is often similar in schools that are identified as needing improvement and schools that meet federal AYP standards. Figures 1 and 2 compare the achievement characteristics of Virginia schools that met AYP and schools that were identified as needing improvement in 2003–2004. There has been at least a 25-percentage-point gap in average reading and mathematics proficiency levels between schools needing improvement and schools meeting AYP in each of the six administrations (1998 to 2003) of the Virginia SOL tests. Nonetheless, both types of schools display similar patterns of improvement in their proficiency levels: Average reading trends look flat from 1998 to 2001 and begin to increase in 2001 for schools needing improvement and in 2002 for schools meeting AYP. Mathematics trends for both types of schools have shown steady improvements in proficiency levels since 1999. While schools identified as in need of improvement and schools meeting AYP appear to make similar improvements in mean proficiency levels over time, there is a persistent gap in mean proficiency levels between these two groups of schools in each of the 5 years. This initial gap in mean proficiency, however, appears to reflect differences in the academic skills of students before they

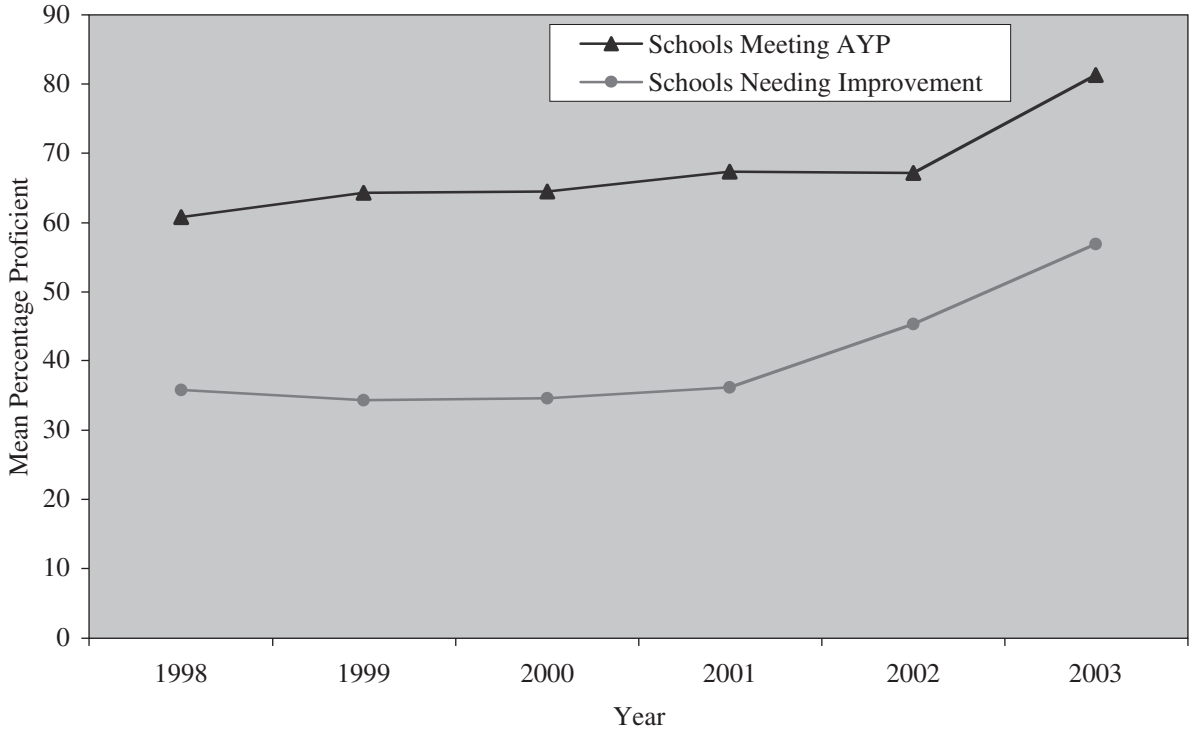

FIGURE 1. *Mean percentage proficient on the Virginia Standards of Learning reading tests in schools meeting adequate yearly progress requirements and in schools needing improvement, 1998 to 2003. From Virginia Department of Education (2004), for public use.*

Note. We use the term "reading" to refer to the Virginia English Language/Reading Standards of Learning test. Calculations are our own. Data are based on schools with valid reading scores in each administration of the Standards of Learning test. The sample sizes are as follows for schools in need of improvement (SINI) and schools meeting adequate yearly progress requirements (AYP): 1998: SINI = 38, AYP = 907; 1999: SINI = 38, AYP = 924; 2000: SINI = 39, AYP = 943; 2001: SINI = 42, AYP = 967; 2002: SINI = 43, AYP = 1,021; 2003: SINI = 45, AYP = 1,106.

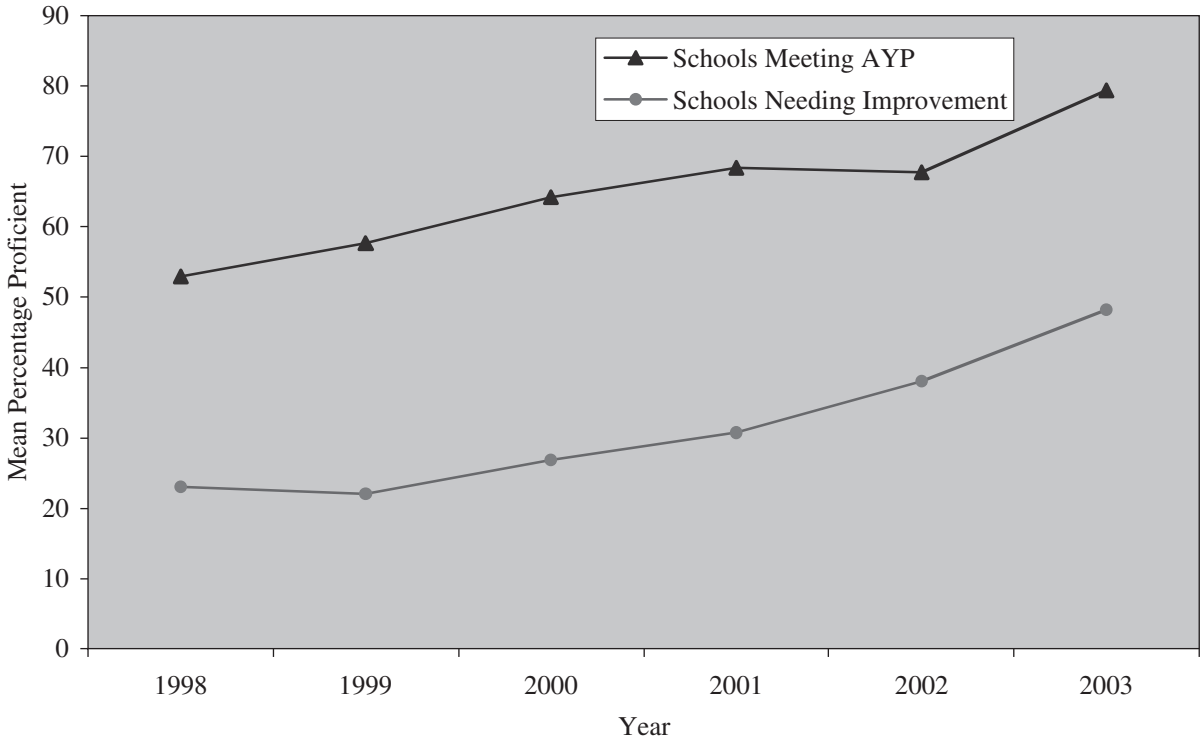

FIGURE 2. *Mean percentage proficient on the Virginia Standards of Learning mathematics tests in schools meeting adequate yearly progress requirements and in schools needing improvement, 1998 to 2003. From Virginia Department of Education (2004), for public use.*

Note. Calculations are our own. Data are based on schools with valid reading scores in each administration of the Standards of Learning assessment. The sample sizes (number of students in sample) are as follows for schools in need of improvement (SINI) and schools meeting adequate yearly progress requirements (AYP): 1998: SINI = 38, AYP = 907; 1999: SINI = 38, AYP = 924; 2000: SINI = 39, AYP = 943; 2001: SINI = 42, AYP = 967; 2002: SINI = 43, AYP = 1,021; 2003: SINI = 45, AYP = 1,106.

entered school. In other words, students in schools identified as needing improvement began with lower average test scores than students in schools meeting AYP. These results suggest that mean proficiency measures are likely to identify many high-poverty schools as in need of improvement even when they show evidence of improving reading and mathematics performance over time.

Because the previous analysis categorized schools on the basis of the NCLB accountability ratings, it does not provide explicit comparisons of mean proficiency levels by school poverty rates. Table 4 displays the mean proficiency level in reading and mathematics in three groups of Virginia schools (low-poverty, middle-poverty, and high-poverty), which are based on the percentage of students eligible for free and reduced-price lunch. The 50% eligibility rate for free and reduced-price lunch is often used as a criterion for identifying "high-poverty schools" (Orfield & Lee, 2005; Puma et al., 1997; Raudenbush, 2004); in Virginia, at least 46% of students in high-poverty schools are eligible for free and reduced-price lunch. The columns in Table 4 display the reading and mathematics proficiency levels in 1998 and 2003, and the difference in proficiency levels between 1998 and 2003 by school poverty levels. During the 1st year of the SOL administrations in 1998, low-poverty schools attained substantially higher mean proficiency levels in reading and mathematics than middle- and high-poverty schools. This fact indicates that a number of low-poverty schools started with an advantage over high-poverty schools. For example, in 1998, 74% of students in low-poverty schools were proficient in reading, as compared with 48% of students in high-poverty schools. Nonetheless, both low-poverty schools and high-poverty schools in Virginia increased the percentage of students scoring at the proficiency level from 1998 to 2003. This suggests that using a single mean proficiency level to measure school performance is biased against high-poverty schools, which enroll students with lower average scores, and that an accountability system based on mean proficiency is likely to over-identify many high-poverty schools as underperforming schools. On the other hand, an accountability system based on percentage change in the proficiency level of students would be more likely to indicate that both high- and low-poverty schools increased student performance on standardized tests.

As was noted earlier, the Virginia accountability system incorporates the performance of all students in computing an overall school proficiency score. Because this is a compensatory accountability system, no single subgroup score can keep an entire school from earning full accreditation. Under NCLB's conjunctive accountability model, on the other hand, the performance of a single subgroup can result in a school's being identified as needing improvement. This increases the probability that a number of high-poverty schools and racially diverse schools will be identified as failing to make AYP, even though many of these schools meet the accreditation standards established by state law. Table 5 shows the demographic characteristics of 352 Virginia schools that earned full state accreditation and failed to make AYP and

Table 4
Mean Proficiency on the Virginia Standards of Learning Assessment (Reading and Mathematics) by School Poverty Level, 1998–2003

School Poverty Level	Reading			Mathematics		
	Mean Proficiency, 1998	Mean Proficiency, 2003	Change in Mean Proficiency, 1998–2003	Mean Proficiency, 1998	Mean Proficiency, 2003	Change in Mean Proficiency, 1998–2003
Low Poverty (0–23%)						
Mean % (*n*)	74 (427)	88 (595)	14 (427)	67 (427)	86 (595)	19 (427)
Middle Poverty (24%–45%)						
Mean % (*n*)	61 (480)	81 (593)	20 (480)	52 (480)	80 (593)	28 (480)
High Poverty (46%+)						
Mean % (*n*)	48 (491)	73 (557)	25 (488)	39 (491)	71 (556)	32 (486)
Total						
Mean % (*n*)	60 (1,398)	81 (1,745)	21 (1,395)	52 (1,398)	79 (1,744)	27 (1,393)

Note: Mean proficiency is the percentage of students scoring at or above the proficient level in reading and mathematics on the Virginia SOL assessment. The change in mean proficiency is the difference between mean proficiency in 2003 and 1998. School poverty level is the percentage of students eligible for free and reduced-price lunch. Data on school poverty levels are from the website of the Virginia Department of Education and are for public use.

the demographic characteristics of 824 schools that met both state and federal performance requirements. The subset of 352 schools that failed to make AYP had a larger percentage of minority and low-income students than did the 824 schools that met AYP. These figures imply that the AYP requirements have a disparate impact on schools serving low-income and minority students that are identified as fully accredited by Virginia policy. Stated simply, schools that failed to make AYP had more subgroup targets than schools that met AYP. In Virginia, a school with 50 or more students in any one subgroup is held responsible for a separate subgroup accountability goal in order to make AYP (Virginia Department of Education, 2003). Table 6 shows that 48% of the 352 schools failing AYP had more than three subgroup targets, as compared with 15% of the 824 schools meeting AYP. It is important to note that 110 of the 352 schools failing AYP receive Title I funds and are thus subject to federal sanctions if they continue to fail AYP in subsequent years. These findings suggest that AYP failure rates are concentrated in schools serving a diverse student population, which are often required to meet multiple accountability targets for students in different economic and racial/ethnic subgroups.

Effects of Subgroup Accountability on High-Poverty Schools and Racially Diverse Schools in California

The analysis of Virginia school data suggests that the NCLB subgroup policy results in a number of high-poverty schools' failing to meet AYP, even when those schools improve their proficiency levels and meet state accreditation standards. It is important to note, however, that subgroup accountability can be incorporated into an accountability system in a way that has less of a disparate impact on high-poverty schools. This can be seen in California, one of a few states that requires schools to meet separate subgroup targets as part of the school accountability policy. However, there are two key differences in the design of the subgroup policy

Table 5
Demographic Characteristics of Virginia Schools Based on Performance Ratings in 2003

Performance Label	Black (%)	Latino (%)	Asian (%)	White (%)	Low-Income (%)
Met Virginia Standards of Accreditation but ailed Federal AYP (*n* = 352)	24	11	7	58	32
Met Virginia Standards of Accreditation and federal AYP (*n* = 824)	16	4	4	76	25

Note. Demographic characteristics include the average percentage of students by race/ethnicity and eligibility for free and reduced-price lunch (low-income). Data on school demographics are from the website of the Virginia Department of Education and are for public use. Calculations are our own.

Table 6
Subgroups in Virginia Schools by Their Performance Rating in 2003

Number of Subgroups	Schools Meeting Virginia Standards but Failing Federal AYP (*n* = 352)	Schools Meeting Virginia Standards and Federal AYP (*n* = 824)
0–2 Subgroups	52	85
3+ Subgroups	48	15

Note. To be included in this calculation a subgroup had to consist of 50 or more students with valid test scores. We counted the number of subgroups that met this criterion in schools meeting AYP requirements and in schools failing to meet AYP requirements. Data on school demographics and performance ratings are from the website of the Virginia Department of Education and are for public use. Calculations are our own.

under California law and federal law. First, California adopted a compensatory accountability system based on a single numeric index called the Academic Performance Index (API). Because the API scores represent a weighted average of student performance on statewide tests primarily in reading and mathematics (California State Department of Education, 2003), low performance in one subject area or by one group of students can be compensated by high performance in a different subject area and by a different group of students. Second, there are differences in the definition of subgroup categories and minimum group size requirements used for federal and state accountability purposes. Because California did not include the scores of students with limited English proficiency and students with disabilities until it was required to do so by the NCLB legislation, the state's API scores are based on the test scores of fewer subgroups than the federal AYP ratings. In addition to differences in subgroup definitions, there are differences in the number of students required for subgroup accountability to take effect. California law defines "numerically significant" as a subgroup that constitutes at least 30 students and 15% of the school's total population, or 100 students regardless of their percentage of the total enrollment (Betts & Danenberg, 2002). For AYP, a subgroup is counted for school accountability if it represents 50 students and 15% of the students with valid test scores, or 100 students in the overall school enrollment. The higher minimum group size criterion used for NCLB is intended partly to reduce AYP failure rates in racially diverse schools. Nonetheless, because NCLB relies on a conjunctive accountability system, low performance by a single subgroup can result in an entire school's failing to make AYP.

To illustrate the consequences of the NCLB subgroup accountability policy in California schools, we examined AYP failure rates in schools with varying numbers of subgroup accountability targets for the 2002–2003 school year.[2] As shown in Figure 3, AYP failure rates increased as schools were held accountable for meeting additional subgroup targets. In schools with only one subgroup target, only 22% of the schools failed to make AYP. However, over 50% of schools with four or more subgroup targets failed to make AYP. For example, 61% of schools with five subgroup targets and 75% of schools with six subgroup targets failed to make AYP.

California has a number of Title I schools that are meeting their API targets but are identified as failing to meet federal AYP requirements. In many instances, AYP failure rates are concentrated in high-poverty and racially diverse schools that are meeting state performance goals in reading and mathematics. Table 7 compares the demographic characteristics of a subset of 403 schools that met their California API growth targets but failed to make AYP, and 4,512 schools that met both state and federal performance requirements. The most striking difference between schools that met AYP and failed to make AYP is in their demographic characteristics, rather than their academic performance, as measured by California's Academic Performance Index. In particular, Latino students and low-income students constitute more than 70% of the student enrollment in the 403 schools that exceeded their California API targets but failed to make the federal AYP goal. However, schools that made AYP had a relatively smaller percentage of both Latino students (33%) and low-income students (41%). There are also clear differences in the number of subgroup targets in both groups of schools. Table 8 (page 12) shows that 59% of schools meeting AYP had three or more subgroup targets, while virtually all (97%) of the 403 schools that failed to make AYP were required to meet three or more subgroup accountability targets. Because 370 of the 403 schools failing AYP are Title I schools, they are subject to school improvement sanctions starting with the requirement to offer transfers and supplemental educational services to their students. An additional 52 of the 403 schools were in their 3rd year of program improvement and required to undergo corrective action in 2004–2005. These school improvement remedies do not seem justified in Title I schools that are making clear progress toward performance goals established by state law.

Conclusion

Few Americans disagree with the ultimate objective of the No Child Left Behind Act—to eliminate achievement disparities in reading and mathematics by the 2013–2014 school year. To realize this objective, some proponents of the NCLB accountability system (Chubb, 2005; Haycock & Wiener, 2003) argue that all schools should be held to a uniform performance goal and that strong accountability pressures are needed to accelerate the achievement of minority and low-income students. However, our analysis of the NCLB accountability system in six geographically diverse states shows that the use of mean proficiency has a disparate impact on schools serving low-income children and subgroup accountability rules can over-identify racially diverse schools as failing to make AYP. Our research parallels findings of other researchers and policy analysts, who suggest that the AYP requirements are unlikely to improve student achievement if mean proficiency is the primary indicator for measuring the performance of schools and subgroups of students (Linn, 2003;

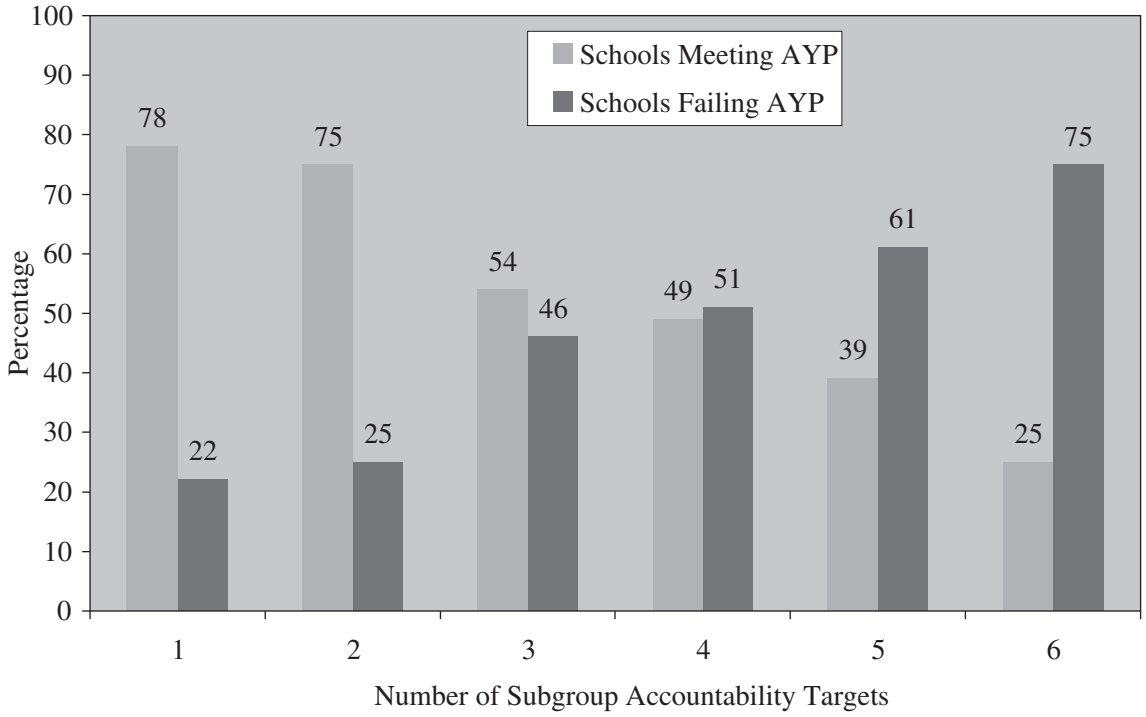

FIGURE 3. *Percentage of California schools meeting adequate yearly progress requirements by the number of subgroup accountability targets in reading, 2003. From California Department of Education (2004a), for public use.*

Note. We use the term "reading" to refer to California's English Language Arts test. Calculations are our own. Sample sizes (number of students in sample) for schools meeting annual yearly progress requirements by the number of subgroup categories: 0 subgroups ($n = 564$), 1 subgroup ($n = 674$), 2 subgroups ($n = 749$), 3 subgroups ($n = 1{,}170$), 4 subgroups ($n = 1{,}029$), 5 subgroups ($n = 277$), 6 subgroups ($n = 49$). Sample sizes (number of students in sample) for schools not meeting annual yearly progress requirements by the number of subgroup categories: 0 subgroups ($n = 1{,}006$), 1 subgroup ($n = 187$), 2 subgroups ($n = 244$), 3 subgroups ($n = 1{,}007$), 4 subgroups ($n = 1{,}084$), 5 subgroups ($n = 436$), 6 subgroups ($n = 149$).

Novak & Fuller, 2003; Raudenbush, 2004). Because Title I schools failing AYP for 2 or more years are subjected to mandatory sanctions, multiple sources of information on student achievement are needed to more accurately determine the performance of these schools and their students.

Multiple indicators of school performance could address two design flaws in the current definition of AYP. The first design flaw in the NCLB accountability system is the use of mean proficiency as the primary measure for determining whether schools are making adequate yearly progress. Supplementing mean proficiency with a measure of improvement could provide a more accurate assessment of school quality by isolating the impact of schools on improvements in standardized test scores over several consecutive years. Data from Virginia suggests that improvement in the percentage of students attaining proficiency in reading and mathematics is similar, on average, for high-poverty schools and low-poverty schools. Despite the progress made by both types of schools, the initial gap in mean proficiency levels persisted for

Table 7
Demographic Characteristics of California Schools Based on Performance Ratings in 2003

Performance Label	Black (%)	Latino (%)	Asian (%)	White (%)	Low-Income (%)
Met California Academic Performance Index (API) but failed Federal AYP ($n = 403$)	7	71	5	14	76
Met California Academic Performance Index (API) and Federal AYP ($n = 4{,}512$)	16	4	4	76	25

Note. Demographic characteristics include the average percentage of students by race/ethnicity and eligibility for free and reduced-price lunch (low-income). Percentages for race/ethnicity do not add up to 100% because our analysis did not include Native American, Filipino, and Pacific Islander students. Data on school demographics and performance ratings are from the website of the California Department of Education and are for public use. Calculations are our own.

Erratum: Click here for corrected table

Table 8
Subgroups in California Schools by Their Performance Rating in 2003

Number of Subgroups	Schools Meeting California API and Federal AYP (*n* = 4,512)	Schools Meeting California API and Federal AYP (*n* = 403)
0–2 Subgroups	41	3
3+ Subgroups	59	97

Note. Data on school demographics and performance ratings are from the website of the California Department of Education and are for public use. Calculations are our own.

6 years and reflected differences in the academic skills of students from the 1st year of Virginia's testing program. This tendency toward identifying high-poverty schools as failing AYP stems largely from selection bias: High-poverty schools serve students with mean test scores that are lower than those of students in low-poverty schools. Raudenbush (2004) has argued that "high-stakes decisions based on mean proficiency are scientifically indefensible" and that rewarding "schools for high mean achievement is tantamount to rewarding those schools for serving students who were doing well prior to school entry" (p. 35). In other words, differences in school mean proficiency alone are not valid indicators of school effectiveness because they reflect large differences in the academic skills and socioeconomic backgrounds of students before they enter school.

The second design flaw in the NCLB accountability system is the subgroup accountability policy, which requires each subgroup of students within a given school to meet a separate test score target. Consequently, the NCLB subgroup policy puts racially diverse schools at greatest risk of failing AYP. Our analyses of 2003–2004 assessment data revealed a subset of 352 Virginia schools and 403 California schools that failed to make AYP solely because they were required to meet three or more subgroup targets. These schools failed to make AYP despite showing evidence of success on performance measures established by state policy, such as the criteria for earning full accreditation in Virginia and for meeting the Academic Performance Index score in California. The disparate impact of AYP policy on high-poverty schools with diverse student enrollments has prompted some state education leaders to question the fairness of NCLB's school performance ratings and their usefulness for improving achievement in the lowest-achieving schools. For example, the California state superintendent argued that the 403 schools that met state accountability goals "are clearly on the right track, yet they did not make adequate yearly progress. . . . Under NCLB, these schools will be viewed in the same 'failing' category as schools not meeting [California] targets and clearly needing intervention" (California State Department of Education, 2004b). In Virginia, both U.S. Senators and several Congressmen expressed concern that schools meeting the state's accreditation standards were failing to make AYP because of its one-size-fits-all approach to measuring school performance (Robelen, 2005). These federal legislators were concerned that labeling Virginia schools that had earned full accreditation under state policy as failing under NCLB caused confusion among educators and parents and undermined political support for the NCLB accountability system (Allen, 2005).

The current approach to measuring adequate yearly progress has contributed to the political opposition to NCLB. This opposition to the NCLB accountability system is likely to persist as proficiency targets are raised and more schools that meet state accountability goals are identified as failing AYP and are consequently required to implement federal sanctions. Using multiple indicators of school performance could help to minimize errors in classifying the performance of schools, particularly as proficiency targets are raised. Indeed, mean proficiency should be viewed as only one of many performance indicators to be used in determining whether schools are contributing to student achievement and whether subgroups of students are making academic progress in reading and mathematics. If the NCLB accountability system rewarded schools for improving proficiency levels over time and meeting state-mandated performance goals, a number of high-poverty schools would be identified as improving student achievement rather than failing to make adequate yearly progress. The principle of using multiple criteria should guide efforts to reform NCLB by encouraging policymakers to adopt a broader set of outcome measures for school accountability, including information on improvement in proficiency rates and state classifications of school performance. Incorporating multiple indicators in the NCLB accountability system could enhance the validity of AYP as a measure of school performance and strengthen political support for the federal legislation.

NOTES

We would like to thank the three anonymous reviewers and Stafford Hood for their excellent suggestions and critical feedback, which helped us to improve the clarity of this manuscript and sharpen the focus of the argument.

[1] These sanctions apply only to Title I schools. Title I schools failing to make AYP for 2 consecutive years are identified for their 1st year of school improvement and must (a) offer all students the option to transfer to another school, and (b) develop a 2-year school improvement plan. In addition, these schools are eligible to receive technical assistance from the state. If a school fails to make adequate yearly progress for 3 consecutive years, students are eligible for supplemental educational services. If a school fails AYP for 4 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 to make AYP for 5 consecutive years, it enters its 4th year of school improvement and can be restructured, taken over by the state or a private management contractor, converted to a charter school, or reconstituted with a new staff (NCLB, 2001 Sec. 1116 [b][1–8]).

[2] We excluded Native American, Filipino, and Pacific Islanders in our subgroup analyses for California schools 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.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33,* 4–14.

Allen, G. (2005, May 22). Undermining school progress. *Washington Post,* B7.

Betts, J. R., & Danenberg, A. (2002). School accountability in California: An early evaluation. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 123–198). Washington, DC: Brookings Institution Press.

Borman, G. D., Stringfield, S. C., & Slavin, R. E. (Eds.). (2001). *Title I, compensatory education at the crossroads.* Mahwah, NJ: Lawrence Erlbaum.

Bryk, A. S., & Raudenbush, S. (1988). On heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104,* 396–404.

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

California State Department of Education. (2004a). *Adequate yearly progress (AYP) data files.* Retrieved March 21, 2004 from http://ayp.cde.ca.gov/datafiles.asp

California State Department of Education. (2004b). *Double API growth schools.* Retrieved May 11, 2004, from http://www.cde.ca.gov/eo/co/co/apigrowth.asp

Chubb, J. E. (Ed.). (2005). *Within our reach: How America can educate every child.* Lanham, MD: Rowman & Littlefield.

Chubb, J., Linn, R., Haycock, K., & Wiener, R. (2005, Spring). Do we need to repair the monument? Debating the future of No Child Left Behind. *Education Next, 5,* 8–19.

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

Finn, C. E., & Hess, F. M. (2004). On leaving no child behind. *The Public Interest, 157,* 35–56.

Haycock, K., & Wiener, R. (2003). *Adequate yearly progress under NCLB.* Paper presented at the National Center on Education and the Economy Policy Forum, Implementing the No Child Left Behind Act, Washington, DC.

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

Kane, T. J., & Staiger, D. O. (2003). Unintended consequences of racial subgroup rules. In P. E. Peterson & M. R. West (Eds.), *No child left behind? The politics and practice of school accountability* (pp. 152–176). Washington, DC: Brookings Institution Press.

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

Lee, J. (2004). How feasible is adequate yearly progress (AYP)? Simulations of school AYP "uniform averaging" and "safe harbor" under the No Child Left Behind Act. Retrieved May 1, 2005, from http://epaa.asu.edu/epaa/v12n14/

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

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

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

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, DC: National Academy Press.

Mosteller, F., & Moynihan, D. P. (Eds.). (1972). *On equality of educational opportunity.* New York: Random House.

Murphy, J. T. (1971). Title I of ESEA: The politics of implementing federal education reform. *Harvard Educational Review, 41,* 35–63.

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

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

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425.

Novak, J. R., & Fuller, B. (2003). *Penalizing diverse schools? Policy brief 03-4.* Berkeley, CA: Policy Analysis for California Education.

Orfield, G. (1996). The growth of segregation. In G. Orfield & S. E. Eaton (Eds.), *Dismantling desegregation* (pp. 53–72). New York: New Press.

Orfield, G., & Lee, C. (2005). *Why segregation matters: Poverty and educational inequality.* Cambridge, MA: The Civil Rights Project, Harvard University.

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.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service.

Robelen, E. W. (2005, May 11). Virginians offer bill on NCLB flexibility. *Education Week,* p. 29.

Rose, L. C., & Gallup, A. M. (2003, September). The 35th annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan,* 41–52.

Rudalevige, A. (2003). No child left behind: Forging a congressional compromise. In P. E. Peterson & M. R. West (Eds.), *No child left behind? The politics and practice of school accountability* (pp. 23–54). Washington, DC: Brookings Institution Press.

Sunderman, G. L., Kim, J. S., & Orfield, G. (2005). *NCLB meets school realities: Lessons from the field.* Thousand Oaks, CA: Corwin Press.

U.S. Department of Education. (2001). *No Child Left Behind.* Washington, DC: U.S. Department of Education, Office of the Secretary.

U.S. Department of Education. (2002). *The condition of education 2002* (NCES #2002-025). Washington, DC: U.S. Government Printing Office.

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

Virginia Department of Education. (2004). *Virginia school report card.* Retrieved March 4, 2004, from http://www.pen.k12.va.us/VDOE/src/index.shtml

AUTHORS

JAMES S. KIM is an Assistant Professor of Education at the University of California, Irvine, Department of Education, 2001 Berkeley Place, Irvine, CA 92697; jamesk@uci.edu. His research interests include federal education policy and the use of experimental methods to assess the effectiveness of compensatory education programs.

GAIL L. SUNDERMAN is a Senior Research Associate in K–12 education for the Civil Rights Project at Harvard University, 125 Mt. Auburn Street, Cambridge, MA 02138; glsunderman@yahoo.com. Her research focuses on educational policy and politics and urban school reform, including the development of education policy and the impact of policy on the educational opportunities for at-risk students.

Manuscript received May 19, 2004
Revision received February 21, 2005
Accepted July 10, 2005