Fifty Ways to Leave a Child Behind:
Idiosyncrasies and Discrepancies in States' Implementation of *NCLB**

Elizabeth Davidson
Teachers College, Columbia University
ekd2110@columbia.edu

Randall Reback **(presenter)**
Barnard College and ISERP, Columbia University
rr2165@columbia.edu

Jonah Rockoff
Columbia Business School and NBER
jr2331@columbia.edu

Heather L. Schwartz
RAND Corporation
hschwart@rand.org

Abstract

The *No Child Left Behind* (NCLB) *Act* required states to adopt accountability systems measuring student proficiency on state administered exams. The federal legislation contained several strict requirements for NCLB implementation, such as escalating student proficiency targets that reach 100 percent proficiency by 2014. But it also gave states considerable flexibility to interpret and implement components of NCLB. Using a data set we constructed, this paper is the first national study examining which schools failed during the early years of NCLB and which performance targets they failed to meet. We explore how states' NCLB implementation decisions affected their schools' failure rates, which ranged from less than 1 percent to more than 80 percent across states. Wide cross-state variation in failure rates resulted from how states' decisions interacted with each other and with school characteristics like enrollment size, grade span, and ethnic diversity. Subtle differences in policy implementation led to dramatic differences in measured outcomes.

---

* We thank participants at the 2013 AEFP conference for helpful suggestions. A longer version of this paper is available as *National Bureau of Economic Research Working Paper #18988 (April 2013)*.

I. Introduction

The *No Child Left Behind Act* (NCLB) requires states to construct school accountability systems using standardized tests to measure student proficiency rates in math and English Language Arts (ELA). A school fails to make Adequate Yearly Progress (AYP) if proficiency rates fall short of that year's targets. Using a newly-available data set, this paper is the first national study examining which schools failed during the early years of NCLB and which performance targets they failed to meet. We also examine the school characteristics and state policy implementation decisions that drove differences in failure rates across states. We find that wide cross-state differences in failure rates were largely the result of subtle differences in states' own NCLB rules. A common misconception regarding the wide variation in AYP failure rates across states is that these differences were driven by more obvious state policy differences, such as the difficulty of the exam questions and the proficiency standards. A better understanding of how subtle policy differences influenced schools' ratings during the early years of NCLB may inform current efforts to reform NCLB and other school accountability programs.

II. NCLB Overview

A school's performance rating under NCLB is based on student proficiency rates on statewide tests, student participation rates on those tests, and an additional state-selected indicator of student performance.[1] Both the campus as a whole and various student subgroups—racial/ethnic subgroups, students eligible for free/reduced priced lunch, students with limited English language proficiency, and disabled students—must meet all of the performance targets

[1] We provide a brief overview of NCLB in this section and refer the reader to the U.S. Department of Education's *Desktop Reference* (2002) and to Manna's *Collision Course* book (2010) for more details on NCLB policies. Manna also provides revealing anecdotes concerning the challenges faced by states and schools in implementing these policies.

for the school to make AYP.[2] Many states already had their own testing and accountability systems prior to NCLB, and so the impact of NCLB could depend on whether students were already being tested under similar accountability systems (Dee and Jacob, 2011; Dee & Jacob, forthcoming).

The three core mandatory elements of NCLB pertain to annual testing of virtually all public school students in certain grade levels and subjects, an increasing bar for student proficiency on these tests, and annual determinations of school performance with consequences for schools that fail to make AYP. NCLB required states to administer baseline student exams in the spring of 2002 and to adopt school accountability systems for the school year 2002-03. States selected their own exams and defined proficiency on those exams. States then determined a schedule for the percentage of students who must meet proficiency each year, with targets increasing annually up to a mandated 100% target for 2014. States could set different benchmarks by grade level and by subject area, but not by student subgroup. To prevent schools from strategically exempting low-performing students from taking exams, NCLB dictates that student subgroups are required to meet a 95% participation rate on both math and ELA exams. The final category of school performance is the state-selected "other" academic indicator. NCLB rules allowed for flexibility in states' selection of elementary and middle schools' other indicators, and most states used attendance rates. NCLB rules required that states use graduation rates for high schools' other indicator.[3]

[2] Students are counted in all subgroups to which they belong. For example, a Hispanic student who is limited English proficient and eligible for free lunches will contribute to eight different proficiency rates—the campus-wide group, the Hispanic subgroup, the limited English proficient subgroup, and the free/reduced priced lunch subgroup proficiency rates in math and ELA. Subgroup proficiency rates only influence the school's AYP rating if there are sufficient numbers of students enrolled at the school (and meeting the "continuous enrollment" definition described elsewhere in the paper).

[3] Initially, NCLB permitted states to use their own formulae for calculating graduation rates. In December 2008, the U.S. Department of Education announced that all states must use a standardized four-year graduation rate formula.

In addition to the stigma of failing to make AYP, there are additional consequences for schools serving low-income populations that receive funding under the federal Title I program. Students at failing Title I schools have the opportunity to transfer to non-failing schools within the same district. After consecutive years of AYP failure, these schools' students from low-income families are entitled to use school funds to purchase private tutoring services (called "supplemental education services"). If these schools fail to make AYP for several years, then they are subject to closure or restructuring.

Beyond these core requirements, there are several key areas where states have latitude in calculating AYP. We summarize them here and provide further detail in the sections that follow. The first area relates to acceptable adjustments to student proficiency rates under the law. Even if a subgroup's or school's performance falls below the proficiency target for the given school year, the school may still make AYP because NCLB allows states to employ various statistical techniques and contingencies to adjust proficiency rates.[4] Two types of adjustments permitted under NCLB are the application of confidence intervals and the use of "safe harbor." Confidence intervals provide leniency around proficiency rate targets to account for small numbers of tested students. They lower a student group's effective proficiency targets based on the number of tested students in that group at that school—the smaller the group, the larger the confidence interval. "Safe harbor" rules offered leniency to schools that missed proficiency targets but had students make large gains in proficiency rates from the previous year. To make

---

The U.S. DOE requested states implement the new formula as soon as possible but required states to comply by 2010-2011 (U.S. DOE, 2008).

[4] Beyond the formal NCLB rules, states also allowed school districts and schools to submit appeals of schools' AYP ratings. Acceptable grounds for appeal varied by state. For example, in Colorado, schools could successfully appeal AYP failure if the sole reason for failure was the performance of the disabled subgroup and this subgroup did meet its targets in another year. In several states, (e.g., Iowa and Michigan), schools could appeal by retroactively exempting students from contributing to participation rates if the students had experienced significant medical emergencies.

AYP under the safe harbor rule, states typically require a 10% reduction in the fraction of students failing to reach proficiency.

The second area where states have latitude is determining which students count towards the accountability system. In the initial years of implementation, not all states applied consistent definitions of special needs categories exempted certain students from the general standardized test. However, the U.S. Department of Education later issued exemption rules to close loopholes related to testing of disabled students. But several other discrepancies remain. For example, not all states hold the same student racial and ethnic categories accountable under AYP. In addition, states determine how long students must be enrolled in the same school for their test performance to contribute to schools' AYP determinations. These "continuously enrolled students" comprise the denominator of the participation rate calculation. A state with a very strict definition of continuous enrollment only counts students enrolled at their schools for one calendar year prior to testing. More commonly, states count students who were tested in the spring and had been enrolled at their schools since late September or October. Schools could also exempt students from contributing to participation rates if the students experienced significant medical emergencies. To protect student anonymity and avoid using unreliable measures of subgroup performance, states also had to establish a minimum group size for a subgroup to count toward their school's NCLB rating. Most states chose a minimum subgroup size between 30 to 40 students, but the range extended from 5 students to 100 students. In some states, minimum group size was a function related to school population. For example, California's subgroups were held accountable if they either had 100 tested students or at least 50 tested students that composed at least 15% of the schools' total tested population.

A third, often-overlooked area of flexibility is the coverage of testing across grade levels and methods of aggregating performance across grade levels. Although tested grade levels became more standard as of 2006-2006 when states were required to test students in grades 3 through 8 and in one high school grade[5], the aggregation of scores has not become standardized. For schools that serve multiple tested grade levels, states could decide whether to aggregate statistics across these grade levels or treat each grade separately. States like Washington treated each grade separately, so that a school with both a 4th grade and 7th grade would need students from each of those grades to exceed proficiency targets. This could make it more challenging for that school to make AYP. Yet Washington also treated each grade separately when counting the number of tested students for determining whether subgroups are accountable and for applying confidence interval adjustments, which would lead to fewer accountable subgroups and more generous confidence interval adjustments.

III. NCLB Data

NCLB has greatly expanded the amount of student performance data available to researchers and the public, though dissemination of data has been uneven across states. To promote studies of NCLB, we approached each of the 50 states individually in an attempt to form the most complete school-level data set concerning the early years of NCLB. We used a combination of methods to obtain the most comprehensive and accurate data possible—primarily requesting data directly from state education departments and downloading data from state

[5] Before 2002-2003 to 2004-2005, states were required to test in at least one elementary grade, at least one middle school grade, and at least one high school grade. Consequently, tested grade levels varied across states during the first few years of NCLB. On the one extreme, states like Maryland tested in all grades 3 through 8 for AYP determinations. On the other extreme, states like New Jersey only tested grades 4, 8, and 11 up until 2004-2005.

websites. The resulting data and our documentation of sources are publicly available.[6] For the school years 2002-2003 and 2003-2004, we filled in otherwise missing data with information provided by the American Institutes for Research (2005) and the Council of Chief State School Officers (2005). For 2004-2005, we use school and subgroup proficiency target data from the American Institutes for Research (2005). The data include school-level AYP determinations and the subcomponents for these determinations.

IV. Descriptive Evidence on Failing Schools

Looking nationwide from 2003 to 2005, there are clear observable differences between AYP failing and non-failing schools (Table 1). AYP failing schools were more likely to have higher total student enrollments, to have larger enrollments of poor and minority students, and to be designated as Title I schools. On average, schools that failed all three years had nearly double the percentage of students eligible for free and reduced-priced lunch as schools that made AYP all three years. Failing schools also have fewer teachers per student and are disproportionately located in urban school districts. Middle schools and high schools fail far more frequently than elementary schools.

Figure 1 reveals that most schools failed to make AYP due to proficiency rate requirements as opposed to participation rates. In 2005, only about 4% of failing schools would have made AYP if not for their participation rates. This rate was substantially lower than in the prior two years, suggesting that schools took action to ensure that sufficient numbers of students

[6] Data for the first two years of NCLB are currently accessible from our "No Data Left Behind" website at http://www7.gsb.columbia.edu/nclb/.

were tested.[7] Among schools failing to make AYP due to low proficiency rates, there was a slightly greater tendency to fail to meet ELA targets than math targets. Most commonly, failing schools had groups of students not meeting targets in both subjects.

While schools were potentially accountable for many student subgroups, the rate at which different subgroups caused schools to fail AYP varied widely. Such differences could simply be due to whether a subgroup was large enough to be held accountable. Figure 2 shows the percentage of schools where various subgroups counted toward AYP in 2004, as well as the rates at which these subgroups failed to make AYP. White and economically disadvantaged subgroups were held accountable in about 60% and 50% of schools, respectively, while fewer than 5% of schools had a Native American subgroup held accountable.

However, conditional on being accountable, subgroup failure rates varied considerably. White and Asian subgroups rarely failed, while more than half of all accountable Native American and disabled subgroups failed to meet proficiency targets. Disabled subgroups were also the most likely to be the only subgroup failing their schools' proficiency targets: 40% of accountable disabled subgroups were the only group to fail to meet targets at their schools.

V. Cross-State Differences in Failure Rates

Figure 3 illustrates the wide variation in states' AYP failure during the first three years of NCLB. In the first year of AYP designations (2003), 32% of the nation's schools failed AYP, but failure rates ranged from 82% in Florida to 1% in Iowa. The national failure rate declined to 26% by 2005, but failure rates ranged from 66% in Hawaii to 2% in Oklahoma. Failure rates changed substantially over time in some states. Alabama's failure rate jumped from 4% in 2003

[7] Participation data are not available for as many states in 2003 and 2004 as in 2005. When we restrict the sample to the 31 states with data available for all three years, then we observe a downward trend in the fraction of schools failing only due to participation: from 17% in 2003 to 14% in 2004 to 5% in 2005.

to 68% in 2004.[8] Tennessee's failure rate declined from 47% in 2003 to 7.6% in 2005. Failure rates by school level also varied substantially within some states. For example, only 11% of Georgia's elementary schools failed to meet AYP, yet 72% of its high schools failed. Similarly, only 20% of West Virginia's elementary schools failed yet more than 80% of its high schools failed. Reback, Rockoff, & Schwartz (forthcoming) document how a sizable fraction of schools that did not make AYP in their own states would have very likely made AYP in many other states.

A common misconception is that this wide variation in failure rates is due to cross-state differences in student proficiency rates. In reality, states' school failure rates are not strongly related to their students' performance. Figure 4 illustrates the lack of a strong relationship between school failure rates and student proficiency rates, showing student performance on states' math exams for the spring of 2004 against their states' school failure rates. This weak relationship arises because states determined NCLB proficiency targets based on their own pre-NCLB student proficiency rates. In essence, states were grading their own schools on state-specific curves, with varying starting points and trajectories for school performance targets. For example, Iowa set 2003 proficiency targets at 64% in math and 65% in ELA, while Missouri chose 8.3% and 18.4%, respectively.

Based on linear regressions that correspond to Figure 4, a one percentage point increase in state math proficiency rates is associated with only a statistically insignificant 0.1 percentage point decline in the fraction of a state's schools making AYP.[9] Even states with similar starting

[8] In 2002-2003, Alabama had an interim accountability system that used students' grade-level, not subgroup-level, norm-referenced scores to determine school-level AYP status. By 2003-2004, Alabama transitioned to a NCLB-compliant accountability system.

[9] The relationship with state ELA proficiency rates is larger: a 0.2 percentage point decline in the fraction of schools making AYP, significant at the .10 level. These relationships overall are weak and statistically insignificant. If we regress states' school AYP failure rates on quadratic terms for their states' proficiency rates in each subject (i.e.,

points had dramatically different rates of schools failing AYP. For example, proficiency targets in Louisiana and Florida differed by less than 7 percentage points, but their 2003 school failure rates differed by more than 75 percentage points.

VI. Explaining Cross-State Variation in Failure Rates

Various dimensions of NCLB implementation led to the wide variation in school AYP failure rates.[10] No individual state policy decision appears to be the primary culprit. Instead, failure rates appear to be the result of interactions among several decisions and states' school characteristics (e.g., enrollment size, grade spans, ethnic diversity of students). Given that we only have a sample of 50 states and a host of potentially important explanatory variables, there are insufficient degrees of freedom to tease out the relative importance of state policy variables via regression analysis. To examine the nature of these complex interactions, we instead describe five categories of policy decisions that we have identified as having substantial impacts on some states' school failure rates. We provide examples of states where failure rates were strongly influenced by these decisions. The first of these categories covers implementation errors that were rectified within the first couple of years of NCLB, but the remaining categories encompass policy decisions that continue to affect school failure rates.

1. *A few states initially deviated from NCLB rules.*
    a. *Calculations.* Iowa continued to develop its AYP formula and data collection processes throughout the initial two years of NCLB. Using proficiency rate and

four independent variables total), the R-squared is .13 but the adjusted R-squared is only .05. The joint significance level of these estimated coefficients is only .20.

[10] To determine each state's confidence intervals, safe harbor policies, and other AYP formulae choices, we referred to their approved state accountability workbooks. We obtained the workbooks from http://www2.ed.gov/admins/lead/account/stateplans03/index.html in January of 2007. Where possible, we selected criteria that applied to the 2003-2004 school year. However, as the workbooks were updated sometimes annually and often overwrote prior versions, we are not always able to determine when states adopted their criteria. For example, many states began to apply a 75% confidence interval to safe harbor determinations in 2005-2006.

participation rate data we retrieved from Iowa's Department of Education website, we applied Iowa's AYP formula and found higher failure rates than the state's official published rates.[11] In 2003 and 2004, respectively, 20% and 3% of Iowa's schools made AYP even though they had at least one accountable subgroup missing the 95% participation target.[12] Iowa did have an appeals process by which schools can petition to have up to 1% of students excused from participation due to illness, but the reported participation rates were often too low to have warranted a successful appeal. Data disaggregated by grade level is unavailable for Iowa, but we can examine proficiency rates for the 90% of Iowa's schools that served only one tested grade level.[13] Among these schools in 2004, 27% of schools that Iowa labeled as making AYP had either: (a) a subgroup with a participation rate below 95%, or (b) a subgroup with a proficiency rate ineligible for safe harbor and too low to meet the highest possible confidence interval upper bound.[14]

b. *Alternative Assessments.* Because disabled subgroups' performance were often the only reason for a school failing to make AYP, states' policies toward disabled subgroups have substantial ramifications. NCLB requires states to incorporate in AYP determinations nearly all special education students' scores on regular, grade-level assessments in AYP determinations. Student scores on alternative assessments can account for no more than 1% of a school's total scores. Texas state officials petitioned to "phase-in" the 1% rule over time, but the U.S. DOE denied their request. In 2003, the Texas State Education Agency ignored the U.S. DOE's ruling and approved the appeals of 1,718 schools whose special education

[11] During the summer of 2004–the months when state officials typically make AYP determinations – the state official responsible for AYP determinations suffered an injury that required a leave of absence (Deeter, personal communication, 3/5/13). This disruption and subsequent understaffing may have led to inconsistencies in Iowa's AYP determinations and may partially explain why Iowa's failure rates were extraordinarily low: less than 1% in 2003 and less than 5% in 2004.

[12] In 2004, Iowa used a uniform averaging procedure for both its proficiency and participation rates. If either the 2004 proficiency (participation) rates or the average of the 2003 and 2004 proficiency (participation) rates were greater than or equal to the proficiency target (95%), the subgroup met the proficiency (participation) target.

[13] In 2003 and 2004, Iowa tested students in grade 4, 8, and 11.

[14] This 27% estimate is actually conservative because we lack data on the size of Iowa's student subgroups. We apply the confidence interval formula by setting the subgroup size to 30, the minimum size for holding a subgroup accountable in Iowa. The actual, larger N's would yield smaller confidence intervals, so we may be overstating the number of subgroups that should have made AYP.

subgroup failed due to NCLB's 1% rule. These approvals prevented the failure of 22% of Texas schools (Hoff, 2005). In 2004, the U.S. DOE issued new guidance allowing states to petition to raise the 1% limit; in 2007, the U.S. DOE raised this limit from 1% to 2% (U.S. DOE, 2007).

c. *Applying a large confidence interval to safe harbor calculations.* NCLB gives states the option of applying safe harbor exceptions, as well as a further option to apply a 75% confidence interval to safe harbor calculations. Prior to 2005, Louisiana applied a 99% rather than a 75% confidence interval to its safe harbor calculations. This added increment helped more than 62% of otherwise failing economically disadvantaged subgroups, 79% of otherwise failing Black subgroups, and 90% of otherwise failing disabled subgroups avoid failing status.[15] Applying such a wide confidence interval adjustment to a safe harbor rule even allows some subgroups to make AYP when their proficiency rates *fell* instead of rose from the prior year. For example, the 31 fourth graders at McDonogh Elementary School #7 in Orleans Parish, LA, had a proficiency rate of 20% in ELA on state exams in 2002, which fell to 16.1% for the fourth graders in the same school in 2003. This 2003 performance failed to meet both the AYP ELA target of 36.9% and the lower target established by the confidence interval adjustment. To qualify for safe harbor without a confidence interval adjustment, the fourth grade group would need a 28% proficiency rate in 2003, representing a 10% reduction in the prior year's 80% failure rate. Louisiana's 99% confidence interval applied to this 28% target, however, set the safe harbor target rate at 7%, meaning the fourth grade 2003 proficiency rate could have met Louisiana's safe harbor criteria even if its proficiency rate was as low of 7%. The extremely generous confidence intervals applied to the safe harbor rule allowed McDonogh to make AYP even though its proficiency rate had actually declined by 4 percentage points.

[15] Reported figures are for math performance in 2003. The analogous figures for ELA performance are 49%, 57%, and 90%, respectively.

2. *States use more and less generous confidence interval adjustments*. States varied in the generosity of the confidence interval rules they adopted—ranging from no confidence intervals to 90, 95, or even 99%. States can reduce school failure rates by using larger confidence interval adjustments. Twenty-three states opted to use the maximum 99% confidence intervals. This typically meant that they used a 2.33 critical value, meaning a subgroup would still make AYP if their proficiency rate was within 2 times the standard deviation of the target proficiency rate. Yet failure rates in states with 99% confidence intervals were not substantially different from those in the fourteen states using 95% confidence intervals; in fact, the average state failure rate across 2004 and 2005 was slightly higher for the states using 99% confidence intervals (24% versus 21%).[16] The interaction of the other AYP decisions about continuous enrollment, minimum subgroup size, tested grade levels, and baseline proficiency rates helps to explain this counterintuitive result.

At the other end of the spectrum, four states did not employ any confidence interval adjustment at all—Florida, Ohio, South Carolina, and Virginia—and this dramatically increased their school failure rates as a result. The average failure rate in these states was 57% in 2003 and 44% in 2004. Florida identified over 80% of its schools as failing AYP in 2003. If Florida had instead applied even a 95% confidence interval that year, we estimate that 14% of its schools failing to meet proficiency targets would have instead made AYP.[17] Michigan applied 99% confidence interval adjustments but only for schools with very small campus-wide enrollments. If Michigan had instead applied 99% adjustments to all of its schools in 2004, we estimate that the percent of its schools failing to meet at least one proficiency target would have declined from 19% to 5%.

Some states altered their school failure rates by adjusting confidence interval policies over time. During the first two years of NCLB, South Carolina did not employ confidence interval adjustments on either absolute subgroup proficiency rates or safe

[16] For these calculations, we only include states that used standard confidence interval adjustments applied to both student subgroups and the overall student population.

[17] Florida also had low cutoffs for minimum subgroup size. Their limited English proficient, disabled, and Black subgroups had relatively low proficiency rates and were frequently held accountable: in 2003, these groups were accountable for math performance in 27%, 80%, and 68% of schools respectively. Florida's schools thus failed frequently and only 11% of them had at least one subgroup pass via safe harbor.

harbor calculations. In 2005, South Carolina amended its accountability system to include a one standard error band adjustment (i.e., a 68% confidence interval adjustment), and its school failure rate declined by ten percentage points that year.

Confidence intervals applied to safe harbor are another important source of cross-state variation in failure rates. Polikoff & Wrabel (forthcoming) describe how the number of schools making AYP due to safe harbor has increased over time in California, one of several states applying a 75% confidence interval to its safe harbor calculations. The vast majority of states allow at least some form of safe harbor, so cross-state differences are less about the presence of safe harbor and more about the generosity of the specific safe harbor policies.

3. *Some states adopt homogenous targets across grade levels whereas others do not.* As mentioned earlier, states were allowed to set grade-specific, subject-specific proficiency rate targets or could set uniform targets across grade levels and subjects. In most states, high school student proficiency rates were lower than those in younger grade levels. Because proficiency targets were based on pre-NCLB performance levels, states setting uniform targets may have thus been setting up relatively easy targets for elementary and middle schools to reach—particularly if high school students' proficiency rates lagged far behind. Texas and Pennsylvania provide examples of states with this policy and situation. In 2002, the proficiency rates in both Texas and Pennsylvania were at least 7 percentage points greater in elementary schools than in high schools for both ELA and math. These states' decision to use uniform targets across grade levels led to low failure rates among elementary schools. For Texas in 2004, only 1% of elementary schools failed to make AYP, 17% of high schools failed, and the overall failure rate was 6% of schools. Similarly, for Pennsylvania, only 7% of elementary schools failed to make AYP, 27% of high schools failed, and the overall failure rate was 15% of schools.

   Setting a more easily obtained proficiency rate target for elementary and middle schools relative to high schools can lower states' school failure rates for both computational and meaningful reasons. On the purely computational side, high schools are larger and less numerous than elementary schools, so a relatively low elementary school failure rate means a low proportion of *schools* failing AYP even though the

proportion of *students in schools* failing AYP may be much higher. But on a more substantive note, given the safe harbor policy, having fewer schools close to the margin for meeting their student proficiency rate targets can decrease school failure rates. Schools that expect to perform close to their proficiency rate targets do not benefit from a safe harbor policy—if their proficiency rates improve from the prior year than they would already be meeting their proficiency targets without using safe harbor. Safe harbor is more likely to enable schools to make AYP if schools' proficiency rates are nowhere near the targets to begin with. So, all else equal, states will have lower school failure rates if they have more (elementary and middle) schools that will easily meet their proficiency targets even if they also have more (high) schools that are nowhere near these targets, since some of these (high) schools might still meet AYP via safe harbor.

South Carolina was operating an interim accountability system in the initial year of NCLB that provides a counter example to Texas and Pennsylvania. South Carolina applied pre-NCLB proficiency rates of students in grades 3 to 8 to elementary, middle, and high schools, because South Carolina had not yet calculated high school proficiency rates for a sufficient number of prior years. Fewer students scored proficient or above in high schools than in elementary or middle schools, so applying the grades 3-8 proficiency rate as a baseline caused 97% of South Carolina's high schools to fail AYP in 2003. When separate targets were established for high schools in 2004, the high school failure rate decreased to 52%.

4. *States established different minimum subgroup sizes and held a different number of subgroups accountable.* The all or nothing nature of the AYP designations increases the risk of failure for schools with greater numbers of accountable student subgroups (Kane and Staiger, 2002, 2003; Simms, 2013). Within states, schools with a greater number of accountable subgroups were indeed more likely to fail AYP. Across states, there is a mild correlation between schools' average number of accountable student groups and their failure rates. Figure 5 displays this comparison for 2004. If we regress failure rates on the number of accountable student groups and this variable squared, then this produces an R-squared of less than .07 and the joint significance is .23.

   But Figure 5 also reveals that this relationship would be stronger if not for a few

outliers—the low failure rates in Louisiana, Montana, and Texas. With these three outlier states omitted, the R-squared from the quadratic term regression jumps to .14, with a joint significance of .05.[18] The other policy implementation decisions described above created exceptionally low failure rates in these three states. Louisiana had low cutoffs for minimum subgroup size and thus had a larger number of accountable subgroups per school, but used wide confidence intervals that, in combination with small subgroup sizes, made the effective proficiency target quite low. Texas used a uniform proficiency target across grade levels, resulting in extremely low failure rates among its elementary and middle schools. Montana did not use any minimum subgroup size, so subgroups would technically be held accountable even if there was only one student in that group. Montana's small schools and 95% confidence interval policy, however, meant that subgroups were so small that they would make AYP even with few students passing.

Because disabled subgroups' performance was often the only reason for a school failing to make AYP, one might expect states' policies toward disabled students to influence their schools' failure rates. The fraction of schools with accountable disabled subgroups will depend not only on states' minimum subgroup size rules but also on how they allocated disabled students across schools. School failure rates were initially higher in states with larger fractions of schools with accountable disabled subgroups. If we regress state failure rates on a quadratic for the fraction of schools with disabled subgroups accountable for math performance in 2003, then the R-squared is .13, with joint significance of .09 and adjusted R-squared of .08. But this relationship disappeared by 2004: the R-squared declined to .02, the joint significance was .70, and the adjusted R-squared was negative. States with higher fractions of disabled subgroups tended to mitigate this effect by having more generous confidence interval adjustments. In 2004, five of the eight states with the highest fractions of schools with disabled subgroups held accountable for math performance used 99% confidence interval adjustments.

5. *States defined continuous enrollment differently.* Five states—Hawaii, Illinois, Iowa, New Jersey, and Wisconsin—used starting dates for continuously enrolled students that

[18] The adjusted R-squared increases from .02 to .10 when these three states are omitted.

precede September of the school year of the testing.[19] In these states, students who have transferred schools prior to the first day of the school year will not affect their schools' AYP determinations. Two of these states, Hawaii and Wisconsin, chose early enrollment cutoff dates because they test students during fall months. If mobile students tend to be relatively low achieving, or if school districts tend to strategically wait to enroll students at particular schools (Jenning and Crosta, 2011), then these long required enrollment windows would make it easier for schools to make AYP. Aside from Hawaii, which had a high 2003 failure rate due to low participation rates and low proficiency rates in the disabled and limited English proficient subgroups, one may speculate that these five states would have had much higher failure rates if they used post-September enrollment cutoffs, since the fraction of students excluded from the accountable pool was sometimes quite high. In Wisconsin, for example, 14% of 4th grade students, 10% of 8th grade students, and 8% of 10th grade students were enrolled during test administration in November of 2003 but did not contribute to their schools' proficiency rate calculations, because they had not been enrolled in the same school since late September of 2002.

VII. Discussion

The early years of NCLB provide an important example of how variation in state policy implementation can cause a federal law to have very different consequences across the country. Discrepancies in states' AYP formulae teach us that details have important ramifications. Complex and off the radar of all but the most embedded policymakers and researchers, esoteric differences in rules had substantive impacts on schools due to the escalating sanctions under NCLB. Purposefully or not, some states took advantage of loopholes that made it much easier for schools to meet targets. Variation in these rules has only increased in recent years, as some states have received waivers allowing their schools to avoid failure designations even if their students do not reach 100% proficiency by 2014 (Riddle & Kober, 2012; U.S. Department of

[19] We thank Jennifer Jennings and Heeju Sohn for providing information on states' rules for continuous enrollment and testing dates, collected from state government websites.

Education, 2012). These waivers are idiosyncratic to each state, so that cross-state variation in the minutia of accountability policy rules is as complicated and important as ever (Polikoff et al., 2013).

While flexibility may be a positive aspect of NCLB or other school accountability systems, many of the discrepancies in states' NCLB rules reflect arbitrary differences in statistical formulae rather than substantive policy disagreements. When states and districts design test-based accountability policies, schools would be best served by a consistent set of directions about acceptable statistical practices and common definitions. Formulae for issues like continuous enrollment and safe harbor adjustments, if used, could be standardized. Even after statistical definitions are standardized, school accountability policies could still provide states and districts with discretion in their *substantive* choices of how to measure school effectiveness and which sanctions or rewards to attach to performance outcomes. Ideally, consequences for schools in an accountability system should be linked to student learning rather than the idiosyncrasies of state rules. This ideal might be better served if the federal government offered states a selection from a menu of accountability systems, while maintaining precise definitions and formulae within each of these systems.

References

American Institutes for Research (2005). National AYP and Identification Database (NAYPI). Washington, D.C. Data retrieved via webcrawl on November 12, 2008 from: http://www.air.org/publications/naypi.data.download.aspx.

Dee, T. & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management* 30(3), 418-446.

Dee, T., Jacob, B., & Schwartz, N.L. (Forthcoming). The effects of NCLB on school resources and practices. *Education Evaluation and Policy Analysis*. forthcoming, published online on Dec. 28, 2012.

Deeter, Tom. (2013). Iowa Department of Education. Personal communication via phone on 3/5/13.

Jennings, Jennifer & Crosta, Peter. (2011). The Unaccountables. paper presented at the 2011 conference of the Association for Education Finance and Policy.

Kane, Thomas J., and Douglas Staiger. (2003). "Unintended Consequences of Racial Subgroup Rules" in Paul E. Peterson and Martin R. West (eds.) No Child Left Behind? The Politics and Practice of Accountability. Washington, DC: Brookings Institution Press.

Kane, Thomas J., and Douglas Staiger. (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures" *Journal of Economic Perspectives 16*, 91-114

Ladd, H. & Lauen, D. (2010). Status Versus Growth: The Distributional Effects of Accountability Policies. *Journal of Policy Analysis and Management*. 29(3): 426-450.

Linn, R. L. (2005). "Conflicting Demands of *No Child Left Behind* and State Systems: Mixed Messages About School Performance." *Education Policy Analysis Archives*, 13(33). Retrieved May 31, 2006, from http://epaa.asu.edu/epaa/v13n33/.

Manna, P. (2010). *Collision Course: Federal Education Policy Meets State and Local Realities*.

Polikoff, M., and Wrabel, S. (forthcoming). When is 100% not 100%? The Use of Safe Harbor to Make Adequate Yearly Progress.

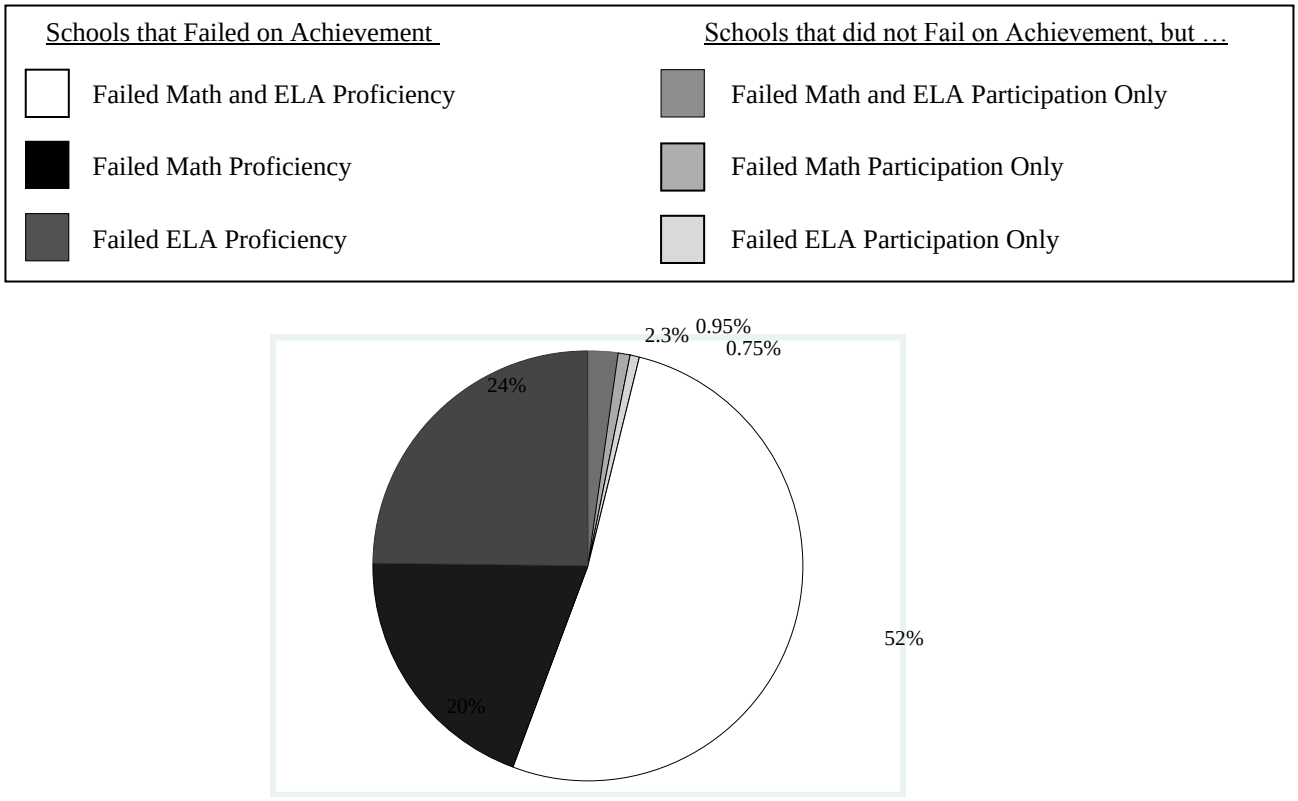
Polikoff, M., McEachin, A., Wrabel, S., and Duque, M. (2013). The Waive of the Future: School Accountability in the Waiver Error. Paper presented at the 2013 conference of the Association for Education Finance and Policy.

Reback, Randall, Rockoff, Jonah, & Schwartz, Heather L. (forthcoming). Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind. Forthcoming in the *American Economic Journal: Economic Policy*.

Reback, Randall, Jonah E. Rockoff, Heather S. Schwartz, and Elizabeth Davidson (2011), "Barnard/Columbia No Child Left Behind Database, 2002-2003 and 2003-2004," http://www.gsb.columbia.edu/nclb

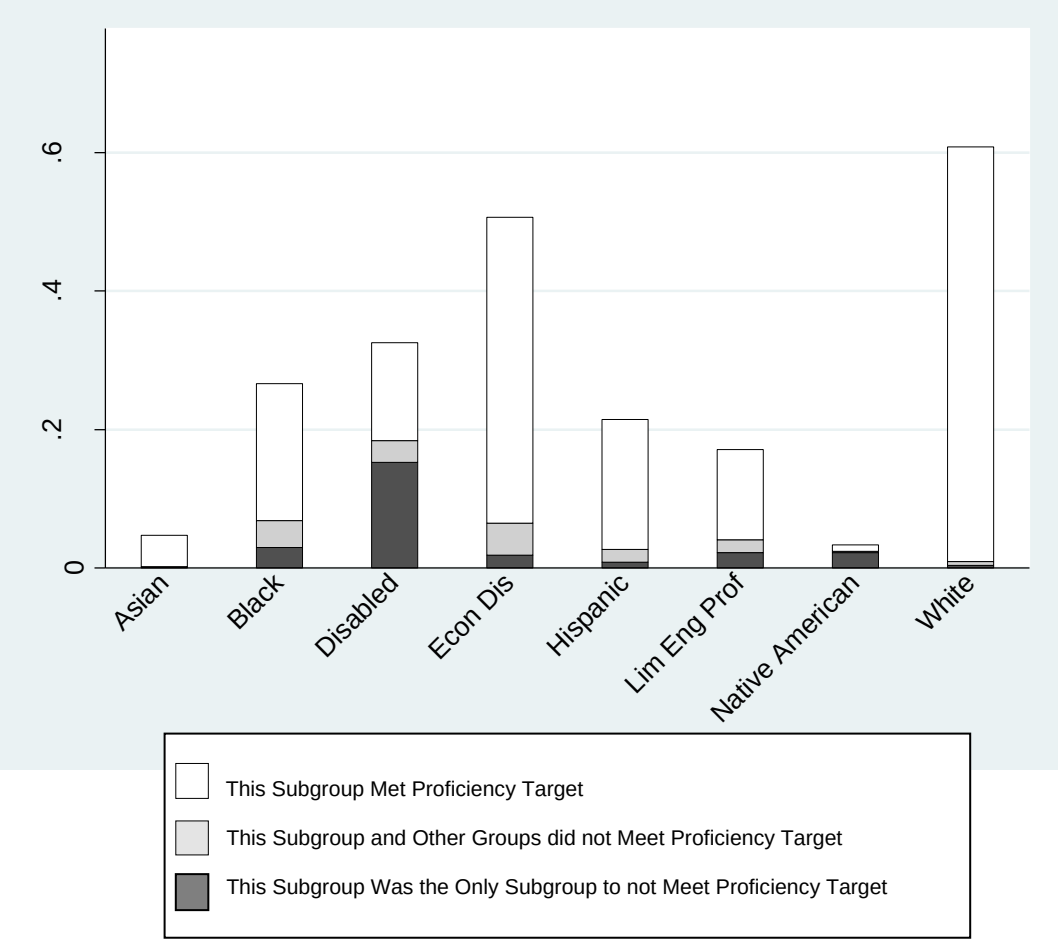
Riddle, W., & Kober, N. (2012). What Impact will NCLB Waivers have on the Consistency, Complexity and Transparency of State Accountability Systems" Center on Education Policy. Washington, D.C.

Figure 1: Percent of Schools that Failed by AYP Component, 2005



Notes to Figure 1: We exclude New York, Oklahoma, and Wyoming as these states are missing participation indicators in 2005.

Figure 2: Subgroup Accountability and Likelihood of Failure in Math, 2004



Notes to Figure 2: The figure is based on 46 states with available data. Iowa, North Dakota, Nebraska, and New Mexico are missing subgroup-level AYP data in 2004. The overall height of each column represents the fraction of schools where the subgroup's performance was held accountable.

Figure 3: Distribution of State Failure Rates, 2003 – 2005

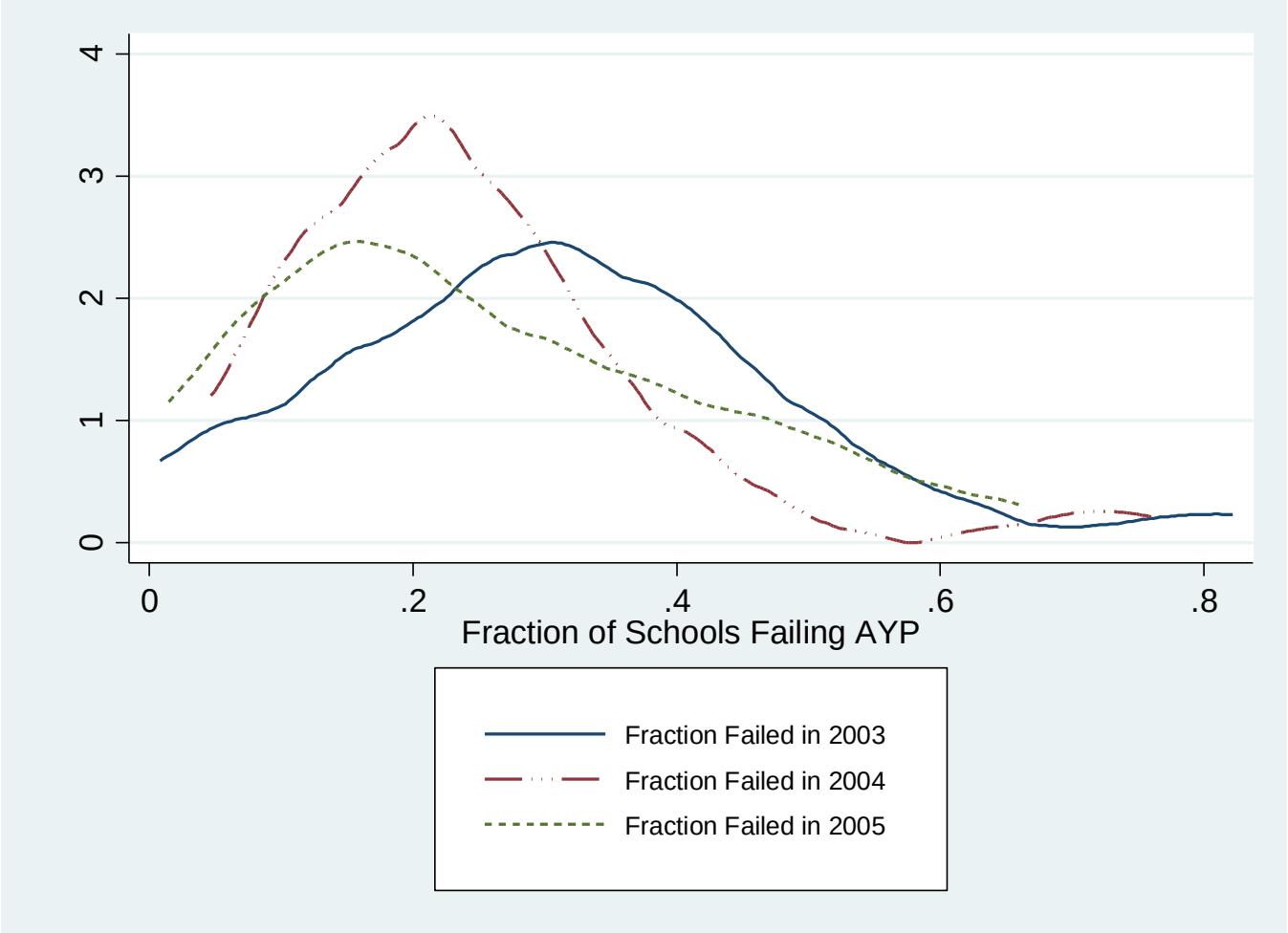
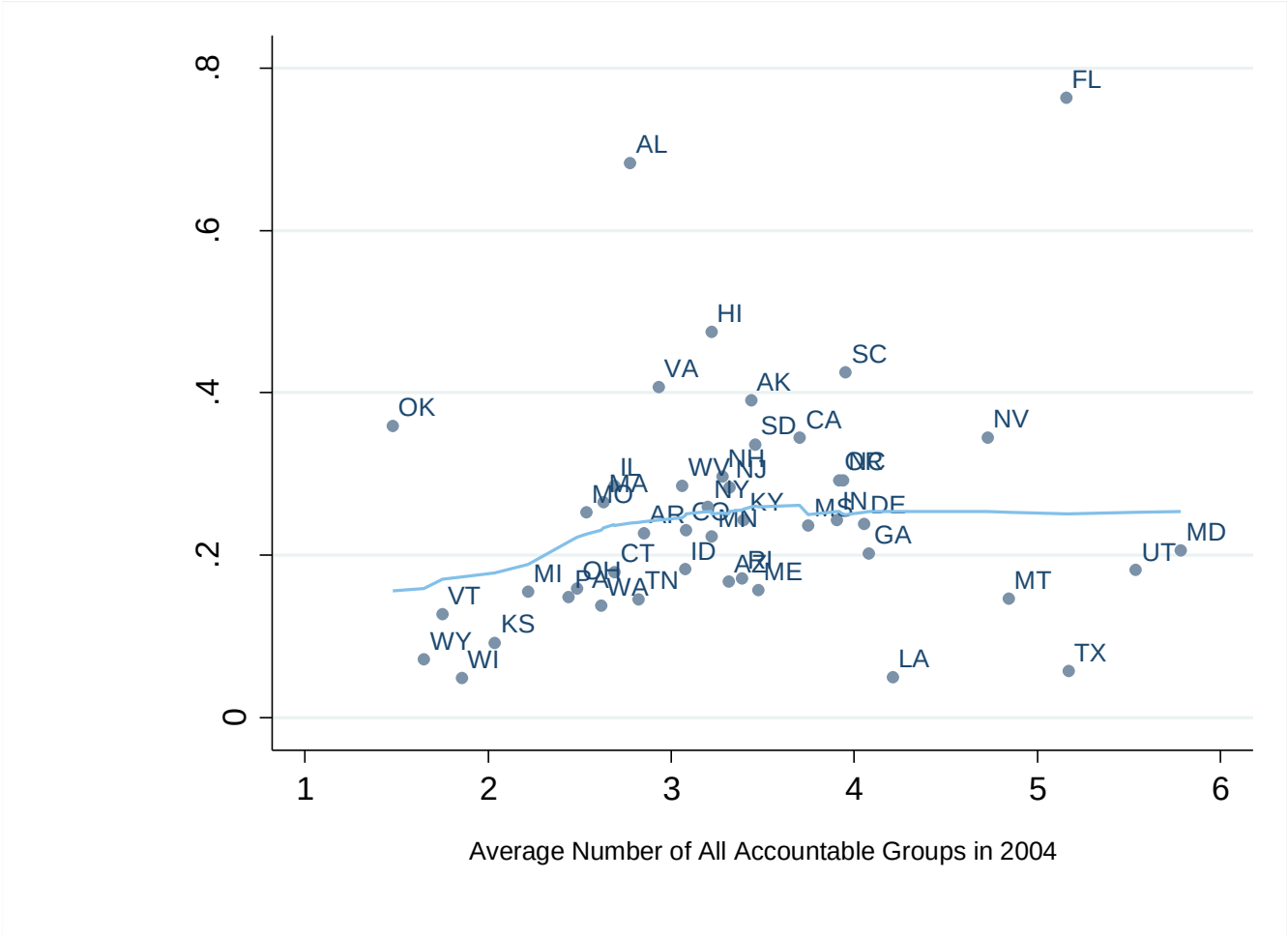


Figure 4: School Failure Rates vs. State Proficiency Rates in Math, 2004

Notes to Figures 4: N = 46 states in math and English Language Arts (ELA). Alabama, Nebraska, and New Hampshire are missing proficiency rates. Vermont reports a performance index in lieu of proficiency rates. When we aggregate proficiency rates to the state level for the x-axis, we weight schools by their number of tested students; in 11 states, we use schools' total school enrollment as reported in the Common Core of Data as a proxy for the number of students tested.

Figure 5: School Failure Rates vs. Average Number of Accountable Groups in Schools, 2004



Notes to Figure 5: Based on 46 states with available data. Iowa, North Dakota, Nebraska, and New Mexico are missing subgroup-level proficiency data in 2004. Accountable groups include both student subgroups and the overall student population. For each state, we take the average of the number of accountable groups for math achievement and the number of accountable groups for ELA achievement. For states that hold schools accountable separately for the grade-level performance of student subgroups, we accordingly treat each subgroup-by-grade-level as a separate group.

Table 1: Characteristics of Schools by Whether They Failed to Make AYP

	2003-2005		
	Failed all three years	Failed at least once	Never failed
Number of Schools	9,382	37,909	42,883
Average Enrollment	891	681	469
Student/Teacher Ratio	17.6	16.5	15.7
Percent of Students…			
Eligible for Free/Reduced Lunch	55.0%	49.5%	34.1%
White	39.3%	52.1%	73.9%
Black	29.9%	23.3%	9.9%
Hispanic	23.8%	18.3%	11.4%
Asian	4.0%	3.4%	3.4%
Percent of Schools…			
Eligible for Title I	67.9%	61.0%	44.9%
Serving Primary Grades	32.8%	46.7%	71.5%
Serving Middle Grades	35.2%	25.7%	14.2%
Serving High Grades	31.9%	27.6%	14.3%
Located in City	41.2%	31.1%	18.3%
Located in Suburb	32.8%	30.5%	33.9%
Located in Town or Rural Area	24.4%	33.6%	46.7%

Notes to Table 1: The data on school characteristics are from the Common Core of Data, 2001-2002. For schools in Tennessee, data on student ethnicity comes from 1998-99 instead of 2001-2002 and data on free/reduced price lunch eligibility is unavailable.