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
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* (NLCB) *Act* required states to adopt accountability systems measuring student proficiency on state administered exams. Based on statewide student test score performance in 2002, states developed initial proficiency rate targets and future annual benchmarks designed to lead students to 100% proficiency on state exams by 2014. Any year a school fails to meet these targets, either across all students or by various subgroups of students, the school is does not make Adequate Yearly Progress. While the federal government's legislation provided a framework for NCLB implementation, it also gave states flexibility in their interpretation of many NCLB components, and school failure rates ranged from less than 1% to more than 80% across states. In this paper, we explore how states' NCLB implementation decisions affected their schools' failure rates. Wide cross-state variation in failure rates resulted from how states' decisions (e.g, confidence intervals applied to proficiency rates, numerical thresholds for a student subgroup to be held accountable) 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.

This research project was made possible by funding from the Institute for Education Sciences and the Spencer Foundation, as well as seed grants from the Columbia University Institute for Social and Economic Research and Policy and Barnard College, and support from the Paul Milstein Center for Real Estate at Columbia Business School. The authors are solely responsible for any opinions or errors in the paper. We thank participants in the APPAM/INVALSI/UMD 2012 conference in Rome, Italy, "Improving Education through Accountability and Evaluation," for their helpful comments. We also thank participants in the 2013 conference of the Association for Education Finance and Policy.

I. Introduction

The United States has historically relied almost exclusively on local and state governments to monitor the educational needs of its citizens. In the latter part of the 20th century, however, the federal government increased its role in education with the No Child Left Behind Act of 2001 (NCLB). NCLB requires states to construct accountability systems that use standardized tests to measure student and subgroup proficiency rates in math and English language arts (ELA). NCLB originally required states to create a timeline of annual achievement benchmarks that gradually increased to 100% of students scoring proficient or higher on state exams by 2014. NCLB holds schools accountable for the proficiency rate of their total student population and the proficiency rates of various student subgroups: students from low-income families, students with disabilities, students with limited English proficiency, and students in various racial/ethnic groups. A school fails to make Adequate Yearly Progress (AYP) if any one subgroup fails to meet its yearly proficiency rate target in either math or ELA (English Language Arts).

In addition to the potential stigma of failing to make AYP, schools receiving federal Title I funds (because they serve students from low-income families) face defined consequences. Students at failing Title I schools are afforded the opportunity to transfer to other, non-failing schools within the same district. After consecutive years of AYP failure, these schools' students from low income families are also 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 (removal of the principal, etc.).

From NCLB's inception, federal policymakers avoided a "one size fits all" policy and encouraged states to adapt NCLB guidelines to meet the demands of their particular contexts. Rod Paige, U.S. Secretary of Education in 2002, wrote in a letter to Virginia legislators, "The law and subsequent regulations were intentionally written...to provide maximum flexibility to states while at the same time working toward

academic achievement for all children" (Hoff, 2004b). Consequently, NCLB guidelines, while detailed, provided significant leeway for state interpretation. NCLB allowed states to choose students' ELA and math exams, define proficiency levels on those exams, set yearly performance benchmarks, determine the minimum number of students that must belong to a subgroup for that subgroup to be accountable under AYP rules, select grade levels to test, and incorporate safeguards that provide flexibility in determining schools' ratings.

The early years of NCLB thus provide an important example of how variation in state policy implementation can cause a federal law to have very different consequences across the country. Figure 1 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.

Contrary to what one might expect, this wide variation in failure rates is not strongly related to student performance on states' high-stakes exams. Figures 2a and 2b illustrate this relationship for student performance on states' math and ELA exams for the spring of 2004. Based on corresponding linear regressions, a one percentage point increase in state math proficiency rates is associated with only a 0.1 percentage point decline in the fraction of a state's schools making AYP, while a one percentage point increase in state ELA proficiency rates is associated with only a 0.2 percentage point decline in this fraction. The former estimate is statistically insignificant, while the latter is significant at the .10 level.[1]

States' school failure rates are not strongly related to their students' performance because states determined early NCLB proficiency targets based on state-specific pre-NCLB student proficiency rates. State were thus grading their own schools on state-specific curves, with varying starting points and trajectories for school performance targets. These policy implementation differences, rather than real

[1] If we regress states' school AYP failure rates on quadratic terms for their states' proficiency rates in each subject (i.e., 4 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.

differences in student achievement or exam difficulty, caused the wide variation in states' school failure rates. As policymakers continue to debate reforming school accountability systems, it is critical to understand how the first-ever national school accountability law led to such different outcomes across states.

This paper describes which schools failed during the first three years of NCLB and which performance targets they failed to meet. We also examine the implementation decisions and school characteristics that drive differences in failure rates across states. We find that the wide cross-state differences in failure rates were largely the result of subtle differences in states' own NCLB rules. Complex and off the radar of all but the most embedded policymakers and researchers, these esoteric rules have substantive impacts on schools due to the escalating sanctions under NCLB. While some policymakers might argue that state flexibility is a positive aspect of NCLB, many of the discrepancies in states' rules reflect arbitrary differences in statistical formulae rather than substantive policy disagreements. Consequently, states' AYP failure rates may reflect policy idiosyncrasies more than true differences in educational performance. While the U.S. Department of Education issued guidance to rectify some of the most glaring discrepancies from the early years of NCLB, most of the variation in states' rules remains to this day. Attention to the ramifications of these remaining differences can help improve future federal and state school accountability policies by standardizing certain statistical practices and closing loopholes.

The paper is organized as follows. Section II provides an overview of NCLB rules and describes variation in states' implementation. Section III describes challenges faced by states in implementing NCLB, and Section IV discusses the nature and availability of school-level NCLB data. Section V describes the most common targets that schools failed to meet under NCLB. Section VI explores the extent to which variation in policy implementation caused the wide cross-state differences in schools' failure rates. Section VII concludes with a brief discussion of the implications of these findings for local and state policymakers developing school accountability systems and for federal policymakers seeking to reform NCLB.

II. NCLB Overview

Passed with significant bi-partisan support in Congress, NCLB re-authorized the Elementary and Secondary Education Act of 1965 and was the federal government's biggest foray into K-12 education since the original version of that Act. NCLB was signed by President George W. Bush in January of 2002, and it required states to administer baseline student exams in the spring of 2002 and to adopt school accountability systems for the school year 2002-03. NCLB required significant state investment to establish or overhaul school accountability systems and, for the first time, states risked forfeiting federal funding if they failed to comply.

Measuring Performance

School performance under NCLB is based on three distinct measures: student performance on a state-selected exam, rates of participation on those exams, and an additional state-selected indicator of performance. Both the campus as a whole and various student subgroups—students with disabilities, limited English proficient, white, Black, Asian, Hispanic, American Indian—must meet each of these three measures. Here we describe each of these measures and the flexibility given to states to determine them.

States selected their exams and defined proficiency on those exams with complete flexibility. States then determined a schedule for the percentage of students who must meet proficiency each year from 2003 to 2013; a target of 100% proficiency is required for 2014. NCLB specified that states must set their 2003 benchmark as the higher value of (a) the 2001-2002 statewide proficiency rate of the lowest performing student subgroup or (b) the 2001-2002 proficiency rate at the 20th percentile school, weighted by student enrollment, when schools are sorted by performance.[2] The higher value was typically (b), so that a state's performance benchmarks were aligned to 2002 test scores such that, if 2003 test scores resembled 2002 test scores, then approximately 20% of the state's students attended schools whose average proficiency rate in

[2] These rules are stated in No Child Left Behind (2002), Sec. 1111 (b)(2)(E)(ii). They are summarized by the U.S. Department of Education's (2002) desktop reference manual.

2003 was below the proficiency target. States could set different benchmarks by grade level and by subject area but not by student subgroup.

Even if a subgroup's or school's performance falls below the proficiency target, the school may still make AYP because NCLB allows states to employ various statistical techniques and contingencies to adjust proficiency rates. 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. States varied in the generosity of the confidence interval rules they adopted—ranging from no confidence intervals to 90, 95, or even 99%. The larger the confidence interval applied, the easier it is for a subgroup to make AYP. For instance, in Alaska, the 2003 ELA proficiency target was 64%, but a subgroup with 20 students would only have to reach 39% proficiency in order to meet AYP due to the state's 99% confidence interval.

Another source of variation in the relative difficulty of making AYP are "safe harbor" rules designed to reward schools that missed proficiency targets but had students make large gains in proficiency rates from the previous year. To make AYP under the safe harbor rule, states typically require a 10% reduction in the fraction of students failing to reach proficiency. Some states also apply confidence intervals to safe harbor rules, making it easier for schools to meet these alternative standards for annual growth in proficiency rates (Spellings, 2005). Importantly, safe harbor measures are not typically based on changes in the performance of the same cohort of students, but rather the change over time in the average performance of a subgroup (e.g., Asian students in 4th grade). Consequently, proficiency rate changes from one year to the next may reflect the composition of students rather than instructional quality.

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. While participation rate calculations have fewer contingencies than proficiency calculations, 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.

The final category of school performance is the state-selected "other" academic indicator. NCLB rules allow for flexibility in states' selection of elementary and middle schools' other indicator. Most states decide to hold these schools accountable for their attendance rates or, less commonly, performance on an additional exam. For high schools, NCLB rules require states use the high school graduation rate but permit states to use their own definition of that metric.[3] Unlike student proficiency rate targets, performance targets for the state-selected "other" indicator are not required to increase over time.

Defining Student Groups

Schools' performance measures are based on both the overall student population *and* various subgroups of students: racial/ethnic subgroups, students eligible for free/reduced priced lunch, students with limited English language proficiency, and disabled students.[4] Students are counted in all subgroups to which they belong. For example, a Hispanic student who is an English Language Learner and eligible for free

[3] In December 2008, the U.S. Department of Education announced that all states must use a standardized four-year graduation rate formula. 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] The most common subgroups are Asian/Pacific-Islander, African-American, Economically-Disadvantaged, Limited English Proficient, Disabled/Special Education, Hispanic, Native American, and White students.

lunches will contribute to eight different proficiency rates—the campus-wide group, the Hispanic subgroup, the English Language Learner subgroup, and the free/reduced priced lunch subgroup proficiency rates in math and ELA. But these subgroup proficiency rates only influence the school's AYP rating if there are sufficient numbers of students who meet the continuous enrollment definition within the school.

To protect student anonymity and avoid using unreliable measures of subgroup performance, states 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.

NCLB also allowed for wide differences in how states administered tests across grade levels and how they aggregated results across grade levels. States were required to test students in grades 3 through 8 and in one high school grade by the 2005-2006 school year, but before this they were only 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 when it introduced 3rd grade exams and 2005-2006 when it tested grades 3-8. For schools that serve multiple tested grade levels, states could also 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. On the other hand, 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.

This policy treatment would lead to fewer accountable subgroups and more generous confidence interval adjustments.

Appeals Process

Beyond the formal NCLB rules, states 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.

III. State Implementation Challenges

The complexity of NCLB rules and large data demands taxed state education agencies' capacity to implement NCLB. Some states had difficulty interpreting NCLB guidelines, meeting federal deadlines, or reconciling their existing state accountability plans with the new federal requirements. State policy-makers reported that they lacked the necessary guidance from the U.S. Department of Education (USDOE) to make informed implementation decisions (Minnici & Hill, 2007; Olson & Robelen, 2002). Many states reported insufficient technical expertise and financial resources to meet NCLB's substantial data collection and reporting requirements and to support under-performing schools (Center on Education Policy, 2004; Goertz, 2005). NCLB was different from previous legislation because the law was enacted without an implementation schedule. States had only a matter of months to build their new accountability system – "even before the [US] department [of education] completed the regulation-writing process" (Manna 2006, p. 482; Erpenbach, et al., 2003). Most states immediately submitted waivers for exemption from some NCLB components (Manna, 2006).

Most states created a Technical Advisory Committee to advise their state department of education on an appropriate confidence interval, minimum subgroup size, how to phase in their testing system, and how to account for small schools whose subgroups all fall below the minimum size. States also invested in resources to ensure that principals and teachers understood testing protocols and NCLB rules and regulations. In addition, school districts had to process data into school report cards and make those report cards publicly available. Estimates suggest these expenditures far exceeded the additional revenue provided by the federal government after the passage of NCLB (Goertz, 2005; Driscoll & Fleeter, 2003).

These capacity constraints were compounded by the variation in states' decision-making structure. States' organizational hierarchy has multiple layers – the governor and legislature, boards of education and/or education agencies, and chief state school officers/superintendents. As NCLB implementation decisions vary by state, so do the parties responsible for making those decisions, and power struggles within states may have complicated implementation (Manna, 2003). Brewer, Kileen, and Welsh (forthcoming) provide Oklahoma as a case study in how resource constraints, technical capacity limitations, and political forces limited the implementation of a meaningful state accountability system under NCLB.

To ease the implementation process, states were hoping for flexibility on their interpretation of NCLB guidelines. Soon after the law went into effect, at least fourteen states petitioned the government for increased flexibility or exemption from certain NCLB statutes (Schemo, 2004).[5] These states argued that, under NCLB's structure, most of their schools would be labeled as failing in only a few years. Some states also felt their existing accountability systems met the intent of the law, and they should only be required to make minor modifications to these systems. For example, Louisiana argued that their existing plan had closed achievement gaps between white and black students, and was better than NCLB at identifying low-performing schools (Olson, 2002).

[5] Utah was the only state that seriously threatened to ignore NCLB altogether. Utah's Congressional Panel on Education voted to propose a bill to the state House of Representatives that would deny the federal order to implement NCLB and thus reject federal Title I funds (Hoff, 2004b). Ultimately, Utah complied with NCLB.

After the USDOE refused states' requests to maintain or modify their existing state accountability systems, several states operated their state and federal accountability systems simultaneously. Often, states faced difficult decisions about which system deserved priority, and schools received "mixed messages" concerning their performance ratings (Hoff, 2004a). In Colorado, nearly half of the state accountability system's "low" rated schools made AYP; of the schools that ranked "high," over 13% failed AYP (Linn, 2005). North Carolina's AYP school failure rate was about 45% in 2003, but only about 2% of schools failed the states' own accountability system in that same year (Ladd & Lauen, 2010).

IV. 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. NCLB's reporting requirements applied differently to states and school districts. State were only required to publish: (a) annual statewide report cards that reported statewide proficiency rates—both overall and disaggregated by student subgroups, and (b) annual lists of "schools needing improvement," the schools receiving Title I funds that failed to make AYP. Districts were required to "prepare and disseminate annual report cards that include information on student achievement for the district and for each school (U.S. DOE, 2002, 4)." Someone seeking data on school-level NCLB performance outcomes across multiple school districts might have to obtain that information separately from each school district. In practice, many states directly published school-level report cards or maintained these data in publicly available spreadsheets, saving their districts from this reporting requirement. Since states were not required to do this, however, the content in these reports and data sets varies across states. By putting such lenient reporting requirements on states, Congress arguably missed an important opportunity to create a nationwide data set with accurate, standardized NCLB performance data.

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 websites. The resulting data and our documentation of sources are publicly available.[6] For the 2002-2003 and 2003-2004 school years, 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).

V. Descriptive Evidence on Failing Schools

There are clear observable differences between AYP failing and non-failing schools if we look nationwide from 2003 to 2005 (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 3 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 were tested.[7] Among schools failing to make AYP

[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/.

[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.

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 4 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, failure rates were quite varied across subgroups. Conditional on being accountable, the 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.

VI. Explaining cross-state differences in failure rates

With these school-level sources of failure in mind, we next investigate the reasons for the wide variation in AYP failure rates across states. Basing benchmarks on prior student performance levels created wide variation in proficiency targets across states. For example, Iowa set 2003 proficiency targets at 64% in math and 65% in ELA, while Missouri chose 8.3% and 18.4%, respectively. Even states with similar starting 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. These disparities occurred because proficiency targets were only one of many elements that influenced failure rate determinations.

Failure rates changed substantially over time in some states. Alabama's failure rate jumped from 4% in 2003 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.

Various dimensions of NCLB implementation led to variation in school AYP failure rates.[9] No individual state policy decision appears to be the primary culprit; 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 had substantial impacts on some states' school failure rates. 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. Many of these processes relied heavily on Iowa school districts to provide accurate student counts and performance data. During the summer of 2004– the months when state officials typically make AYP determinations – the state

[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] 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.

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. Using proficiency rate and participation rate data we retrieved from Iowa's Department of Education website, we applied Iowa's AYP formula and found higher failure rates. First, consider the participation rates. 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.[10] 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. In 2003, we find that 42 (4%) of Iowa's schools made AYP even though their overall student participation rate was below 94 percent. Next, consider the proficiency rates. Because data disaggregated by grade level is unavailable for Iowa, we use proficiency rates for the 90% of Iowa's schools that served only one tested grade level.[11] 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.[12]

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 USDOE denied their request. In 2003, the Texas State Education Agency ignored the USDOE's ruling and approved the appeals of 1,718 schools whose special education subgroup failed due to NCLB's

[10] 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.

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

[12] This 27% estimate is actually conservative because we lack data on the size of Iowa's student subgroups. We apply the following confidence interval formula to subgroups' proficiency rates (X): . Here, N = 30, Iowa's minimum subgroup size for holding the subgroup accountable. The actual, larger N's would yield smaller confidence intervals, so we may be overstating the number of subgroups that should have made AYP.

1% rule. These approvals prevented the failure of 22% of Texas schools (Hoff, 2005). In 2004, the USDOE issued new guidance allowing states to petition to raise the 1% limit; in 2007, the USDOE raised this limit from 1% to 2% (U.S. Department of Education, 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.[13] 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.

2. *Confidence Interval Adjustments*. 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 (see footnote 12 for an example). Failure rates in states with 99% confidence intervals were not substantially

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

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 percent).[14] The interaction of the other AYP decisions about continuous enrollment, minimum subgroup size, tested grade levels, baseline proficiency rates helps to explain this counterintuitive result.

At the extreme, 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 has 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.[15] 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 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.

---

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

[15] Florida also had low cutoffs for minimum subgroup size. Their LEP, 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.

3. *Homogenous Targets Across Grade Levels.* 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. On a more meaningful 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 and provides an example at the other end of the spectrum. 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. *Minimum Subgroup Size and the Number of Accountable Subgroups.* 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.[16] 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

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

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. *Definition of Continuous Enrollment.* Five states—Hawaii, Illinois, Iowa, New Jersey, and Wisconsin—use starting dates for continuously enrolled students that precede September of the school year of the testing.[17] 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 will 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 LEP 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

There is a long tradition of local control over education, and NCLB's reach into state and local education policy has received criticism from policymakers and educators. Current research evidence on the

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

overall effects of NCLB and other school accountability programs, however, tends to suggest small positive or at least neutral effects along most measured outcomes. Estimated effects of accountability pressure on student achievement tend to be positive, though high-stakes exam performance may improve for some students more than others (e.g., Booher-Jennings, 2005; Krieg, 2008; Reback, 2008; Neal & Whitmore Schanzenbach, 2010) and gains on low-stakes exams may be relatively modest (Jacob, 2005; Dee and Jacob, 2011; Reback, Rockoff, & Schwartz, 2013). Potentially harmful effects of NCLB pressure may include changes in teachers' sorting across public schools and exits from the profession (Feng, Figlio, & Sass, 2010; Reback, Rockoff, & Schwartz, 2013), principals' sorting across schools (Li, 2012), and the amount of time devoted to non-tested subjects (Dee, Jacob, & Schwartz, forthcoming; Reback, Rockoff, & Schwartz, 2013).[18] These responses are particularly problematic if they disproportionately hurt the schools most likely to fail under accountability systems. Yet most studies have found positive net effects, at least in the short run, for student performance on high-stakes exams after schools receive unfavorable ratings (Figlio & Rouse, 2006; Chakrabarti, 2007; Rouse et al., 2007; Chiang, 2009; Rockoff & Turner, 2010).[19]

The current policy trend is towards expanding rather than contracting school accountability policies. States and districts have been rapidly adopting new accountability policies that apply directly to teachers, principals, and superintendents. Many of these policies beg the same questions that NCLB does—e.g., which students should count towards ratings, which grade levels should be tested, which subjects and topics should be tested, what are acceptable adjustments to make to accountability ratings.

---

[18] Other unintended effects of school accountability systems can include cheating among teachers (Jacob & Levitt, 2003), re-classification of students (into special education programs, etc.) to remove them from the testing pool (Figlio & Getzler, 2006; Cullen & Reback, 2006), strategic re-assignment of students to schools in order to prevent them from being continuously enrolled (Jennings & Crosta, 2011) and changing school disciplinary practices (Figlio, 2006). Another effect of school accountability systems, possibly intended, is to alter school expenditure levels. Dee, Jacob, and Schwartz (Forthcoming) find that states lacking strong school accountability systems prior to NCLB used increases in local and state education expenditures to reduce gaps in their teachers' salaries compared to other states after NCLB was adopted.

[19] Not all studies have produced such encouraging findings. For example, Sims (2012) finds that students in California subsequently perform worse if their schools faced a greater likelihood of failure due to student subgroups barely large enough to be held accountable under California's minimum subgroup size rules.

Discrepancies in AYP formulae teach us that details have important ramifications. They also reveal that, purposefully or not, some states took advantage of loopholes that made it much easier for schools to meet targets. 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 should be standardized.

As Congress considers reauthorizing NCLB, there is tension between adopting a truly standardized national school accountability policy and allowing states to experiment with their own policies. As we have documented in this paper, the rules applied to schools under NCLB were far less uniform than many observers would expect. 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 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).

In our view, allowing subtle and often arbitrary differences in state policies is undesirable when under the guise of a uniform national school accountability system. Ideally, the rewards and consequences in a school 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 of several types of accountability systems, while maintaining exact definitions and formulae *within* each of these systems.

When considering alternative systems of school accountability, various advocates have proposed a number of policy features to improve upon the original version of NCLB. There is currently limited research evidence concerning the relative merits of these policy options. Rather than advocating for any particular approach, we mention three commonly-proposed alternatives that one might expect to be part of the debate for the reform of school accountability systems. If any of these are offered to states as options, then we would

hope that federal policy would later refine states' options as additional research evidence emerges on these new systems and as educators' experiences come to light.

First, some advocates argue that revised accountability systems should ensure that schools face more continuous incentives to improve a wider array of outcomes for students. New accountability systems might provide schools with incentives to improve student performance regardless of students' prior achievement levels and the scaling of their test scores.[20]

Second, others argue that accountability systems might extend beyond the use of standardized student test scores by incorporating inspectorate visits (Ladd, 2010; Hussain, 2012) or other means of assessing learning environments. These assessments of educational processes, as opposed to educational outcomes, could either be merged with high-stakes assessments or done separately as a means of assisting schools with their self-evaluations.[21]

Third, some advocates argue for reforming the labeling of school ratings and the consequences for schools that fail to meet targets. The current consequences are mostly punitive in terms of their impact on school resources, such as allowing students to transfer out of the school or to use school funds for private tutoring services. These advocates point to programs like School Improvement Grants which dedicate additional resources to schools that are in danger of closure due to repeated low performance.[22]

Whatever its form, a successful school accountability system should have some key qualities. It should encourage educators to align their instruction with important learning standards, foster the adoption of instructional techniques that produce measurable results in student learning, improve the quality of educators

[20] See Barlevy and Neal (2010) for a detailed discussion and analysis of one example of such a system. They propose a system for rating teacher performance, but a similar system could instead be used to assess school performance.
[21] Critics of these approaches argue that process measures are relatively costly and/or that attaching high stakes to process measures could corrupt their validity and undermine their utility as a diagnostic tool (Hamilton et al., 2013).
[22] School Improvement Gants originated in the Elementary and Secondary Education Act and were greatly expanded by the 2009 American Recovery and Reinvestment Act (Hurlburt et al, 2011).

entering and remaining in schools, and inform educators and the public about how well students are learning various subjects and skills. Given the inherent complexities in measuring student learning, a school accountability system should be broadly conceived in terms of performance goals yet sensibly standardized in terms of how performance is measured.

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.

Barlevy, G. & Neal, D. (2012). Pay for Percentile. *American Economic Review* 102(5), 1805-31.

Brewer, D., Killeen, K., & Welsh, R.O. (forthcoming). The Role of Politics and Governance in Educational Accountability Systems. *Education Finance & Policy*, forthcoming.

Center on Education Policy. (2004). From the Capitol to the Classroom: Year 2 of the No Child Left Behind Act. Washington, D.C.

Chakrabarti, R. (2007). Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida. *Federal Reserve Bank of New York Staff Reports*, no. 306.

Council of Chief State School Officers (2005). School Data Direct. Washington, D.C. Data retrieved on December 2, 2007 from: http://www.schooldatadirect.org.

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.

Driscoll, W., & Fleeter, H. (2003). *Projected costs of implementing the federal "No Child Left Behind Act" in Ohio.* Prepared for the Ohio Department of Education. Columbus, OH: Levin, Driscoll & Fleeter.

Erpenbach, W. J., Forte-Fast, E., and Potts, A. (2003). *Statewide Educational Accountability Under* NCLB*: Central Issues Arising from an Examination of State Accountability Workbooks and U.S. Department of Education Reviews Under the* No Child Left Behind Act of 2001. Washington, D.C.: Council of Chief State School Officers (CCSSO). Retrieved May 31, 2006, from: http://www.ccsso.org/content/pdfs/StatewideEducationalAccountabilityUnderNCLB.pdf.

Feng, L., Figlio, D., and Sass T. (2010) "School Accountability and Teacher Mobility," NBER Working Paper 16070.

Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics 90*, 837-851.

Figlio, D. & Getzler, L. (2006). Accountability, ability, and disability: Gaming the system? In T. Gronberg & D. Jansen (Eds), *Advances in Applied Microeconomics*, *14*.

Figlio, D. & Rouse, C. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics 90*, 239-255.

Goertz, M.E. (2005). Implementing the No Child Left Behind Act: Challenges for the States. *Peabody Journal of Education*, 80(2), pp. 73-89.

Hamilton, L., Schwartz, H., Stecher, B., and Steele, J. 2013. Improving Accountability through Expanded Measures of Performance. *Journal of Educational Administration* 51(4).

Hoff, D.J. (2004a). "Accountability Conflicts Vex Schools." *Education Week,* 23 (26), 1.

Hoff, D. J. (2004b). Utah panel votes to quit No Child Left Behind Act. EducationWeek, p. 11.

Hoff, D. J. (2005). Texas stands behind own testing rule. *Education Week,* pp. 1, 23.

Hurlburt, S., Le Floch, K.C., Therriault, S.B., and Cole, S. (2011). *Baseline Analyses of SIG Applications and SIG-Eligible and SIG-Awarded Schools* (NCEE 2011-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hussain, I. (2012). Subjective Performance Evaluation in the Public Sector: Evidence From School Inspections. Unpublished Manuscript. University of Sussex.

Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5-6), 761-796.

Jacob, B. & Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics 118*(3), 843-877.

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

Krieg, John M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy* 3(2): 250-281.

Ladd, H.F. (2010). Education Inspectorate Systems in New Zealand and the Netherlands. *Education Finance and Policy* 5(3), 378-392.

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. (2003). States as Middle Managers in the American Policy System: Control, Management, and the Challenge of Leaving No Child Behind. Paper presented at: The Association for Public Policy Analysis and Management. November, 2003. Washington, D.C.

Manna, P. (2006). Control, persuasion, and educational accountability: Implementing the No Child Left Behind Act. *Educational Policy* 20(3): 471-494.

Minnici, A., & Hill, D. (2007). Education Architects: Do State Education Agencies Have the Tools Necessary to Implement NCLB?. Center on Education Policy. Washington, D.C.

No Child Left Behind Act of 2001. (2002). Pub. Law No. 107-110.

Olson, L., & Robelon, E. (2002). Frustration grows as states await 'adequate yearly progress' advice. *Education Week*: Vol. 21, Issue 42, pp. 1-41.

Olson, L. (2002). States anxious for federal guidance on yearly progress. Education Week, pp. 1, 14, 15.

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. (2013). Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under *NCLB*. mimeo, Columbia University.

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.

Rockoff, Jonah H., and Turner, Lesley J. (2010). Short run impacts of accountability on school quality. *American Economic Journal: Economic Policy, 2(4): 119-147.*

Rouse, C., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *National Bureau of Economic Research*, working paper 13681.

Schemo, D. (2004). Fourteen states ask U.S. to revise some education rules. *New York Times.* Retrieved on January 24, 2012 from: http://www.nytimes.com/2004/03/25/education/25CHIL.html

Sims, D. (2013). Can failure succeed? Using racial subgroup rules to analyze the effect of school accountability failure on student performance. *Economics of Education Review 32*: *262-274.*

Spellings, M. (2005). No Child Left Behind: *A Road Map for State Implementation*. United States Department of Education. Washington, D.C.

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

U.S. Department of Education. (2002). Office of Elementary and Secondary Education, *No Child Left Behind: A Desktop Reference*, Washington, D.C., 2002.

U.S. Department of Education. (2007). State and Local Implementation of the No Child Left Behind Act: Volume V—Implementation of the 1 Percent Rule and 2 Percent Interim Policy Options. Retrieved on April 4, 2013 from: http://www2.ed.gov/rschstat/eval/disadv/nclb-disab/nclb-disab.pdf.

U.S. Department of Education. (2008). High School Graduation Rate: Non-Regulatory Guidance. Retrieved on March 5, 2013 from: http://www2.ed.gov/policy/elsec/guid/hsgrguidance.pdf.

U.S. Department of Education. (2012). President Obama: Our Children Can't Wait for Congress to Fix No Child Left Behind, Announces Flexibility in Exchange for Reform for Ten States [Press Release]. Retrieved from http://www.ed.gov/news/press-releases/president-obama-our-children-cant-wait-congress-fix-no-child-left-behind-announc

Figure 1: Distribution of State Failure Rates, 2003 – 2005

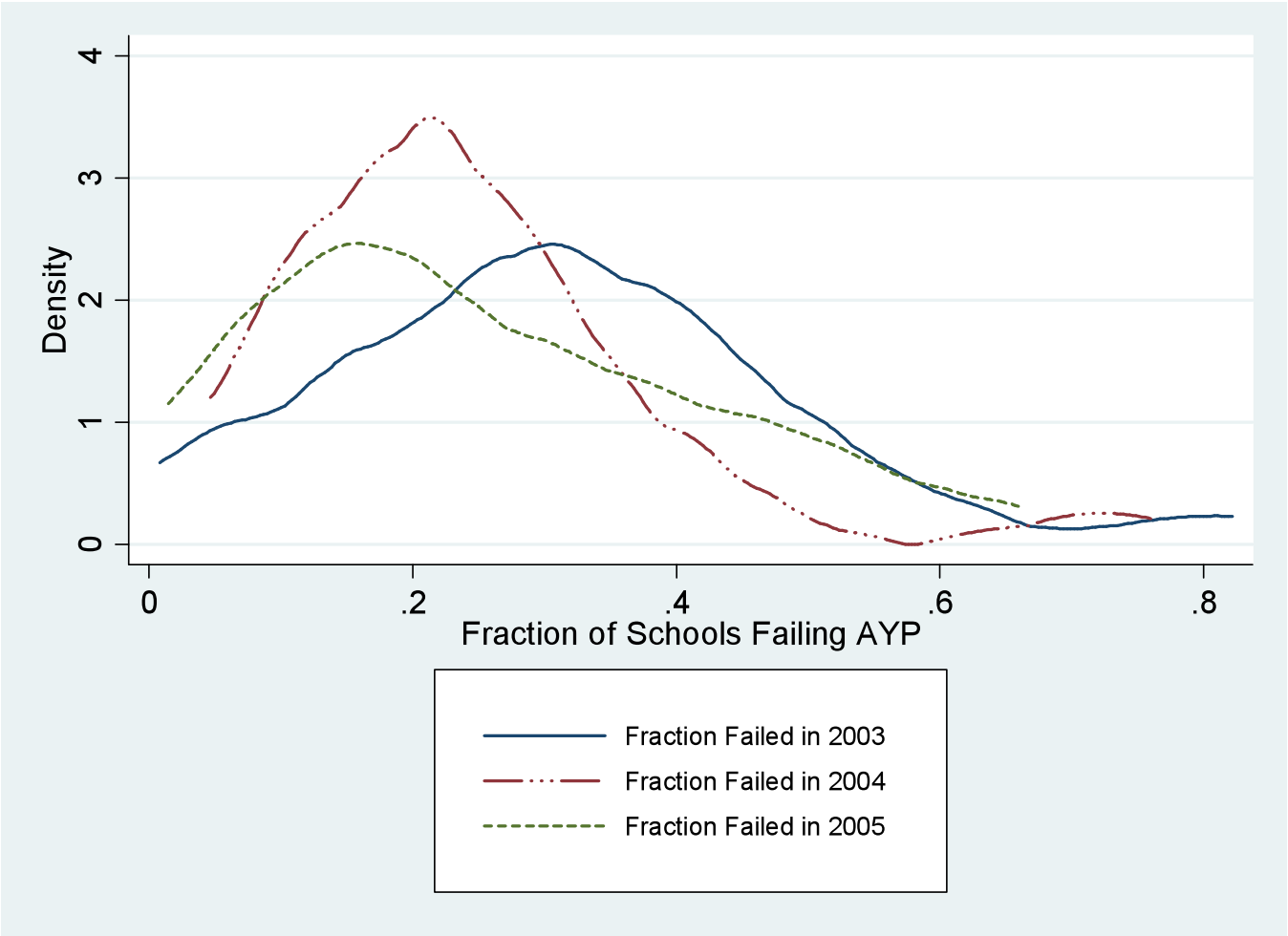
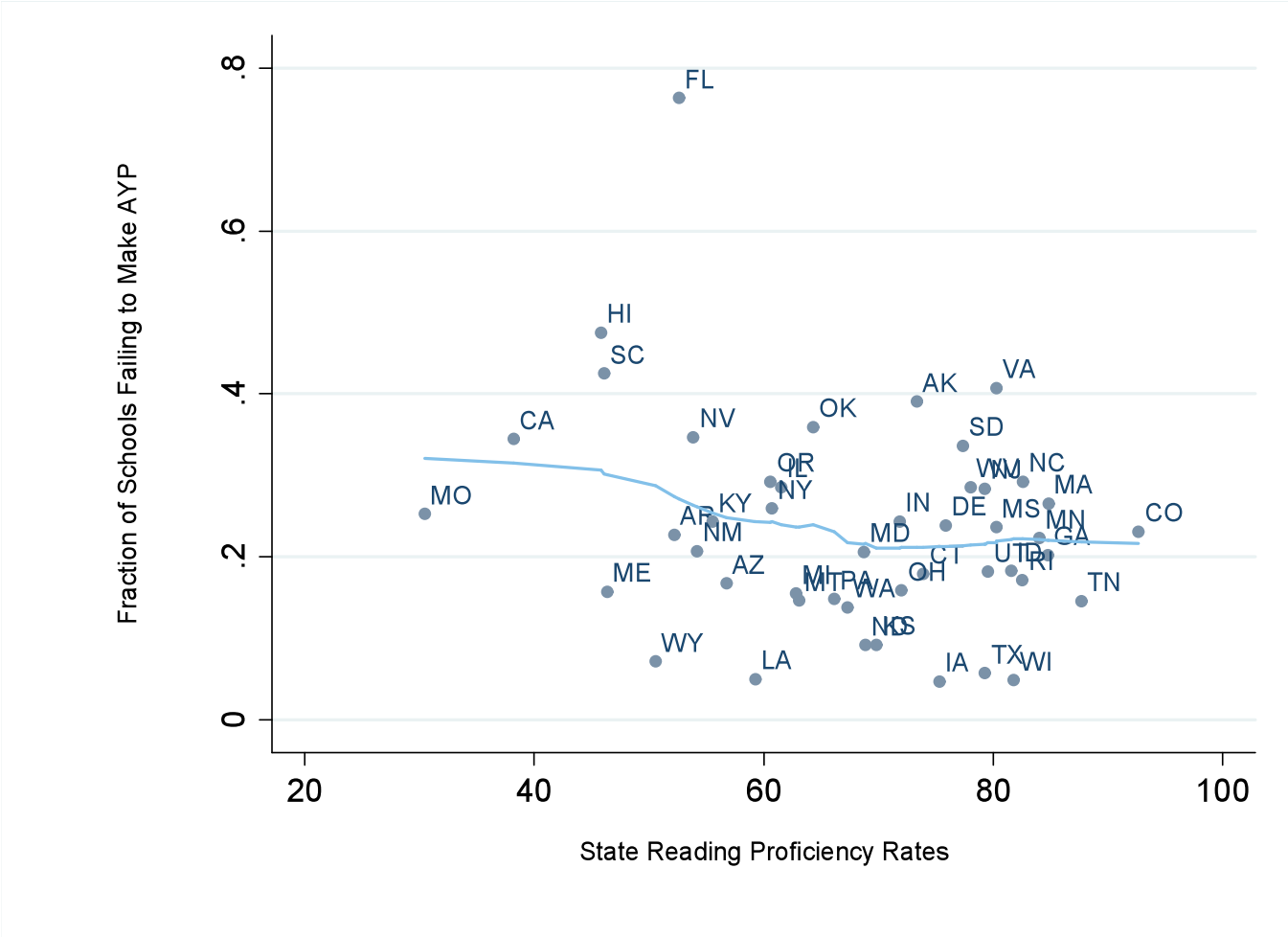


Figure 2a: School Failure Rates vs. State Proficiency Rates in Math, 2004

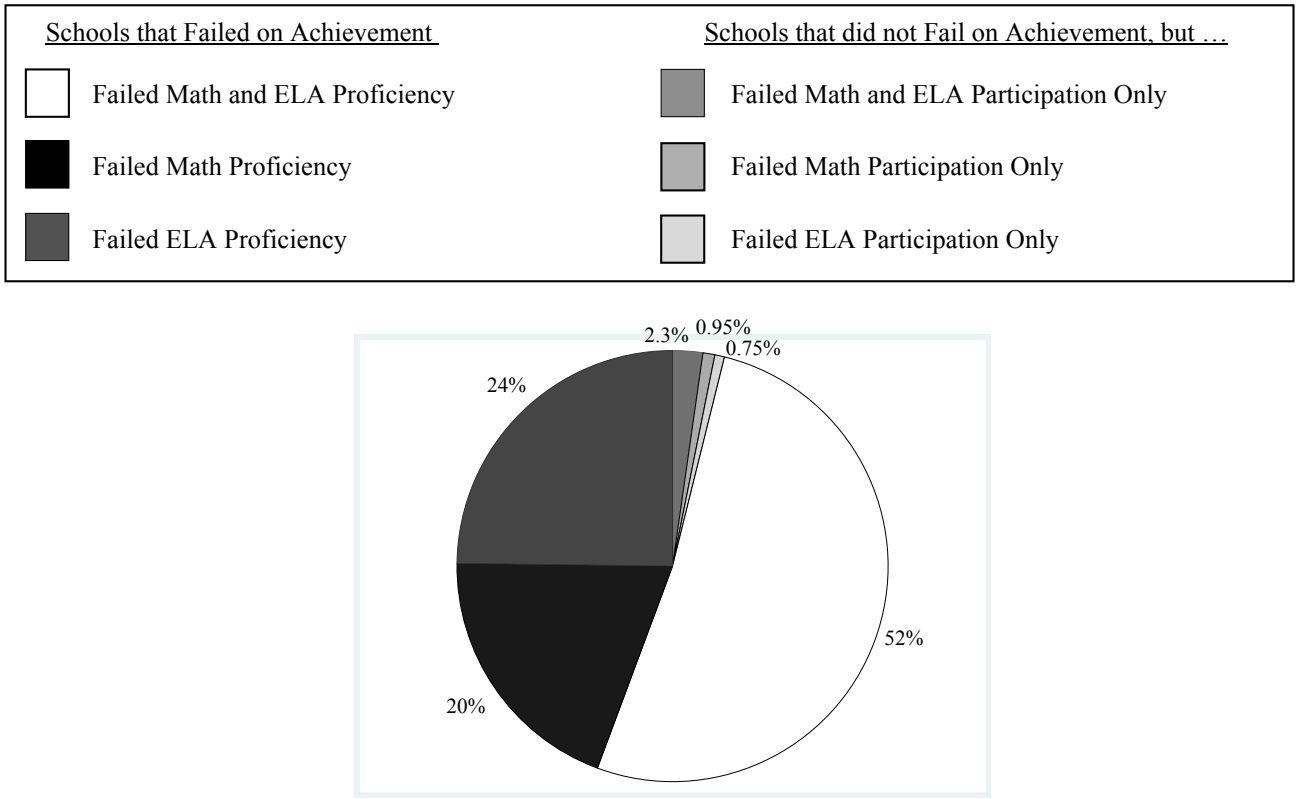
Notes to Figures 2a and 2b: N = 46 states in math and English Language Arts (ELA). Alabama, Nebraska, and New Hampshire are missing proficiency rates. Vermont reports a performance index in lieu of proficiency rates. When we aggregate proficiency rates to the state level for the x-axis, we weight schools 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 2b: School Failure Rates vs. State Proficiency Rates in English Language Arts, 2004



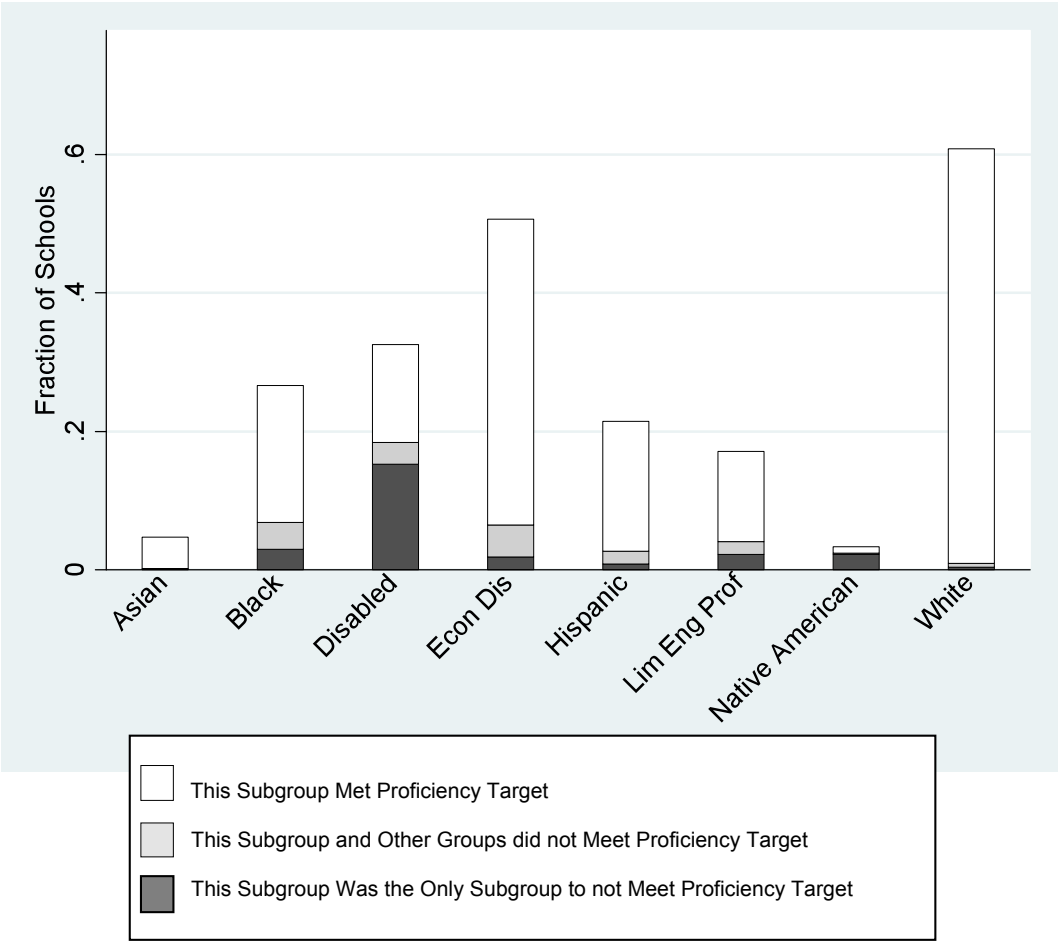
Note: see notes on previous page.

Figure 3: Percent of Schools that Failed by AYP Component, 2005



Notes to Figure 3: We exclude New York, Oklahoma, and Wyoming as these states are missing participation indicators in 2005.

Figure 4: Subgroup Accountability and Likelihood of Failure in Math, 2004



Notes to Figure 4: 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.

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.