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

Elizabeth Davidson[1], Randall Reback[1], Jonah Rockoff[1], and Heather L. Schwartz[2]

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% 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 were related to their schools' failure rates, which ranged from less than 1% to more than 80% 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 may cause dramatic differences in measured outcomes.

Keywords: accountability; descriptive analysis; educational policy; educational reform; performance assessment

The American public education system has had a history of strong local community control of public schools. U.S. public schools are predominantly funded through a combination of state and local tax revenues. Since the Elementary and Secondary Education Act of 1965, the federal government has been supplementing these revenues by awarding funds to states for allocation to public schools serving students from low-income families. These federally funded revenues are known as "Title I funding." Cohen and Moffit (2009) describe how the first 35 years of Title I funding included several rounds of debates concerning schools' flexibility in using their Title I funds and whether the impacts of Title I funds on student achievement should be evaluated. Interest in preserving America's history of strong local control of schools sometimes clashed with desires to attach strings to Title I funding to increase its efficacy as a poverty reduction program. Local control generally won the day during those first 35 years. The relative size of the Title I program did not grow (3.6% of public school revenues in 1969–1970 and only 2.6% in 1999–2000[1]), nor did a national system emerge to evaluate whether Title I funds were improving student achievement. States and school districts arguably had greater incentives to monitor the fiscal compliance of their Title I funds than to assess whether these funds were going to the most productive outlets (Gordon & Reber, in press). Several federal administrations during that time encouraged the adoption of national standards and the development of state accountability systems for schools, but these were voluntary (Manna, 2006).

In 2001, the reauthorization of the Elementary and Secondary Education Act marked the single greatest expansion of the federal role in education policy since the original 1965 act (Manna, 2010). This reauthorization, known as the No Child Left Behind (NCLB) Act, broke new ground by mandating schools be held accountable for their students' achievement as a condition of states' receipt of Title I funds. 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. This AYP determination was based not only on the proficiency rates of schools' general student populations but also on the proficiency rates of various ethnic and categorical subgroups of students, such as students from low-income families.

NCLB changed education policy by leveraging Title I funds to compel states to develop standardized testing systems for assessing student proficiency levels. NCLB increased the size of the Title I program—from roughly $8.5 billion appropriated in 2000–2001 to $13.6 billion appropriated in 2005–2006 (U.S. Department of

[1]Columbia University, New York, NY
[2]RAND Corporation, New Orleans, LA

Educational Researcher, Vol. 44 No. 6, pp. 347–358
DOI: 10.3102/0013189X15601426

Education, 2001, 2006). Title I funding remained equivalent to only about 3% of total public school operating expenditures, though this percentage remains much higher for some school districts than others. NCLB did not establish an evaluation system for the impact of Title I funds, or funding in general, on student achievement. One of the few direct changes to the use of Title I funds was to allow students from low-income families to purchase after-school tutoring services (called "supplemental education services"), by redirecting Title I funds away from schools that had failed to make AYP. Rather than holding states or schools accountable for the use of Title I funds, NCLB forced states to hold schools accountable for their students' proficiency rates.

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. For example, states could choose their own exams and set definitions of proficiency on those exams. 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 & Jacob, 2011; Dee, Jacob, & Schwartz, 2013).

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. Whereas previous studies have examined states' and schools' implementation of NCLB (Hamilton et al., 2007; Manna, 2006, 2010; Srikantaiah, 2009), these studies each examine a limited number of states or localities. No prior study has used national data to examine the link between states' initial NCLB implementation decisions and their schools' ratings. In so doing, our work provides a concrete example of the effects of the expanding federal role within education (Cohen & Moffit, 2009; Henig, 2013; Manna, 2006, 2010).

Using a newly assembled national data set, we investigate the following questions:

1. Which types of schools failed during the early years of NCLB? How are student demographics, school grade levels, and schools' urbanicity related to failure rates?
2. Which performance targets did schools fail to meet? Did schools frequently fail due to the performance of one student subgroup alone?
3. What explains cross-state differences in school failure rates? Are these differences associated with student demographics or with specific state policy implementation decisions?

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 wide variation in AYP failure rates across states is that this variation was driven by more obvious state policy differences, such as the difficulty of the exam questions and the proficiency standards. In fact, school failure rates are only weakly related to student proficiency rates. 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. Even if states are given wide flexibility in the design of their accountability and testing systems, policymakers may wish to remove loopholes that create disparate standards for schools via haphazard differences in rules and calculation methods. Flexibility need not come at the cost of transparency.

Our paper proceeds as follows. The next two sections provide an overview of NCLB policies and describe our data. The subsequent analysis describes which types of schools most frequently failed and which performance targets they failed to meet. We then describe cross-state variation in school failure rates and explore reasons for this variation. The concluding section briefly discusses the implications of our findings for current policy decisions.

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.[2] 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 students with disabilities—must meet all of the performance targets for the school to make AYP.[3]

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 the fraction of students demonstrating 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 2002–2003. 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.[4]

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 nonfailing 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 (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 three 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.[5] 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.[6] 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 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 toward the accountability system. In the initial years of implementation, not all states applied consistent definitions of special-needs categories exempted from the general standardized test. However, the U.S. Department of Education (DOE) later issued exemption rules to close loopholes related to testing of students with disabilities. But several other discrepancies remain. Not all states hold the same racial and ethnic subgroups of students separately accountable for meeting proficiency rate targets; for example, Asian American students might be a separate category in one state but not in another. 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" are the denominator of the participation rate calculation. A state with a very strict definition of continuous enrollment counts only 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, the proficiency rate of a student subgroup affected its school's AYP determination only if the number of students in that subgroup exceeded a specific threshold. States had flexibility in choosing that minimum subgroup size threshold. Most states chose a minimum subgroup size between 30 and 40 students, but the range extended from five 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 had either 100 tested students or at least 50 tested students who composed at least 15% of the schools' total tested population.

A third, often-overlooked area of flexibility is which grade levels of students were tested and the methods of aggregating performance across grade levels. Although tested grade levels became more standard as of 2005–2006, the aggregation of scores across tested grade levels within a school was not.[7] For schools that served multiple tested grade levels, states could decide whether to aggregate statistics across all of the tested grade levels or to consider the student proficiency levels of each grade separately. For example, in a state, like Washington, that considered each tested grade's proficiency level separately, both fourth graders and seventh graders in a hypothetical school would each need to exceed proficiency targets, making it more likely the school could fail AYP. However, other state AYP criteria pertaining to minimum subgroup size and confidence intervals could offset that challenge. Specifically, Washington counted the number of tested students in each grade separately to determine the size of the confidence interval to apply to that grade level's proficiency rate. This means the respective confidence intervals for fourth-grade and for seventh-grade proficiency rates were more generous than a confidence interval applied to a proficiency rate that pooled fourth and seventh graders. It is also more likely that the number of fourth graders or seventh graders, when considered separately, would fall below Washington's minimum subgroup size threshold, rendering fourth- or seventh-grade proficiency rates inapplicable to a school's AYP determination.

NCLB Data

NCLB has greatly expanded the amount of student performance data available to researchers and the public, although 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 websites.

The resulting school-level data set includes school-level AYP determinations and the subcomponents for these determinations. Our variables include indicators of whether the school as a whole and each individual student subgroup made AYP, school- and subgroup-level average student proficiency rates on state assessments, and the number of students tested in the school and in each student subgroup. 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). The resulting data and our state-by-state documentation of sources are publicly available.[8] For 2004–2005, we use school and subgroup proficiency target data from the American Institutes for Research (2005).

Descriptive Evidence on Failing Schools

Looking nationwide from 2003 to 2005, there were clear observable differences between AYP failing and nonfailing 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 3 years had nearly double the percentage of students eligible for free and reduced-priced lunch as schools that made AYP all 3 years. Failing schools also had fewer teachers per student and were disproportionately located in

Table 1
Characteristics of Schools by Whether They Failed to Make Adequate Yearly Progress, 2003 to 2005

Characteristic	Failed all 3 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
Percentage 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
Percentage 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

Note: The data on school characteristics are from the Common Core of Data, 2001–2002. For schools in Tennessee, data on student ethnicity come from 1998–1999 instead of 2001–2002, and data on free/reduced-price lunch eligibility are unavailable. Aside from the percentage of students who are Asian, all differences in means between the second and third columns are statistically significant at the .01 level.

urban school districts. Middle schools and high schools failed far more frequently than elementary schools.

Most schools failed to make AYP due to proficiency rate requirements as opposed to participation rates. The majority of failing schools had groups of students not meeting proficiency rate targets in both subjects. In 2005, 52% of failing schools missed proficiency rate targets in both subjects, 24% of failing schools missed ELA proficiency rate targets only, 20% of failing schools missed math proficiency rate targets only, and the remaining 4% of failing schools satisfied all of their proficiency rate targets but not their participation rate targets. The number of schools failing due to participation alone was substantially smaller in 2005 than in the prior 2 years, suggesting that schools took action to ensure that sufficient numbers of students were tested.[9]

Although 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 have been due to whether a subgroup was large enough to be held accountable. Figure 1 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. The total height of each bar illustrates the fraction of schools where that subgroup's proficiency rate counted toward the AYP determination, and the shaded areas of the bars represent the fraction of schools where that subgroup failed to make AYP. White and economically disadvantaged subgroups were held accountable in about 43% and 37% of schools, respectively, while fewer than 4% of schools had a Native American subgroup held accountable.

However, conditional on being accountable, subgroup failure rates varied considerably. Figure 1 reveals that White and Asian subgroups rarely failed, whereas more than half of all accountable Native American subgroups and students-with-disabilities subgroups failed to meet proficiency targets. The students-with-disabilities subgroup was also the most likely to be the only subgroup failing their schools' proficiency targets: 57% of accountable students-with-disabilities subgroups were the only group to fail to meet targets at their schools.

Cross-State Differences in Failure Rates

Figure 2 illustrates the wide variation in states' AYP failure during the first 3 years of NCLB. Figure 2 is a density plot, the continuous version of a histogram, so the area under the curve represents the proportion of states falling in various ranges of values for the fraction of their schools failing to make AYP. For example, in the 1st year of AYP designations (2003), approximately 40% of states had AYP failure rates between 20% and 40%.[10] That same year, 32% of the nation's schools failed AYP, but failure rates ranged from 1% in Iowa to 82% in Florida. The national failure rate declined to 26% by 2005, but failure rates ranged from 2% in Oklahoma to 66% in Hawaii.

Failure rates changed substantially over time in some states. Alabama's failure rate jumped from 4% in 2003 to 68% in 2004.[11] 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 in 2003, yet 72% of its high schools failed. Similarly, only 20% of West Virginia's elementary schools failed in 2003, yet more than 80% of its high schools failed.

A common misconception is that this wide variation in failure rates resulted from cross-state differences in the proportion

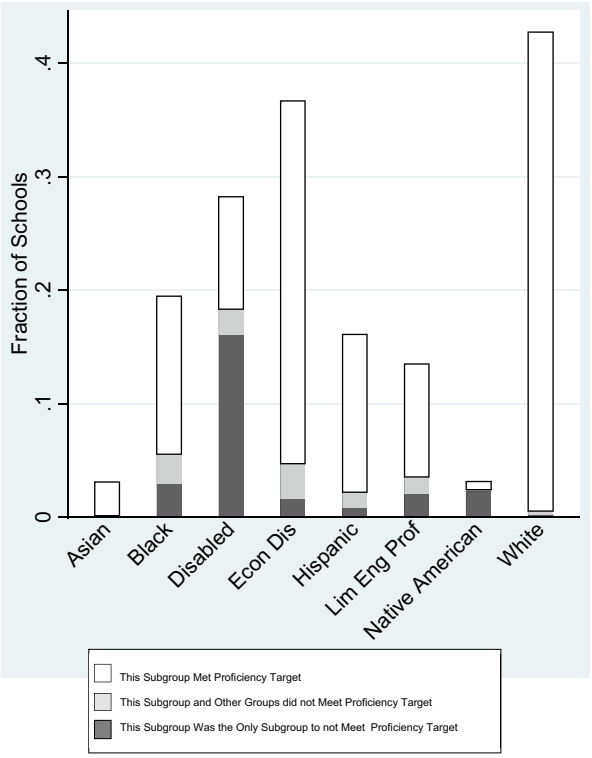

FIGURE 1. *Subgroup accountability and likelihood of failure in math, 2004*
The total height of each bar illustrates the fraction of schools where that subgroup's proficiency rate counts towards the adequate yearly progress (AYP) determination, and the shaded areas of the bars represent the fraction of schools where that subgroup failed to make AYP. The figure is based on 46 states with available data. Iowa, North Dakota, Nebraska, and New Mexico are missing subgroup-level proficiency data in 2004.

of students identified as proficient. In reality, states' school failure rates were not strongly related to their students' performance. Figure 3 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 the states' school failure rates. Based on corresponding linear regression, a 1 percentage point increase in state math proficiency rates is associated with only a statistically insignificant 0.05 percentage point decline in the fraction of a state's schools making AYP.[12] 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 schools on a curve, with state-specific curves based on the starting points and trajectories for proficiency targets. For example, Iowa set 2003 proficiency targets at 64% in math and 65% in ELA, whereas 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. Reback, Rockoff, and Schwartz (2014) 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.

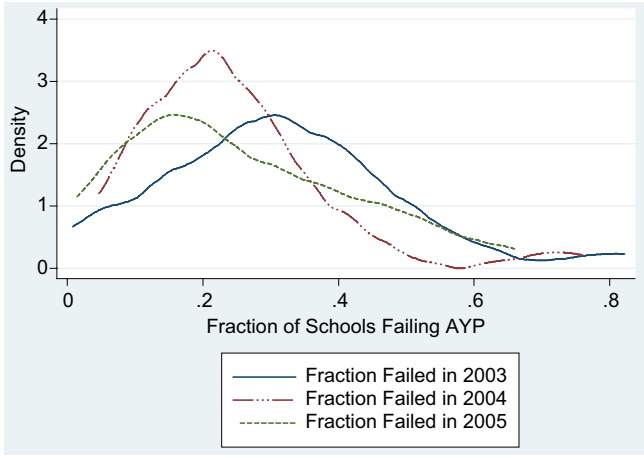

FIGURE 2. *Distribution of state failure rates, 2003 to 2005*

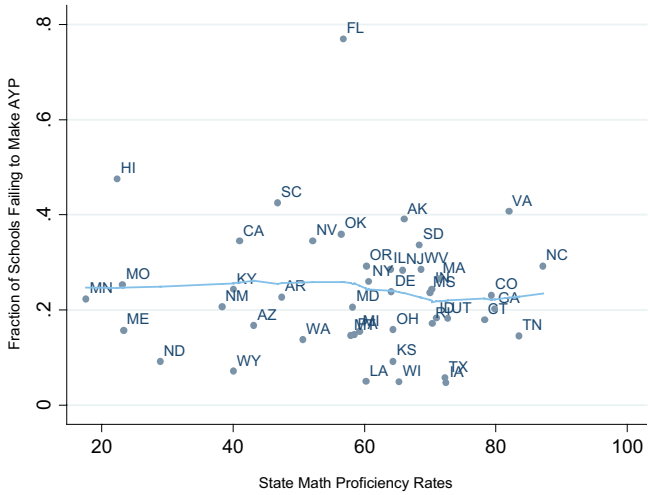

FIGURE 3. *School failure rates versus state proficiency rates in math, 2004*
N = 46 states. Alabama, Nebraska, and New Hampshire are missing proficiency rate data. Vermont reports a performance index in lieu of proficiency rates. When we aggregate proficiency rates to the state level for the *x*-axis, we weight schools by their number of tested students. For 12 states that failed to report the number of tested students by school, we use schools' student enrollment in tested grades as reported in the Common Core of Data as a proxy for the number of students tested.

Explaining Cross-State Variation in Failure Rates

Various dimensions of NCLB implementation contributed to the wide variation in school AYP failure rates.[13] No individual state policy decision appears to have been the primary culprit. Instead, failure rates appear to have been influenced by interactions among several decisions and states' school characteristics (e.g., enrollment size, grade spans, ethnic diversity of students). Given that we have only 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

Table 2
States' Early Policies for Determining AYP, With States Sorted by the Fraction of Schools Failing to Make AYP in 2004

State	% of Schools Failing to Make AYP in . . .			Confidence Intervals Applied to Proficiency Rates During Early Years of NCLB	Grades Tested in 2004		Average No. of Student Subgroups Contributing Proficiency Rates Toward AYP Ratings, 2004[a]
	2003	2004	2005		Math	ELA	
Iowa	0.8%	4.7%	7.3%	98%	4, 8, 11	4, 8, 11	N/A
Wisconsin	4.5%	4.8%	2.4%	99%	4, 8, 10	4, 8, 10	1.9
Louisiana	6.4%	5.0%	17.2%	99%	4, 8, 10	4, 8, 10	4.2
Texas	8.2%	5.7%	11.6%	95%	3–8, 10	3–8, 10	5.2
Wyoming	15.1%	7.1%	18.9%	95%	4, 8, 11	4, 8, 11	1.7
North Dakota	31.7%	7.8%	11.5%	99%	4, 8, 12	4, 8, 12	N/A
North Carolina	9.1%	9.1%	9.2%	99%	3–8, 10	3–8, 10	3.9
Kansas	29.3%	9.2%	8.8%	99%	4, 7, 10	5, 8, 11	2.0
Vermont	12.7%	12.7%	3.3%	99%	4, 8, 10	2, 4, 8, 10	1.8
Washington	22.0%	13.8%	19.4%	99%	4, 7, 10	4, 7, 10	2.6
Tennessee	46.7%	14.6%	7.6%	95%	3, 5, 8, HS[b]	3, 5, 8, HS[b]	2.8
Montana	20.1%	14.6%	6.1%	95%	4, 8, 10	4, 8, 10	4.8
Pennsylvania	35.5%	14.8%	19.3%	95%	5, 8, 11	5, 8, 11	2.4
Michigan	24.0%	15.5%	7.7%	None[c]	4, 8, 11	4, 7, 11	2.2
Maine	26.5%	15.7%	26.5%	95%	4, 8, 11	4, 8, 11	3.5
Ohio	24.2%	15.8%	24.2%	None	4, 6, 9	4, 6, 9	2.5
Arizona	23.3%	16.7%	13.2%	99%	3, 5, 8, 10	3, 5, 8, 10	3.3
Rhode Island	31.1%	17.1%	11.4%	None[d]	4, 8, 11	4, 8, 11	3.4
Connecticut	14.7%	17.9%	20.4%	99%	4, 6, 8, 10	4, 6, 8, 10	2.7
Utah	35.8%	18.2%	13.1%	99%	3–8, 11	3–8, 10	5.5
Idaho	35.3%	18.2%	42.8%	None[d]	3, 4, 7, 8, 10	3, 4, 7, 8, 10	3.1
Georgia	36.2%	20.2%	18.1%	95%	3–8, 11	3–8, 11	4.1
Maryland	35.2%	20.6%	23.1%	99%	3–8, 10	3–8, 10	5.8
New Mexico	20.7%	20.7%	52.6%	99%	4, 8, 11	4, 8, 11	N/A
Minnesota	7.8%	22.3%	13.1%	95-99%	3, 5, 7, 11	3, 5, 7, 10	3.2
Arkansas	22.2%	22.7%	42.5%	95%	4, 6, 8, HS[e]	4, 6, 8, 11	2.9
Colorado	37.6%	23.1%	27.5%	95%	5–10	3–10	3.1
Mississippi	23.1%	23.6%	11.8%	99%	3–8, 10[f]	3–8, 10	3.7
Delaware	54.0%	23.8%	26.4%	98%	3, 5, 8, 10	3, 5, 8, 10	4.1
Kentucky	40.7%	24.3%	25.7%	99%	5, 8, 11	4, 7, 10	3.4
Indiana	23.2%	24.4%	40.8%	99%	3, 6, 8, 10	3, 6, 8, 10	3.9
Missouri	48.3%	25.2%	34.8%	99%	4, 8, 10	3, 7, 11	2.5
New York	25.9%	25.9%	18.7%	90%	4, 8, HS[g]	4, 8, HS[g]	3.2
Massachusetts	44.3%	26.5%	29.0%	95%	4, 6, 8, 10	3, 4, 7, 10	2.6
New Jersey	42.4%	28.4%	37.8%	95%	4, 8, 11	4, 8, 11	3.3
West Virginia	40.5%	28.5%	16.9%	99%	3–8, 10	3–8, 10	3.1
Illinois	32.4%	28.6%	26.3%	95%	3, 5, 8, 11	3, 5, 8, 11	2.7
Nebraska	52.6%	29.2%	42.6%	95%	4, 8, 11	4, 8, 11	N/A
Oregon	29.7%	29.2%	32.6%	99%	3, 5, 8, 10	3, 5, 8, 10	3.9
New Hampshire	31.4%	29.6%	46.8%	99%	3, 6, 10	3, 6, 10	3.3
South Dakota	33.6%	33.6%	13.8%	99%	3–8, 11	3–8, 11	3.5
California	45.9%	34.4%	38.8%	99%	2–8, 10[h]	2–8, 10[h]	3.7
Nevada	42.6%	34.5%	60.0%	95%	3, 5, 8, 11	3, 5, 8, 11	4.7

(continued)

Table 2 (continued)

State	% of Schools Failing to Make AYP in . . . 2003	2004	2005	Confidence Intervals Applied to Proficiency Rates During Early Years of NCLB	Grades Tested in 2004 Math	ELA	Average No. of Student Subgroups Contributing Proficiency Rates Toward AYP Ratings, 2004[a]
Oklahoma	22.8%	35.9%	1.5%	95% for campus-wide group only[d]	3, 5, 8, HS[i]	3, 5, 8, HS[i]	1.5
Alaska	57.7%	39.0%	40.9%	99%	3–10	3–10	3.4
Virginia	40.5%	40.7%	24.9%	None	3, 5, 8, HS[j]	3, 5, 8, HS[j]	2.9
South Carolina	79.7%	42.5%	51.7%	68% starting in 2005	3–8, 10	3–8, 10	4.0
Hawaii	60.6%	47.5%	65.9%	68%	3, 5, 8, 10	3, 5, 8, 10	3.2
Alabama	4.2%	68.3%	46.7%	99%	4, 6, 11	4, 6, 8, 11	2.8
Florida	82.2%	76.3%	64.0%	None	3–10	3–10	5.2

Note. AYP = adequate yearly progress; ELA = English language arts; NCLB = No Child Left Behind; HS = high school.

[a]The number of subgroups reported here are averaged across math and ELA.

[b]Proficiency and participation rates are based on the cohort of students enrolled in Algebra I and English II courses, which may be taken at varying grade levels in high school.

[c]Only very small schools were allowed to use confidence interval adjustments.

[d]Although these states did not use confidence interval adjustments for subgroups, they used relatively large minimum required subgroup sizes.

[e]To calculate proficiency and participation rates, Arkansas officials combine students' assessment results on End-of-Course (EOC) exams in Algebra I and Geometry.

[f]To calculate proficiency and participation rates, Mississippi matches 10th-grade students with their Algebra I test scores whether or not they take the exam in 10th grade.

[g]Proficiency and participation rates are based on the cohort of students who are in courses culminating in state math and ELA high school Regents Exams.

[h]California requires all 10th-grade students to take the California High School Exit Exam.

[i]Proficiency and participation rates are based on the cohort of students who are in courses culminating in State math and ELA End-of-Instruction exams.

[j]Proficiency and participation rates are based on the cohort of students who are in courses culminating in math and ELA EOC exams.

nature of these complex interactions, we instead describe five categories of policy decisions that we have identified as having had 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. We focus on examples below, and Table 2 provides some relevant policy information for all 50 states. The states in Table 2 are sorted in ascending order by the percentage of schools failing to make AYP in 2004.

Five Categories of Policy Decisions That Impact School Failure Rates

A few states initially deviated from NCLB rules

Calculations. Iowa continued to develop its AYP formula and data collection processes throughout the initial 2 years of NCLB. 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 than the state's official published rates.[14] 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.[15] 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 are unavailable for Iowa, but we can examine proficiency rates for the 90% of Iowa's schools that served only one tested grade level.[16] Among these schools in 2004, 27% of schools that Iowa labeled as making AYP should not have made AYP by our calculations due to either (a) a subgroup with a participation rate below 95% or (b) a subgroup with a proficiency rate too low to meet the required targets, even after considering safe harbor and the most generous possible confidence interval adjustment.[17]

Alternative assessments. Because the students-with-disabilities subgroups' performances were often the only reason for a school's failing to make AYP, states' policies toward these subgroups have substantial ramifications. NCLB requires states to incorporate 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 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% (Title I, 2007).

Applying a large confidence interval to safe harbor calculations. NCLB gives states the option of applying these safe-harbor calculations as well as a further option to apply a 75% confidence interval to safe-harbor calculations. Fourteen states incorporated this safe-harbor confidence interval as allowed. Louisiana and Massachusetts, however, applied confidence intervals that were more generous than allowed: Louisiana employed a 99% confidence interval, and Massachusetts employed a 95% confidence interval. In Louisiana, this added increment helped more than 62% of otherwise failing economically disadvantaged subgroups, 79% of otherwise failing Black subgroups, and 90% of otherwise failing students-with-disabilities subgroups avoid failing status.[18] 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 No. 7 in Orleans Parish, Louisiana, 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 the 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.

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. As shown in Table 2, 23 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 its 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 14 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% vs. 21%).[19] 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.[20] Michigan applied 99% confidence interval adjustments but only for schools with very small campuswide enrollments. If Michigan had instead applied 99% adjustments to all of its schools in 2004, we estimate that the percentage 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 2 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 the proportion of schools failing to make AYP in South Carolina promptly fell by 10 percentage points from the prior year.

Confidence intervals applied to safe harbor were another important source of cross-state variation in failure rates. By 2004, all but one state—Alabama—employed safe-harbor calculations.[21] Yet 16 states applied confidence intervals to their safe harbor calculations, and the other states did not.[22] As discussed above, Louisiana and Massachusetts applied improperly large confidence intervals to safe-harbor calculations, whereas 14 other states applied the permitted 75% confidence interval to safe harbor calculations. Polikoff and Wrabel (2013) describe how the number of schools making AYP due to safe harbor has increased over time in California, one of the states applying a 75% confidence interval to its safe-harbor calculations.

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. Twenty-three states employed this policy.[23] Of these, Texas and Pennsylvania provide examples of states with lagging high school proficiency rates. 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, then 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 counterexample to Texas and Pennsylvania. South Carolina applied pre-NCLB proficiency rates of students in Grades 3 through 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 proficiency rate for Grades 3 through 8 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%.

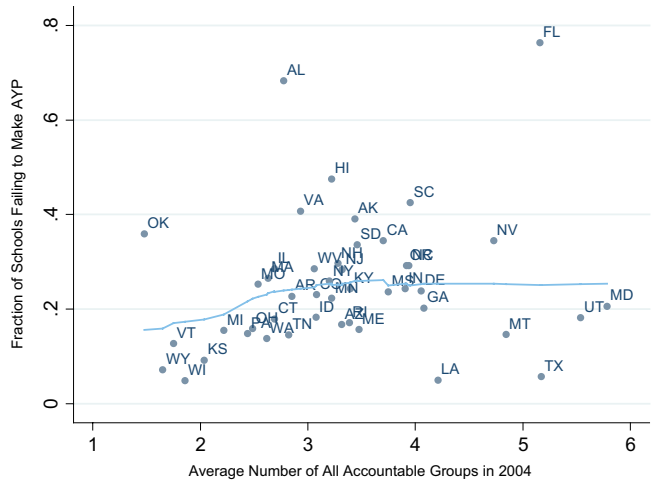

FIGURE 4. *School failure rates versus average number of accountable groups in schools, 2004*
Based on 46 states with available data. Iowa, North Dakota, Nebraska, and New Mexico are missing subgroup-level proficiency data in 2004. Accountable groups include both student subgroups and the overall student population. For each state, we take the average of the number of accountable groups for math achievement and the number of accountable groups for English language arts 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.

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 & Staiger, 2002, 2003; Sims, 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 4 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 4 also reveals that this relationship would have been 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.[24] 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 Louisiana 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 were only one student in that group. However, Montana's small schools and 95% confidence interval policy meant that subgroups were so small that they would make AYP even with few students passing.

Because the performance of the students-with-disabilities subgroup was often the only reason for a school's failure to make AYP, one might expect states' policies toward this subgroup to influence their schools' failure rates. The fraction of schools with accountable subgroups will depend not only on states' minimum subgroup size rules but also on how they allocated students with disabilities across schools. School failure rates were initially higher in states with larger fractions of schools with accountable students-with-disabilities subgroups. If we regress state failure rates on a quadratic for the fraction of schools where these subgroups were 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 accountable students-with-disabilities 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 holding these subgroups accountable for math performance used 99% confidence interval adjustments.

States defined continuous enrollment differently. Five states—Hawaii, Illinois, Iowa, New Jersey, and Wisconsin—used starting dates for continuously enrolled students that precede September of the school year of the testing.[25] In these states, students who have

transferred schools prior to the 1st 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 tended to be relatively low achieving or if school districts tended to strategically wait to enroll students at particular schools (Jennings & Crosta, 2011), then these long required enrollment windows would make it easier for schools to make AYP. Hawaii already had a high failure rate in 2003 in spite of their early enrollment cutoff date, due to low participation rates and low proficiency rates in the subgroups of students with disabilities and students with limited English proficiency. The other four states may have had much higher failure rates if they had 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 fourth-grade students, 10% of eighth-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.

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. DOE, 2012). These waivers are idiosyncratic to each state, so that cross-state variation in the minutiae of accountability policy rules is as complicated and important as ever (Polikoff, McEachin, Wrabel, & Duque, 2014).

Although 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 may be best served by a consistent set of directions about acceptable statistical practices and common definitions. The federal government could convene a panel of experts or commission a professional association, such as the American Statistical Association, to provide guidance on sound statistical practices related to confidence interval setting, safe-harbor exceptions, and minimum subgroup sizes. Formulae for these procedures, if used, could then be standardized. These formulas themselves attempt to adjust evaluation to treat schools in a fair and just manner. Standardizing rules for exceptions and adjustments does not eliminate this quest for fairness. Rather, using uniform accounting practices might promote transparency and better insulate state accountability systems from the political whims of governors and state legislatures. Although our own analysis does not investigate whether arbitrary differences across states were harmful, we are hard-pressed to think of a compelling reason why citizens should prefer these arbitrary differences in accounting.

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.

NOTES

This research project was made possible by funding from the Institute for Education Sciences (No. R305A090032) and the Spencer Foundation (No. 200900082) 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 three anonymous referees and participants in the 2013 conferences of the Association for Education Finance and Policy and the Association for Public Policy and Management. Data used in this paper are available for download from the Barnard/Columbia No Child Left Behind Database, http://www.gsb.columbia.edu/nclb.

[1]Calculated based on statistics reported by the U.S. Department of Education (1994, 2001, 2012).

[2]We provide a brief overview of No Child Left Behind (NCLB) in this section and refer the reader to the U.S. Department of Education's (DOE; 2002) *Desktop Reference* and to Manna's (2010) *Collision Course* book for more details on NCLB policies. Manna also provides revealing anecdotes concerning the challenges faced by states and schools in implementing these policies.

[3]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 campuswide group, the Hispanic subgroup, the limited-English-proficient subgroup, and the free/reduced-priced lunch subgroup proficiency rates in math and English language arts (ELA). Subgroup proficiency rates influence the school's adequate yearly progress (AYP) rating only if there are sufficient numbers of students enrolled at the school (and meeting the *continuous enrollment* definition described elsewhere in the article).

[4]Initially, NCLB permitted states to use their own formulae for calculating graduation rates. In December 2008, the U.S. DOE announced that all states must use a standardized 4-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).

[5]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 subgroup of students with disabilities and if 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.

[6]The confidence interval adjustment lowers the target from p to $p - [\sqrt{\frac{p(1-p)}{n}} * C]$, where p is the unadjusted proficiency rate target in decimal form, n is the number of students contributing to the proficiency rate, and C is the critical value for the specified confidence interval, such as 1.96 for a 95% two-sided confidence interval. For example, in Alaska, the 2003 ELA proficiency target was 64%, and the state used a 99% confidence interval adjustment. An Alaskan student subgroup with 20 students would only have to reach 36% proficiency that year to make AYP, because $.36 = [.64 - [\sqrt{\frac{.64(1-.64)}{20}} * 2.575]$, where 2.575 is the critical value for the 99% confidence interval.

[7]As of 2005–2006, states were required to test students in Grades 3 through 8 and in one high school grade. Before this, 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.

[8]Data for the first 2 years of NCLB are currently accessible from the Barnard/Columbia No Child Left Behind Database at www.gsb.columbia.edu/nclb (Reback et. al, 2011).

[9]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 3 years, then we observe a downward trend in the fraction of schools failing due only to participation: from 17% in 2003 to 14% in 2004 to 5% in 2005.

[10]This is found by calculating the area under the solid curve, which equates to approximately two units on the *y*-axis multiplied by 0.2 (= 0.4 – 0.2) units on the *x*-axis.

[11]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 an NCLB-compliant accountability system.

[12]The relationship with state ELA proficiency rates is also statistically insignificant and small, only a 0.16 percentage point decline in the fraction of schools making AYP. If we regress states' school AYP failure rates on quadratic terms for their states' proficiency rates in each subject (i.e., four independent variables total), the *R* squared is .07 but the adjusted *R* squared is only .02. The joint significance level of these estimated coefficients is 0.56.

[13]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.

[14]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 (T. Deeter, personal communication, March 5, 2013). 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.

[15]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.

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

[17]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.

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

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

[20]Florida also had low cutoffs for minimum subgroup size. Their subgroups for limited-English-proficient students, students with disabilities, and Black students 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.

[21]Alabama employed safe-harbor adjustments in 2005.

[22]The postal abbreviations for these sixteen states are AK, CA, CT, DE, KS, LA, MA, ME, MO, NJ, NV, PA, SD, UT, WI, and WY.

[23]The postal abbreviations for the 23 states with homogenous targets across grade levels are AK, CA, CT, DE, FL, HI, ID, IL, IN, LA, MA, MO, MT, NH, NY, OK, OR, PA, SC, TN, TX, VA, and WI.

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

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

REFERENCES

American Institutes for Research. (2005). *National AYP and identification database (NAYPI)*. Washington, DC: Author. Retrieved November 12, 2008, from http://www.air.org/project/national-ayp-and-identification-database

Cohen, D. K., & Moffitt, S. L. (2009). *The ordeal of equality: Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

Council of Chief State School Officers. (2005). *School data direct*. Washington, DC: Author. Retrieved 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. (2013). The effects of NCLB on school resources and practices. *Education Evaluation and Policy Analysis*, *35*(20), 252–279.

Gordon, N., & Reber. (in press). The quest for a targeted and effective Title I ESEA: Challenges in designing and implementing fiscal compliance rules. *RSF: the Russell Sage Foundation Journal of the Social Sciences*.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., . . . Barney, H. *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation.

Henig, J. (2013). *The end of exceptionalism in American education: The changing politics of school reform*. Cambridge, MA: Harvard Education Press.

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

Jennings, J., & Crosta, P. (2011, March). *The unaccountables*. Paper presented at the annual conference of the Association for Education Finance and Policy, Seattle, WA.

Kane, T. J., & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*, 91–114.

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

Manna, P. (2006). *School's in: Federalism and the national education agenda*. Washington, DC: Georgetown University Press.

Manna, P. (2010). *Collision course: Federal education policy meets state and local realities*. Thousand Oaks, CA: CQ Press.

Polikoff, M., McEachin, A., Wrabel, S., & Duque, M. (2014). The waive of the future: School accountability in the waiver error. *Educational Researcher*, *43*, 45–54.

Polikoff, M., & Wrabel, S. (2013). When is 100% not 100%? The use of safe harbor to make adequate yearly progress. *Education Finance and Policy*, *8*(2), 251–270.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*, 207–241.

Reback, R., Rockoff, J., Schwartz, H.L., & Davidson, E. (2011). *Barnard/Columbia No Child Left Behind Database, 2002–2003 and 2003–2004*. Retrieved from 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?* Washington, DC: Center on Education Policy.

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.

Srikantaiah, D. (2009). *How state and federal accountability policies have influenced curriculum and instruction in three states: Common findings from Rhode Island, Illinois, and Washington*. Washington, DC: Center for Education Policy.

Title I–Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA); Final rule. 72 Fed. Reg. 17,748-17,781 (April 9, 2007) (to be codified at 34 CFR pts. 200 and 300).

U.S. Department of Education. (1994). *Biennial evaluation report FY 93-94: Chapter 101, Education of disadvantaged children (Chapter 1, ESEA) formula grants to local education agencies*. Retrieved from https://www2.ed.gov/pubs/Biennial/101.html

U.S. Department of Education. (2001). *Digest of education statistics, Table 369*. Retrieved from https://nces.ed.gov/programs/digest/d01/dt369.asp

U.S. Department of Education. (2002). *No Child Left Behind: A desktop reference*. Retrieved from https://www2.ed.gov/admins/lead/account/nclbreference/reference.pdf.

U.S. Department of Education. (2006). *Digest of education statistics, Table 368*. Retrieved from https://nces.ed.gov/programs/digest/d06/dt06_368.asp

U.S. Department of Education. (2008). *High school graduation rate: Non-regulatory guidance*. Retrieved from http://www2.ed.gov/policy/elsec/guid/hsgrguidance.pdf.

U.S. Department of Education. (2012). *Digest of education statistics, Table 205: Summary of expenditures for public elementary and secondary education, by purpose, 1919–20 through 2009–10*. Retrieved from https:/nces.ed.gov/programs/digest/d12/tables/dt12_205.asp

AUTHORS

ELIZABETH DAVIDSON is an independent education consultant in New York and PhD candidate in the economics of education at Teachers College, Columbia University; *ekdavidson@gmail.com*. Her research focuses on school accountability and school closure, and her consulting work focuses on data analysis and teacher and school evaluation.

RANDALL REBACK, PhD, is an associate professor in the Barnard College Economics Department, 3009 Broadway, New York, NY 10027; *rr2165@columbia.edu*. His research focuses on behavioral responses to K–12 education policies, such as accountability programs, school choice programs, and schools' health and mental health services.

JONAH ROCKOFF, PhD, is an associate professor at Columbia Business School, 3022 Broadway New York NY 10027; *jr2331@gsb.columbia.edu*. His research focuses on the organization and management of public schools.

HEATHER L. SCHWARTZ, PhD, is a policy researcher at RAND Corporation, 450 Poydras St., Suite 1400, New Orleans, LA, 70130; *hschwart@rand.org*. Her research focuses on education and housing policies intended to narrow the achievement gap between children from low- and high-income families.

Manuscript received October 21, 2013
Revisions received September 11, 2014; May 1, 2015; and July 15, 2015
Accepted July 16, 2015