Were All Those Standardized Tests for Nothing?

The Lessons of No Child Left Behind

Thomas Ahn and Jacob Vigdor

May 2013

AMERICAN ENTERPRISE INSTITUTE

A Report Prepared for the National Research Initiative

Executive Summary

The No Child Left Behind Act of 2001 (NCLB) introduced the first nationwide annual standardized testing requirement for students in grades 3 through 8. The law officially expired in 2007, and there is little or no legislative momentum to reauthorize it now. Should NCLB be thought of as a well-intentioned initiative that failed? Or did it make some progress in its stated goal of improving academic achievement, particularly for disadvantaged students?

This paper reviews the basic structure of the school incentives introduced by NCLB, as well as research and data from North Carolina public schools on the effect of these various sanctions on student learning. Among our main findings:

- Evidence indicates that school accountability systems in general, and NCLB in particular, have beneficial systemic effects on standardized test scores. The overall effects are modest; however, accountability systems are complex policies that may entail a mix of beneficial and harmful elements. The most critical question is not whether NCLB worked, but which components worked.
- Schools exposed to punitive NCLB sanctions, or the threat of sanctions, tend to outperform nearly identical schools that barely avoided them. Studies come to varying conclusions regarding differential effects by subject.
- Most of the individual sanctions in the NCLB regime—including offering students transfers, tutoring, or modest "corrective actions"—appear to have had no effect.
- Schools forced to undergo restructuring under NCLB posted significant improvements in both reading and math scores, suggesting that leadership change is an essential component of reform in persistently low-performing schools.
- While a pure focus on proficiency can lead to scenarios where schools divert resources from higher- or lower-performing students, complementary policies focusing on those students appear to mitigate the risk substantially.
- State and local initiatives have taught us much about promising strategies for offering schools incentives to improve student performance. NCLB encouraged a bottom-up approach to some extent, but in the final analysis did not go far enough. In imagining

"accountability 2.0," evidence indicates that a series of modifications to the NCLB approach would improve the system:

- **Focus on test-score gains, not levels.** Many states have already moved toward implementing "value-added" systems that more directly measure the progress that students make while enrolled in a school. Value-added measurement is not a panacea; it is less transparent than simple proficiency measures and introduces tricky questions about what to expect from disadvantaged students. But the benefits of using it outweigh the costs.
- **Incentivize schools, not teachers.** Teacher-level value added cannot be measured for most public school teachers today, and for elementary teachers it takes multiple years of test score data to form a reliable picture of performance. Rewarding teachers for value added introduces incentives to avoid cooperating and to engage in zero-sum competition for better students. School-level incentives avoid these problems and have been shown to be equally, if not more, powerful in shifting behavior.
- **Intervene with, rather than summarily fire, underperforming teachers.** Recent studies have documented effective ways of delivering performance-improving feedback to teachers on the basis of classroom observations. Recent proposals to systematically fire underperforming teachers assume the existence of a "reserve army" of competent teachers; it is not clear that any such pool exists.
- **Move local autonomy even further.** NCLB relied on schools to figure out how to improve performance on their own but retained a top-down incentive structure. States and districts can play a much greater—and potentially much more effective—role in crafting rewards or punishments for schools.

- It has been six years since NCLB expired, and there appears to be little or no momentum to reauthorize it at this time. Nonetheless, the school accountability movement is alive and well, as evidenced by the federal Race to the Top initiative and countless other state and district initiatives to more carefully scrutinize the return on investments of public dollars in the K–12 education system. The lessons learned from NCLB and other first-generation accountability systems promise to make these new efforts more productive.

Introduction

The No Child Left Behind Act of 2001 (NCLB) was a piece of legislation with a clear vision. It imagined a world where every student in the United States—particularly those belonging to historically disadvantaged groups—met basic standards of literacy and numeracy. It imagined a path to that world that did not require spending significantly more money, but rather focusing on directing resources more efficiently toward this basic goal. In a foreshadowing of the era of "big data," it introduced a national mandate for standardized testing, imagining test scores as the key to understanding and addressing the shortcomings of the American education system.

Recognizing the underlying diversity of America's schools, it did not come with an instruction manual, leaving the task of figuring out how to accomplish the law's goals to states, districts, schools, and individual teachers. In place of explicit instructions, the law introduced a series of punitive sanctions for schools that failed to meet performance targets. It combined autonomy with accountability, a combination which has been lauded as highly effective in international comparisons.[1]

A dozen years after Congress passed the law, and a half dozen after it officially expired, postmortem analyses overwhelmingly suggest NCLB has been a failure. The goal of bringing all children up to the proficiency standard by 2014—even those with extremely severe learning disabilities—has been dismissed as unrealistic, even though most states have adopted proficiency standards far weaker than those put forth by the US Department of Education. The Obama administration gave states the option of waiving the universal proficiency target in 2010, and the majority of states have exercised that option. Researchers have exposed some of the drawbacks of encouraging schools to focus on universal proficiency, documenting tendencies for schools to divert resources from more advanced students.[2]

The system of sanctions designed to spur schools to action has been derided as toothless and ineffective. Federal initiatives in education policy have instead moved on, with the Obama administration championing its Race to the Top (RTTT) initiative rather than expending political capital on NCLB. Importantly, the RTTT initiative represents a move away from the philosophy that autonomy and accountability can foster school improvements without significant new funding.

No Child Left Behind was a complicated law, constituting 670 pages of fine print. To ask the simple question of whether the law succeeded or failed is to ignore a potential wealth of information regarding the specific provisions of the law. Even if the goal of 100 percent proficiency by 2014 has been abandoned, significant scope remains for the law's provisions to improve public education. Even with little to no momentum to renew the legislation in its entirety, federal and state governments, as well as individual school districts responsible for most policy decisions within their borders, stand to benefit from a comprehensive analysis of where the law succeeded and failed.

Many of the important lessons of NCLB have yet to be learned. Students educated under the law's provisions continue to filter into colleges and the labor force, and an assessment of the long-run impact can only occur once we observe the end product of the education system—skilled and self-actualized adults. This brief reviews the evidence to date on the short- to medium-term impacts of the law.

Theoretically, the possible impacts of NCLB on education quality would be attributable to one of two mechanisms:

1. The basic goal setting of NCLB, coupled with the mere threat of punitive sanctions, might have impacted education quality even in schools never actually exposed to the sanctions. This would be most likely to hold in situations where NCLB occasioned a relatively low-cost refocusing of school efforts, or where schools perceived the costs of being sanctioned as high enough to justify more intensive interventions in their practices.
2. The punitive sanctions themselves might have improved education quality. The NCLB sanctions were not designed to be socially wasteful; rather, they were intended to force schools to take an escalating series of actions that could, at least in theory, help students. The sanctions, described in more detail later in the paper, range from offering students free transfers to other public schools to restructuring low-performing schools.

Researchers have devoted some attention to both possible mechanisms. After reviewing the basic structure of NCLB in the next section, we examine the evidence on both topics.

The Structure of NCLB

NCLB, introduced as legislation in 2001 and signed into law in early 2002, was first implemented in the 2002–03 academic year. Among other things, the law required states to assess student performance in every public school receiving federal funding, using standardized reading and math tests in grades 3–8. States were expected to establish a series of objectives leading to the overriding goal of 100 percent proficiency by 2013–14. Proficiency on standardized tests, along with other measures such as high school graduation rates, were to be used to determine whether individual schools made adequate yearly progress (AYP) toward the overriding goal in both subjects. States were permitted to implement their own tests and to choose their own definition of proficiency. Research has established that states tended to select proficiency standards well below what the federal government employs on its National Assessment of Educational Progress (NAEP) tests and that standards vary widely across states.[3]

Importantly, the determination of AYP was required to incorporate not only the overall performance of a school's student body but also measures specific to the performance of students belonging to a series of subgroups, including students in particular racial categories, free or reduced-price lunch participants, special education students, and students with limited English proficiency. As long as the number of students belonging to one of these groups exceeded a state-defined threshold, the AYP designation for the entire school rested on the performance of the group. This forced schools to focus on traditionally disadvantaged groups and created a scenario where more diverse schools faced more hurdles to establish AYP.

States were permitted to introduce two forms of flexibility in the AYP determination process. First, the law acknowledged that standardized tests are imperfect statistical measures of proficiency and permitted states to grant AYP status to schools where student performance was close enough to the proficiency rate target that one could not reject the statistical conjecture that poor performance was a result of bad luck on test day. The so-called confidence-interval exemption resulted in particular leeway for schools working with small subgroups that barely met the state standards for incorporation into the AYP determination process.

Second, states were permitted to give credit to schools that made large strides toward meeting proficiency targets, even if they did not actually meet the targets. The safe-harbor exemption

applied, for example, to schools serving students with a low average level of kindergarten readiness. These students might have too much ground to make up to reach the proficiency standard by grade 3, but if the school was successful in bringing them toward that standard at a rapid pace, the AYP designation would be authorized.

Sanctions for Failure to Make AYP. The NCLB Act offered only a token amount of guidance to states, districts, and schools regarding how to meet the law's provisions. States were required to ensure that all students were taught by "highly qualified" teachers, though it left the important issue of defining the term to the states. Rather than provide a series of prescriptions, the law incorporated a series of punitive sanctions for schools that failed to meet state-specific AYP standards. The basic structure of the sanction regime is described in table 1.

Table 1. The NCLB Sanction Regime

Number of consecutive years missing AYP in same subject	Sanction
1	None; placed on watch list and must develop school improvement plan.
2	District must offer transfers (with transportation) to higher-performing public schools in the same district. School listed as "needs improvement."
3	District must offer supplemental education services to students qualifying for free or reduced-price lunch
4	School must undertake "corrective action." Corrective actions may include staff/leadership changes, curriculum changes, instructional time changes, or appointment of outside advisers.
5	School must formulate a restructuring plan.
6	School must implement the restructuring plan. Restructuring must involve either conversion to a charter school, replacement of the principal and most staff, state takeover, contracting with another entity to manage the school, or similar major changes to school governance.

Source: Derived from T. Ahn and J. Vigdor, "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina" unpublished manuscript, 2013.

Schools that consistently meet the AYP standard for their state in both reading and math are exempt from all sanctions. In schools that fail to meet the standard in one subject or the other, the consequences depend on that school's history of success or failure.

For schools with at least a two-year track record of making AYP in both reading and math—and for all schools in the first two years of the NCLB regime—few consequences exist for an initial failure to make AYP. Schools are placed on a "watch list" and required to develop a school improvement plan. They are exposed to the threat of further sanctions if they subsequently fail to make AYP in the same subject.

Once placed on the "watch list," schools can be removed, and other sanctions canceled, if they make AYP for two consecutive years.

Should a school fail to make AYP a second time in the same subject, the district is required to offer students the option to transfer to a higher-performing school within the district, with free transportation. In theory, the transfer offer could improve student outcomes by providing parents with more information about choice alternatives or opening up a route to enrollment in certain schools other than residence in the associated catchment area. In practice, behavioral economics research shows that individuals have a strong tendency to select the default option when presented with a complicated choice; thus the offer of a transfer might not matter much if the student's default school option remains the same. In practice, the take-up rate of transfers under NCLB has been low, though research has shown that simple informational interventions—like sorting a list of school alternatives by test scores rather than alphabetically—can make a difference.[4]

After a third failure, the school must offer its free or reduced-price lunch students supplementary education services (SES), tutoring typically provided by third-party vendors, at the district's expense. These supplemental services could, theoretically, directly lead to improved test-score performance, so long as districts select competent vendors and find ways to cover the cost that do not entail a countervailing negative impact.

Following a fourth failure in the same subject, schools are required to take "corrective action," selected from an eclectic menu of reforms, ranging from changes in the curriculum to the hiring of outside advisers. It is hard to assess the potential impact of this sanction; one might expect schools to opt for the least invasive option on the menu, even if it happens to be the least effective.

The fifth and sixth consecutive failures require the formulation and implementation of a restructuring plan, respectively. If other reforms are thought of as band-aids, restructuring represents major surgery. Restructuring plans once again can be constructed from a menu of options that are more consequential this time around, ranging from reconstituting the school as a charter school to accepting a state takeover to replacing the principal and most of the faculty. Once again, this is a sanction of highly uncertain impact. Restructuring could be a source of significant improvement in persistently low-performing schools, depending on the quality of the personnel brought in.

From a strictly rationalist perspective, one need not believe that these sanctions would yield improvements in education to think NCLB would be a beneficial law. One need only believe that these sanctions would be costly to those who operate schools. Were that the case, school officials would have a natural incentive to find their own path to school improvement. Nonetheless, it is clear these sanctions were not chosen to be purely punitive; in each case, there is at least some argument that students could benefit. At the same time, in each case there is reason to doubt that such benefits would be realized.

Early Concerns about NCLB. Test-based accountability systems such as NCLB have been criticized for a number of reasons. Testing of only mathematics and reading creates an incentive to deemphasize the instruction of other subjects. Emphasis on testing could lead some teachers to "teach to the test." Proponents might counter that tests can be designed to emphasize exactly those subjects that society most wants students to learn.[5] Research has also shown that attaching high stakes to standardized tests has led school administrators to engage in a variety of tactics that boost test scores without necessarily enhancing lifelong learning, such as increasing the calorie content of school meals during testing week.[6]

Beyond these general concerns, the specific structure of NCLB spawned strong reactions. The use of unfunded punitive sanctions in failing schools was considered by some to be a recipe for exacerbating inequality. Proponents argued, with a basis in evidence, that low-performing schools often have no shortage of dollars flowing through them—only a shortage of wise decisions about how to spend those dollars.

The use of proficiency, rather than a value-added measure of school performance, created additional concerns. Because so much of the groundwork for mathematics and reading is laid by parent(s) prior to the student's first day of kindergarten, the school does not have complete (or even, some would argue, much) control over a student's proficiency in these subjects. As a result of large differences in financial, educational, and time resources of parents, students across the state and even within a school will show up to school with widely varying readiness levels.

While measuring value added is conceptually preferable to proficiency, computing a value-added statistic for a school requires longitudinal education data, which to this day do not exist in some states. The establishment of longitudinal data sets has been opposed in some cases by those who fear their use for the purposes of evaluating schools and teachers. In more recent years, the Obama administration has offered further incentives for state education agencies to resolve these political difficulties and create systems for tracking student performance over time. In 2001, however, requiring value-added data was a nonstarter.

The federal government left the setting of the actual standards of what qualifies as "proficient" to the states. Given the punitive nature of the sanctions associated with failing to push students above the proficiency threshold, states had very clear incentives to set a low bar. In addition, some states set forth a plan to reach 100 percent proficiency entailing small strides in early years followed by great leaps approaching 2014. It is clear that at least some of these states realized the unrealistic goal of NCLB would force the federal government to revisit the issue sometime before the 2014 deadline and gamed the system accordingly.

Laudable Components of the Law. The No Child Left Behind Act did incorporate some praiseworthy design elements. First, it exposed schools that appeared to be doing a good job of educating their students in aggregate but were letting small numbers of disadvantaged students fail. One of the emphases of NCLB was that disadvantaged students had to be counted as separate subgroups. These students could not be swept up into an average measure for the whole school, but every individual subgroup's proficiency rates had to be reported. If a single subgroup failed to make AYP, the entire school was labeled as failing. This forced schools to change the way they allocated resources to educate students who were falling through the cracks because their bad academic results could not be hidden within the aggregate numbers. Of course, a similar result could have been obtained with an appropriately designed value-added system,

which would also have exposed schools that accomplished strong results primarily by starting with advantaged students and doing little for them. But within the data constraints of the era, the subgroup focus was a reasonable compromise.

As we have noted, NCLB was premised on the notion that significant improvements in public education could be achieved without the commitment of significant additional funds. It was inspired by a voluminous amount of data collected over many years documenting little correlation, either in cross-section or time series, between education spending and education outcomes. The question of whether money matters in education is still controversial; one might argue that money matters when spent wisely, but the degree of inherent wisdom in the system is a matter of some dispute. Regardless of one's prior beliefs on this question, one can be grateful for the opportunity to collect empirical evidence as to whether interventions designed in this manner can yield meaningful improvements.

At this point, a dozen years after the legislation was introduced, we have evidence at our disposal.

The Systemic Effects of NCLB

No Child Left Behind might have improved educational outcomes "without firing a shot," so to speak, by encouraging schools, districts, and states to improve practice to avoid punitive sanctions. To the extent they responded to this encouragement, the law can be credited with improving education even without imposing any specific sanctions.

In practice, discerning whether NCLB had systemic effects is a difficult task. Basic time-series NAEP data suggest some improvement in test scores over the past decade, but the key question for policy evaluation is what the test score trend would have been in the absence of the law. Sophisticated studies pay special attention to this counterfactual question.

Eric Hanushek and Margaret Raymond address the counterfactual question by using the fact that prior to NCLB, many states had already begun to institute their own accountability systems.[7] Although not a direct test of NCLB—indeed, the study was published at a point where very little post-NCLB data existed—the study offers an opportunity to test the underlying hypothesis that

autonomy and accountability yield superior results. Using this strategy, the researchers estimated academic growth in the National Assessment of Educational Progress (NAEP) test , a low-stakes reading and mathematics test administered to a sample of students nationwide at 4th, 8th, and 12th grades.

The authors found that states that instituted accountability systems that attached consequences for schools not making the grade experienced larger test-score gains compared to states that had adopted accountability systems that merely reported the testing results. Focusing on ethnicity, the researchers found that white, black, and Hispanic students all experienced gains in test scores, and although the academic achievement gap between whites and Hispanics narrowed, it widened between whites and blacks.

Thomas Dee and Brian Jacob expanded on the basic Hanushek/Raymond research design, treating the introduction of NCLB as a "natural experiment" that would be expected to have a disproportionate impact in states that did not already have their own accountability systems.[8] This quasi-experimental research design would determine the systemic effect of NCLB under the assumption of no effect whatsoever in states with existing systems. If NCLB did, in fact, have an impact in those states, the research design would underestimate the total impact.

Dee and Jacob report relatively large gains in mathematics scores for the average 4th graders and larger impacts for low-income, black, and Hispanic students. They also found that high-achieving students were not harmed by the adoption of NCLB as found by other researchers, addressing the concern raised by Derek Neal and Diane Schanzenbach in their study of an accountability program in Chicago Public Schools.[9] Dee and Jacob report no effects of NCLB on reading scores, consistent with a host of education policy evaluations which find stronger effects in math than in reading.

Other studies find impacts of differing magnitudes (and some even negative results), but the general consensus in the literature seems to be that NCLB resulted in modest increases in mathematics test scores and little to no impact for reading test scores.

These papers represent solid efforts to address a difficult program. By focusing on NAEP test scores, rather than state assessments, they address concerns that accountability produces illusory effects on test scores achieved by gaming the system rather than improving education. The

NAEP is a low-stakes test from every school's perspective. The NAEP is limited in the sense that it is administered not every year, nor to every student, rendering more detailed analysis impossible.

Though they may represent the best efforts undertaken to date, these efforts to determine the systemic effects of NCLB have limitations worth emphasizing. State-level accountability systems vary in many ways, and although both studies are cognizant of this variation, they cannot hope to provide a complete understanding of which elements of a system are most responsible for generating beneficial effects. Hanushek and Raymond divide states into "report card" and "consequence" accountability categories but do not distinguish across the broad variety of potential consequences.[10] Dee and Jacob also do not account for the fact that their control-group states are all experiencing different levels of accountability prior to NCLB. Control-group states with weaker accountability systems may in fact have been better served by NCLB because NCLB sanctions had more bite than the state system.

In spite of these concerns, it is important to emphasize the main conclusion of research undertaken to date: accountability systems have been associated with improvements in student performance, even on low-stakes tests. The American evidence is corroborated by international studies documenting that nations combining strong accountability with local autonomy tend to outperform others on international student assessments.[11]

Why do accountability systems work? Given the complexity of systems like NCLB, there are numerous theoretical arguments. Further analysis of the impact of NCLB sanctions promises to help distinguish among them.

The Effects of NCLB Sanctions and Threats of Sanctions: History Matters

This section summarizes the results of a study we have undertaken using administrative data on the entire population of students enrolled in North Carolina public schools between 2002–03, the first year of implementation of NCLB, and 2010–11.[12] In this study, we examine the impact of each NCLB sanction, and just as important, of the threat of these sanctions, on schoolwide test-

score gains in mathematics and reading. As we have described, NCLB sanctions and threats escalate as schools continue to fail to make AYP in a single subject in successive years.

The simplest method of inferring the impact of any given sanction or threat would be to compare exposed and unexposed schools at a point in time, to determine whether the former set of schools posts significant test-score improvements relative to the latter. The concerns with such a strategy would be that passing and failing schools are likely to differ systematically in ways that are difficult to fully undo in one year. We would expect to see sanctioned schools performing worse—after all, that is what led them to miss AYP in the first place.

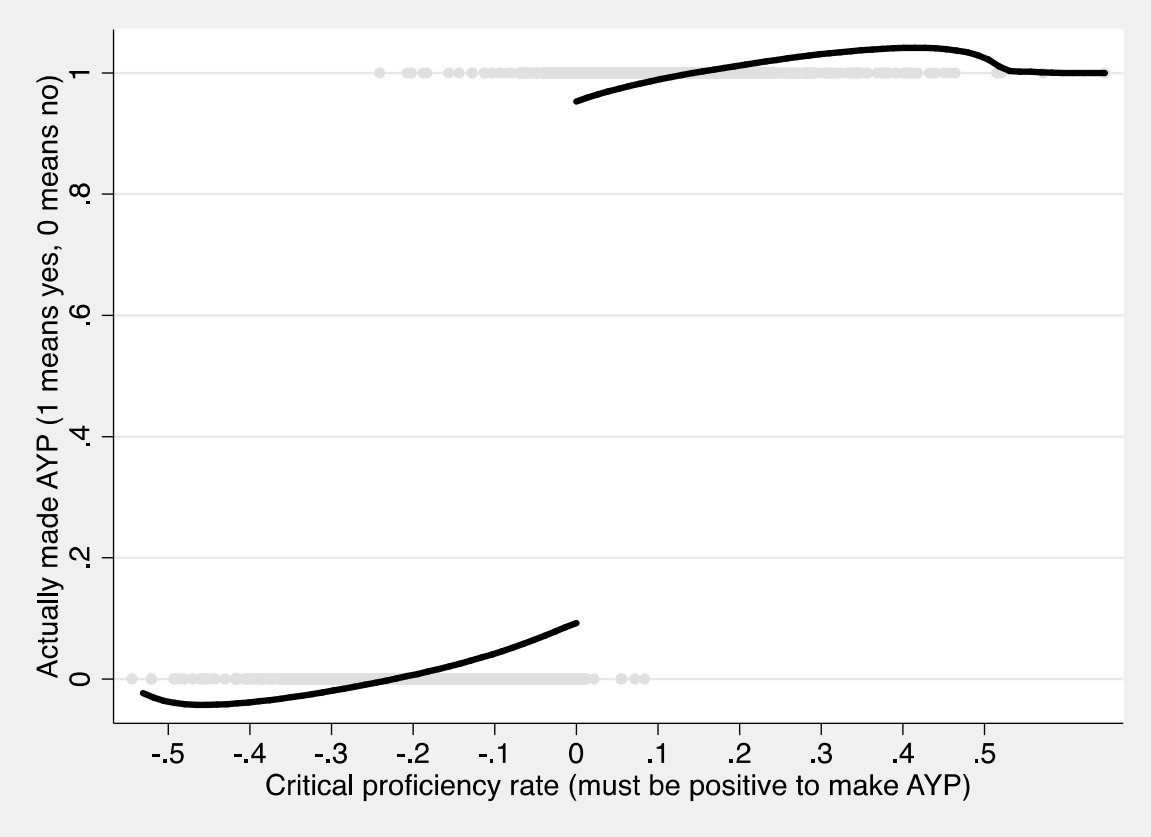
To infer the impact of NCLB sanctions, we employ a simple yet sophisticated strategy known as regression discontinuity design. This strategy can be employed in scenarios where exposure to a sanction is reserved for those schools that miss a strict cutoff. In such a scenario, one can make the argument that schools just above and below the cutoff are not all that different from one another, yet they are treated very differently by the accountability policy. The NCLB sanction regime offers us a perfect opportunity. For example, in schools required to obtain a 65 percent proficiency rate to meet AYP, our strategy entails comparing the schools that scored 64.9 percent to those at 65.1 percent. The difference between the two may be as little as one student making or missing the proficiency level—a nearly arbitrary distinction seeing that idiosyncratic factors such as student illness or classroom distractions could affect performance on a given day.

Given the nature of the NCLB accounting system, particularly the use of the safe-harbor and confidence-interval exemptions we have described and the requirement that individual subgroups also meet the proficiency threshold, our strategy is somewhat more complicated than a simple comparison of schools with near-identical proficiency rates on either side of a fixed cutoff. We describe our actual strategy in some detail in the paper on which this brief is based, but it can be summarized as follows: Each school, in each year, has a set of predetermined thresholds—for the school and each subgroup therein—that mark the minimum qualification for AYP. These predetermined thresholds are a function of the school's size and past performance, incorporating both the safe-harbor and confidence-interval exemptions.

Beginning with knowledge of where these thresholds are, we can then examine a school's performance in a given year to determine which schools just barely qualified or missed

qualification. We then compare the two sets of schools in the following year to determine whether the schools exposed to a sanction performed better or worse in terms of improving student test scores than those who barely escaped exposure to the same sanction.

Figure 1. Likelihood of a School Making AYP as a Function of Proficiency Rates



Source: North Carolina Department of Public Instruction data; authors' calculations.

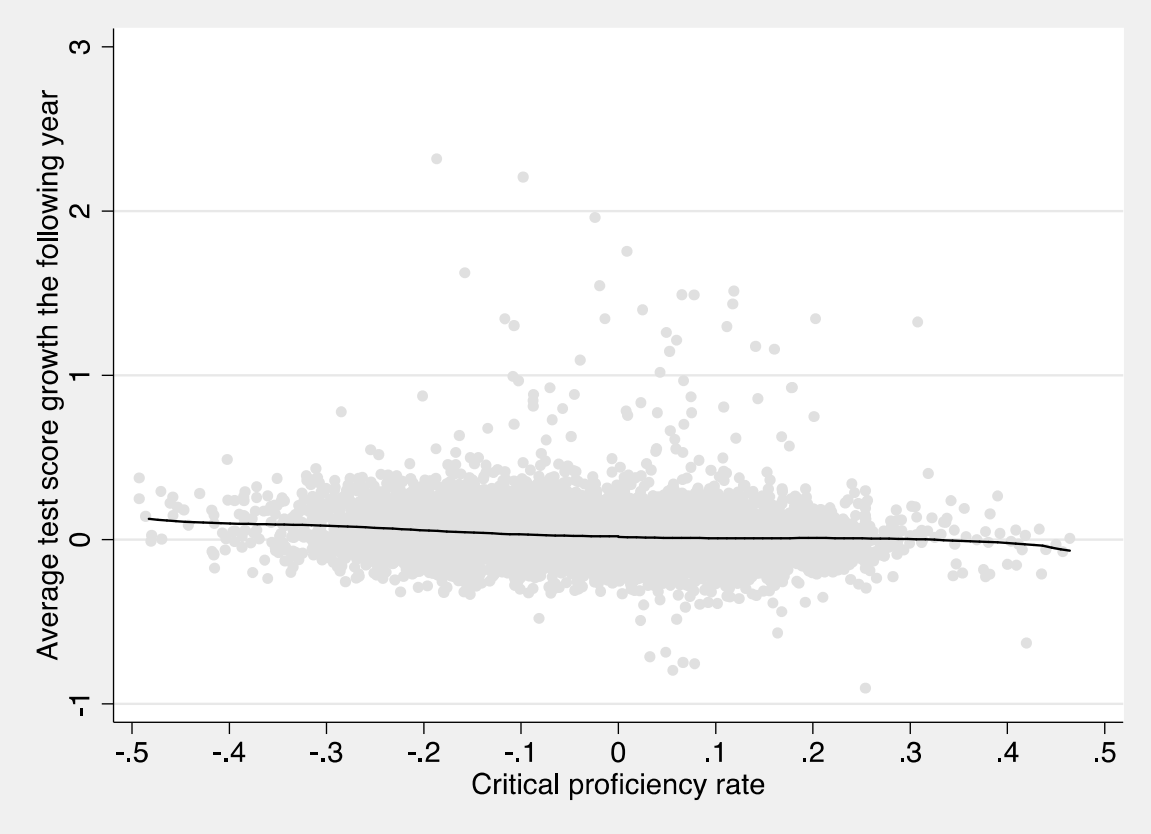
Figure 1 documents that our strategy works. It shows the relationship between the "critical proficiency rate," a variable that, when equal to or just above zero, signifies that a school has barely qualified for AYP, and the actual administrative coding of AYP status. The black line is a statistical estimate of best fit; the gray dots represent the actual underlying data. We do an imperfect job of matching AYP coding, in part because determination of AYP depends on factors other than test scores—including test participation rates and graduation rates for high

schools. Nonetheless, it is clear that schools on the right side of zero in this chart are treated very differently from schools immediately to the left of zero, on average.

To document the impact of NCLB sanctions, we will show graphs very similar to figure 1, plotting test-score measures on the vertical axis. These graphs enable us to determine whether schools exposed to a particular sanction or threat because they have just missed the AYP cutoff experience superior test-score growth in the following year. These graphs can also give us a general sense of what happens in schools that fall far above or below the threshold, but our claim to causality is weaker when examining those schools.

Our initial analysis combines observations on all schools in all years, examining the impact of failure to make AYP in a given year on performance in the subsequent year, ignoring the actual sanction imposed. Figure 2 shows that reading test-score growth is nearly identical in schools that make and miss AYP—there is barely any break in the best-fit line at the threshold. Across the board, change in test-score growth is rarely very far from zero, indicating that students tend not to make radical progress relative to grade-level norms in a single year. There is a slight downward tilt to the best-fit line depicted in figure 2. Although this could be read to indicate that schools with very low proficiency made more progress than those with high proficiency, this particular pattern could be a statistical artifact of mean reversion.

Figure 2. Effect of Failure to Make AYP on All Schools, All Years, in Reading

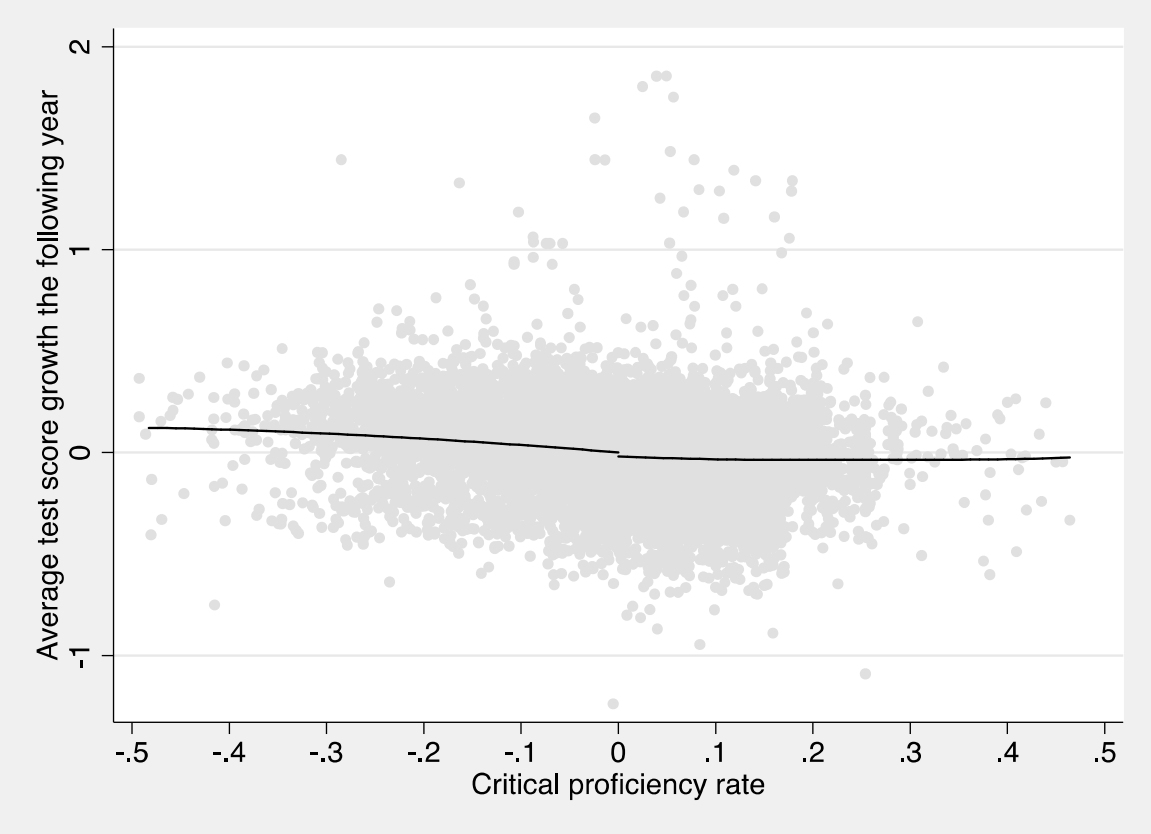


Source: North Carolina Department of Public Instruction data; authors' calculations.

Figure 3 shows that math scores rise slightly more quickly in schools that miss AYP. The magnitude of the effect, in terms typically used in the economics of education literature, is about 2 percent of a standard deviation. This is equivalent to saying that an "average" student—one performing at the 50th percentile—would be expected to improve to nearly the 51st percentile in a school that just missed AYP. Although small in magnitude, this result is statistically significant.

Once again, there is also a detectable tilt to the best-fit line, indicating that the worst-performing schools improved the most, perhaps because of mean reversion. These results echo those found in studies of systemic effects, indicating that the act of missing the AYP threshold spurs on at least some positive change in schools. As the evidence so far indicates an impact on math but not reading, we focus on math scores exclusively from this point onward.

Figure 3. Effect of Failure to Make AYP on All Schools, All Years, in Math



Source: North Carolina Department of Public Instruction data; authors' calculations.

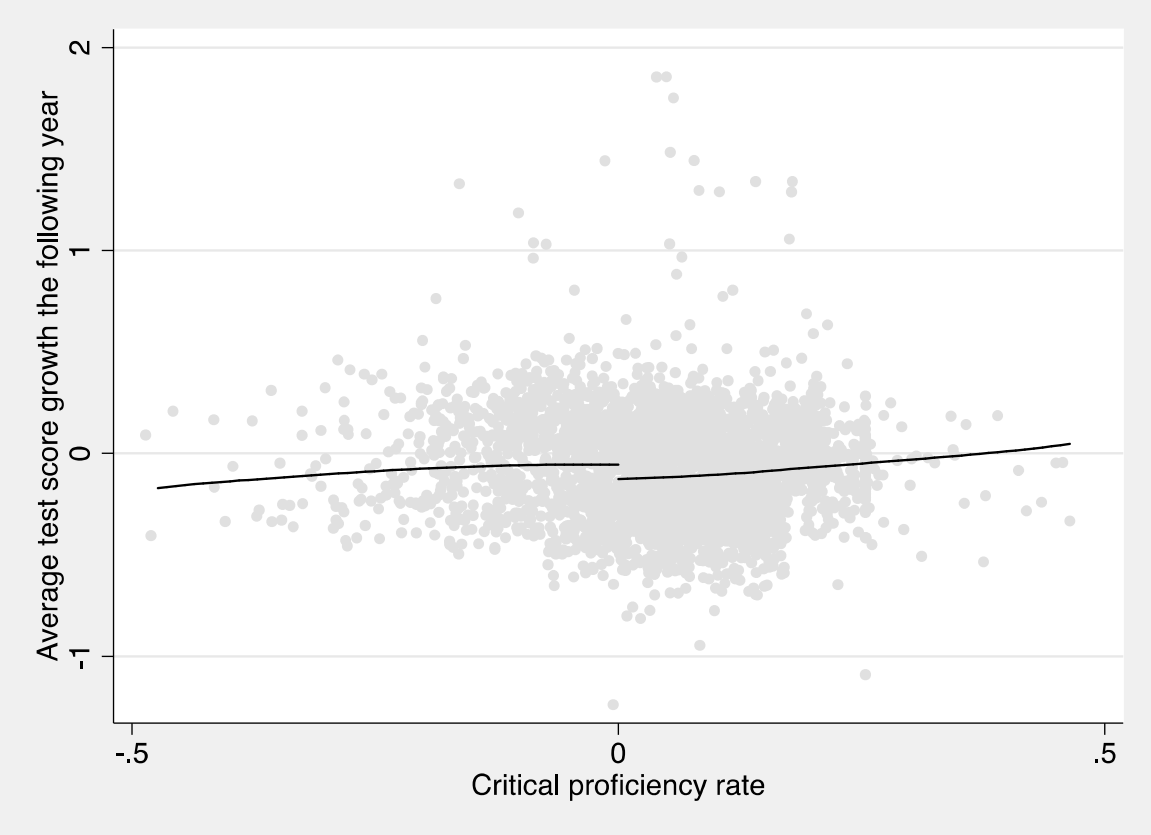
The data shown here provide evidence consistent with existing literature but do nothing to tell us which sanctions or threats matter. When we divide schools by the AYP history to see the impact of sanctions and the threat of sanctions, an interesting and cautiously encouraging pattern emerges.

For schools that had no previous history of problems in exceeding the state's threshold, their first failure to make AYP imposes carries no direct penalties. We might expect a strong policy response in these schools, however, because they have been placed at risk of sanction in the event they fail a second time. Figure 4 shows that this threat of sanctions is enough to get schools to use current resources more effectively or exert more effort to increase math standardized test scores.

The magnitude of the math score increase, on the order of 5 percent of a standard deviation, is on par with much costlier educational interventions, such as class size reductions as studied in the

Tennessee STAR experiment.[13] In the year after a school misses AYP for the first time, an average-performing student would be boosted from the 50th to the 52nd percentile. It is worth emphasizing again that no actual sanctions have been imposed on these schools—this intervention was virtually costless. The shock of failing for the first time is enough to induce marginal schools to exert themselves to achieve better results in the next year. Research by Rajashri Chakrabarti corroborates these results, finding similar patterns in a regression discontinuity-based study of Wisconsin public schools.[14]

Figure 4. Effect of Failure to Make AYP on Schools with a "Clean" History



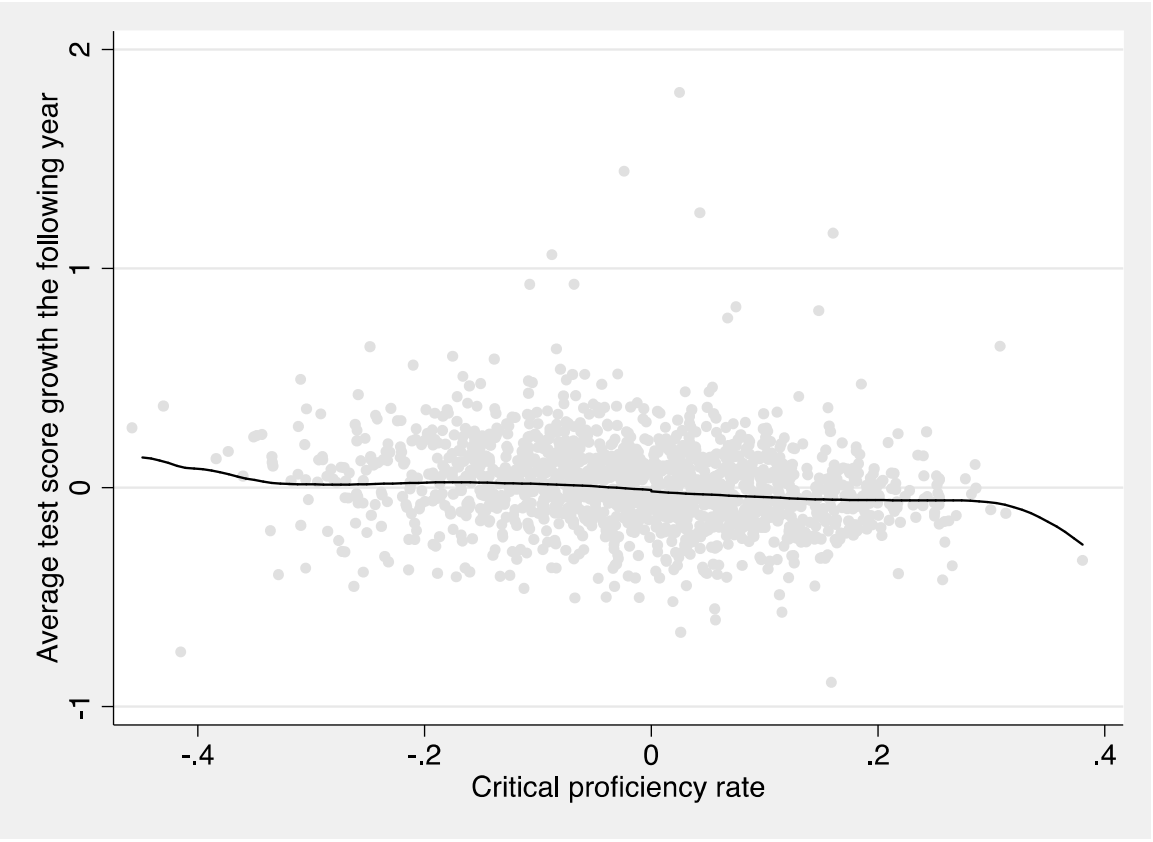
Source: North Carolina Department of Public Instruction data; authors' calculations.

It is interesting to note that the best-fit lines have a slight upward tilt to them in this graph. Among schools with no prior AYP failures, those that experience the worst failures are more likely to see continued failure the following year. Moreover, the lines fall below zero: the average school with a clean AYP history falls below average in the year after analysis here. We would expect both patterns to occur because some schools that have not run into NCLB trouble

in the past are just lucky; this graph suggests that their luck tends to run out. We note that there are significant effects on reading test scores among these schools as well, though those are smaller in magnitude.

One might expect that if the mere threat of sanction has such a strong impact on schools, the sanctions themselves would matter even more. In fact, with few exceptions, we report no effect from actually being exposed to an NCLB sanction.[15] Figure 5 shows a representative analysis of the impact of failure to make AYP when a school is at risk for the first sanction in the system, offering students transfers to another school. No significant difference exists between schools exposed to this sanction and schools that barely avoid it in terms of subsequent math score performance. There is some evidence of more remarkable effects at the extremes of the distribution, but once again these might simply reflect mean reversion in very high- or very low-performing schools.

Figure 5. Effect of Failure to Make AYP on Schools at Risk for Offering Transfers

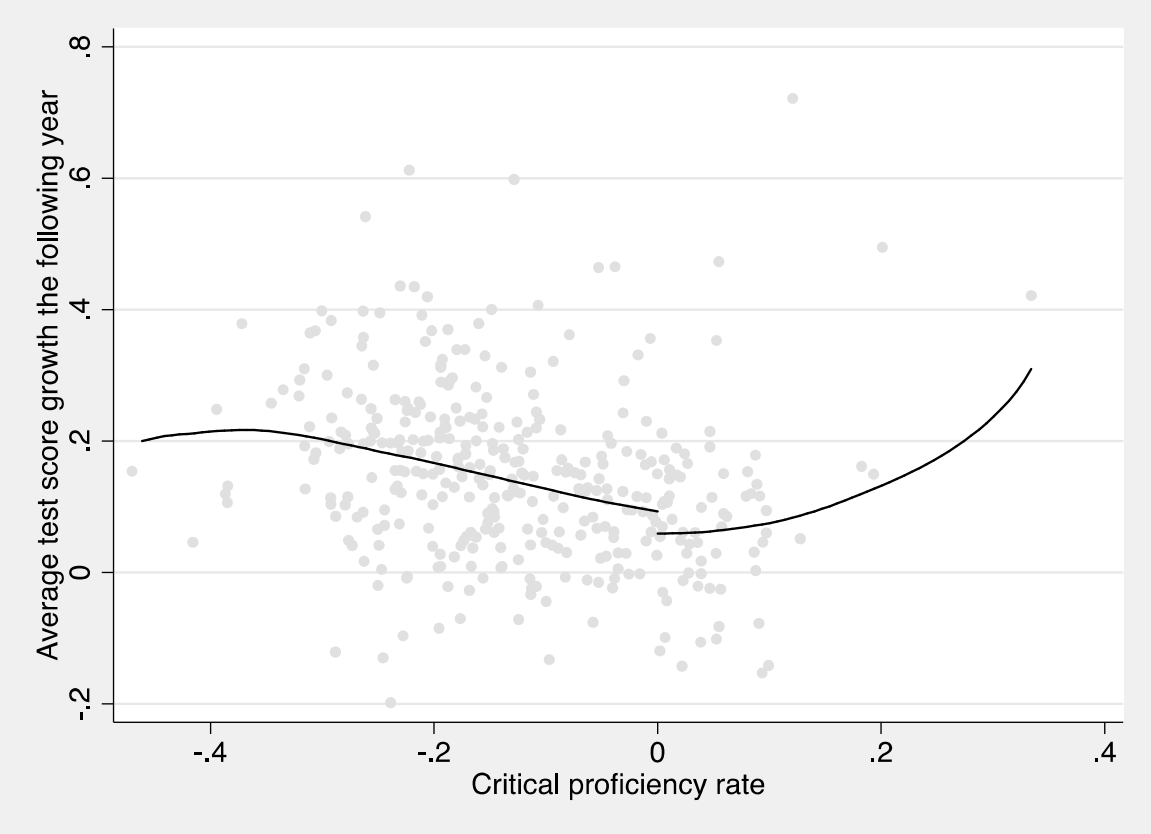


Source: North Carolina Department of Public Instruction data; authors' calculations.

We could draw very similar graphs for other NCLB sanctions—supplemental education services, corrective action, and the formulation of a restructuring plan. We can find no evidence that exposure to any of these sanctions improves student achievement. This may be surprising, particularly in the case of supplemental education services. One might think that providing tutoring services to disadvantaged children would help them. Recall, however, that the SES mandate was unfunded, implying that schools on fixed budgets would need to cancel other programs to pay for the tutoring. It is also unclear whether schools selected SES providers carefully.

Eventually, after six consecutive years of failing to make AYP, schools are exposed to the ultimate sanction—restructuring. Figure 6 shows the consequences of failing to make AYP among schools that will have to implement a restructuring plan as a result. Perhaps unsurprisingly, most of the gray dots, which represent actual schools, fall to the left of zero. Most schools that have failed to raise proficiency rates for five consecutive years fail again when facing the restructuring threat. Among those exposed to the sanction, however, we observe a significant improvement in math test scores. The magnitude of the effect is roughly 5 percent of a standard deviation, and in this case there is a similar-size effect on reading test scores as well. Again, this effect can be interpreted as equivalent to boosting a 50th percentile student to the 52nd percentile.

Figure 6. Effect of Failure to Make AYP on Schools Facing Restructuring Threat



Source: North Carolina Department of Public Instruction data; authors' calculations.

Several other features of this plot are interesting. Unlike most of the other plots, the best-fit lines do not hover around zero. After a year in which they faced the threat of restructuring, most schools exhibit above-normal test score improvements. The improvements tend to be greatest in those schools where the threat of restructuring was carried out—the data points on the left of the chart tend to be higher than those on the right. There is an upward flip to the line on the right-hand side of the chart, which reflects a single data point that might be dismissed as an outlier.

Thus, among the various NCLB sanctions, the one with the strongest evidence of making a difference is the ultimate one, restructuring. As the restructuring sanction entails the replacement of a school's leaders, it points to the quality of leadership—or lack thereof—as a primary factor in poor school performance. This conjecture could help explain the meager results found in analysis of other sanctions. One characteristic of poor leaders may well be that they do not

understand what actions to take when a school is sanctioned or faces a sanction threat. When forced to select a tutoring program, for example, they do not choose wisely.

The notion that school leadership significantly determines success is widely subscribed to, but to this point has been hard to establish in a quantitative study with a strong claim to causal validity. In this case, we can argue that schools near the AYP threshold in figure 6 were effectively randomly assigned to replace their leaders, and those that did experienced significant gains.

One concern about an accountability system like NCLB is that it may incentivize schools to direct resources and effort to students just below the proficiency threshold and away from those at no risk of falling below it (and even, in some cases, those with no chance of reaching it). Neal and Schanzenbach found exactly this pattern in their study of Chicago Public Schools.[16]

Our own analysis of North Carolina public schools under NCLB does indicate that test-score improvements in sanctioned schools were concentrated among those students near the proficiency threshold, but we find no evidence of negative effects on any students. In fact, the positive impacts of restructuring appear to apply up and down the achievement distribution—new leaders find ways to improve the education of all students, not just those at the proficiency threshold.

The absence of adverse impacts for high- or low-performing students may be an artifact of North Carolina's accountability system that existed simultaneously with NCLB. The North Carolina system pays cash bonuses of up to $1,500 to teachers in schools where student test-score growth—not proficiency, but year-over-year improvements—exceeds a predetermined value. Raising proficiency by reducing the quality of education provided to students at the ends of the distribution would place a school at risk of losing actual compensation. We conclude that the phenomenon of leaving high performers behind to make others proficient is not an inherent feature of an accountability system.

Envisioning Accountability 2.0

Six years after the expiration of NCLB, where do we go from here? Contrary to popular opinion, evidence indicates that the act did have positive impacts on schools simply by introducing

consequences for bad performance and additionally by spurring needed leadership change at some of the most troubled schools. At the same time, the NCLB regime was far from perfect. Many of the sanctions it forced on schools appear to be entirely ineffective. Moreover, the focus on proficiency rather than improvement necessarily entailed punishing some schools for working with students who started at a low level.

Over the past dozen years, two innovations in public education have already begun to point the way forward. Neither is perfect; an ideal system would capture the best elements while rethinking more troubling ones.

The first is an emphasis on measuring and rewarding value added—the propensity of a school to improve student performance. As noted above, the primary impediment to introducing this type of reward system a decade ago was the absence of the longitudinal data required to track student performance. Since then, the federal government has provided incentives for the creation of student data systems in states and districts that lack them. Many of the states that requested waivers from NCLB in fact proposed instituting value-added statistics as an alternative means of evaluating teachers and/or schools.

The value-added system has the benefit of more directly reflecting the gains made during the academic year. In its simplest form, it looks at test-score gains from one year to the next to gauge the performance of the teacher or the school. Correction factors attempting to account for advantages or disadvantages a school may face because of resource availability or student body composition can be easily added to value-added models. It can be argued that this system is more "fair," as a school is not penalized if it is located in a district with students who arrive at school with poor preparation.

In addition, the value-added system is fairly inexpensive to implement. As long as the original standardized exam is well-designed and correctly calibrated, the value-added system can piggyback on the existing system while adding little administrative cost or additional time and resource commitment from teachers or administrators.

That said, the value-added system is not without its own faults. As with any accountability system, it is not immune to gaming. Therefore, many of the criticisms levied at NCLB regarding unintended consequences from undesirable responses by teachers and administrators to the

incentives remain valid. In addition, the emphasis on average gains in the school may defeat the one unequivocally positive development from NCLB, the increased focus on traditionally disadvantaged subgroups, as schools face new pressure to focus resources on the students most likely to make the largest test-score gains from year to year.

Although the idea of measuring how much knowledge students gain from year to year is intuitively appealing, there is an inherent problem in test coverage. Not every student in every subject is tested every year. For example, because students in kindergarten (and, in many cases, 1st and 2nd grade) are not administered a standardized exam, there is no previous-year score to assess value-added gains. In middle and high schools, subjects such as social studies and foreign languages are typically not tested. Grades and subjects that are not tested complicate truly evaluating how much a student body has learned since the previous year.

Another concern with value-added systems is that they are less transparent than simple proficiency statistics. Although NCLB has one proficiency target that every school in the state must pass, value-added systems necessarily have individualized targets for each school based on test scores from last year (plus other correction factors). If one of the goals of accountability systems is to disseminate easy-to-understand information about a school to parents to allow them to make informed consumption decisions, increasing the complexity of the accountability system will always be a double-edged sword: more information may help to inform some principals, teachers, and parents but confuse others.

Additionally, the correction factors we have mentioned can themselves be contentious. To "correct" the value-added standard for some schools implies that the bar is lowered for some schools or raised for others. Although it may be easy to argue in broad terms that making disadvantaged schools try to hit the same target as more advantaged schools is not fair, our best description of a disadvantaged school is one which serves a traditionally disadvantaged student population, namely racial minorities. Lowering the bar for schools that serve minority students reintroduces concerns about the "soft bigotry of low expectations" raised by then-candidate George W. Bush in the election campaign that directly preceded the passage of NCLB.

Many of the limitations of value-added systems are particularly acute when systems are intended to provide teacher-level evaluations. North Carolina's growth-based accountability system

illustrates another way forward by focusing on school-level performance data and not worrying about individual teachers. In theory, one worries about the "free-rider" effect when examining group rather than individual incentives. As we have noted in a previous policy brief for AEI,[17] the free-rider effect is, in practice, offset in the education context by a more important force.[18] Group-level rewards create incentives to collaborate, and collaboration turns out to be a potent force in education. When teachers share information about students and instructional strategies, education improves. A hybrid system, then, with NCLB-style school-level incentives but North Carolina–style emphasis on test-score growth, introduces many of the benefits of a complete value-added system without the logistical drawbacks.

In the era before standardized tests, teacher evaluation rested on classroom observation. Although many have complained that these traditional evaluation systems were too lax, these concerns rest largely on the standards applied and not necessarily on the data collection method. When information gleaned from classroom observations is used to actually intervene with underperforming teachers, rather than tucked away in a file cabinet for eternity, significant improvements can result.

Taylor and Tyler evaluate a program in Cincinnati public schools in which master teachers were paid to become full-time classroom observers and observers were tasked with intervening with teachers.[19] Teachers who completed the observation-feedback process showed significant improvement in subsequent performance.

These systems have the obvious benefit of solving many of the problems associated with exclusive reliance on test scores. Ideally, they represent a more holistic evaluation of teachers or schools, allowing us to learn more about their strengths and weaknesses in a much more detailed manner than is plausible from merely observing test scores. Although there have been recent policy proposals to summarily dismiss low-performing teachers, identifying cost-effective means of improving their performance would have the twin benefits of causing less turnover in the system and leading those devoted to teaching to become more successful at it.

Although the benefits are clear, there are two reasons why these systems do not enjoy more widespread adoption. The first is cost. Evaluation and feedback require a large number of trained evaluators, many man-hours to visit schools, and large administrative costs to run the programs.

Taylor and Tyler's analysis suggests that a Cincinnati-style intervention is cost-effective but that many school districts face significant budget limitations.[20]

Related to these issues is the worry about the fidelity of implementation. It may be feasible to recruit and train enough good evaluators to run such programs on a small scale or in a small country, but scaling these programs up to the large-scale state or national level may not work so well. It is difficult to imagine a training program comprehensive enough to take a random person off the street and turn him or her into a competent evaluator who can observe and correct teachers to improve didactic technique. The pool of capable evaluators, which includes current and former teachers, is limited. Without competent, consistent evaluation, the system is bound to create more problems than it solves.

The existence of promising school improvement policies in places like North Carolina and Cincinnati reveal what is probably the most essential flaw in the No Child Left Behind regime: while the act can be praised for leaving the essential work of figuring out how to improve education to states, districts, and schools, it could have taken things a step further. Rather than mandate a series of nationwide sanctions—most of which would ultimately prove ineffective—it could have left it to local jurisdictions to design their own sanction regimes. Had this been done, the nation's "laboratories of democracy" might have produced a wealth of evidence on promising new strategies for school improvement.

Having espoused the worth of experimentation, we must follow with a caution derived from our own results. Permitting schools, districts, and state education agencies to experiment will work best if each of these organizations is led by individuals who think carefully about the design and implementation of education policy. Leaders meeting this description may, in fact, be in short supply.

Notes

1. L. Woessmann, "International Evidence on School Competition, Autonomy, and Accountability: A Review," *Peabody Journal of Education* 82 (2007): 473–97.

2. D. Neal and D. Schanzenbach, "Left Behind By Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics* 92, no 2 (2010): 263–83.

3. V. Bandeira de Mello, "Mapping State Proficiency Standards onto NAEP Scales: Variation and Change in State Standards for Reading and Mathematics, 2005–2009," US Department of Education, 2011.

4. J. S. Hastings and J. M. Weinstein, "Information, School Choice, and Academic Achievement: Evidence from Two Experiments," *Quarterly Journal of Economics* 123, no. 4 (2008): 1373–414.

5. E. Lazear, "Speeding, Tax Fraud, and Teaching to the Test" (working paper no. 10932, National Bureau of Economic Research, Cambridge, MA, 2004).

6. D. N. Figlio and J. Winicki, "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89, no 2–3 (2005): 381–94.

7. E. A. Hanushek and M. E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24, no. 2 (2005): 297–327.

8. T. S. Dee and B. Jacob, "The Impact of No Child Left Behind on Student Achievement," *Journal of Policy Analysis and Management* 30, no. 3 (2011): 418–46.

9. Neal and Schanzenbach, "Left Behind by Design."

10. See, for example, M. Carnoy and S. Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24, no.4 (2002): 305–31.

11. Woessmann, "International Evidence on School Competition, Autonomy, and Accountability."

12. T. Ahn and J. Vigdor, "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina," unpublished manuscript, 2013.

13. A. B. Krueger, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114, no. 2 (1999): 497–532.

14. R. Chakrabarti, "Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition," Federal Reserve Bank of New York Staff Report #525, 2011.

15. Ahn and Vigdor, "The Impact of No Child Left Behind's Accountability Sanctions."

16. Neal and Schanzenbach, "Left Behind by Design."

17. T. Ahn and J. Vigdor, "Making Teacher Incentives Work: Lessons from North Carolina's Teacher Bonus Program," AEI *Education Outlook* (June 2011), www.aei.org/outlook/education/k-12/making-teacher-incentives-work-outlook/.

18. See T. Ahn, "The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation," unpublished manuscript, 2011, http://sites.google.com/site/tomsyahn.

19. E. Taylor and J. Tyler, "Can Teacher Evaluation Improve Learning?" *Education Next* 12, no. 4 (2012): 78–84.

20. Ibid.

About the Authors

Thomas Ahn is an assistant professor of economics at the University of Kentucky. Jacob Vigdor is a professor of public policy and economics at Duke University, a research associate at the National Bureau of Economic Research, and an adjunct fellow at the Manhattan Institute for Policy Research.