Testing and Accountability: What Have We Learned and Where Do We Go?

DEVEN CARLSON, UNIVERSITY OF OKLAHOMA

JANUARY 16, 2018
President Bush’s 2001 signing of No Child Left Behind (NCLB) produced an immediate and dramatic change in the federal government’s relationship with states, districts, and schools across the country. Gone were the days when the federal government doled out dollars to states and districts with relatively few strings attached. Instead, NCLB ushered in an era where—under the threat of withheld Title I dollars—states were required to adopt challenging standards, administer annual reading and math assessments aligned with those standards, and hold schools accountable on the basis of the assessment results. Almost overnight, the notion that standards, testing, and accountability could be combined in an effort to drive school improvement morphed from a diverse set of initiatives in place in about half the states to a federal policy affecting every public school and district across the country.

The fact that NCLB broke so sharply with past federal education policy effectively guaranteed that there would be growing pains as states and districts worked to meet the new federal mandates. The hope was that these pains would ultimately be justified by significant improvements in educational outcomes for students across the U.S. Whether the law fulfilled this justification has been—and remains—a hotly debated topic in the education policy community. Regardless of one’s view of the ultimate wisdom of NCLB, though, there is broad consensus that we have learned a wide range of important lessons about standards, testing, and accountability policy throughout the Bush-Obama educational era.

This chapter explores these lessons. It explores the lessons we have learned about the ability of standards, testing, and accountability to drive academic improvement, which is the primary goal of these policies. Just as importantly, though, this chapter examines the lessons we have learned about the unintended consequences of these policies, such as the nature of states’ responses to federal mandates and accountability pressures. This chapter also explores what we
know today about the politics of testing and accountability policy that we did not yet know in
2001. Finally, the chapter concludes by discussing whether it is reasonable to think we can ever
get testing and accountability policy “right” and by predicting the direction these policies will
take in future years. Prior to delving into those topics, however, the chapter sets the stage by
briefly describing the evolution of federal policy on standards, testing, and accountability.

How Did We Get Here? Testing and Accountability at the Federal Level

Education fell almost exclusively under the purview of the states for the first 150-plus years of
the United States’ existence. States’ control of this issue is attributable to the fact that the U.S.
Constitution does not specify a fundamental right to education—an interpretation cemented by
the Supreme Court’s 1973 decision in San Antonio Independent School District v. Rodriguez—
combined with the 10th Amendment, which reserves powers not explicitly delegated to the
federal government for the states. Thus, through the mid-20th century there had been no
meaningful federal involvement in education in the United States. Passage of the Elementary and
Secondary Education Act (ESEA) of 1965 changed that. Part of the Johnson Administration’s
broader War on Poverty, ESEA was driven by concerns over educational inequity across the
country. This initial federal foray into education was designed to alleviate these inequities by
providing federal dollars to states, which would then distribute this money to districts in an effort
to improve the quality of education available to disadvantaged students. The law contained
nominal reporting requirements for states and districts, but contained no formal testing or
accountability mandates.

Throughout the 1970s, ESEA’s strict focus on promoting educational equity reflected the
prevailing view of the proper federal role in education. Beginning in the 1980s, though, there
was a growing sense that the federal government should promote educational excellence, in addition to equity. Initially, federal efforts to promote educational excellence took place outside of the legislative arena. For example, the 1983 report *A Nation At Risk* (released by a federal commission appointed by President Reagan) recommended that states adopt challenging standards and raise graduation requirements. Although the report caused quite a stir, its recommendations were nothing more than mere suggestions that states were free to take or leave. Similarly, at the 1989 National Education Summit, President George H.W. Bush and the nation’s governors agreed to develop a set of national education goals—one of which involved adopting challenging standards and administering assessments aligned to those standards—and hold themselves accountable for achieving those goals. States were again at liberty to ignore the recommendations of the National Education Summit, but the event is notable for its insertion of testing and accountability into the federal policy conversation.

It was not long before these events began to shape Congressional priorities. In 1994 Congress passed the Goals 2000 legislation, which codified the goals that had emerged from the National Education Summit, including the objective that students demonstrate mastery—via assessments—of challenging curricular content. Although the legislation was careful to note that participation in Goals 2000 was voluntary, it offered states financial assistance for developing and implementing plans for achieving the goals, so states that opted out effectively declined the possibility of millions of federal dollars. Despite the voluntary nature of Goals 2000, the legislation officially put the federal government on record in support of testing and accountability.

The feds reiterated their endorsement of testing and accountability through the 1994 ESEA reauthorization—titled the Improving America’s Schools Act (IASA)—that followed
close on the heels of Goals 2000. Whereas Goals 2000 provided financial incentives for states to adopt challenging standards and administer assessments aligned to those standards, IASA mandated it. In particular, IASA required states to adopt standards by the 1997-98 school year and implement assessment systems by the 2000-01 school year. The law mandated that states assess students in reading and math once in grades 3-5, 6-9, and 10-12. The bill, however, contained no provisions specifying sanctions or rewards on the basis of the assessment results—there was no accountability. Implementation and enforcement of IASA was haphazard, with the U.S. Department of Education routinely issuing waivers to states that were lagging in development of their assessment systems. Unsurprisingly, by the time the 2000-01 school year rolled around, a significant number of states had not developed any sort of meaningful assessment system.

**The No Child Left Behind Era**

The 2000 election changed the politics of education reform at the federal level. President Bush brought with him several years of experience with testing and accountability in Texas, a strong conviction that such was the optimal approach to school improvement, and a Secretary of Education—Rod Paige—who had significant experience with these policies in Houston. With education as one of their top domestic priorities, the Bush administration worked throughout 2001 with leaders of both parties in Congress to reauthorize ESEA. They bill they negotiated, which garnered substantial bipartisan support, pulled hard on the levers of testing and accountability in an effort to reach aspirational learning goals.

The legislation mandated that by the 2005-06 school year states adopt challenging standards in reading and math, and annually test students in grades 3-8 and once in high school
using an assessment aligned to those standards. The law also required states to identify grade-specific scores on those tests—cut scores—that corresponded to proficiency. Students scoring above the relevant cut score would be deemed proficient while students scoring below it would not. NCLB compelled states to annually report the percent of proficient students not only for each school, but also for each of several subgroups within each school. Specifically, similar to IASA, the law called for separate subgroup reporting on the basis of race/ethnicity, eligibility for free or reduced-price lunch, disability status, and English language learner (ELL) status.

Unlike IASA, NCLB did not end with subgroup reporting requirements. The law also set a goal of all students in each school—and in each subgroup within each school—reaching proficiency by the 2013-14 school year. Further, it required states to specify a trajectory that schools must maintain in order to reach the universal proficiency goal. Schools lagging behind that state-specified trajectory were classified as failing to make Adequate Yearly Progress (AYP). This requirement was stringent—a school was classified as failing to make AYP if just a single subgroup fell off the trajectory toward universal proficiency in either reading or math.¹ Schools that received Title I funding and failed to make AYP were subject to a series of sanctions—the severity of which increased with the number of consecutive years that a school failed to make AYP. The sanctions included offering students the ability to transfer to a school in the district that had made AYP, providing supplemental educational services (i.e., tutoring), and taking “corrective action” or “restructuring” with regard to some aspect of school personnel, management, or structure.² Table 1 summarizes the sanctions that schools faced for each consecutive year they failed to make AYP.
Table 1. Sanctions for missing AYP

<table>
<thead>
<tr>
<th>Consecutive year missing AYP</th>
<th>Sanction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>Transfer option</td>
</tr>
<tr>
<td>3</td>
<td>Transfer option; Supplemental educational services</td>
</tr>
<tr>
<td>4</td>
<td>Transfer option; Supplemental educational services; Corrective action</td>
</tr>
<tr>
<td></td>
<td>Transfer option; Supplemental educational services; Corrective action; Restructuring plan</td>
</tr>
<tr>
<td>5</td>
<td>Restructuring plan</td>
</tr>
<tr>
<td></td>
<td>Transfer option; Supplemental educational services; Corrective action; Restructuring</td>
</tr>
<tr>
<td>6</td>
<td>Restructuring</td>
</tr>
</tbody>
</table>

The U.S. Department of Education (USED) was serious about implementation and enforcement of NCLB and took a hard line with states, which were accustomed to the lax enforcement of IASA and thus prone to drag their feet in implementing NCLB. This hard-nosed approach likely contributed to states successfully implementing the testing and accountability provisions of NCLB on the specified timeline. This is not to say that the implementation process was all peaches and cream—there were plenty of hurdles and difficulties—but by the time the 2005-06 school year rolled around, states were largely in compliance with the testing and accountability provisions.

NCLB was scheduled for reauthorization in 2007 but it quickly became clear that the law would not be addressed on that timeline, and implementation of NCLB continued apace. In 2008,
just a few years after the testing and accountability requirements took effect, the U.S. was rocked by the financial crisis and ensuing Great Recession. The severe economic downturn decimated state and school district revenues, and officials began to look for dollars wherever they could. They found a potential funding source in the American Recovery and Reinvestment Act (ARRA), which Congress passed in 2009 as its primary legislative response to the Great Recession. Contained in ARRA were a few lines of statute that provided the Secretary of Education with $4.35 billion to operate a competitive grant program among states based on criteria the secretary “determines appropriate” and to provide grants to “consortia of states to develop academic assessments that are aligned with academic standards.”

Secretary Arne Duncan took advantage of this authority to design the Race to the Top (RTTT) grant competition, which based state award decisions—each worth up to hundreds of millions of dollars—on states’ plans to advance reforms on each of several priorities identified by USED. The priority most relevant here is the adoption of “college and career-ready” standards, and implementation of common, high-quality assessments aligned to those standards. This was a barely-disguised euphemism for adopting the newly-developed Common Core State Standards and participating in one of two consortia awarded RTTT funding to develop assessments aligned to the Common Core. Unsurprisingly, as states began to put together their RTTT applications—46 states and D.C. submitted applications in at least one of the phases—the vast majority of them pledged to adopt the Common Core standards and one of the two assessments intended to complement the standards. Moreover, to demonstrate their commitment to “high standards,” several states pledged to increase the cut scores students would need to achieve to be classified as proficient. In making this commitment, however, these states were
also making it more difficult for their schools to make AYP under NCLB, which continued to be the law of the land.

As the legislative stalemate over ESEA reauthorization continued into 2011, NCLB began to lose legitimacy in the eyes of both policymakers and the public. Among the public, results from the annual PDK/Gallup poll on education revealed that the percentage of the public viewing NCLB unfavorably grew from 27 percent in 2005 to 46 percent in 2010. The growth in public dissatisfaction over this time is partially attributable to the fact that, each year, more and more schools were failing to meet AYP and thus subject to NCLB’s sanctions. Testifying before Congress in the spring of 2011, Secretary Duncan stated that, according to estimates produced by the department he led, 82 percent of schools across the country could fail to make AYP in 2011. Although these comments generated considerable press and attention, they did not spur Congress to reauthorize ESEA. In September 2011 President Obama announced that, due to Congressional inaction, his administration would offer states flexibility with respect to several provisions of NCLB, including the provisions governing AYP status and the corresponding sanctions. As with RTTT, though, it quickly became clear that this flexibility would come with strings attached. In particular, the administration made clear that states would be granted waivers from NCLB only if they promised to work to implement a series of administration priorities.

The priorities most relevant here are the adoption of “college- and career-ready standards,” high-quality assessments aligned to those standards, and accountability systems that provided differentiated levels of recognition and support on the basis of assessment results and other system inputs. In effect, the administration used NCLB waivers as a method for driving their preferred approach on standards, testing, and accountability. The requirements surrounding standards and accountability echoed those emphasized under RTTT, but the accountability
provisions were newly developed for the NCLB waiver process. These provisions required states to design accountability systems with the following features: first, states had to set “ambitious but achievable” achievement objectives in reading and math and, mirroring NCLB, design accountability systems that incorporated reading and math test scores for all students, as well as for student subgroups. Unlike NCLB, though, these systems had to incorporate measures of student test score growth—not just performance levels—into accountability calculations. Second, states were required to group schools into at least four categories on the basis of accountability calculations. This requirement was designed to encourage states to adopt A-F school report cards, or at least a variant thereof. Finally, states had to identify “priority” and “focus” schools, which are defined (respectively) as the lowest-performing five percent of schools in the state, and the ten percent of schools in the state with the largest achievement gaps.

Considered together, the waivers offered by the Obama administration certainly provided flexibility from NCLB, but did not provide states with much flexibility around testing and accountability generally. The waivers effectively replaced the testing and accountability requirements in NCLB with requirements better aligned to the Obama administration’s preferences on the issues. Brushing aside potential concerns about the legality of the waiver process, most states reasoned that continued operation under the outdated requirements of NCLB was so onerous that they would seek a waiver from the law. Indeed, over the next few years, 45 states submitted waiver applications, and USED ultimately approved 43 of these requests. For the 43 states granted NCLB waivers, the promises they made in their application served as the status quo testing and accountability for the next several years—the handful of states not granted waivers continued to operate under NCLB. Conventional wisdom held that this status quo would persist until after the 2016 general election, when a newly elected president and
a Congress facing no imminent electoral pressures might work together on a bipartisan basis to reauthorize ESEA. This conventional wisdom proved wrong, however. It seems that pundits underestimated the breadth of dissatisfaction across the political spectrum with NCLB and the ensuing waiver process—the left was displeased with how the process played out while the right was unhappy that the process had happened at all.

This broad dissatisfaction provided the common ground necessary for strong bipartisan consensus on the latest ESEA reauthorization. Although reaching this consensus took several fits and starts, on December 10, 2015, President Obama signed the Every Student Succeeds Act (ESSA) into law. In doing so, he brought to an end a 15-year era where testing and accountability policy were largely determined in Washington, D.C. Although ESSA maintained annual testing and reporting requirements, it provided states with significant autonomy to determine the design of their accountability systems. On this score, ESSA delivered the flexibility that waivers had been portrayed as providing. In a very real sense, then, ESSA is closer in spirit to IASA than to NCLB. The end of the NCLB era provides a natural point for reflecting on the lessons the past 15 years have taught us. How effective are testing and accountability at driving academic improvement? Do any gains that students and schools realize come at the expense of other knowledge and skills we want schools to impart? How do individuals and institutions respond to accountability pressures? How do the politics of testing and accountability play out? We learned a great deal about these questions throughout the NCLB era, and these lessons which should inform future policy surrounding testing and accountability.
Lesson One: Test-Based Accountability Increases Test Scores in Reading and Math

Research into the effects of accountability systems followed close on the heels of their implementation. Initially, this research focused primarily on the effects of test-based accountability on students’ reading and math test scores—increasing achievement in these subjects was generally seen as the main goal of NCLB, and of test-based accountability more generally. The vast majority of this research focuses on estimating the effect of a single provision of NCLB or a particular aspect of accountability policy design on students’ test scores. A couple rigorous studies, though, attempt to estimate the effects of NCLB writ large on students’ reading and math achievement. These studies use state-level reading and math scores on the National Assessment of Educational Progress (NAEP), which is often referred to as the Nation’s Report Card, from both pre-NCLB and post-NCLB years. Using these data, the studies compare the change in test scores from pre-NCLB years to post-NCLB years for two groups: 1) States that first implemented accountability systems in response to the NCLB mandates, and 2) States and private schools for whom NCLB did not require a meaningful change in accountability requirements. The logic here is that NCLB did not represent a significant change for states that already had accountability systems or for private schools that were not subject to NCLB, which meant that they could serve as a valid comparison group for states where accountability systems were only developed because of NCLB mandates. Together, the preponderance of evidence from these two studies suggest that NCLB increased math achievement at both the 4th and 8th grade levels, but had little effect on reading scores, at least as measured by NAEP.

Along with these studies on NCLB writ large, a large body of work examines the achievement effects of accountability in a manner more modest in scope. This work typically evaluates a single aspect of accountability system design, or a particular provision of NCLB. For
example, several studies estimate the effect of a school receiving an “F” rating, relative to a “D,” on an A-F school rating system. Results from New York City and Florida provide consistent evidence that an “F” rating has a positive effect on students’ test scores in subsequent years, compared to students in “D” rated schools that scored just above the threshold for receiving an “F.” This effect is likely due to the fact that significant sanctions accompany consecutive “F” ratings, and schools typically do all they can to avoid these sanctions.

Finally, a series of studies analyze the achievement effects of accountability using data from states that implemented these policies on their own accord in the years prior to NCLB, comparing their outcomes to those of states that had not implemented accountability systems. These cross-state analyses typically use NAEP scores as their achievement measure and generally find that the adoption of high-stakes testing led to increases in both reading and math scores. Other studies take a similar approach, but limit themselves to studying accountability in a single state or district. Studies of this sort from Chicago and Texas conclude that accountability increased reading and math scores, although work from Charlotte finds no evidence that accountability improved student test performance.

Considered together, the bulk of the evidence emerging from research on test-based accountability points toward these policies increasing test scores in the subjects for which schools are held accountable, typically reading and math. On their face, these achievement increases are a good thing. They are consistent with accountability policies having their intended effect and improving student achievement outcomes. However, the last 15 years have taught us that the interpretation is not that simple, which brings us to Lesson Two.
Lesson Two: It is Not Clear These Achievement Increases Always Corresponded to Actual Learning Gains

There are several reasons that student reading and math scores might increase in response to accountability pressures. First, increased test scores could reflect student learning gains, either generally or in the specific subject(s) being tested. Regardless, these sorts of achievement increases are generally desirable. Second, test score improvements could reflect greater familiarity with the testing process, and this acclimation process produces higher scores, even if the underlying degree of student learning remains unchanged. Third, test score increases could be generated by schools and teachers placing greater emphasis on preparing students to take the test or imparting strategies for maximizing performance on the test, or both. Finally, increased test scores could be a product of manipulation, which could take the form of outright cheating, but also less nefarious forms such as the timing of test administration. The second, third, and fourth reasons for test score gains are unequivocally less desirable than the first one. Unfortunately, most analyses that find test-based accountability to produce achievement gains cannot distinguish between these and other potential reasons for the observed test score gains.

A few studies do shed light on this important issue, though. For two primary reasons, we can be reasonably confident that the set of studies showing accountability to increase NAEP scores reflect true student learning gains. First, the low-stakes nature of NAEP—there are no sanctions or rewards tied to NAEP performance—means that schools typically do not prepare their students for the assessment. The lack of preparation incentives is further reinforced by NAEP’s biannual administration to students in a relatively small number of schools within a state. And the fact that the sample of selected schools changes with each administration of NAEP means that it is relatively rare for any given school to administer the assessment multiple times.
within a short period of time. Together, these considerations render it unlikely that any observed NAEP gains are a product of increased familiarity, explicit test preparation activities, or outright manipulation. Second, NAEP has long been regarded as a high-quality assessment effective at gauging a broad range of student knowledge in a given subject. NAEP’s quality further contributes to an interpretation of increased NAEP scores representing true student learning gains, particularly when considered alongside the low stakes of the assessment and its administration protocol.

The interpretation of studies showing accountability to increase student scores on high-stakes tests is much murkier. In these cases, score gains attributable to intense test preparation activities, increased familiarity with the assessment process, or even manipulation are very real possibilities. A couple studies—conducted in Chicago and Florida—provide direct evidence on this issue. In particular, they analyze how test-based accountability affects student scores on both low- and high-stakes assessments. In Chicago, the study concluded that introduction of the accountability system had a large effect on student performance on the high-stakes exam, but no effect on the test with no stakes attached. The study from Florida determined that accountability had an effect on student performance on both the low- and high-stakes exams, but the effects on the high-stakes exams were 1.5-2 times larger than the effects on the low-stakes exams. Taken as a whole, the evidence suggests that accountability led to true increases in students’ math and reading ability, but that the size of these gains are somewhat smaller than what analysis of high-stakes exams would indicate.
Lesson Three: Reading and Math Gains Often Came at the Expense of Instruction in Other Subjects

NCLB was explicit that states must hold schools accountable on the basis of reading and math scores, but was largely silent on other subjects—it required grade-span testing in science, but attached no rewards or sanctions to test performance in that subject. Thus, schools faced immense pressure to maximize test scores in reading and math, but had few formal incentives—at least from an accountability perspective—to focus on instruction in other subjects.

Almost from the beginning of NCLB implementation, a collection of parents, teachers, and other stakeholders began to express concerns that the emphasis on increasing reading and math scores was crowding out instruction in subjects like art, music, social studies, and even science. Studies of this phenomenon provided evidence that these concerns were well-founded. One particularly well-done study drew a representative sample of districts across three states—California, Pennsylvania, and Georgia—and surveyed superintendents, principals, and teachers in those districts from 2003-2005.10 Across each of the three states, about half of principals reported requiring their teachers to increase the time they spent on tested subjects (i.e., reading and math), with corresponding reductions in non-tested areas. And 40 percent of district superintendents reported that they eliminated one or more programs, such as art or music, in order to increase instruction in tested subjects. Other studies conducted in different contexts echo these conclusions, finding that schools would focus time, effort, and resources on maximizing math and reading instruction at the expense of instruction in other subject areas.11

Even within math and reading, many schools focused instruction primarily on the concepts and content that would appear on the tests. And, to put it generously, many of the tests that states adopted were not of the highest quality, particularly in the early years of NCLB—
these assessments were typically multiple choice and assessed relatively superficial knowledge and skills. So even the extensive math and reading instruction that schools provided was often one particular type of math and reading instruction, a type that was often criticized for failing to promote creative thinking, a deep understanding of the content, and a true love of learning.

There is no denying that developing math and reading skills is an important endeavor—these skills underlie just about every aspect of life to one degree or another. There are valid questions, though, over the ideal balance between developing math and reading skills and meaningfully exposing students to a broad range of subjects. There is certainly an argument to be made that schools focused on math and reading to an insufficient degree in the pre-NCLB era. There is also an argument to be made, however, that NCLB swung the pendulum too far in the other direction, and placed far too much focus on these subjects. Indeed, the near single-minded pursuit of math and reading achievement over the past 15 years has been disconcerting to those who value a balanced and well-rounded education.

Lesson Four: States, Districts, and Schools Respond to Accountability Pressures in Unintended Ways

By design, NCLB provided states with substantial autonomy in developing standards, adopting assessments, and defining the student performance level on those assessments that corresponded to proficiency or mastery of the standards. The hope was that states would adopt rigorous standards that, when mastered upon high school graduation, would set students up for success in postsecondary education or the workforce. Moreover, policymakers envisioned states adopting high-quality assessments and carefully calibrating the level of student performance on these tests
that denoted mastery of the standards. It quickly became clear, however, that the reality of states’ decisions on these issues would not match the original visions.

Out of the gate, some states did adopt reading and math standards that were judged to be rigorous and “world class.” At least as many states, though, did not. It became something of a parlor game among education policy organizations to assess the quality of the standards that states adopted in response to NCLB. For example, the Thomas B. Fordham Foundation issued a series of reports titled *The State of State Standards* where they graded states’ reading and math standards on an A-F scale. The 2005 version of this report assigned only five states an A grade for their reading standards, but gave 8 states a D or an F. In math, just three states earned an A from Fordham’s evaluators, and a majority of states received a D or an F. It wasn’t until the Obama administration incentivized adoption of the Common Core State Standards through RTTT and NCLB waivers that most states adopted the sort of standards that policymakers initially envisioned.

In addition to going their own way on the adoption of standards, a number of states also defined proficiency in a manner contrary to policymakers’ visions. Instead of ratcheting up the level of test performance required for a student to be deemed proficient—to have demonstrated true mastery of the standards—some states actually lowered the proficiency bar in an effort to put their schools in a better position to meet the universal proficiency requirement of NCLB. For example, between 2005 and 2009 South Carolina substantially dialed down the proficiency bar across all grades in both reading and math.\(^{12}\) There is evidence that New York did the same thing, although to a more modest extent. These decisions flew directly in the face of policymakers’ intentions on these issues. In response, the Obama administration again used the RTTT and NCLB waiver application process to incentivize states to ratchet up proficiency
thresholds. They wanted states to define proficiency as originally intended by policymakers, which would convey the “hard truths” that most kids in the United States were not on track to succeed in postsecondary education or the workforce. Wisconsin made such a promise in its unsuccessful RTTT application and, almost overnight, proficiency rates dropped by 30-45 percentage points in the state, depending upon the particular grade and subject. Oklahoma made a similar promise in its NCLB waiver application and from 2016 to 2017 saw declines in proficiency rates on the order of 30-35 percentage points in most subjects and grades.

States were not the only entity that responded to accountability mandates in unintended ways. Districts and schools also exhibited their share of undesirable responses to the pressures created by implementation of high-stakes accountability, including the intense focus on math and reading outlined previously. Initially, the school accountability systems required by NCLB were based almost entirely on the percent of students deemed proficient—those who score above the specified threshold on the exam. Even today, proficiency levels are the primary component of most states’ accountability systems. In addition to having the undesirable property of rewarding or sanctioning schools primarily on the basis of the characteristics of the students they serve, these systems also provide an incentive for schools to focus on students who they believe have a reasonable shot of scoring above the proficiency bar. From an accountability perspective, there is little reason to attend to students who will surely score well on the test, or those who will not approach the proficiency threshold under any circumstances. And indeed, multiple studies provide evidence that accountability leads to a focus on “bubble kids,” those who are thought to be just above or just below the proficiency cutoff.13

Perhaps the most problematic response of districts and schools to accountability pressures has been outright cheating. Over the past 15 years, audits and analyses have uncovered cheating
scandals in several districts, including Columbus, Washington DC, and Atlanta, among others. In Atlanta, schools were facing significant pressure from superintendent Beverly Hall—who had a reputation as a no-nonsense, hard-driving leader—to improve their test performance. A number of schools in the district did improve their test scores, oftentimes substantially. These massive achievement increases gained these schools accolades, but also aroused suspicions—how were these schools able to produce achievement increases of a magnitude almost never seen? This issue caught the attention of the Atlanta Journal-Constitution and they commissioned an analysis of the test score improvements, concluding that the observed score increases were highly improbable. This spurred an investigation by Atlanta Public Schools and—when their investigation proved lackluster—by the Georgia Bureau of Investigation, who uncovered evidence of widespread cheating across the district. This led to the indictment of 35 educators, nearly all of whom either pled guilty or were convicted for their role in the conspiracy. Situations like the one in Atlanta illustrate the potential dark side of accountability policy.

The litany of unintended responses by states, districts, and schools to the pressures created by high-stakes accountability suggest that, at a minimum, a great deal of thought and care needs to go into the design of these systems. It may be the case, though, that even the utmost attention and care will be insufficient to prevent states, districts, and schools from taking actions designed to game the system. The concluding section of this chapter considers this possibility in more detail.

Lesson Five: Unrealistic Expectations and Goals are Counterproductive

At the time NCLB passed in 2001 its goal of universal student proficiency in math and reading was more than a decade off. Policymakers, who are notoriously short-sighted, relished the
immediate soundbite the law provided them—the legislation was going to ensure that each and
every student was ready for college or the workforce—and gave relatively little thought to future
implications. Surely, they thought, NCLB would be reauthorized before this blatantly unrealistic
goal caused any real problems.

Problems started to appear on this score sooner than anticipated, however. Schools could
be labeled as failing to make AYP if just one subgroup fell off the state-specified trajectory
toward universal proficiency by 2014. Consequently, just a year or two into NCLB’s testing and
accountability regime schools started to face the sanctions laid out in Table 1. Predictably, these
failure labels did not always go over well with educators or parents. With respect to educators,
they often felt resentful for their school being labeled as a “failure” for not achieving something
that, by any reasonable standard, was simply not possible to achieve, no matter how much blood,
sweat, and tears they poured into the cause. What kind of law, from their point of view, penalizes
schools and teachers for not doing the undoable? These feelings quickly turned many
educators—a hugely important constituency for the long-term viability of any education
policy—against both NCLB and testing and accountability policy more generally.

Many parents also reacted negatively to the “failure” label being slapped on their school.
Indeed, the past 15 years have taught us that people are fine with a law that tells them that other
peoples’ schools are failing, but are less receptive when the law tells them that their school is a
failure. This is particularly true if they paid hundreds of thousands—or even millions—of dollars
for a house that provided entry to that school. Or, if they walk into that school multiple times per
week and see all the positive things that are happening and sense the teachers’ and principal’s
commitment to the school. In cases like that, whose judgment are people going to believe? The
judgment of a formula designed by Congressional staffers and bureaucrats in Washington D.C.? Or their own eyes and experiences? The answer is pretty clear.

The unrealistic goal of universal proficiency might have instilled folks with some warm feelings and provided policymakers with a nice soundbite, but it also created a lot of problems. And, considered broadly, it is hard to see a case for the benefits of the universal proficiency goal exceeding the costs and problems it created.

*Lesson Six: Accountability Policy Has a Politics Problem*

The blowback from parents and educators described above highlights a broader political lesson that we have learned about testing and accountability policy over the past 15 years: these policies—particularly accountability—do not have any sort of natural or sizable political constituency. Political science teaches us that in order for a policy to have any meaningful staying power, it needs to develop a vocal grassroots constituency who will advocate for it in times of turmoil. As the testing and accountability era has progressed, though, it has become increasingly clear that testing and, particularly, accountability have not developed such a constituency. To the contrary, results from the annual PDK/Gallup poll on education demonstrated that the percentage of the public viewing NCLB unfavorably almost doubled from 2005 to 2010—public disapproval jumped from 27 to 46 percent in just five years. These are clearly not the approval trends of a policy with a large and growing constituency.

When you dig down into the issue, it is not difficult to see why such a constituency has failed to form. Who sees direct and tangible benefits from accountability? Certainly not teachers or principals. Similarly, it is hard to see any direct benefits of accountability accruing to parents or kids. At least testing provides families with information on the academic performance of their
children. More broadly, the annual testing and subgroup reporting requirements shined a light on potential differential performance across demographic subgroups, a tangible benefit that commanded the support of civil rights groups for annual testing. Accountability, on the other hand, does not provide these sorts of perceptible benefits to any particular group.

Accountability supporters often claim that these policies offer clear benefits to taxpayers—they ensure that schools and districts are spending their money well. It may well be the case that accountability provides these benefits to taxpayers. In the scheme of things, though, any such benefits are relatively intangible and so imperceptible that it is difficult to imagine that they motivate or mobilize anyone on their behalf. Put differently, it may be the case that accountability policy contributes to the public good in some way. The public good, however, does not contribute to campaigns, lobby legislators, or vote in elections.

The failure of accountability to generate any sort of meaningful grassroots constituency is primarily attributable to its structural flaws. However, the economic and political contexts that evolved during the mid- to late-2000s generated additional headwinds for the policy. Economically, the Great Recession hit just as NCLB sanctions began to affect schools and districts. This economic downturn decimated education budgets at both the state and district levels, which prevented policymakers from providing schools and districts with any additional money to ease the bite of NCLB sanctions. Accountability is probably best paired with additional resources that schools and districts can use to provide a cushion from policy sanction and to mitigate political opposition. The Great Recession made such a pairing all but impossible.

Politically, education has long been considered an issue with wide bipartisan support at the elite level, as evidenced by NCLB’s passage with broad bipartisan majorities. Recently, though, there are indications that education, particularly at the federal level, is becoming more
partisan in nature. The reasons underlying this shift are multifaceted, complex, and beyond the scope of this chapter. However, the end result is a politics where one party’s support of a given policy position almost certainly corresponds to opposition by the other. Thus, the Obama administration’s wholesale embrace of accountability policy, particularly federal accountability, corresponded to diminished support for the policy among Republican elites. And the fact that this occurred before accountability policy had fully matured and entrenched itself in the federal education landscape left it susceptible to the vagaries of partisan politics. Together, the lack of broad support for accountability at either the grassroots or elite level leaves the policy in a politically precarious situation.

**Lesson Seven: We Would Not Have Learned Many of These Lessons Without the Testing and Reporting Requirements of NCLB**

The defining legacy of the testing and accountability policy of the last 15 years seems likely to be twofold: 1) A shift from measuring inputs to focusing primarily on outcomes, and 2) increased transparency surrounding those outcomes, particularly student achievement.

Throughout much of the 1980s and even into the 1990s most folks in education focused primarily on the resources that were going into schools, and less on the outcomes those resources were generating. If *A Nation At Risk* is generally recognized as the starting point in moving from an input- to outcome-oriented view of education, then it is reasonable to consider NCLB as the capstone of that movement.

Prior versions of the law had certainly addressed the issues of testing and accountability, but had done so in a manner where enforcement was lax and states were given wide discretion in determining how far they wanted to go down those roads. NCLB changed that. As described
earlier, it mandated annual reading and math testing in grades 3-8 and once in high school. These testing mandates were accompanied by significant reporting requirements that were strictly enforced—states, districts, and schools had to report student performance at the subgroup level, they had to report about attendance rates, teacher qualifications, and numerous other topics. Recognizing that some of these reporting requirements were unduly burdensome for schools and districts, it is also true that this information has been instrumental in allowing us to learn the lessons discussed above. It has allowed us to learn about the achievement effects of accountability, institutional and individual responses to accountability pressures, and even about the politics of these issues.

Although there is general agreement that many of the accountability requirements of NCLB were quite ham-handed, there is also a broad consensus that the transparency provisions of the law were one thing policymakers got right. It has illuminated schools that achieved excellence and equity, as well as those that fell short on either, or both, of those dimensions. It has allowed for discussions of what different numbers on different topics mean, whether they are important, and what we should do about them. These sorts of discussions have proved valuable—if occasionally reductive and counterproductive—and contribute to a broad desire to maintain requirements to collect and publish this information. To the degree that constituencies continue to see a benefit of this information—and it seems likely that they will—transparency surrounding educational processes and outcomes will probably be the defining legacy of the testing and accountability era.
Can We Get Testing and Accountability Policy “Right”?

In the wake of all the lessons we have learned about testing and accountability over the past decade-plus, it is reasonable to question whether we can ever design and implement these policies in a manner that commands broad support over the long haul. The experiences over the past 15 years suggest that such a reality seems unlikely, particularly with respect to accountability. For perhaps the most important lesson we have learned over the NCLB era is that we, as a society, do not agree on the purpose or intent of K-12 schooling in the U.S. Or, at least, we do not agree on the ideal balance of the various things we want schools to do. Some folks think schools should primarily focus on developing skills—particularly reading and math—that will set students up for success in postsecondary education or the workforce. Others want schools to expose students to a broad curriculum, focusing less on standardized assessment of student knowledge and skills and more on developing curiosity and instilling a love of learning. These tensions have always existed, but NCLB brought them to the surface by effectively privileging one conceptualization of the purpose of schools over others. Thus, debates that purportedly began over accountability generally mutated, at least to some degree, to a debate over the appropriate purpose and design of schooling.

If we cannot reach consensus on the purpose and design of schooling, then it is silly to think that we could design an accountability policy that commands broad support. As has been the case over the past 15 years, there would likely be a constant stream of arguments and complaints over this provision being too strong or that requirement being irrelevant. More generally, as long as there are substantial differences across society in what we want schools to do—and it is tough to imagine those differences disappearing anytime soon—it is difficult to
imagine an accountability system that adequately balances all stakeholders’ priorities and addresses their concerns.

Even on the off chance that we can reach some sort of consensus on the purpose of schooling, accountability policy would still have a tough row to hoe. The lessons discussed above demonstrate the difficulty of designing an accountability policy that induces all relevant entities—states, schools, and districts—to act in intended ways. It is a near-guarantee that the policy will leave open some loopholes or will create perverse incentives, and these institutions and organizations will exploit these loopholes or respond to the perverse incentive structure.

Furthermore, we have examples of similarly intended policies in other policy domains—domains with narrower purpose than education and broad agreement over that purpose—proving difficult to maintain over time. Take, for example the Program Assessment Rating Tool (PART), which was implemented in 2002 by the George W. Bush administration and designed to rate the effectiveness of each and every federal program. The developers of PART intended for the performance measurement tool to be adapted to each federal program and then be applied within that context. Compared to schooling, the goals of a particular federal program are generally much clearer, and the consensus over these goals is much stronger. Even within this relatively friendly context, PART was in place for only six or seven years before the Obama administration abandoned the initiative. Looked at in this light, perhaps school accountability policy had a good run, particularly given its lack of a natural grassroots constituency.

So where does that leave us? What does the future of testing and accountability policy in the United States look like? Of course, it is difficult to make detailed predictions on this score, but the lessons we have learned over the past 15 or so years do support two more general forecasts. First, broad support for shifting the focus from inputs to outcomes, coupled with the
recognition that information on those outcomes can be quite valuable, renders it likely that the testing and reporting requirements mandated in NCLB—and reiterated in ESSA—will remain in place for the foreseeable future. Indeed, there is widespread recognition that the increased information on student and school performance was not the problem; it was what we did with that information that proved most problematic. This leads to the second prediction, that test-based accountability seems likely to continue the retreat that started with ESSA. In the near term, this retreat could manifest itself by states deciding to throw in the towel on accountability under ESSA—they will follow the letter of the law, but will design systems with fairly meaningless rewards or sanctions. Further down the line, perhaps in the next ESEA reauthorization, this retreat may become codified in statute. In the long term, it seems likely we will get to a point where Congress mandates that states produce information on student achievement—and student outcomes more generally—but leaves it entirely to states, districts, and parents to decide how to use that information.
Although the AYP classification was based primarily on reading and math test results, schools were also required to document that 95% of students across the tested grades—and within each subgroup—took the assessments. Schools also had to meet attendance targets and, at the high school level, graduation rate goals in order to make AYP. The legislation also contained a “safe harbor” exception to the AYP designation. Even if one or more subgroups failed to meet the state-specified trajectory toward universal proficiency, a school could still make AYP if the number of proficient students in each relevant subgroup increased by at least 10 percent each year.


The other two waiver priorities emphasized by USED included “supporting effective instruction and leadership” and “reducing duplication and unnecessary burden.”

The waiver applications from Iowa and Wyoming did not receive approval from USED. California, Montana, Nebraska, North Dakota, and Vermont did not submit waiver applications.


