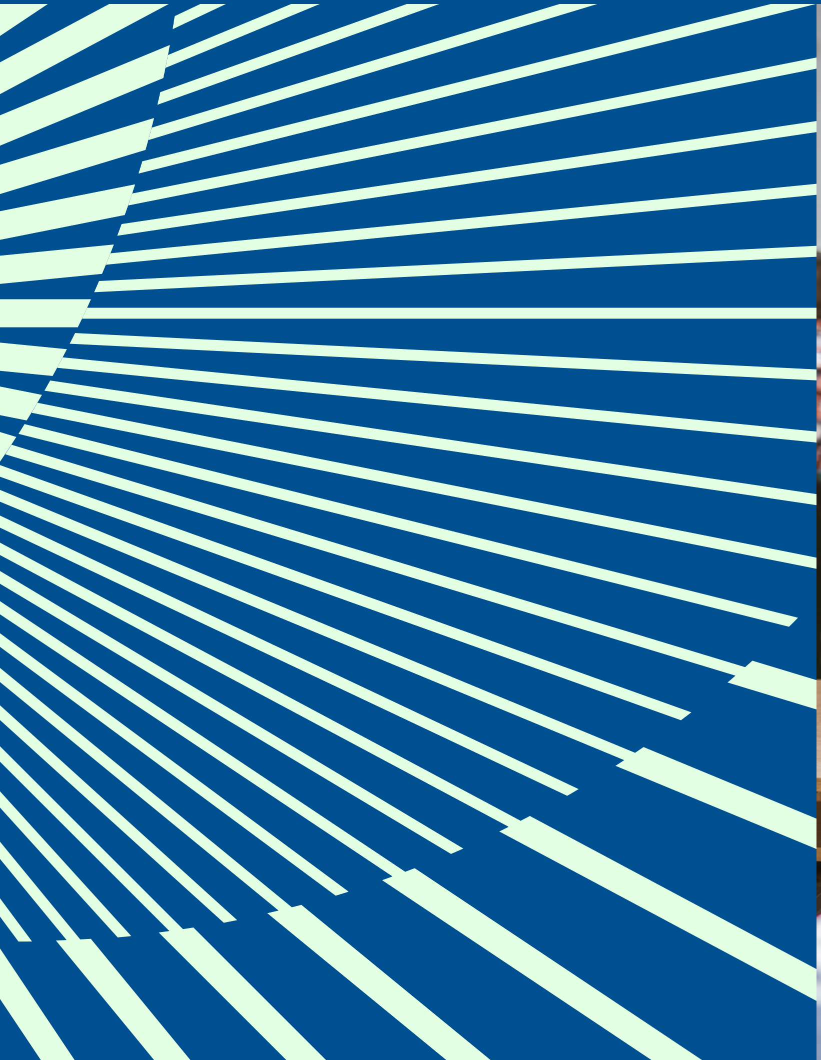
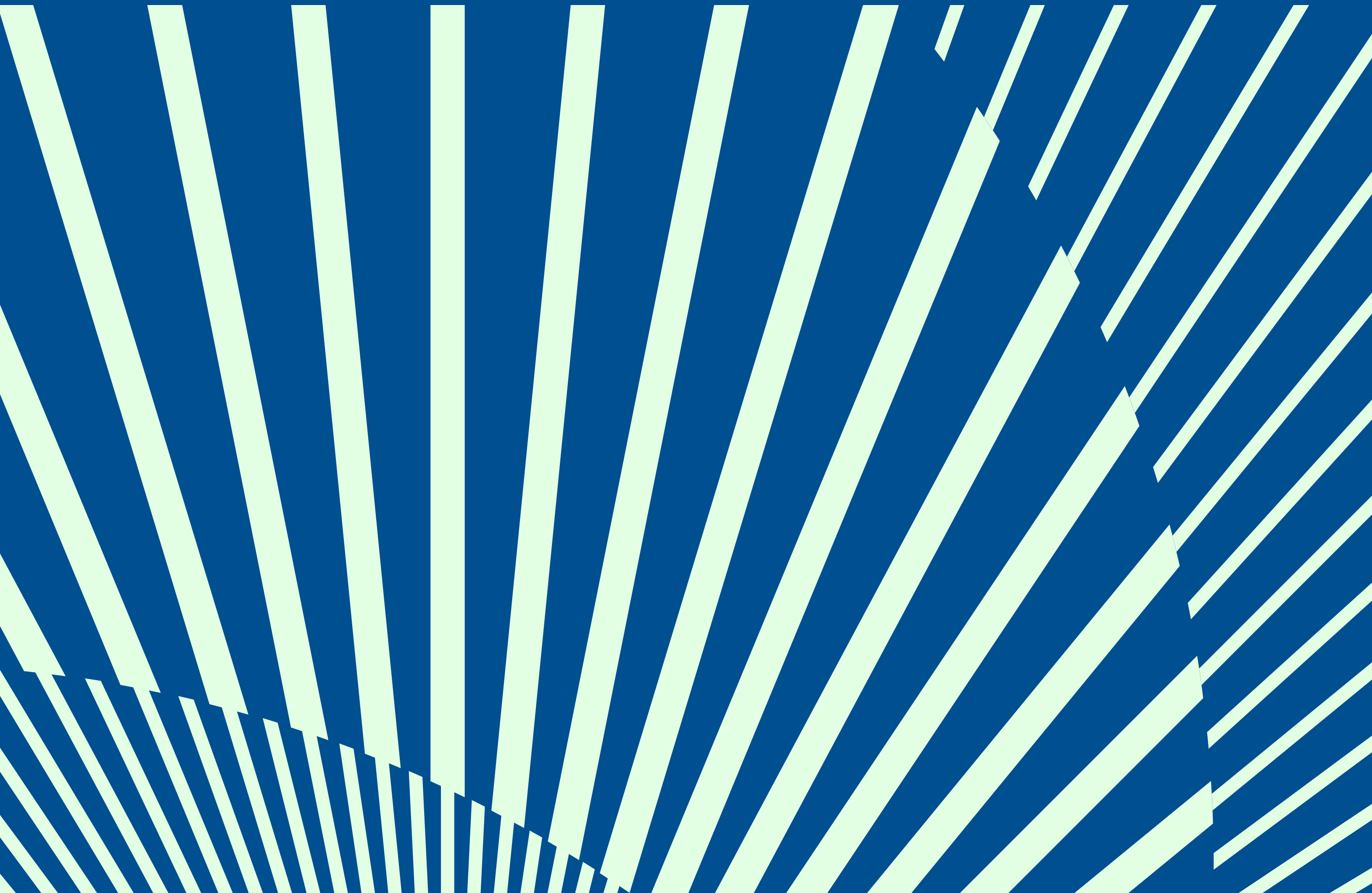
U.S. Chamber of Commerce
Foundation

Looking Back to Look Forward

Quantitative and Qualitative Reviews of the Past 20 Years of K-12 Education Assessment and Accountability Policy



Contents



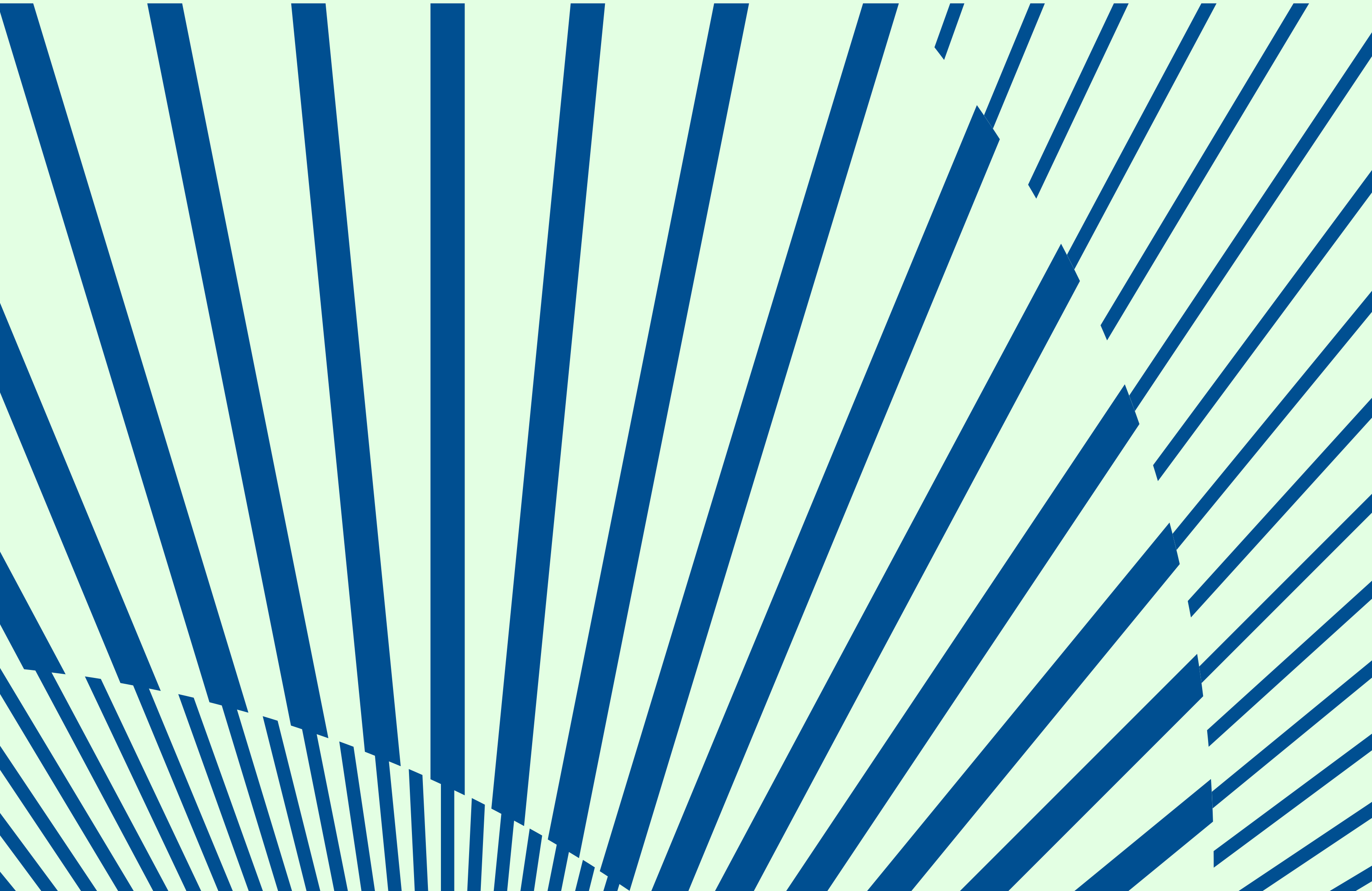
Foreword

What Does Empirical Research Say About Federal Policy From NCLB to ESSA?[1]

Stakeholder Perceptions of the Legacy of 20 Years of Education Data and Accountability Efforts

References

Foreword



To effectively engage in the next era of education policy, the U.S. Chamber of Commerce Foundation and the business community broadly need to know what recent history has taught us. Two years ago, we realized we could not answer this critical question: after the last twenty years of education reform, what has worked and what has not? As tireless advocates for high-quality academic standards, assessments, and accountability as tools for academic achievement, this was a "drop everything" moment.

We decided to embark on an ambitious venture—to create the most comprehensive analysis to date of existing research and qualitative feedback on federal K-12 education policies of the past 20 years from No Child Left Behind (NCLB) to Every Student Succeeds Act (ESSA). The full report following this foreword—a quantitative research review authored by Dan Goldhaber and Michael DeArmond of CALDER (at the American Institutes for Research) and a qualitative analysis authored by Chris Stewart and his team at brightbeam—is the result of our collective effort.

Once we dropped everything to take a hard look at the last 20 years of education in the United States, we discovered that researchers have learned more about what works in education in the last 20 years than we did in the previous 50 years. Few other countries are in the position to use data to compare the efficacy of educational programs, policies, and interventions since most do not test annually in consecutive grades. In our world of business, we long ago learned to pilot ideas, evaluate impact, and improve. Continuing to improve how we collect and analyze data in education should allow schools, districts, and states to do more of this and to better effect. We now understand how important it is for future iterations of federal education policy to incentivize and ensure rigorous data analysis so that we, as a country, can understand which students our schools are serving—and those they are not—and where we need to intervene.

This research is a foundational component of our Future of Data in K-12 Education initiative, a multi-year undertaking to learn about what worked, what didn't, and what are the most promising ways to improve assessment and accountability in America's K-12 public schools. We are fortunate to be joined by a diverse group of exceptional education leaders working alongside us. We are using this effort—and the thoughtful reflections of the many experts and practitioners whom we are grateful to engage—as an opportunity for clear-eyed reflection that will lead to policy recommendations for the future of public education.

Before NCLB, less than half of states had outcomes-based accountability measures for schools based on objective measures of student learning. Only 30 states participated in the National Assessment of Educational Progress. Without consistent, comparable, and disaggregated information on student outcomes, we and employers across the country knew that ensuring our public education system is preparing all students for civic life and the workforce would be impossible.

> "There was the hopeful vision that eventually had to be implemented in the real world. As that happened, complexities arose, which made it easier to start changing the narrative from the hopeful one to a problematic one. There was a real effort to problematize everything having to do with standards, accountability, testing, outcome data—basically NCLB in total. And if your only goal is to make this look like a problem, boy, do you have so much fuel, because you're talking about big complex systems and there is plenty of stuff to make an issue."
>
> —Chris Stewart

NCLB dramatically improved our ability to know the extent to which states, districts, and schools were successfully helping all students meet grade-level standards, which is critical to a quality education for all students. By providing a better picture of student achievement, NCLB also allowed us to understand which innovations in education successfully improve school and life outcomes.

As we look forward to building a new framework for future education policy, what does quantitative research tell us and what do we still not know? According to what Goldhaber and DeArmond identify as the most credible existing studies, we can say the following:

- Disaggregated data and requirements for transparency moved the system to consider the needs of individual student groups, including students of color, students from low-income backgrounds, English learners, and students with special needs, helping to ensure that students—and their families—who our education system had long under-served could not be ignored. States, districts, and schools were no longer able to hide the performance of some students behind an average.
- Student achievement increased due to NCLB-era assessment and accountability policies, especially in math and for student groups that the system had not been serving well.
- Reforms in teacher evaluation and school turnaround initiatives did not consistently improve student outcomes at scale, in part due to significant variation in design and the quality of implementation.

While these findings are important, there are also a host of critical questions we still can't answer because, to our knowledge, sufficient effort has not been dedicated to analyzing available data. As a result, progress toward so much of what we hoped would happen over the last 20 years is, unfortunately, unknown:

- Did schools serving historically under-served students get more money to improve than they otherwise would have? How much more money did the lowest-performing schools receive compared to other schools in their district?
- If identified schools did get more money, what did they do with it?
- How many identified low-performing schools became successful? How many did not? How many of those that did not improve are continuing to enroll students?
- Have states seen improvement in measures other than academics, such as chronic absenteeism or school climate, that ESSA was also intended to elevate?
- If statewide assessments have not improved as much as many of us had hoped, what incentives and policy changes should be implemented to spur more innovative, equitable assessments of student learning?

The reforms our authors tackle here obviously did not occur in a vacuum. NCLB, Race to the Top, the Common Core State Standards, School Improvement Grants, ESSA, and other national policy efforts all took place within a broader push for more transparency and accountability in public education —a push the U.S. Chamber was proud to help lead. Meanwhile, states were busy implementing their own education reforms alongside, but not always directly tied to, federal policy.

While there was certainly a great deal of unity and organization around these principles, there was opposition at every turn, which slowed down, and in some cases, halted progress. Research from our colleagues at brightbeam certainly alludes to, but cannot fully unpack, the repercussions of those that stood in the way of the push for better outcomes for historically underserved students.

We also see in the qualitative analysis that some in the education community—from policymakers to educators to parents—are advocating for changing how we assess learning and hold schools accountable for student achievement. While many still support the tenants of transparency, accountability, and equity, the tools have had unintended consequences upon implementation. The U.S. Chamber Foundation has and will remain a strong supporter of accountability rooted in objective measures of student learning as a means of holding our education system accountable for academic achievement for all students; however, we have heard loud and clear the feedback that many of the assessments we use today don't provide a full picture of what a student knows and can do, can at times cause unnecessary stress for students and teachers, often take too much away from instructional time, and that some contain outdated material that is not culturally inclusive. Further, the promise that new assessments would provide more information to parents is unfulfilled. Results come often too late and are hard to understand, making them less likely to be used for decisions at the school and family level.

"There felt like a lack of awareness around the different systemic challenges and the poverty that impacts our community. There was a constant sense of urgency. I felt like all of our classrooms were in a high pressure situation. We had to perform and produce ... only in retrospect did I realize that we overprivileged the students' ability to perform in different ways."

—Los Angeles educator

When ESSA returned more—but not all—education decision-making to the states the role of national organizations with state-based presence, like ours, became more important. With the education policy debate no longer centered in Washington, DC, there is a great deal of work to be done to analyze and vet the constant supply of school improvement ideas, spread the best ideas across the country, and prevent the resurfacing of strategies that have previously failed to deliver. We want to make it clear to the education community that the U.S. Chamber Foundation takes its role seriously in supporting the proliferation of sound, data-driven policies and practices to the benefit of all students, especially those that the system has always struggled to support effectively, and it starts with a continued commitment to accountability, transparency, and high-quality data in our public schools.

We would like to express our warm appreciation for the individuals that agreed to serve on our Future of Data Working Group. They have all contributed many hours of thoughtful collaboration and insights to this effort. Our intention was never to create a consensus document; therefore, while each member of the working group does not necessarily agree with all the information presented here, all their perspectives have helped to shape how we interpret these findings and what we do with them. Thank you to Cindi Williams, Duncan Robb, and Kate Poteet at HCM Strategists, who are our close partners in this effort.

Cheryl Oldham
Senior Vice President

Caitlin Codella Low
Vice President, Policy and Programs

Kyle Butler
Manager, Programs

U.S. Chamber of Commerce Foundation

Future of Data Working Group

Cheryl A. Oldham, Vice President of Education Policy, U.S. Chamber of Commerce; Senior Vice President, U.S. Chamber of Commerce Foundation

Kaleem Caire, Founder & CEO, One City Schools

Ivan Duran, Ed.D., Superintendent, Highline Public Schools

Denise Forte, President & CEO, The Education Trust

Christy Hovanetz, Ph.D., Senior Policy Fellow, Accountability, ExcelinEd

Lynn Jennings, Ph.D., Senior Director of National and State Partnerships, Education Trust

Robin Lake, Director, Center on Reinventing Public Education

Patricia Levesque, Chief Executive Officer, Foundation for Excellence in Education

Maya Martin Cadogan, Executive Director, Parents Amplifying Voices in Education

Nicholas Munyan-Penney, Assistant Director of Policy, Education Trust

Christine M. T. Pitts, Ph.D., Director of Impact & Communications, Center on Reinventing Public Education

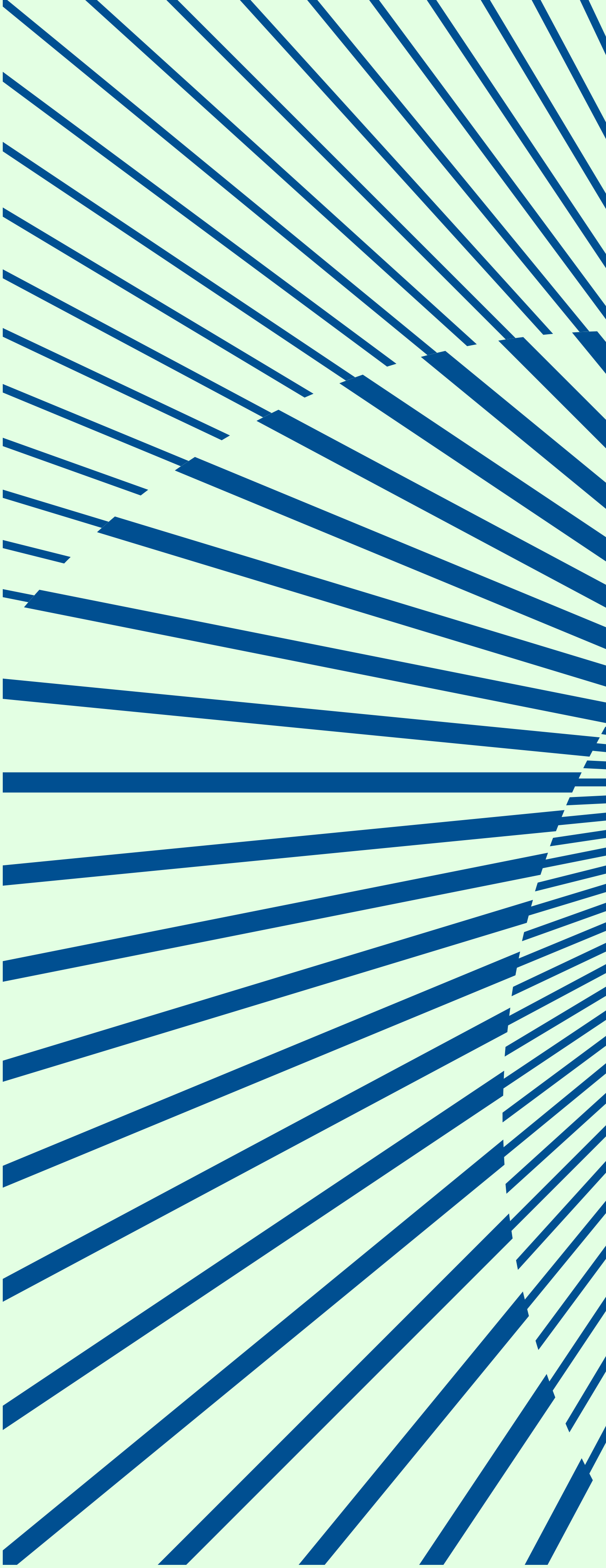
Jennifer Randall, Ph.D., President, Center for Measurement Justice

Rodney Robinson, 2019 National Teacher of the Year and Senior Policy Advisor, Richmond Public Schools

Ian Rosenblum, former Deputy Assistant Secretary for Policy and Programs Delegated the Authority to Perform the Functions and Duties of the Assistant Secretary, Office of Elementary and Secondary Education, U.S. Department of Education

Margaret Spellings, President and CEO, Texas 2036

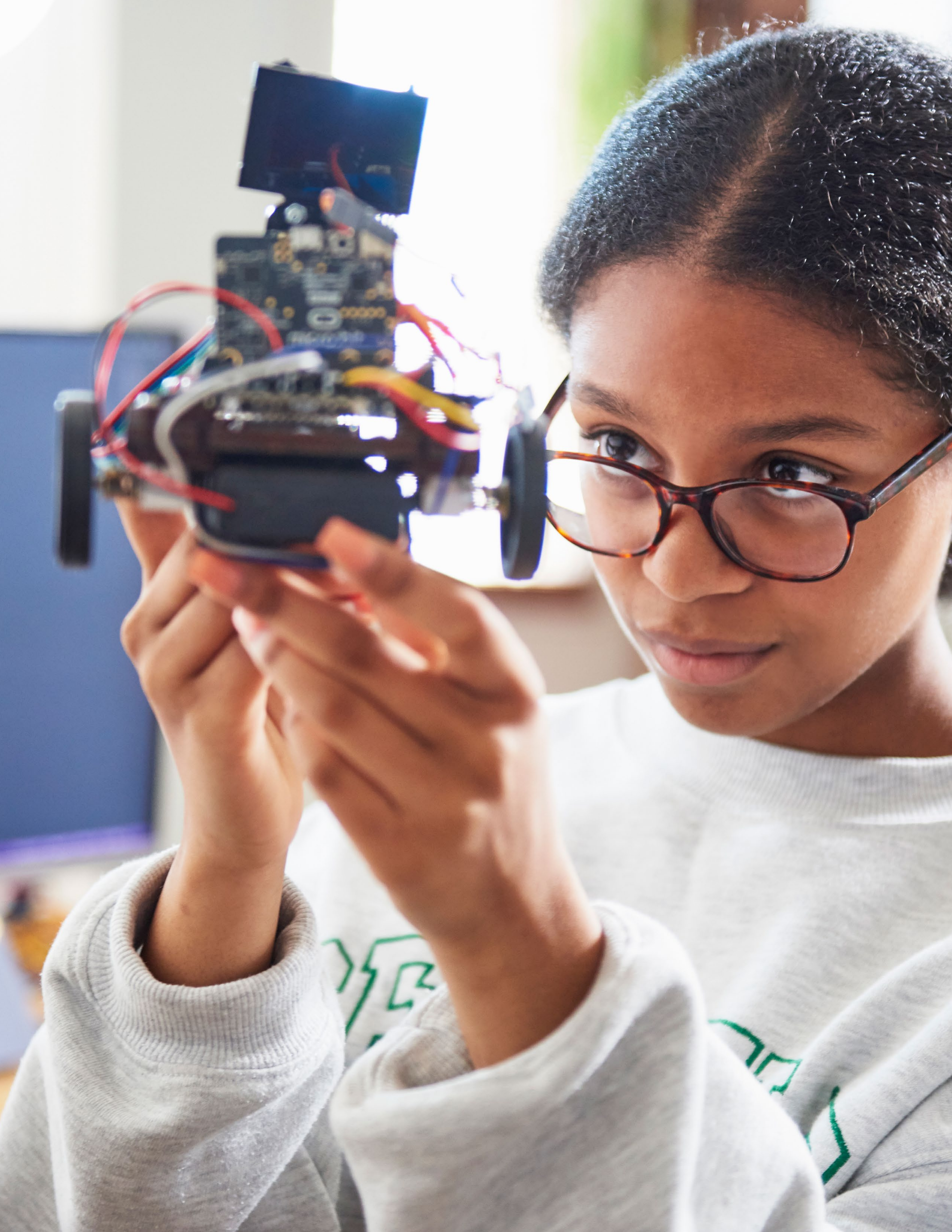
John White, Chief Success Officer, Great Minds



What Does Empirical Research Say About Federal Policy From NCLB to ESSA?[1]

Dan Goldhaber
Michael DeArmond

CALDER Center
American Institutes for Research



1. Introduction

1.1 From Inputs to Outcomes

Between 2001 and 2015, the federal government's role in education policy expanded in new and important ways. For much of the previous 35 years, beginning with the Elementary and Secondary Education Act (ESEA) in 1965, the federal government had focused on promoting equal opportunity and access in the nation's public schools. It funded categorial programs for special populations (e.g., students with disabilities and students living in poverty) and guaranteed civil rights in the courts; decisions about what to teach in classrooms and how to teach were largely left to local schools and school districts (Fuhrman et al., 2007).[1]

But worries about academic achievement in the 1970s and 1980s generated growing concerns about the performance of public schools, famously captured by A Nation at Risk in 1983. These concerns were echoed in other high-profile reports that also called for reform, including reports from the Twentieth Century Foundation (Peterson, 1983), the Carnegie Foundation (Boyer, 1983), and the Education Commission of the States (Education Commission of the States. Task Force on Education for Economic Growth, 1983). The Education Commission of the States' 1983 report, Action for Excellence, captured the mood with a forward titled: "A Conviction that a Real Emergency is Upon Us." Meanwhile, news accounts highlighted stories of students graduating from high school without basic reading skills (Draper, 1987, April 22). Calls for reform were growing (McGuinn, 2006).

In response, policymakers and business leaders began demanding better academic outcomes from all schools for all students. By the early 1990s, these demands were taking shape under the banner of "standards-based reform" in a handful of states like Kentucky and Texas.

Soon, the ideas spread to federal policy. In 1994, the federal government encouraged states to use standards, tests, and accountability to drive school improvement in the Goals 2000: Educate America Act and the Improving America's Schools Act (IASA). By 2001, the landmark No Child Left Behind Act (NCLB) required states to test students in reading and math and hold schools accountable in both subjects, to calculate what it would take for students (including sub-groups of students) to achieve proficiency, and to follow federal timelines for sanctioning underperforming schools.[2]

Given the far-reaching nature of the changes associated with this shift from inputs to outcomes, it can be hard to remember how different things were in the years leading up to NCLB. In 2000, less than half of the states had outcomes-focused accountability systems in place;[3] around only a quarter consecutively tested students in the same subjects between 2nd or 3rd grade and at least 8th grade (Goertz & Duffy, 2001). Before NCLB, most states reported some type of outcome data, but many did not report results for subgroups or underserved students (Government Accountability Office, 2000); and at the time the law passed, participation in the National Assessment of Educational Progress (NAEP)—the Nation's Report Card—was voluntary.[4] In many places, systematically assessing student achievement levels or progress over time was impossible.

By the early 2000s, all that had changed: every state had adopted accountability policies in line with NCLB's emphasis on outcomes and achievement for all schools and students. With this, the expectation that public schools would produce elevated academic outcomes for all students took center stage and the federal government's push for outcomes-based accountability dominated the nation's education policy agenda for the next 15 years.

[1] We would like to thank several reviewers who helped sharpen this paper's findings and conclusions. We received multiple rounds of feedback from members of the Future of Data Working Group as well as guidance from Caitlin Codella Low, Kyle Butler, Cindi Williams, Cheryl Oldham, and Duncan Robb at the U.S. Chamber of Commerce Foundation. We are also thankful for the feedback we received from Matthew Chingos, Marty West, and, especially, Jim Wycoff (who indulged us with multiple conversations about the paper's conclusions). Equia Aniagyei-Cobbold helped copyedit the paper. This paper was funded by the U.S. Chamber of Commerce Foundation, but the findings and conclusions are ours alone and do not necessarily reflect the Chamber's positions or policies, or those of the CALDER Center and the American Institutes for Research. As always, despite all the help we had, fault for any errors or omissions in the paper rests with us alone.

[2] NCLB's approach was based, in part, on education reforms in Texas that had originated under Governor Ann Richards in the 1990s and continued under Governor George W. Bush prior to his successful run for the presidency.

[3] By outcomes-focused we mean a system that judges schools at least in part based on the academic success of the students attending them.

[4] For example, in 1996, 40 states participated in NAEP and met the NAEP reporting requirements in 8th grade math (see National Center for Education Statistics, National Assessment of Education Progress (NAEP): History of Participation of Public Schools. https://nces.ed.gov/nationsreportcard/subject/about/media/naep_history_participation.xls).

1.2 Bipartisan Backlash and a Policy Vacuum

Early on, the ideas associated with the outcomes-focused agenda of the 2000s—standards, tests, and accountability—enjoyed bipartisan support. Leaders from both parties, business, and civil rights organizations supported a more muscular role for the federal government and its emphasis on improving outcomes for all, especially historically marginalized students.[5] The Obama administration continued to promote these ideas and extended the reach of the federal government with its Race to the Top (RTTT) program, pushing outcomes-based accountability into the realm of teacher evaluation.

But after more than a decade of implementation, conservative and liberal critiques of the approach and the federal government's role promoting it grew (Loss & McGuinn, 2018; Hess & McShane, 2018). Partly in response, NCLB's replacement—2015's Every Student Succeeds Act (ESSA)—curtailed some of the federal government's agenda and influence. ESSA preserved NCLB's testing requirements, but it gave states more flexibility on goal setting and accountability. As ESSA's architect, Senator Lamar Alexander, explained, the reauthorization would continue NCLB's "important measures of academic progress of students but restore to states, school districts, classroom teachers, and parents the responsibility for deciding what to do about improving student achievement" (U.S. Senate Committee on Health, Education, Labor & Pensions, 2015). With this, ESSA backed off NCLB and RTTT's approach to outcomes-based accountability, reduced the federal government's role in education policy, and avoided any new federal initiatives. It also left a policy vacuum of sorts, with less policy consensus and greater policy fragmentation than when NCLB was originally enacted (Finn & Hess, 2022).

Like everything else, life under ESSA was upended by the COVID-19 pandemic in the spring of 2020. In the months that followed, unprecedented disruptions to schooling were followed by social mobilization over police violence against Black people; debates over vaccines, masks, and critical race theory; and concerns about student well-being, teacher burnout, and school staffing. As the education system continues to grapple with the academic and social consequences of the pandemic, what comes next, in both policy substance and for federalism, remains far from clear.

In the months and years ahead, policymakers will face a host of difficult and puzzling decisions as they work to fill the vacuum left in the wake of NCLB. The most important ones will be about supporting students' academic and social recovery from COVID-19. But leaders will also confront related policy decisions about the federal government's role, performance accountability, and ESSA reauthorization.

Although current debates seem to have moved on from the outcomes-focused agenda of the 2000s and 2010s, the era holds important lessons that can inform what comes next. Many of the era's key lessons are about the politics of performance accountability (Hess & McShane, 2018) and the limits of federal power (Cohen & Moffitt, 2009). But the era also holds lessons about whether outcomes-based policies can drive improvements in student achievement (Dee & Jacob, 2011). In the pages that follow we consider this last set of lessons by reviewing what we know and do not know about the impact the outcomes-based policies in the 2000s and 2010s had on student achievement nationwide and what it means going forward.

1.3 What Does the Empirical Evidence Say About Impact?

As the education system starts to emerge from the turmoil of the COVID-19 pandemic and policymakers begin to grapple with ESEA reauthorization, they should consider what rigorous empirical evidence says about the outcomes-focused policies of the early and mid 2000s and whether they improved student achievement. We can review such evidence, in part, thanks to two byproducts of the outcomes-based policies themselves: new quantitative data on students and schools, and an increased interest in sophisticated research designs for making causal inferences (Schneider et al., 2007).[6] Both developments drove a wave of rigorous research that can help decision makers think critically about NCLB-era policies and their likely impact. Our modest hope is that this paper encourages some of that critical thinking and helps clarify the stakes involved; we know that decision makers must weigh a range of information—not just research—as they consider the trade-offs and interests involved in what comes next in education policy. We should also note that because we focus on research that rigorously assesses the effects of policies, we do not cover some policies that lack enough high-quality research to help us rule out competing explanations about their impacts.

[5] See McGuinn (2006) for more on the constituencies that coalesced around the law and the politics surrounding its passage.

[6] During this period of reform the federal government created the Institute of Education Sciences (IES) to promote and pursue rigorous scientific research in education. The advent of year-over-year testing data significantly increased the extent to which researchers were able to study education policy and interventions, including the research conducted by winners of the 2021 Nobel Prize in Economic Sciences, Joshua Angrist, David Card and Guido Imbens (See Barnum 2021, October 13).

The policies we review in the paper were part of an alphabet soup of initiatives (NCLB but also RTTT, School Improvement Grants [SIG], and Common Core State Standards [CCSS]). Although varied, all these initiatives were grounded in a basic logic of outcome-focused accountability. To different degrees, they specified who was accountable (schools, teachers), to whom (the government, families), for what outcomes (standards and test results), and with what consequences (sanctions, rewards, and extra support). They assumed that a combination of standards, tests, information, and incentives would drive improved outcomes for students. Because of our interest in the shift from inputs to outcomes, we leave out some important, concurrent initiatives that were not primarily about outcomes-based accountability or whose effects cannot be separated from larger policy programs (e.g., We do not discuss Reading First, a federal program under NCLB that mandated and supported reading instruction using phonics, phonemic awareness, vocabulary, reading fluency, and comprehension in kindergarten through third grade).[7]

Before we describe how we will proceed, it is worth acknowledging up front that some—perhaps many—readers will be skeptical about the evidence in this paper because it emphasizes standardized tests as the primary outcome of interest. Some readers will be skeptical about the technical properties of standardized tests (e.g., worries about item bias or construct bias). Others are likely to have concerns that tests, especially when used for accountability, create incentives to narrow the curriculum or to teach to the test.[8] Regardless of where one stands, annual testing remains an unmistakable part of the policy landscape and, as noted above, remains a requirement under ESSA. Tests were the touchstone of the policies at issue in this paper. We believe asking if the policies improved test scores, as they hoped to, is a first-order question for understanding their impact. But even if we take tests as a given, we still need to ask what they tell us. That is, on their own terms, do tests provide useful information about school performance? If empirical evidence suggests that outcomes-based policies had positive test impacts, do those impacts matter?

With these questions in mind, Section 2 begins with a closer look at standardized tests and what they tell us about other important student outcomes. On balance, we find that student achievement scores on standardized tests, despite their limitations, are useful predictors of later life outcomes and as indicators of school performance.

In other words, asking whether the outcomes-based policies of the 2000s and 2010s improved student test scores tells us something about how these policies affected meaningful outcomes for students. Next, in Section 3, we set the stage further by reviewing national trends in student achievement using the National Assessment of Educational Progress (NAEP), the Nation's Report Card. Then, in Section 4, we review empirical research related to the nationwide impact of test-based accountability, teacher evaluation (accountability for individual performance), school turnarounds (consequences for low performance), and standards (goal setting and assessment). As those parentheticals suggest, some of these policies encompassed the full logic of outcomes-based accountability (e.g., test-based accountability under NCLB and teacher evaluation under RTTT), whereas others offer a more partial view (e.g., turnarounds and standards). After summarizing the nationwide results, Section 5 takes a deeper look at some of the mechanisms that arguably mediate these policies' effects on student achievement: money, teachers, and information. Because the federal government relied on states and districts to make decisions and act on behalf of all the policies in question, Section 6 considers how variation in state and local implementation may have influenced policy impacts. Finally, in Section 7, we conclude with a summary and what our review implies for the future.

In the end, there is no doubt that the 15 years of education policy that NCLB set in motion fundamentally shifted the nation's understanding of public schools and, specifically, what constitutes a "good" school. After 2001, test-based outcomes, including outcomes for historically marginalized and underserved student groups, occupied public and policy attention like never before. In the pages that follow, we tell a nuanced story of impact that goes above and beyond the attentional shift towards outcomes: we find that these policies improved test-based outcomes for some students in some subjects but fell short of their grand ambitions to improve outcomes in all schools for all students. The combination of standards, tests, information, and incentives—and the federal government's role encouraging them—clearly changed the nation's schools. Although this paper does not focus on implementation, our discussion in Section 6 suggests the reforms' problems, in addition to reflecting the backlash associated with their grand ambitions, arguably stemmed more from the timing and quality of their implementation than from flaws in their underlying logic. The NCLB era assembled a potent set of policy ideas and ingredients. Rather than abandoning them, federal policymakers should work on revamping and refining them.

[7] For research on Reading First, see Gamse et al (2008).

[8] These are legitimate concerns. Interested readers can consult a range of evidence on whether teachers responded to the pressure from test-based accountability policies by focusing effort on tested subjects at the expense of non-tested subjects. In multiple studies, for example, elementary teachers reported spending more time on reading instruction and less time on history and science in response to NCLB (Dee et al., 2013; Smith & Kovacs, 2011; West, 2007). National data suggest that teachers in schools facing NCLB sanctions (on the margin of meeting AYP) were slightly less likely to report having taught science and social studies lessons in the prior week than non-threatened schools (Reback et al., 2014) and that teachers spent more time on science instruction where science test achievement was incorporated into states' accountability formulas (Judson, 2013). The consequences of these responses on student learning in non-tested subjects are not clear, however. Some evidence suggests that NCLB did not lead to significant changes in course taking in non-academic subjects, such as music (Elpus, 2014), or changes in student learning in low-stakes subjects or student reports about enjoyment of learning (Reback et al., 2014). Other evidence suggests that some teachers responded to accountability pressure by adopting narrow test-prep strategies that inflated scores (Koretz, 2017; Wong et al., 2003) and, in some cases, by outright cheating on tests (Fantz, 2015; Jacob & Levitt, 2003; Sass et al., 2015) or by excluding low-performing students from testing to boost school ratings (Cullen & Reback, 2006; Figlio, 2006; Figlio & Getzler, 2006). Although accountability pressure led some schools to manipulate their data in ways that call into question state results for high-stakes decision making, these behaviors do not undermine conclusions suggested by the NAEP reviewed later in the paper.

2. What Do Standardized Tests Tell Us About Later Outcomes?

State standardized achievement tests were a central component of the federal government's outcomes-based policies of the 2000s and 2010s. Policymakers and education leaders used tests to monitor the system's progress toward performance goals and to inform decisions about accountability, including interventions in low-performing schools. Under NCLB, states were required to test students annually in reading and math for Grades 3–8 and once in high school. The law also required states to report on the percentage of students who were "proficient" (overall and for subgroups of students) (Stullich et al., 2007).[9] As noted above, this emphasis on tests continued with the RTTT competition during the Obama administration. RTTT, for example, promoted the use of test-based growth measures in teacher evaluations. The SIG Program also relied on tests to identify low-performing schools for intervention. As noted earlier, even with ESSA's recent scaling back of the federal role, NCLB's testing requirements remain.

Although largely uncontroversial among policymakers when NCLB passed, standardized tests and test-based accountability have come under serious criticism.[10] Critics argue that test-based accountability narrows the curriculum and focuses teachers and students on low-value test-taking skills (e.g., Koretz, 2017). Others make stronger critiques related to the design and purpose of testing as they relate to racial and social justice.[11] Tests clearly do not tell us everything there is to know about how schools and teachers contribute to student learning. A host of important phenomena not covered by tests—for example, the contributions schools make to students' social and emotional development—arguably matter for success in life as well (we return to this issue at the end of Section 4).

We want to acknowledge all these concerns about tests, even if responding to them is beyond the paper's scope. Given our focus, the most immediate question about tests is whether test scores, taken on their own terms, measure something important that is predictive of better outcomes later in life for students. After all, if standardized tests—the cornerstone of test-based accountability—are not measuring skills and knowledge connected to students' longer term life outcomes, the incentives created by outcomes-focused policies in the 2000s and 2010s were arguably misaligned with what most people would agree is one of the school system's ultimate goals: preparing students for success after K-12.

2.1 Do Tests Predict Later Success?

At first glance, the answer here seems clear: Students who do better on tests also tend to do better in college and work. Ample evidence suggests that test scores predict a range of student outcomes after high school (Hanushek, 2009).[12] Heckman et al. (2006), for example, find that test scores are significantly correlated with educational attainment and labor market outcomes (e.g., employment, choice of occupation, work experience) and negatively correlated with risky behaviors (e.g., teenage pregnancy, smoking, illegal activity). More recently, Lin et al. (2018) find that cognitive skills at the end of high school are associated with rising labor market returns as people age. But, as the well-worn adage says, correlation is not causation.

[9] As we note later in Section 6, the law left both the design of the tests and the determination of what constituted proficiency to the states.

[10] For example, see Strauss 2020. Tests have come under fire from both teachers' unions (Taylor & Rich, 2015) and researchers (Hitt et al., 2018). For a more extensive discussion, see Goldhaber and Özek (2019).

[11] For example, Kendi (2016) argues that standardized tests are racist tools, born of early 20th-century eugenics and White supremacy, that effectively maintain the racial hierarchy in the United States. "Standardized tests," Kendi writes, "have become the most effective racist weapon ever devised to objectively degrade Black minds and legally exclude their bodies." (Why the academic achievement gap is a racist idea. Black Perspectives, October 20, 2016. https://www.aaihs.org/why-the-academic-achievement-gap-is-a-racist-idea/).

[12] See also Chamberlain (2013), Chetty et al. (2014b), Heckman et al. (2006), Mulligan (1999), Murnane et al. (2000), Lazear (2003), Lin et al. (2018), and reviews by Hanushek (2009) and Watts (2020).

The relationship between test scores and life outcomes obviously reflects other, unmeasured student characteristics (e.g., diligence) and conditions (e.g., family characteristics, the systemic effects of racism and social inequality, adverse environmental exposures [such as lead]). There is no doubt that a range of individual, social, and environmental factors affect students' opportunities and success in school and life. The question is, do the learning outcomes measured by tests tell us anything else?[13] If a school improves student learning as measured by test scores, should we conclude that it is helping improve its students' opportunities later in life? This is a key question for test-based accountability. Answering it is not easy. The underlying issue is that non-school factors that are unobserved to researchers, such as the degree to which students receive encouragement in the home, may influence both test scores and lead to better adult outcomes.

Researchers have generally relied on quasi-experimental methods to investigate the causal effects of interventions on learning (which is not observed) as measured by their effects on test scores (which are observed). The question is whether test scores capture the longer run effects of changes in learning independent of other factors that also affect longer run outcomes (Athey et al., 2019). Researchers have tried to tackle this problem by examining the causal effects of other educational inputs on both test scores and later life outcomes. The reasoning in this approach is that if we find that these inputs have effects in the same direction on both tests and later life outcomes, then there is plausible evidence that test scores are measuring something that affects later outcomes.

Unsurprisingly, not all these types of studies reach the same conclusion. Some studies of the effects of school choice, for example, find that effects on test scores and later outcomes do not point in the same direction (Hitt et al., 2018). However, a larger body of studies on teachers (Chetty et al., 2014b), peers (Chetty et al., 2011), small class sizes (Dynarski et al., 2013), finance reform (Jackson et al., 2016; LaFortune et al., 2018), and some school choice programs (Angrist et al., 2016; Dobbie & Fryer, 2015) support the idea that there is a causal link between what test scores measure and life outcomes. Dobbie and Fryer (2015), for example, exploit oversubscribed charter schools to compare lottery winners and losers and find that attending a high-performing charter school increases test scores and college attendance and decreases the likelihood of risky behavior. Elsewhere, researchers have used the plausibly random timing of school finance reforms (often driven by lawsuits) to assess the effects of funding on short- and long-run student outcomes. Using this approach, Jackson et al. (2016, 2021) and LaFortune et al. (2018) find that changes in school spending affect test scores and adult outcomes.[14] Consistent with these quasi-experimental results, Dynarski et al. (2013) leverage data from an actual experiment (the famous Tennessee Student-Teacher Achievement Ratio [STAR] class size project) and find that the experiment's test score effects were an excellent predictor of postsecondary outcomes.[15]

Widespread critiques of testing notwithstanding, these studies suggest that the measured outcomes on standardized tests do indeed capture aspects of student learning that matter for both assessing student progress in school and predicting student success after schooling ends.[16] Research aside, this claim has some face validity: To the extent that your experiences later in life depend on literacy, numeracy, or some other specialized knowledge (e.g., chemistry), measures of your knowledge and skills in school should tell us something—again, not everything—about the opportunities you may have later in life.

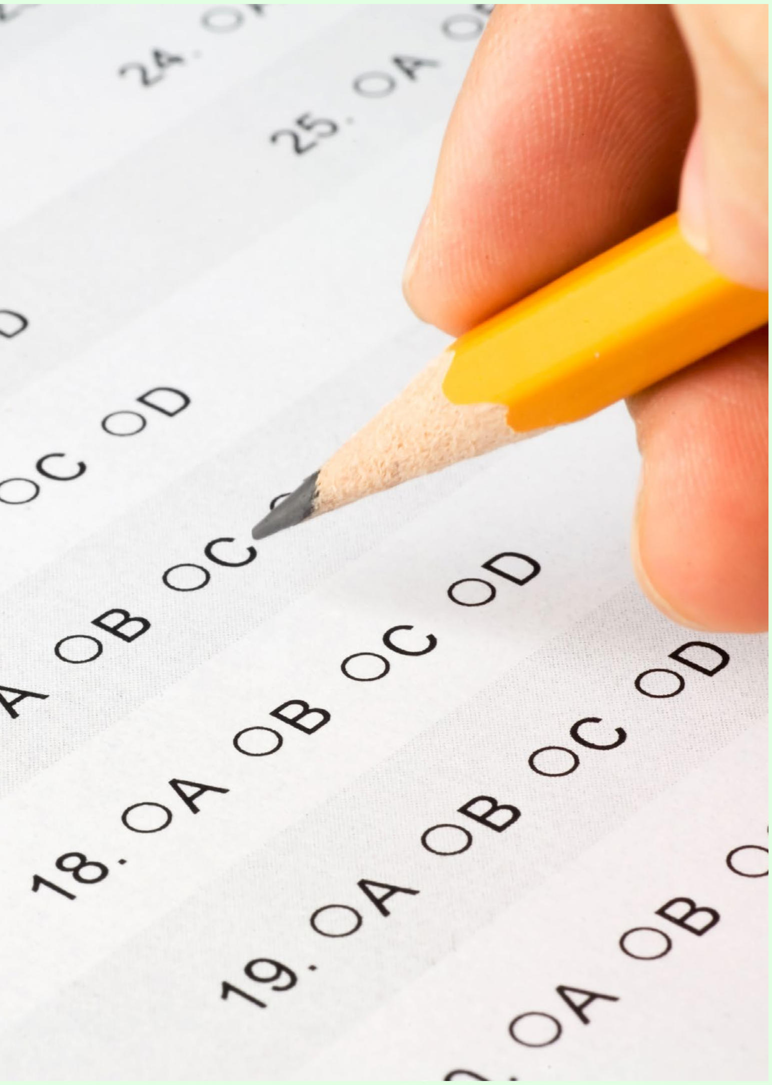
[13] It is worth noting that a range of evidence suggests that school grades predict college success above and beyond college-readiness tests like the SAT and ACT (See Chingos 2018). But outcomes-based accountability policies have rightly avoided using grades to measure school performance given their manipulability and the risk of perverse incentives.

[14] For a useful summary, see Table 1 of Jackson and Mackevicius (2021).

[15] This does not mean, however, that interventions that fail to impact test scores necessarily will fail to have long-term impacts—the field still needs more research on the long-term impacts of interventions that, in the short run, do not appear to improve test scores.

[16] Although we believe the weight of the evidence shows test scores to be causally linked to later life outcomes, interventions that affect test scores will not always lead to changes in adult outcomes and vice versa. For examples of interventions that affect test scores but not adult outcomes, see Greene (2016) and Hitt et al. (2018). In addition, test scores themselves often understate the total impact of successful interventions (Jackson & Mackevicius, 2021).

3. How Have Achievement Results Changed Over Time



Given that test scores tell us something about how well schools are preparing students for success in the future, what can we make of the performance of the nation's schools over time? The best evidence on this question comes from the National Assessment of Educational Progress (NAEP).[17] Known as the Nation's Report Card, the NAEP is the largest and longest-standing account of students' academic performance nationwide.

Overall, NAEP's long-term math and reading trends suggest that American schools have seen a long period of sustained and gradual progress.[18] Figure 1 shows the NAEP's long-term average trends in math (Panel A) and reading (Panel B) for students aged 9, 13 and 17. Overall, Figure 1 shows that results have increased in reading and math achievement for 9- and 13-year-olds since the 1970s. However, the results for 17-year-olds have remained relatively flat. Panels A and B in Figure 1 also reveal different trajectories in math and reading achievement. In math, the results for 9- and 13-year-olds are flatter in the 1970s and then trend upward through 2012. In reading, the results for 9- and 13-year-olds increased in the 1970s, stagnated in the 1980s and 1990s, and then increased again through 2012. Again, the results for 17-year-olds in both subjects are flatter than they are for younger students. However, some analysts (Barnum, 2022, March 31) speculate that the flat results for older students underestimate their achievement gains over time because the composition of test takers has changed thanks to rising graduation rates (Murnane, 2013) (see Figure 2).[19]

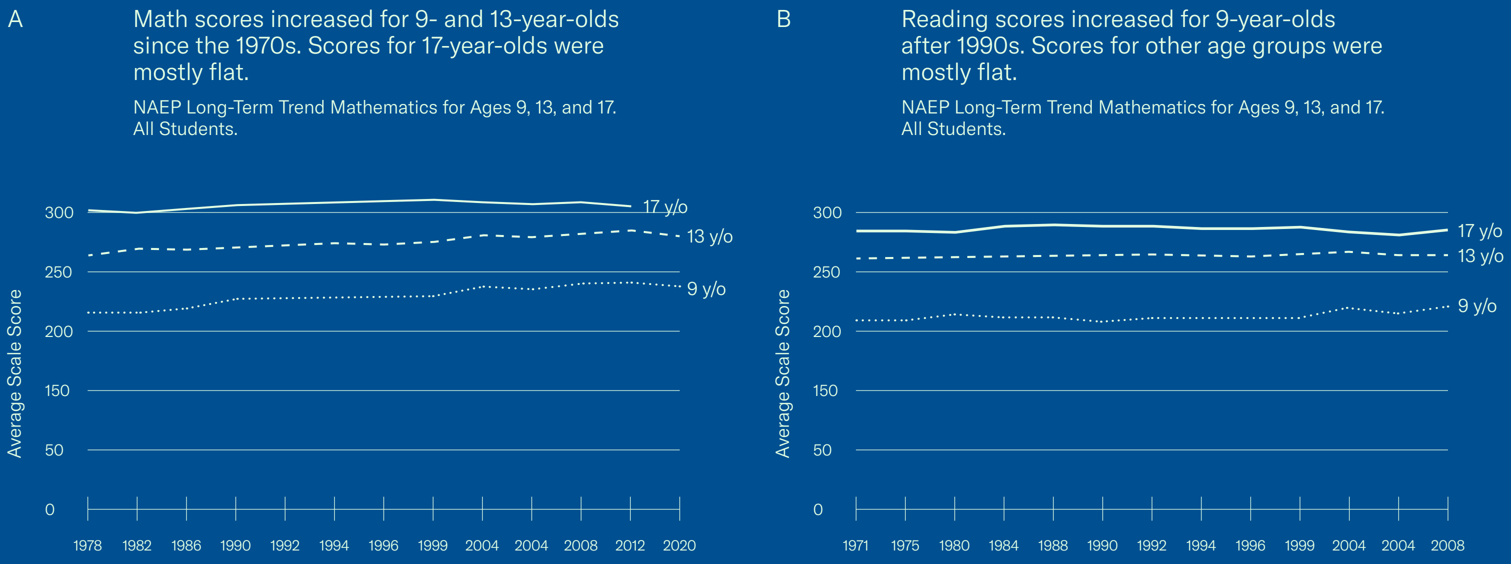
The long-term NAEP trends also tell us something about differences in achievement across student groups. Figure 3 disaggregates long-term NAEP results by student race and ethnicity.

[17] Variation between and within state assessments over time means state assessments have significant limitations when it comes to describing national trends (Backes et al., 2018).

[18] The NAEP long-term trends differ from the main NAEP assessments in terms of the length of the data panel, the frequency of the assessment (the long-term trends are typically administered every 4 years as opposed to main NAEP, which is typically administered every 2 years) and somewhat in terms of the content of the assessments. The content of the NAEP long-term trends has remained relatively stable over time to provide for long-term comparability, whereas the content in main NAEP is updated to reflect changes in curriculum over time. For more details, see https://nces.ed.gov/nationsreportcard/about/ltt_main_diff.aspx. Note that NAEP long-term trends data were not available for 17-year-olds in 2020 because of the COVID pandemic (due to the timing of the assessment for 17-year-olds).

[19] Indeed, over such a long time frame, the composition of all age groups has changed, suggesting these longer-run changes in achievement trends should be interpreted with a healthy dose of caution.

Figure 1 Long-Term NAEP Trends in Math and Reading by Age



A

Math scores increased for 9- and 13-year-olds since the 1970s. Scores for 17-year-olds were mostly flat.

NAEP Long-Term Trend Mathematics for Ages 9, 13, and 17. All Students.

B

Reading scores increased for 9-year-olds after 1990s. Scores for other age groups were mostly flat.

NAEP Long-Term Trend Mathematics for Ages 9, 13, and 17. All Students.

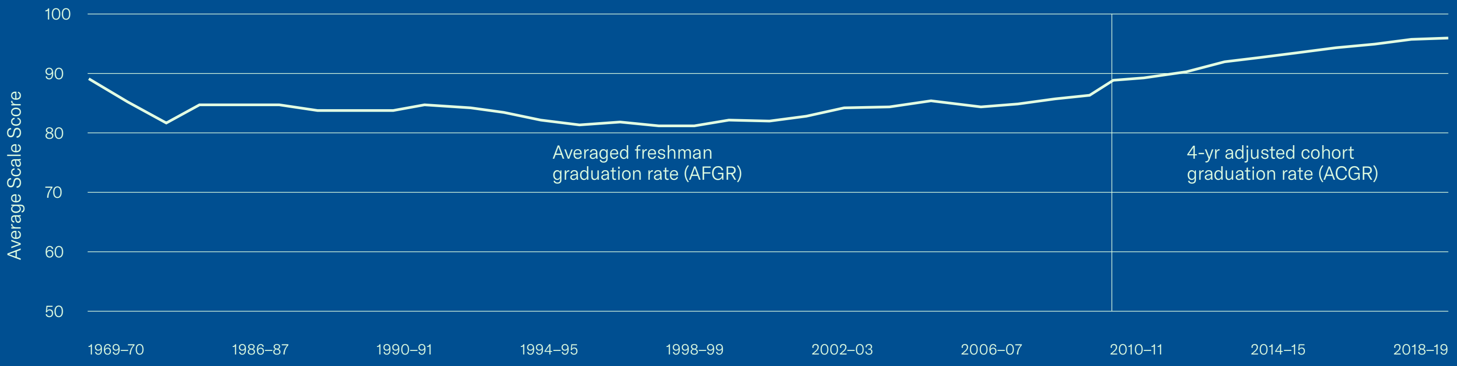
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

Figure 2 Graduation Rates 1969–2018, Using Two Measures

High school graduation rates show increasing trend overtime.

High school graduation rates 1969-2008 (AFGR) and 2010-2018 (ACGR).



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

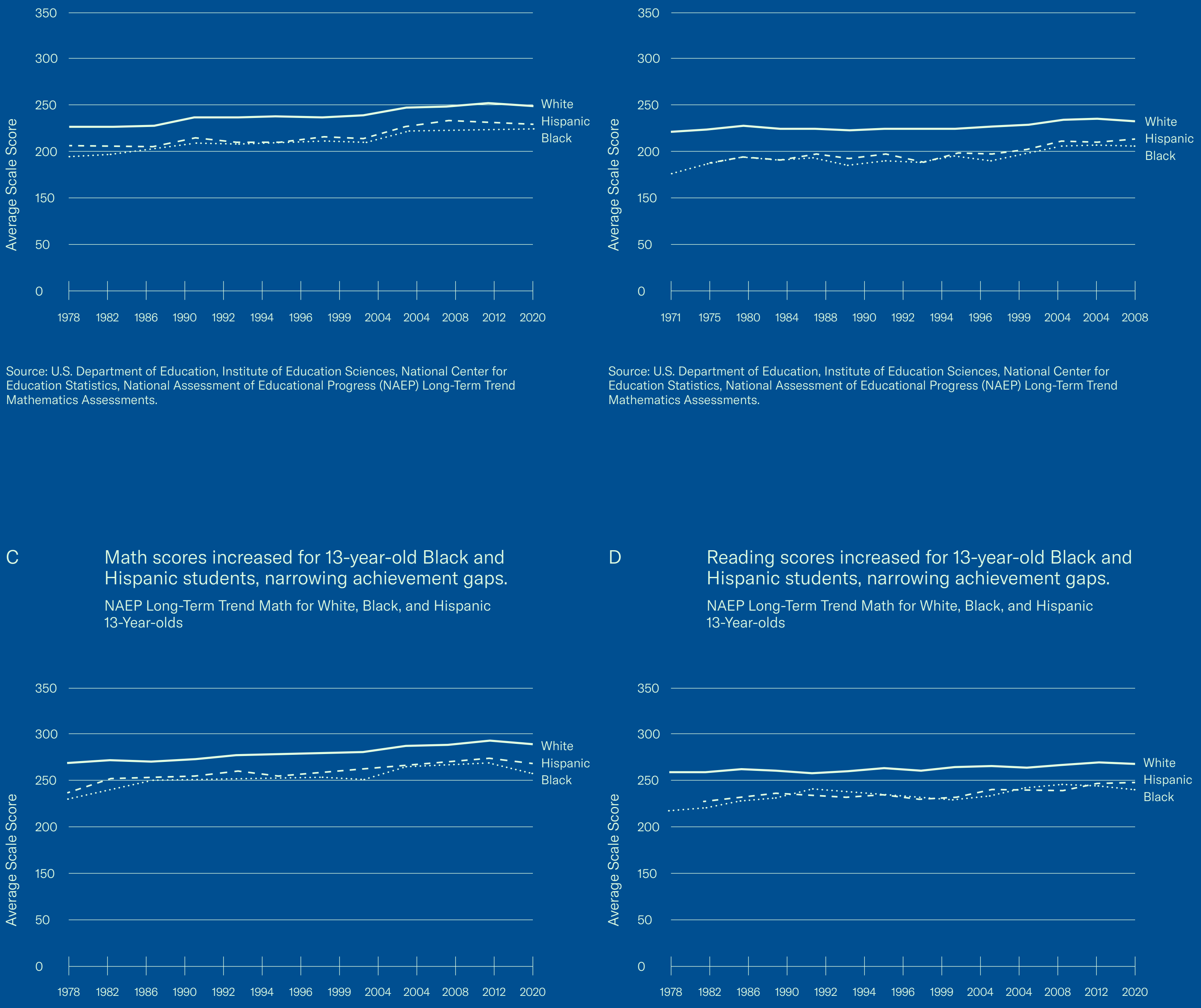
These charts show that gains in NAEP math scores increased the most for younger Black and Hispanic students. More broadly, Shakeel and Peterson (2022) use data from 7 million test takers from 1971 to 2017 (using the NAEP and other tests that are comparable over time) and conclude,

> The median rate of progress made by the average Black student [between 1971 and 2017] exceeds that of the average white student by about 10 percent of a standard deviation per decade in both reading and math. Over 50 years, that amounts to about two years' worth of learning, or about half the original learning gap between white and Black students. The disproportionate gains are largest for students in elementary school. They persist in middle school and, in diminished form, through the end of high school.

Although not shown here, scores have also increased over time for students who perform at the bottom end of the test distribution (Shakeel & Peterson, 2021). In sum, the long-term gains on the NAEP reveal a decades-long narrowing of test score achievement gaps between underserved groups (e.g., students of color, lower achieving students) and more advantaged groups (e.g., White students, higher achieving students) (Hanushek et al., 2020; Hashim et al., 2020; Reardon, 2011).[20]

[20] There is disagreement about the extent to which socioeconomic achievement gaps have closed over time; Reardon (2011) suggests that socioeconomic gaps have widened, whereas Hashim et al. (2020) and Hanushek et al. (2020) find that they have shrunken or remained about the same.

Figure 3 Long-Term NAEP Trends in Math and Reading by Race/Ethnicity



A

Math scores increased for 9-year-old Black and Hispanic students, narrowing achievement gaps.

NAEP Long-Term Trend Math for White, Black, and Hispanic 9-Year-olds

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

B

Reading scores increased for 9-year-old Black and Hispanic students, narrowing achievement gaps.

NAEP Long-Term Trend Math for White, Black, and Hispanic 9-Year-olds

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

C

Math scores increased for 13-year-old Black and Hispanic students, narrowing achievement gaps.

NAEP Long-Term Trend Math for White, Black, and Hispanic 13-Year-olds

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

D

Reading scores increased for 13-year-old Black and Hispanic students, narrowing achievement gaps.

NAEP Long-Term Trend Math for White, Black, and Hispanic 13-Year-olds

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Long-Term Trend Mathematics Assessments.

3.1 NAEP Trends are Clear but Their Cause Isn't

Some observers look at the NAEP trends shown in Figures 1 and 2 and conclude that the rise in results in the 2000s shows that outcomes-based accountability reforms, particularly NCLB, were a success (e.g., NAEP scores rise; NCLB gets credit, 2007). But others conclude that the pace of NAEP gains dropped off in the 2000s because of NCLB, suggesting that the law was a failure (e.g., Fuller et al., 2007; Neill, 2007). Such debates over the NAEP continued when the latest data showed achievement gains stalling and, in some cases, reversing sometime between 2012 and 2020.[21] For the first time in NAEP's history, the average reading and math scores of 13-year-olds declined in 2019; the results were so worrying that public officials questioned their accuracy.[22] Some observers blamed the drop on the federal government's retreat from NCLB and outcomes-based accountability policies (Kress, 2021). Others explained it by pointing to the Great Recession and its impact on education funding and poverty (Forte, 2021; Petrilli, 2021). Still others speculated that the drop stemmed from a misalignment between NAEP and the CCSS (Griffith, 2021; Polikoff, 2021).

We do not think these debates can resolved based on NAEP achievement trends because the trend data have serious limitations for assessing policy impacts.[23] The primary problem is that it is difficult if not impossible to convincingly disentangle changes in NAEP (or any other measure of progress) from other concurrent changes in the economy or broader social forces that could have caused the observed trend over the same period. Changes in family incomes, parental education, nutrition, or the social safety net, for example, could all contribute to the trends we see.

If we want to assess policy impacts—policy X produced outcome Y—we need a more rigorous approach, one that considers other factors that could have produced that outcome in the absence of the policy. As we noted in the introduction, thanks to new data and research, evidence of this kind exists for several strands of the outcomes-based policies of the 2000s and 2010s. By trying to take possibly confounding causes into account (e.g., the characteristics of students and other concurrent events), this evidence helps shed some light on the extent to which outcomes-based policy ideas were effective levers for increasing student achievement. In the next section, we begin looking at what this kind of evidence has to say in more detail.

[21] It is important to note that these shifts occurred prior to the COVID-19 pandemic. A closer look at the data shows that these average declines were driven mostly by drops among already low-scoring students (the results for high-scoring students were flat). In other words, the gaps in achievement that had narrowed in prior decades were widening.

[22] An official at the National Center for Education Statistics quoted in U.S News & World Report stated, "I asked them to go back and check because I wanted to make sure [the results were accurate]. I've been reporting these results for years—for decades—and I've never reported a slide like that." (Camera, 2021, October 14).

[23] The congressionally mandated assessment of Title I (Stullich et al., 2007) showed changes in NAEP tests over time and in relation to NCLB but explicitly counseled against interpreting any changes over time as being a causal reflection of NCLB.

4. Evidence Connecting Outcomes-Based Reforms to Student Achievement

As we noted in the introduction, the decade-plus of reform that NCLB set in motion saw a range of outcomes-based accountability policies in education. From the original NCLB legislation, to RTTT, SIG, and the CCSS, these policies assumed, to some degree, that a combination of standards, tests, information, and incentives (and/or additional targeted resources) would drive improved outcomes for students.

This logic was epitomized by NCLB. In order to access Title I funding, NCLB required states to adopt standards in reading and math as well as set a goal that all students would be "proficient" in those subjects by 2014. It required states to administer math and reading tests annually for students in Grades 3–8 and once in high school. States had to report the results of these tests by school and for subgroups of students, reflecting the law's interest in improving the achievement of all students.[24] States also had to project what it would take for schools to reach universal proficiency by 2014 and monitor how well schools performed against those projections by reporting whether they were making Adequate Yearly Progress (AYP) toward universal proficiency. Finally, NCLB created performance incentives by mandating an escalating series of sanctions for schools that did not meet AYP. These sanctions included allowing the school's students to transfer to a higher performing school, providing supplemental services (e.g., tutoring), and implementing organizational and programmatic reforms (e.g., "corrective action" and "restructuring").[25]

As we note in Section 6, even with these requirements, states and local school districts retained, under United States' decentralized education system, a great deal of discretion over translating these requirements into reality (e.g., standards setting, defining proficiency, implementing interventions etc.).

Much of the accountability logic associated with NCLB under the Bush administration carried forward with RTTT during the Obama administration. Part of 2009's American Recovery and Reinvestment Act (ARRA), the RTTT was a $4.3 billion grant competition for the states that judged proposals based on whether they included a host of preferred policy commitments. Like NCLB, these preferred policy commitments included standards and assessments focused on students' academic learning (e.g., RTTT's emphasis on "college and career ready").[26] RTTT's approach to teachers encouraged evaluation systems that measured teacher performance, in part, by student growth on standardized tests.[27] RTTT also rewarded grant proposals for including interventions in low-performing schools. These reforms were reinforced by the Obama administration's NCLB waiver process, which allowed states to avoid NCLB's accountability provisions (e.g., "corrective action") by making policy commitments like those associated with RTTT.[28] The Obama administration's SIG Program also provided $3 billion to states to intervene in their lowest performing schools using one of four approved intervention models (see Section 4.3).

[24] These subgroups included special education, English language learners, low-income students, and racial minorities.

[25] These "corrective" organizational and programmatic reforms could include replacing school staff who are relevant to the failure to make AYP, adopting a new curriculum based on state standards and providing appropriate professional development for all relevant staff, significantly decreasing management authority at the school level, appointing an outside expert to advise the school on its progress toward making AYP based on its school plan, or extending the school day or school year. The "restructuring" sanctions included a range of governance reforms (e.g., conversion to a charter school, state takeover) and the reconstitution of school staff.

[26] More specifically, RTTT created an incentive for states to adopt the CCSS and aligned assessments. See LaVenia et al., 2015.

[27] In contrast to RTTT's emphasis on teacher performance, NCLB's highly qualified teacher (HQT) provisions had called on districts to comply with new rules and regulations for what constituted a "highly qualified" teacher: having a Bachelor of Arts degree from a 4-year institution, being fully certified, and demonstrating subject area knowledge. To encourage the use of HQT standards in licensure requirements, NCLB required states to report on the HQT status of its teacher workforce and notify parents when their children did not have access to an HQT.

[28] During this same period, the federal government also promoted market-based accountability by encouraging the growth of the charter school sector (Egalite, 2018). Charter schools are public schools of choice that operate with expanded flexibility outside of the traditional school district system. In exchange for expanded flexibility, charter schools face a combination of market accountability (as schools of choice, families can vote with their feet if they are unhappy) and government accountability (charter schools must periodically renew their contracts with authorizing oversight agencies) (Richmond, 2022). Although charter schools were arguably not a centerpiece in NCLB, the law created new support and funding for charter schools by boosting investment in the Clinton-era Charter Schools Program (CSP) and by creating the Credit Enhancement for Charter School Facilities Program. Both programs saw additional funding during the Obama administration. Beyond these two programs, RTTT also promoted charter schools by giving extra points to grant applicants that lifted caps on charter enrollment, equalized funding to charter and traditional schools, and allowed districts to authorize charters. Readers interested in comparisons between charter schools and traditional public schools can consult a wide literature including observational studies (CREDO, 2009; Cremata et al, 2013) as well as so-called lottery studies that use oversubscribed admissions lotteries to approximate a randomized control trial (E.g., Gleason et al, 2010; Angrist et al, 2010, 2012, 2013; Dobbie & Fryer, 2011, 2013; and Hoxby et al, 2009). On balance this literature finds few differences between the outcomes of students in charter schools and traditional public schools in general but better outcomes in urban areas for Black, Latinx, and low-income students who enroll in charter schools compared to traditional public schools (See Cohodes & Parham, 2021 for a recent review).

These intervention models echoed the organizational and programmatic reforms associated with NCLB but came with additional resources and targeted a smaller share of low-performing schools (Burns & Strunk, 2021). In each case, an underlying logic combined some blend of standards, tests, information, and incentives.

As this quick summary suggests, the federal government supported a range of complex outcome-focused policy ideas over the last 20 years. Gauging their impact is challenging because these policies often applied to all schools, teachers, or students at once, making useful comparisons difficult. Moreover, as we note later, their implementation relied on multiple levels of government, from the federal government to states to local school districts and schools. So, assessments of their net effect necessarily gloss over significant variation across jurisdictions.

At the same time, advances in data and research have made it possible to assess the impact of some of these policies in rigorous ways that were less common before NCLB. Using new data and sophisticated methods, researchers have leveraged quasi-experimental strategies (like the methods discussed in Section 2 about the predictive power of tests) to draw comparisons by looking at everything from the timing of reforms, the classification of schools and students into interventions based on some threshold score, and the contrast between federal mandates and policies that existed in states prior to the federal mandates. National evidence of this sort is available for two policies that encompass the full logic of outcomes-based accountability: test-based accountability and teacher evaluation reform. Rigorous research is also available on two important components of outcomes-based accountability: school turnaround efforts and college and career-ready standards. In the following sections, we review what this evidence tells us about the impact these initiatives had on student test scores nationwide.

4.1 Test-Based Accountability

Before we turn to nationwide evidence on test-based accountability, we can find some clues about possible effects by looking at research on states that had test-based accountability policies before NCLB was signed into law. Much of this pre-NCLB research is encouraging. Carnoy and Loeb (2002), for example, find evidence that states with "stronger" accountability systems in the mid-to-late 1990s saw greater gains in math achievement on the NAEP. To measure the strength of state accountability systems, they create an index based on whether states require student testing and performance reporting, impose sanctions or rewards based on test performance, and require students to pass exams to graduate from high school. On achievement, Carnoy and Loeb (2002) find statistically significant test gains in states with stronger accountability systems for Black, Hispanic, and White students in Grade 8 and for Black and Hispanic students in Grade 4 (the Grade 4 results were positive but not statistically significant for White students). In a similar study, Hanushek and Raymond (2005) examine the relationship between pre-NCLB policies and NAEP performance and find positive effects for states whose systems attached consequences to school performance (e.g., state takeovers of struggling schools).[29] Unlike Carnoy and Loeb (2002), Hanushek and Raymond's study of "consequential accountability" finds larger effects for Hispanic students relative to White students and smaller (but not statistically significant) effects for Black students.

Subsequent literature has built off these early findings by examining the effects of NCLB nationwide. If there is a seminal study here, it is Dee and Jacob (2011). Following Hanushek and Raymond, Dee and Jacob contrast achievement trends in states that had consequential accountability policies pre-NCLB with states that adopted consequential accountability after (and because of) NCLB.[30] If test-based accountability has an impact on student test achievement, this comparison should reveal differences between these two types of states.[31] In line with pre-NCLB studies, Dee and Jacob (2011) find large and significant effects of NCLB on test scores of elementary students in math (an effect size of .23 by 2007, or about the equivalent of roughly half a school year worth of learning).[32] These gains are large and across the entire test distribution; the effects are especially large for students at the lower end of the test distribution.[33]

Dee and Jacob (2011) also find differential impacts across student groups, an important finding given the law's focus on reducing inequity. Their study suggests, for example, that elementary math gains are larger for Black and Hispanic students than for White students, and larger for students who are eligible for subsidized lunch compared to those who are not. In contrast to the elementary math results, the estimated effects they find in reading at the elementary level were positive but small and not statistically significant. In Grade 8, the results were small, negative, and not statistically significant. The authors also explore whether NCLB's emphasis on math and reading had negative effects on learning in science (this is related to the concerns about narrowing the curriculum we raised in the introduction). They suggest that these concerns are overstated, concluding that "NCLB did not have an adverse impact on student performance in science as measured by the NAEP" (p. 442).

Studies that use alternative methods to assess the effects of NCLB find similar results. Wong et al. (2015), for example, contrast test achievement in public schools and Catholic schools before and after NCLB and between high- and low-standard states.[34] Again, using NAEP tests as an outcome, they find that across "three contrasts, two grade levels, and three types of causal hypothesis, the math data are totally coherent with NCLB being effective" (p. 266). Like Dee and Jacob (2011), their analyses find much weaker evidence that NCLB affected reading, where no national contrasts are statistically significant.

There are also studies that use international comparisons of tests taken across multiple countries to examine relative country performance before and after NCLB (e.g., Darling-Hammond, 2007; Dee & Jacob, 2010).[35] Although these studies reach different conclusions about whether U.S. trends are up or down relative to other countries, none suggest large or sharp deviations from trends before or after NCLB implementation. These cross-country comparisons are potentially suggestive of the impact (or lack of impact) of national policies. But we agree with Dee and Jacob's (2010) caution about drawing strong inferences from these types of comparisons given factors in other countries that might influence trends over time in student achievement but are hard to measure.

[29] Hanushek and Raymond (2005) classify states based on whether they provide public report cards about student test performance or, in addition to this, attach consequences such as monetary rewards to schools or school personnel for high performance, potential state takeover of struggling schools, and additional choice options for students attending schools judged to be failing. The authors also note that most states did not end up imposing consequences, so the assessment was generally about the threat of consequences for performance. They do not break the results out separately for math and reading NAEP tests.

[30] Dee and Jacob (2011) use Hanushek and Raymond's definition of consequential accountability (see Footnote 17).

[31] More specifically, Dee and Jacob test whether there is a deviation from the prior achievement trends in the states that had to change their accountability systems because of NCLB compared to the states that already had consequential accountability systems before NCLB. In theory, the latter group was less affected by NCLB's implementation than the former. They identify 25 states that implemented consequential accountability systems because of NCLB. Although the authors provide a lot of evidence that their findings are robust to various threats to interpreting the findings as causal, it is also the case that the statistical approach to assessing interventions based on changes over time in program adoption has changed significantly in recent years. We are not sure whether the findings hold up to the recently developed model specification checks (e.g., Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021). This same caution applies to assessments of other large-scale interventions, such as the Common Core that we discuss in Section 4.5.

[32] The effect sizes are large relative to many educational interventions that assess student test outcomes (Kraft, 2020). Translating these effect sizes into typical weeks or months of learning gains is tricky because the measured learning (in standard deviation terms) varies by grade. That said, Lipsey et al. (2012) report that math gains in elementary grades is in the range of .50 standard deviations.

[33] The authors also find positive for Grade 8 math students, although they are not statistically significant at conventional levels (p value of 0.12; levels of 0.10 and 0.05 are usually reported as significant and highly significant, respectively).

[34] The arguments here are that (a) NCLB should have a direct effect on public schools but not private schools and (b) states with high standards for student proficiency (more difficult tests or higher test score cutoffs for proficiency) are more likely to have to make changes to their education systems because they face sanctions under NCLB for having students not reach proficiency.

[35] These include cross-country comparisons on assessments from the Program for International Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS).

Although not national in scope, several other studies provide insight into the effects of test-based accountability by focusing on students who are on the threshold of state proficiency standards, so-called "bubble" students (e.g., Ballou & Springer, 2017; Krieg, 2011; Neal and Schanzenbach, 2010; Reback, 2008; Springer, 2008) and schools on the margin of sanctions (e.g., Ahn & Vigdor, 2014; Reback et al., 2009). This literature generally finds that students and schools in these positions see the most test score gains (for an exception, see Ballou & Springer, 2017).[36] Such results are consistent with the idea that performance-based incentives may have the biggest effects on students and schools performing at the threshold between receiving and not receiving a sanction.

In summary, the weight of the empirical evidence suggests that, overall, NCLB's test-based accountability reforms led to increases in test-based measures of student achievement in math, particularly for elementary-level students. We find less consistent evidence that NCLB had positive effects on reading scores. These conclusions are broadly consistent with several other research reviews.[37]

4.2 Teacher Evaluation Reforms and Related Policies

As we noted earlier, RTTT pushed the idea of outcomes-based accountability into the realm of teacher evaluation during the Obama administration. This push was spurred, in part, by compelling empirical evidence about the impact of teachers on student achievement and how it varies significantly from one teacher to another (e.g., Rivkin et al., 2005). We discuss this evidence more extensively in Section 5.2. For now, it is worth noting that simulations based on this literature suggested that if schools used information about teacher effectiveness to make personnel decisions, it could lead to large improvements in achievement test scores (e.g., Goldhaber & Hansen, 2010) and longer-term outcomes (Chetty et al., 2014b). Meanwhile, related research highlighted some of the problems with business-as-usual teacher evaluations, which typically do little to differentiate teachers based on performance (Weisberg et al., 2009). Based in part on both sets of findings, the federal government pushed teacher evaluation reforms through the RTTT.[38] The government's push for evaluation reform was also part of the NCLB waiver process, which required similar reforms as a condition for receiving a waiver.

Between 2009 and 2015, most states revised their evaluation systems to increase performance accountability for teachers. The number of states that required student learning (often measured by student growth on standardized tests) as part of their teacher evaluation systems jumped from 15 to 43 states by 2015 (Walsh et al., 2017). Most of these states (more than 80%) tied evaluation ratings to decisions about teacher professional development. A smaller share (approximately 60%) tied evaluations to high-stakes decisions (e.g., employment eligibility), and a still smaller share (20%) tied evaluations to teacher pay. In some cases, low ratings could theoretically lead to dismissal (Steinberg & Donaldson, 2016). To teachers, these shifts appeared to be credible threats to pay and employment (Donaldson & Papay, 2015); in practice, consequences for performance were rarely applied (Aldeman, 2017).

Compared to the number of studies on test-based accountability, we are aware of only one study that attempts to assess the federal push for teacher evaluation reform nationwide (Bleiberg et al., 2021). The authors of that study use nationwide data on student performance (state and NAEP assessments from 2009 to 2018), high school graduation, and college enrollment to assess whether teacher evaluation reforms had an impact on achievement at the district level.[39] Exploiting variation in the timing of state adoption of evaluation reforms, they find no evidence that evaluation reform impacted any of those outcomes, even ruling out the possibility of small effects associated with evaluation reform.[40] As we describe in Section 6, however, there is evidence from district-level studies that some evaluation reforms significantly contributed to students' test achievement (e.g., Dee & Wyckoff, 2015). In the next two sections, we turn to evidence about two policies associated with outcomes-based accountability: school turnarounds and standards.

[36] Several studies (e.g., Davidson et al., 2015; Wei, 2010, 2012) have also assessed whether NCLB's subgroup reporting requirements (the size of student subgroups that require states to report about student performance) and state decisions about how to implement those requirements affect achievement. Although there is evidence that state subgroup reporting requirements do influence what schools report, there is little evidence that state thresholds for subgroup reporting are consistently related to student test achievement in subgroups.

[37] See Figlio and Loeb (2011), Dee and Jacob (2011), Burns and Strunk (2021), and Polikoff and Korn (2018).

[38] Specifically, the RTTT application guidance noted that a competitive state proposal would include teacher evaluations that used multiple measures (including student achievement growth), rate teachers on a scale that had multiple categories, and use the teacher ratings to inform high-stakes personnel decisions. Again, these teacher evaluation reforms were conceptually different from HQT because they relied on judgments of teacher performance as opposed to teacher credentials and compliance (see footnote 25).

[39] The statistical approach to assessing interventions based on changes over time in program adoption has changed significantly in recent years. This paper uses the more recently developed statistical approaches to assess the robustness of the empirical findings.

[40] Specifically, they have power to rule out effects as small as 1.5% of a standard deviation on test scores and changes in high school graduation and college enrollment larger than 1 percentage point.

4.3 School Turnarounds

A key aim of NCLB and subsequent federal policy was to drive improvement in schools with a track record of low performance. As already noted, NCLB tried to reach this aim by calling on states to sanction schools that missed their AYP targets. These sanctions ranged from giving families the option to transfer to a different school if their school missed AYP for two years in a row, to offering supplemental services in schools that missed AYP three years in a row, to whole-school interventions for schools that missed AYP six years in a row, including restructuring or closure. In practice, it appears that many schools in corrective action or restructuring simply carried forward actions they were engaged with prior to being identified for corrective action, rather than adopting new ones; indeed, the GAO estimated that 6% of schools identified for sanctions under NCLB took none of the required actions (Government Accountability Office, 2007).

In 2009, the Obama administration changed the consequences facing underperforming schools in several ways when it dramatically expanded the federal SIG Program.[41] Rather than focus on all the schools that failed to meet AYP, the SIG Program focused only on the lowest 5% of performers. The SIG Program required participating schools to use one of four approved interventions (described below), instead of applying a series of escalating sanctions triggered by performance thresholds. Most significantly, SIG included additional resources to support school improvement. Beginning with an initial investment of $3.5 billion, the federal government would eventually invest $7 billion in the program (Ginsburg & Smith, 2018). For individual schools, this could translate into 3-year grants of up to $2 million per year (Dee, 2012).

Like some of NCLB's sanctions, SIG's four intervention models called for changes in school personnel, programs, and oversight. In order of forcefulness, the intervention models included (a) the "transformation" model, which involved replacing the school leader, adopting new employee evaluation systems, and extending learning time; (b) the "turnaround" model, which included many of the same requirements as the transformation model but added staff reconstitution (i.e., replacing teaching staff); (c) the "restart" model, which required the district to hand over the school's operation to a contractor (e.g., a charter school); and (d) the "closure" model, which required the district to close the school and transfer its students to another, higher performing school (GAO, 2012). Approximately three fourths of SIG schools opted for school transformation, arguably the model that requires the least change (GAO, 2012; Ginsburg & Smith, 2018).

As Dee (2012) summarizes, prior empirical research about the prospects of school turnarounds like the ones envisioned by SIG was relatively thin at the time of the program's launch. Related evidence about the promise of other whole-school interventions, Dee notes, provided "relatively little encouragement that these [whole-school] initiatives can be effective at scale" (p. 11). True to form, the evaluation commissioned by the U.S. Department of Education famously found that the SIG Program did not achieve its goals for school improvement (Dragoset et al., 2017). Dragoset et al. (2017) concluded that the SIG intervention models "had no significant impacts on math or reading test scores, high school graduation, or college enrollment" (p. ES3). Researchers looking at school turnaround efforts in North Carolina reached similar discouraging conclusions (Heissel & Ladd, 2018). But methodological criticisms of the national evaluation (mainly having to do with sample size and selection) and results from other states and localities paint a more mixed picture (Ginsburg & Smith, 2018). Given the state of the literature, making strong claims about SIG results nationwide is fraught. In Section 6 we consider localized studies that suggest some turnaround interventions—especially more ambitious and well-funded ones—can lead to improvement.

[41] Such whole-school improvement efforts predated both NCLB and SIG. See Bloom et al., 2001.

4.5 Standards

Under NCLB, all states were required to adopt content standards in reading and math (and, eventually, science). Because NCLB left it to states to choose their own standards and tests, the content and quality of what states adopted varied widely (Carmichael et al., 2010; Porter et al., 2009). Partly in response to the inconsistency and shortcomings of state-developed standards under NCLB, the National Governors Association and the Council of Chief State School Officers launched the CCSS in 2009 to create "college- and career-ready" standards that could be shared across states (Polikoff, 2014). The CCSS not only aimed to raise the bar on standards;[42] they also had the potential to create economies of scale (e.g., on assessments and teaching materials) (*Why are the Common Core State standards important?* N.d.).

Under the Obama administration, the federal government encouraged the spread of the CCSS in several ways. First, as noted earlier, the RTTT program encouraged states to adopt the CCSS by rewarding grant applicants for taking steps to adopt "standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy" (U.S. Department of Education, 2009, p. 2). Second, RTTT supported CCSS by offering grants to create CCSS-aligned assessments. Eventually, these grants went to two test-development consortia: the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Third, the adoption of "college- and career-ready expectations for all students" was one of four requirements in the Obama administration's NCLB waiver process for states seeking flexibility on NCLB requirements (e.g., flexibility on the timeline for determining AYP and improvement actions in schools that fail to make AYP) (U.S Department of Education, 2012).

The government's push for states to adopt college- and career-ready standards, boosted by philanthropic support,[43] was effective (LaVenia et al., 2015). By 2011, all but five states had signed on to the CCSS and joined one or both assessment consortia (Jochim & McGuinn, 2016). Less than a decade after they launched, however, approximately half of the participating states withdrew from the assessment consortia (Jochim & McGuinn, 2016) or were considering it (Bidwell, 2014b). Subsequent accounts of the CCSS and its impact have generally focused on the politics that engulfed the effort (Bidwell, 2014a; Hess & McShane, 2018; Jochim & McGuinn, 2016; Loveless, 2018; Marchitello, 2014). Political dynamics notwithstanding, researchers who have tried to assess the impact of the CCSS on student achievement have also encountered a story fraught with challenges.

Setting aside descriptive survey research that captures district and teacher perceptions of CCSS implementation (Bay-Williams et al., 2016; Rentner & Kober, 2014) and perceptions of impact (Scholastic & the Bill & Melinda Gates Foundation, 2014), attempts to directly assess the impact of the CCSS on student achievement are few and far between. As Polikoff (2017) summarizes, part of the issue is that efforts to assess the impact of the standards face a host of methodological challenges. Researchers must grapple with when, exactly, the CCSS "treatment" started, as states rolled out the standards over several years. In addition, comparisons are challenging because most states adopted the standards in the same year, and the states that did not adopt them are a small group that likely differs systematically from those that did. Finally, the content on the default national assessment researchers use to assess the standards' impact nationwide (NAEP) is not perfectly aligned with the CCSS. All these things make it difficult to assess the causal effects of the CCSS on student learning.

Keeping these challenges in mind, the few studies that do attempt to look at the relationship between CCSS and student learning have found little evidence to suggest that adopting the new standards improved student learning. Descriptive studies suggest that states with stronger CCSS implementation do not perform better on the NAEP than weaker implementers (Loveless, 2015, 2016). A comprehensive assessment of standards from the Center on Standards, Alignment, Instruction, and Learning (C-SAIL), a multiyear research center funded by the U.S. Department of Education to assess the implementation and impact of college- and career-ready standards, reached similar conclusions. The two C-SAIL studies that look at the connection between standards and student learning found little evidence that adopting standards (Song et al., 2022) or standards-aligned content (Smith et al., 2021) improves student test scores (indeed, researchers found some evidence of small negative effects on test scores).[44]
A complementary study of CCSS's impact using NAEP found negative spillover effects on achievement in non-tested subjects, especially for students from historically marginalized groups (Arold & Shakeel, 2021).[45] An earlier study found a positive relationship between CCSS adoption and American College Testing (ACT) scores in Kentucky, but the researchers found that these positive effects preceded CCSS adoption, perhaps reflecting a ramp-up effect (Xu & Cepa, 2015).
More recently, a study found small positive effects in math for advantaged students associated with the initial adoption of the standards (Bleiberg, 2021).

[42] Analyses suggest that the CCSS were both more rigorous and coherent than the state standards they aimed to replace and that differences between the CCSS and existing standards varied by state (Polikoff, 2012; Schmidt & Houang, 2012).

[43] See Dillon (2009) and Layton (2014).

[44] As we noted earlier, the statistical approach to assessing interventions based on changes over time in program adoption has changed significantly in recent years. We are not sure whether the findings hold up to the recently developed model specification checks (e.g., Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021).

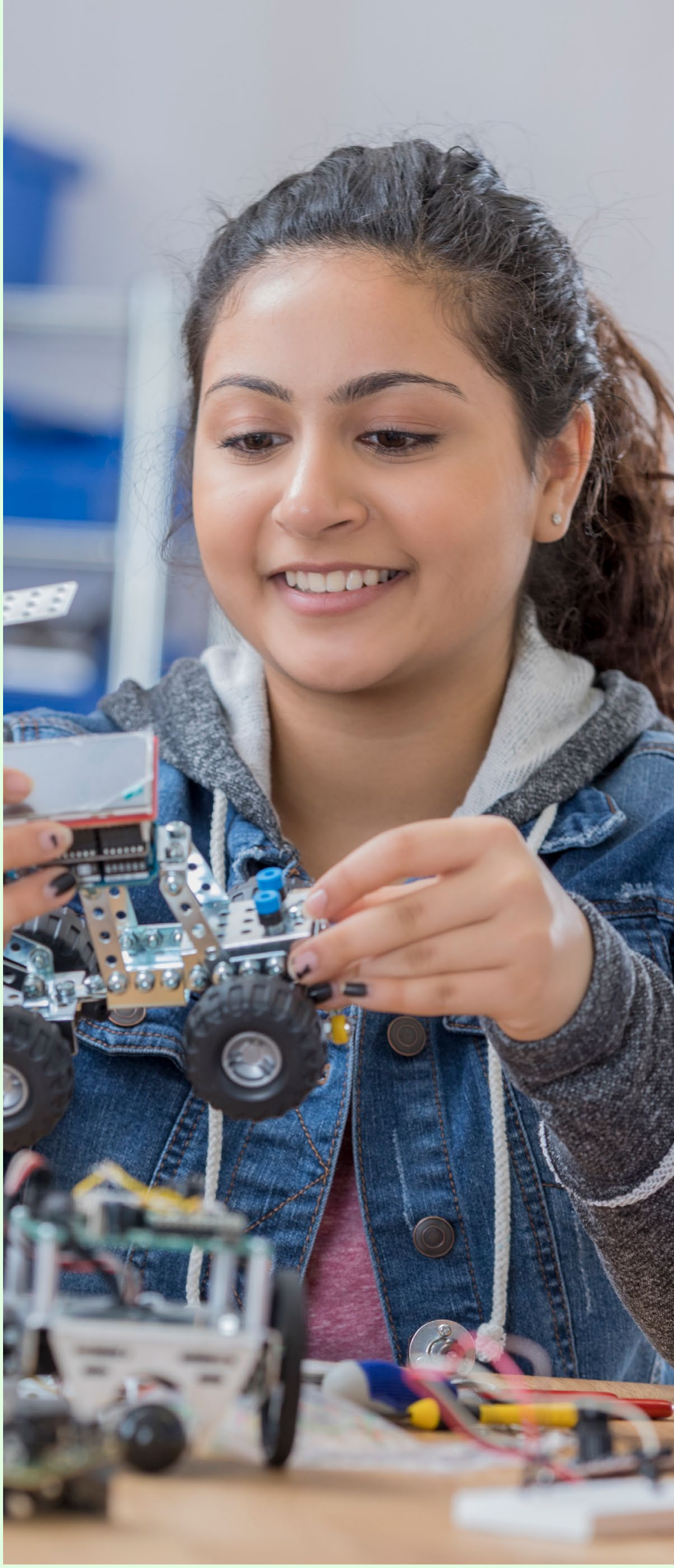
[45] Arold and Shakeel (2021) argue these negative effects were a function of the reduced instructional time teachers spent on non-tested subjects under the CCSS.

The latest paper from researchers at C-SAIL (Polikoff et al., 2022) concluded that, in the seven years since states adopted college- and career-ready standards, "the adoption of more rigorous standards did not improve student achievement overall in mathematics or ELA" (p. 3).[46]

4.6 Summing Up

On balance, (perhaps unsurprisingly) evidence about the nationwide impact of the federal push for outcomes-based accountability is mixed. Several studies suggest that NCLB improved math scores for younger students, particularly historically marginalized students. But the law did not meet its lofty goals at scale. A smaller body of evidence suggests that the translation of outcomes-based accountability to teacher evaluations had no national effects; the results of school turnarounds and standards are similarly dim nationwide. But as we note in Section 6, rigorous evidence from individual school districts provides a more mixed and, in some cases, hopeful picture; in other words, these ambitious reform ideas still hold promise even though scaling them nationwide was challenging.

The preceding sections suggest, by omission, some important gaps in the evidence. As we noted at the start of the paper, standardized tests do not tell us everything we need to know about schools or students. The research we reviewed remind us, for example, that we do not have many insights from rigorous studies about how, if at all, any of the era's outcomes-based policies impacted non-test outcomes, such as students' social and emotional wellbeing or civic participation. We also have less information about how these policies affected the distribution of resources and outcomes in non-tested grades and outcomes for students who were sometimes excluded from the accountability system because of minimum cell size requirements (students with disabilities, for example).[47]
In both cases, the problem is as much a lack of data as a lack of research. We return to the question of what we do not know but should in the conclusion. For now, the rest of this paper covers a range of issues that provide a more nuanced understanding of these mixed findings, beginning with some of the mechanisms that might explain why these reforms did (and did not) improve student outcomes.



[46] In the end, these results may reflect the reality that, like much in education, standards depend on a complex chain of logic that not only includes aligned assessments but also teacher understanding, instructional materials, and teacher instructional practice. These final, classroom-based steps in the causal chain between standards and student learning were often "zones of wishful thinking" in the reforms of the late 2000s and mid-2010s (Hill and Celio [1998] coined the useful phrase "zones of wishful thinking" to identify actions or conditions that policies need to succeed but do not address directly).

[47] Though see Henry et al, 2022 on non-tested early grades.

5. Mechanisms That May Translate Reform Into Changed Outcomes

As the logic of accountability policy described in Sections 1 and 4 suggests, the effects of key reforms in the 2000s and 2010s generally relied on a combination of standards, tests, information, and incentives to motivate schools to improve. But this logic does not spell out how, more specifically, schools and other stakeholders might respond to information and incentives to improve. Like other questions raised in this paper, elaborating the underlying mechanisms here is a challenging and nebulous problem. Although the various psychological mechanisms at play are important (Lerner & Tetlock, 1999), in this section we take a more modest approach and briefly explore three factors operating under the surface that could partly explain the results of outcomes-based accountability: money, teachers, and information.

5.1 Money

For a long time, the standard view about the relationship between increases in funding and student outcomes was mixed, if not pessimistic (e.g., Burtless, 1996; Hanushek, 2003).[48] Recent evidence provides a more optimistic view. Identifying the effects of spending on student outcomes based on spending increases tied to school finance reforms (often court mandated), researchers have found clear evidence that spending impacts outcomes. For example, increases in spending have been found to improve students' short-run NAEP test achievement as well as longer term outcomes, such as high school graduation, college-going and completion, the number of years of schooling completed, wages, and the likelihood of being in poverty as an adult (Candelaria & Shores, 2017; Hyman, 2017; Jackson et al., 2016; LaFortune et al., 2018).[49]

To the extent that accountability policies were accompanied by or prompted additional funding, it might help explain some of their impact. As it turns out, there is evidence that the reforms of the era were associated with increases in funding. Dee et al. (2013), for example, find that the implementation of NCLB led to an increase in spending of approximately $600 per pupil, on average across the U.S., primarily because of increased state and local spending. These increases generally translated into higher teacher compensation (we will return to this issue when we consider teachers as an improvement mechanism in Section 5.2).

In line with the government's interest in improving results for historically underserved students, federal funding to high-poverty school districts and schools also increased during this time. Chambers et al. (2009), for example, find that the highest-poverty districts and schools received increased Title I allocations under NCLB. However, in the schools with the highest poverty levels, these increases did not translate into more funding ***per student***. That is, in part, because the number of low-income students in these schools also increased during the same period. By contrast, they find that per-pupil funding for low-income students increased in schools with lower levels of poverty (where more funding was available for fewer students) (Chambers et al., 2009).

[48] How systems spend money and under what conditions matters. For example, a lot of research suggests benefits of class size reduction based on the Tennessee STAR experiment (e.g., Chetty et al., 2011; Krueger, 1999). But these reforms have not yielded the expected results when taken to scale (Chingos, 2013), possibly because cost constraints limited the size of class size reductions at scale or because of offsetting labor market consequences (such as the redistribution of teachers across student groups (Jepsen & Rivkin, 2009).

[49] See also Card and Payne (2002) for evidence showing that reductions in spending inequality reduce test score (the Scholastic Aptitude Test [SAT]) outcome gaps between students from families whose parents have more or fewer years of education and Jackson et al. (2021) showing that spending cuts connected to the Great Recession had negative impacts on student outcomes.

As a result, we are left with two conclusions about federal funding: first, funding increased during the early 2000s and was targeted to high poverty districts; second, funding ***per student*** in the highest-poverty schools did not appear to increase much, thanks to the growth and distribution of low-income students and how districts allocated funds across schools. It is worth noting that increases in funding (or lack thereof) matter not only because of their association with student outcomes; they matter because the bulk of funding in education is spent on human capital (approximately 80% of current K-12 public school expenditures were on employee salaries and benefits, and approximately two thirds of this is spent on instructional staff).[50] And, as we explain in the next section, human capital is perhaps the most powerful school-based lever we have for improving student outcomes.

5.2 Teachers

RTTT's push for teacher evaluation was predicated on the idea that teachers are the ultimate mechanism for student learning and that status quo policies were misaligned with and underinvested in promoting teacher quality.[51] A few bright spots aside (discussed more in Section 6), RTTT's evaluation reforms failed to live up to their promise and, in the bargain, fed a political backlash against the era's reform agenda. That backlash persisted even though fears that evaluation reforms would lead to a wave of unjust teacher dismissals (e.g., Berliner, 2013; Darling-Hammond et al., 2011, 2012) did not materialize in practice. Relatively few teachers, in the end, were rated as low performing (Kraft & Gilmour, 2017) or dismissed for poor performance, even under evaluation systems that are deemed high-stakes systems (Aldeman & Chuong, 2014; Dee & Wyckoff, 2017).[52]

Even with the general failure of teacher evaluation reform under RTTT, there is evidence that the kinds of information about teacher performance promoted by RTTT can affect personnel decisions in schools. Rockoff et al. (2012), for instance, conducted an experiment in which some principals were provided information about the value-added of their teachers (i.e., their contribution to student test progress) in their schools and others were not. They found that the principals who had access to value-added information changed their evaluations of teachers to align more closely with the value-added evidence on their teacher effectiveness. They also found that the value-added information led to more low-value-added teachers to leave their schools, which in turn led to higher student math test achievement (but not English test achievement) the subsequent year.

In a closely related study, Loeb et al. (2015) evaluated a change to teacher tenure in New York City that included information about teachers' value-added. Principals before and after the reform were asked to recommend teachers for tenure, denial of tenure, or a probationary period. The reform of the tenure system provided principals with information about teacher value-added and guidance about tenure recommendations. It also required principals to provide a rationale for any recommendations that differed from district guidance (e.g., district guidance was to support tenure for teachers found to be highly effective over the 2 previous years and deny it for those found to be ineffective in the 2 previous years). Loeb and colleagues found that the reform led to the extension of probationary periods for ineffective teachers who were later found to be more likely to leave their schools.

The larger point is that policies that reshape the teacher workforce have significant potential to impact student learning. Besides the intuition that teachers matter to student learning, we have a wealth of new information confirming that teachers are a key mechanism for change. The most striking empirical finding here is the extent to which teachers who appear to be similar on the surface (e.g., their degree or experience level) can have starkly different impacts on student learning (e.g., Aaronson et al., 2007; Rivkin et al., 2005). Researchers have used sophisticated techniques to show that a teacher's contribution to student test achievement (e.g., their value-added) is highly predictive of their students' future test scores (e.g., Goldhaber and Hansen, 2013; McCaffrey et al., 2009) and on adult outcomes (Chetty et al., 2014b).[53] Although there are differences in the average value-added of teachers across schools (Sass et al., 2012)—teachers in higher poverty schools tend to have lower value-added scores (Goldhaber et al., 2015, 2018)—researchers consistently find that most of the variation in teachers' value-added scores occurs within schools (Koedel et al., 2015).[54]

[50] Authors' calculations based on Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "National Public Education Financial Survey," 2009–2010 and 2018–2019. See Digest of Education Statistics, 2021, Table 236.60.

[51] For example, the teacher characteristics used to determine compensation (e.g., degree level) or used under federal law to determine teacher qualifications (certification under NCLB's HQT provision) are not aligned with other measures of teacher quality. There is some evidence that teachers having a degree in their subject predicts their performance; however, most rigorous studies find limited evidence that advanced degrees matter generally. And although the certification status of teachers has been found to be predictive of teacher performance in some states (note that what is required for certification is determined at the state level and hence varies across states), certification or route into the classroom is not generally predictive either (e.g., Boyd et al., 2007; Goldhaber, 2019; Goldhaber & Brewer, 1997, 2000).

[52] As a concrete example, Dee and Wyckoff (2017) report that approximately 4% of Washington, DC, teachers were dismissed under IMPACT, the district's high-profile and high-stakes teacher evaluation system; the effects of IMPACT are discussed more extensively in Section 6.

[53] Although questions have been raised about whether value-added estimates are unfair measures of teacher contributions ("biased") (e.g., Rothstein, 2009), both experimental (e.g., Bacher-Hicks, 2014) and quasi-experimental (e.g., Chetty et al., 2014a) tests of value-added suggest that bias, should it exist, is quite small. See Koedel et al. (2015) for a review.

[54] There is also evidence that measures other than just value-added, are inequitably distributed across students with higher poverty and students of color less likely to be taught by higher quality teachers (e.g., Goldhaber et al., 2015; Lankford et al., 2002). For estimates of the consequences of inequitable distribution of teachers across student subgroups, see Goldhaber et al. (2022).

The quality differences captured by a teacher's valued-added scores can have serious consequences for students. Estimates vary from study to study, but a change in teacher value-added effectiveness of 1 standard deviation—that is, the difference between having an average teacher and a teacher at the 84th percentile of the distribution—is generally estimated to increase students' test score achievement by .10 to .30 standard deviations. At the upper end, these effects are the equivalent of approximately 5 months of typical student learning (Goldhaber & Startz, 2017), which is significantly larger than the effects associated with reducing class sizes by 10 students (Rivkin et al., 2005).[55] Recent research also suggests that teachers significantly vary across a host of non-test outcomes as well. Teachers vary by their effects on student absences, suspensions, grades, and how many students advance to the next grade (e.g., Backes & Hansen, 2018; Jackson, 2018; Kraft, 2019). Teacher contributions to these non-test outcomes and their value-added on tests tends to be positively but weakly correlated, suggesting teachers frequently excel in some dimensions of their job more than others.

A second important finding in the literature is that teacher quality is at least somewhat malleable, and therefore efforts to improve teacher and teaching quality make sense.[56] For example, a considerable amount of evidence shows that teacher performance improves with experience (e.g., King-Rice, 2013; Rockoff, 2004). It is also clear that the pace of teacher improvement varies across schools (Kraft & Papay, 2014). Given evidence that teachers learn from one another, that improvement can be encouraged through carefully designed feedback systems (Jackson & Bruegmann, 2009; Papay et al., 2020), and that principals play an important role in creating environments conducive to professional growth, it is unsurprising that different school contexts are associated with different rates of improvement.[57] The extent to which outcomes-based policies were able to leverage teachers for improvement, either through development or staffing, may help explain what they were able (or unable) to accomplish for student learning.

[55] Estimates of the impact of teacher effectiveness on students' adult outcomes allow for the quantification of the dollar value of having more effective teachers. Specifically, Chetty et al. (2014b) connected estimates of teacher value-added to earnings information (from the internal review service) of those students at age 28 and find that a 1 standard deviation increase in teacher value-added (in a single grade) is estimated to increase annual earnings by 1.3%. This may sound like a small effect, but over the course of a student's life, this is estimated to amount to approximately $39,000. The economic consequences of this are enormous given that teachers instruct numerous students: Chetty et al. estimate that the economic value to future student earnings of replacing a teacher at the bottom of the value-added distribution (bottom 5%) with an average teacher on a classroom of students is between $185,000 and $250,000 (depending on the precise assumptions). This is the present discounted value of the impact of higher value-added on students' lifetime earnings.

[56] Although teachers can improve, it is worth noting that the best predictor of out-year performance of teachers (as measured by value-added) is their first-year performance. Importantly, first-year teachers who are low performers are unlikely to catch up with higher performing peers. In both math and ELA, teachers in the highest quintile category of effectiveness in their first year in the classroom are more effective than teachers in any of the lower quintiles who have 4 years of teaching experience (Atteberry et al., 2015). So, in addition to improvement on the job, better teacher hiring and evaluations like those envisioned by RTTT remain high-leverage policy tools. On hiring, contrary to perennial media reports of teacher shortages, far more people are prepared to teach each year than there are available teaching slots (Cowan et al., 2016). School systems generally have a fair amount of choice over which teacher applicants are selected for teaching jobs (e.g., James et al., 2022), although this varies considerably across school systems and subjects (Dee & Goldhaber, 2017). Importantly, a relatively new body of evidence shows that districts can influence the quality of their workforces through their applicant screening processes (Bruno & Strunk, 2019; Goldhaber et al., 2017; Jacob et al., 2018; Sajjadiani et al., 2019).

[57] Unfortunately, the most ubiquitous strategy for trying to improve the performance of teachers—professional development—has been shown to have limited effects, at least when measured by large-scale rigorous studies that assess teachers' impacts on students (e.g., Garet et al., 2011; Glazerman et al., 2010; Jacob et al., 2017); see Hill et al. (2013) for a review. For a more optimistic take on the value of professional development, in particular evidence that smaller scale interventions and teacher coaching have positive effects, see reviews by Kraft et al. (2018) and Lynch et al. (2019). The results are similarly disappointing for far less frequently used strategies such as pay incentive systems that tie teacher bonus pay to student test achievement (e.g., Glazerman & Seifullah, 2012; Marsh et al., 2011; Springer et al., 2010). For a review of evidence suggesting that pay for performance can positively impact student achievement, see Pham et al. (2021).

5.3 Information

So far, our discussion of information has focused on the use of outcome data by education leaders to hold educators accountable (for example by identifying schools for SIG or making decisions about teacher development and employment). And, as we noted above in reference to “bubble kids” (those on the margin of reaching state proficiency), there is evidence that the information inherent in tests did affect the distribution of student achievement. But the required disclosure of information associated with the era’s reforms may also have impacted school improvement in other ways, such as by prompting families to pressure public officials for improvements (e.g., via advocacy or voting behavior) or by exiting low-quality schools.[58]

A range of studies suggest that people outside of schools responded to performance information during this period in different ways. For example, researchers have found that school performance ratings influence voting behavior and exiting failing schools (Holbein & Hassell, 2018), families’ satisfaction with their schools (Jacobsen et al., 2013), donations families make to schools (Figlio & Kenny, 2009), and even housing prices (Black & Machin, 2011; Collins & Kaplan, 2022; Figlio & Lucas, 2004).

Indeed, the evidence is quite clear that the public provision of information about school performance matters to constituencies outside of schools. Figlio and Lucas (2004), for example, study the effect of school report card ratings in Florida on housing prices and find that the housing market values an initial “A” rating at 19.5% more than a “B” rating (they find that these effects attenuate with subsequent ratings).[59] On voting, Holbein and Hassell (2018) find that voter turnout in school board elections in North Carolina increased by 3 percentage points for Black and White families when their school failed to make AYP under NCLB. Elsewhere, Hastings and Weinstein (2007) find that 16% of families in the Charlotte-Mecklenburg School District who received a letter indicating their school was underperforming under NCLB took advantage of the law’s choice provision by moving to a higher performing school (specifically one that performed 1 standard deviation higher than their prior school).

However, other evidence on school choice finds that families consider performance information in combination with other types of information, including school location and demographics, when assessing school desirability (Schneider & Buckley, 2002).

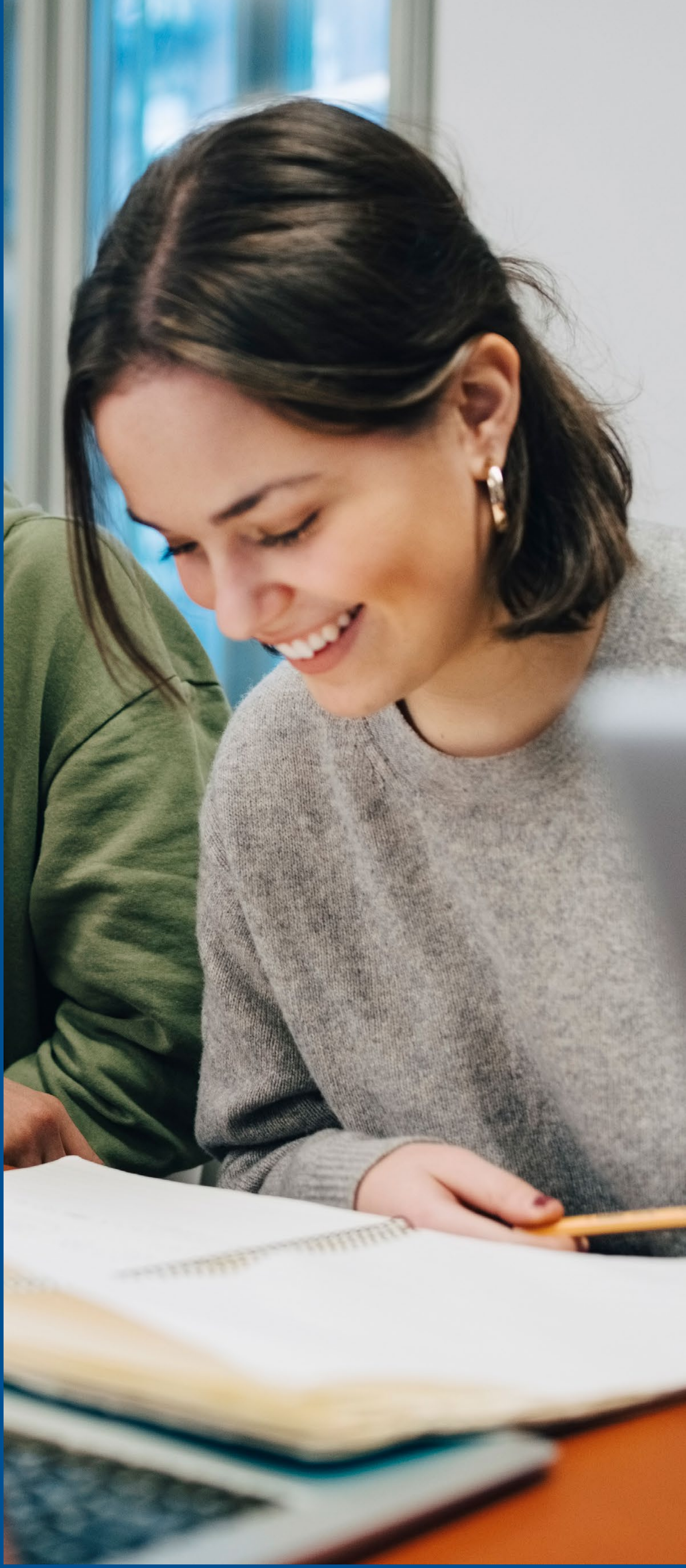
Importantly, studies also reveal that the ability of families to act on information, especially when it comes to school choice, varies. Hastings and Weinstein (2007), for example, note that whether families respond to performance information by moving to a higher performing school depends, in large part, on the availability of better performing schools (also see Denice & Gross, 2016). Hastings and Weinstein (2007) also find that families are responsive to simplified information about school performance (e.g., a single-sheet summary). The clarity of information affects its impact. And although Holbein and Hassel (2018) find that Black and White families are both more likely to vote in school board elections, they also find that White families are more likely to exit failing schools than Black families, who may have fewer options. In each of these cases, the point is that although families respond to information, using information depends on a host of other factors, including family resources (e.g., time, transportation, social networks), available options, and other information.

[58] In addition, the required disclosure of information about different subgroups of students under NCLB highlighted inequities in the system in new and powerful ways (Jacob, 2017) and, as we have noted a few times, prompted a new wave of education research (Fixing No Child Left Behind, 2015).

[59] As for within school responses, West and Peterson’s (2006) study of Florida’s report card ratings find that the stigma of low ratings was associated with increased test scores in the year after the schools received “F” grades. They contrast Florida’s accountability regime with NCLB's and conclude “An accountability system [like NCLB] that identifies problems with many schools, while giving few sanctions or incentives to improve, appears unlikely to be much consequence [for student performance].” (p. C57)

Given that families and teachers appear to respond to information, there is one other point worth making about the performance information used in the 2000s and 2010s and its potential as an accountability mechanism. Although the performance information disclosed under NCLB (e.g., AYP and proficiency rates) provided signals about how well students were doing in school, the consensus is that it did not provide good information about how much schools *contributed* to these outcomes.[60] As Brian Gill (2021) explains:

> A measure can be diagnostic for one purpose and non-diagnostic for another. For example, a low rate of proficiency in grade 3 reading suggests that students need additional support to read proficiently. It does not necessarily mean the school is underperforming in serving its students, because they might be learning rapidly from a very low starting point. Conversely, a high rate of proficiency does not necessarily mean a school is enhancing students' learning, if they started out as high performing. Assessing whether a school is underperforming requires isolating its contribution from factors outside its control, thereby assessing whether students would do better if they were at a different school.



We return to this problem—and the underlying issue of information use by different stakeholders for different purposes—in the conclusion. For now, on balance, we can only speculate that the effects of accountability reforms were likely mediated by the provision of extra resources, the increased attention paid to teacher quality, and the increased availability of public performance information. As we note in the next section, an arguably more powerful factor influencing how these reforms played out was the decisions that the local actors made while implementing them.

[60] In 2016, one of the authors (Goldhaber) co-signed a letter written by Morgan Polikoff to the U.S. Department of Education arguing against the use of proficiency rates to measure school performance. See https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/

6. Implementation Failures and Bright Spots

Any understanding of the effects of the NCLB-RTTT era reforms must consider the varied implementation of the reforms across states and school districts. NCLB may have marked an expansion of federal influence on education policy, but as we noted earlier, the law's ultimate implementation depended on decisions by leaders in states and local school districts (Manna, 2010). For example, NCLB left it to states to choose their own tests and identify the scores that would indicate "proficiency" on those tests (Le Floch et al., 2007). Likewise, NCLB deferred to states on the passing scores teachers needed to demonstrate subject matter knowledge under the law's Highly Qualified Teacher (HQT) provisions as well as how much credit to give teachers for prior experience (Birman et al., 2009). Of course, states were not completely on their own; the federal government issued a range of regulations and non-regulatory guidance to support the law's implementation.[61] But given the decentralized educational governance system in the United States, the overall effects of the initiatives nevertheless reflect a range of state and local implementation decisions.

With that in mind, this section takes a brief look at how some local implementation decisions influenced policy impacts.[62] By implementation decisions, we mean decisions made at the organization level (e.g., where to set cut scores) rather than the day-to-day implementation decisions made by frontline workers. Frontline behaviors clearly affect how policy gets translated into practice (Lipsky, 1980) but are beyond the scope of this paper.[63] While federal policy laid the foundation for outcomes-based accountability with its requirements for testing and consequences for performance, states and districts determined what these requirements and related actions looked like for teachers, students, and schools.

For example, it is well known that state-level decisions about tests and accountability led to wide variation across states in the early implementation of NCLB. Davidson et al. (2015), for example, examined how decisions by states about technical issues, such as the group sizes that made schools accountable for a subgroup's performance, where to set proficiency targets, and how to define "continuous enrollment," could influence school performance ratings in arbitrary ways. Davidson et al. (2015) find that these idiosyncratic decisions by states meant that AYP determinations were not strongly related to school proficiency rates (e.g., in some cases, the use of generous confidence intervals in safe harbor calculations meant that schools could make AYP even as their proficiency rates declined from year to year). In another national study, Reback et al. (2014) identified cases where schools that made AYP in their home state would be unlikely to make it in other states based on the same performance because of differences in how states calculated performance. State discretion on standards under NCLB also produced wide variation in what students were expected to know and be able to do (Carmichael et al., 2010; Porter et al., 2009), motivating, in part, the move toward the CCSS.

As these examples suggest, states and localities made decisions under NCLB that affected the impact of these reforms. Rather than review how such decisions impacted all the reforms reviewed in Section 4, we explore these dynamics further by looking at the implementation of two reforms—teacher evaluation and school turnarounds—to highlight how differences in implementation can produce different results.

[61] The U.S. Department of Education issued detailed guidance on the use of state assessment systems in schools and districts and a range of other topics (For example, see U.S. Department of Education, 2003, 2007).

[62] In another domain—prevention and health programs for children and adolescences—authors of a seminal review (Durlak & DuPre, 2008) underscore the importance of implementation, noting: "Data from nearly 500 studies evaluated in five meta-analyses indicates that the magnitude of mean effect sizes are at least two to three times higher when programs are carefully implemented and free from serious implementation problems than when these circumstances are not present" (p. 340). The importance of implementation has also been documented when it comes to education-specific interventions (e.g., Lendrum & Humphrey, 2012).

[63] Still, it is worth noting that for some of the initiatives we discuss in this section, key design details are determined by state-level policymakers, but others, such as teacher evaluation, involve a second layer of decentralization: As we suggest in our discussion of the Intensive Partnership Initiative (IPI) and DC's IMPACT reforms, teacher evaluation processes can look quite different across districts (e.g., Cowan et al., 2022; Jackson & Cowan, 2018).

6.1 A Tale of Two Teacher Evaluation Reforms

Although some readers will not associate teacher evaluation with the marque reforms of the era, teacher evaluation is a good case to explore how implementation decisions can lead to varied effects. To begin, teacher evaluation has the potential to be a high-leverage reform: Given the central role teachers play in promoting student learning (see Section 5.2), policies designed to increase teacher quality are a potentially powerful improvement mechanism. Second, teacher evaluation reform is an interesting case because rigorous research suggests that the results of these reforms varied considerably across different contexts. As we noted in Section 4.2, these reforms were deemed a failure at the national level (Bleiberg et al., 2021). But there is also good evidence that they succeed in some cases (Dee & Wyckoff, 2015). Indeed, Bleiberg et al. (2021) find that evaluation reforms in a small set of districts did significantly raise student achievement even though the national results showed no effect.[64] To make our discussion more concrete, we focus on two cases of evaluation reform: the Intensive Partnership Initiative (IPI) funded by the Bill & Melinda Gates Foundation (BMGF) and the IMPACT evaluation reform in the District of Columbia Public Schools (DCPS).

Launched in 2009, the BMGF's IPI was intended to change how three school districts and four charter management organizations (CMOs) evaluated teachers. The initiative sought to provide these systems with high-quality measures of teacher effectiveness that they could then use to inform a range of decisions, including decisions about the provision of targeted professional development, rewards, and career advancement, and, in some cases, employee dismissals (for more information, see Stecher et al., 2018). Ultimately, BMGF invested $215 million in the initiative on top of federal and local funding associated with the reforms (Kane, 2018).

Launched in the same year as the IPI, DCPS's IMPACT system was a high-profile, high-stakes teacher evaluation reform. IMPACT is arguably the longest-standing and best-studied teacher evaluation system in the nation (Toch, 2020).[65] Under the reform, teachers are evaluated based on classroom observations, value-added measures of student learning, and a principal's assessment of teachers' contributions to the school community (the weights of these components depend on the year and availability of data on student learning). Teachers who receive the lowest possible rating— "very ineffective"—are dismissed immediately. This turns out to be a very small group of teachers. Those who are rated "minimally effective" have 2 years to improve or face dismissal. Teachers rated "highly effective" receive a bonus; teachers rated "highly effective" for 2 or more consecutive years receive large increases in base pay (Dee & Wyckoff, 2017).

[64] Based on prior evidence and reviews of system-level evaluation, Bleiberg et al. (2021) focused their investigation of bright spots on the Dallas Independent School District, Denver Public Schools, the DCPS, Newark Public Schools, and the states of Tennessee and New Mexico.

[65] IMPACT has evolved somewhat (e.g., in terms of how different information about performance informs teacher summative evaluations), but the main components of the evaluation and the stakes of evaluation ratings have remained intact for more than a decade. For more on specific changes, see Dee et al. (2021).

In many ways, IPI and IMPACT embraced the same logic of change: If schools have and use better information about teacher effectiveness to inform decisions about retaining and rewarding teachers, teacher quality will increase, and students will benefit. But the two efforts played out in ways that produced different results.

6.1.1 What Were the Results?

An analysis of IPI by Stecher et al. (2018) found that participating systems followed some aspects of the reform (e.g., changing the frequency of evaluations and the types of information that fed into evaluations). Participating systems also reported using their new evaluation data to inform human resource-related decisions about teachers (e.g., type of professional development, compensation, dismissal). Nevertheless, Stecher et al. (2018) found little evidence that the IPI led to changes in student tests or graduation rates compared to similar systems that did not participate in the initiative.

Evaluations of IMPACT show that DCPS is using the reform as designed. For example, teachers identified as minimally effective and under threat of dismissal are far more likely (approximately 50%) to leave the district voluntarily or improve. DCPS teachers on the verge of receiving a large permanent salary increase as the result of a highly effective rating are more likely to increase their level of effectiveness in the next year. Research by Dee and Wyckoff (2015) suggests that IMPACT worked as intended, attributing these teacher behaviors to the financial and job-threat incentives embedded in the system.

IMPACT is not, however, without its critics. For instance, some argue that IMPACT has led to teacher attrition levels that are detrimental to student achievement (D.C. Board of Education, 2021, March 17); this is a legitimate concern, given that teacher turnover is generally found to have negative impacts on student test achievement (e.g., Ronfeldt et al., 2013). But, teacher turnover, particularly in schools serving historically marginalized students, can also have beneficial effects on student achievement (Adnot et al., 2017). In the case of DCPS, there is evidence that schools replaced departing teachers with more effective teachers (James and Wyckoff, 2020), thanks, in part, to improved human resource management practices (Jacob et al., 2018). Moreover, follow-up studies show that IMPACT continued to have positive impacts on the quality of teachers a decade after the new system was initially implemented (Dee et al., 2021). Evidence of the reform's positive impact is reflected not only in local assessments but also on the NAEP (Dotter et al., 2021).[66]

DCPS is not alone in successfully implementing evaluation reform. Taylor and Tyler (2012) studied reforms to the evaluation system in Cincinnati Public Schools and found that teachers were more effective, as measured by their value-added impact on students, during the year they received an evaluation. Cincinnati's teachers were also found to be more effective in the years after their evaluation, suggesting that the feedback they received about their performance had longer lasting impacts on their productivity. Elsewhere, studies of evaluation reform in Chicago Public Schools show student achievement benefits (Sartain & Steinberg, 2016, 2021; Steinberg & Sartain, 2015). The positive results in Chicago appear to be driven, at least in part, by a significant increase in the exit of low-performing teachers who were replaced by significantly better teachers.[67]

6.1.2 Why Did Results Differ?

To be clear, no single factor explains IPI's failure or IMPACT's success. Like all complex policies, the implementation of these initiatives depended on dynamic interactions between the policies, the people charged with implementing them, and the places where they occurred (Honig, 2006). There are no credible quantitative studies akin to the research reviewed elsewhere in this paper that can untangle these factors. Still, what we know from these cases suggests some factors that have important implications for understanding accountability reforms and their prospects for success.

Like everything in education, both cases of evaluation reform were embedded in larger systems. Accordingly, their successful implementation depended in part on how well the reforms were integrated into related policies and routines in the broader system. In the case of IPI, there is evidence that key parts of the reform that called for systems-level integration did not happen. Stecher et al. (2018), for instance, report that IPI sites struggled to connect evaluation reforms to decisions about professional development. As a result, professional development was unaligned with teachers' individual needs that were identified in their evaluations.[68] More broadly, the lack of alignment between teachers' needs and what they receive during professional development may explain the mixed record of professional development more generally (see Footnote 54). Even more important, although all IPI teachers received evaluations, their evaluations often did not differentiate between teachers based on performance (e.g., very few IPI teachers were judged ineffective). Finally, the IPI sites provided weak financial incentives by providing small performance bonuses to large proportions of teachers.[69]

[66] Dotter et al. (2021) find that reforms in Washington, DC, were associated with large (approximately one third of a standard deviation in math) significantly increased NAEP achievement in both math and reading in Grade 4 and math in Grade 8 (the Grade 8 reading results were generally positive, but statistically insignificant, in the years in which the district is seen as having implemented reforms).

[67] Cullen et al. (2021) also find evidence (based on an assessment of teacher evaluation reform in the Houston Independent School District) that evaluation can induce the retention of highly effective teachers and the attrition of very ineffective teachers, but the authors conclude that the magnitude of these effects on the composition of the teacher workforce in Houston is too small to be able to detect impacts on student achievement. See also similar evidence about differential teacher retention connected to evaluation reform in Tennessee (Rodriguez et al., 2020).

[68] Principals in IPI sites reported making recommendations about professional development based on needs identified in evaluations, but teachers were not required to participate in suggested areas of training, and there was no monitoring of whether professional development received by teachers was aligned with identified needs.

[69] For instance, most sites adopted annual bonuses in the range of $500 to $3,000 for awarded teachers, and in many sites more than 90% of eligible teachers received a bonus (see Tables 7.1–7.4); across sites and years it was typical for more than 90% of teachers to be classified as effective or above (see Figure 3.1), and only approximately 1% of the teacher workforce was dismissed (in 2015–2016). As Stecher et al. (2018) sum up: "… the sites did not implement … the initiative as fully as the developers might have expected. For example, all teachers received TE [teacher effectiveness] ratings, but very few teachers were classified as ineffective; the sites struggled to deliver evaluation-linked PD [professional development]; they offered relatively small performance-based bonuses to relatively large proportions of eligible teachers; and although they created some specialized leadership roles, none created fully developed CLs [career ladders]." (pp. 487–488)

By contrast, IMPACT is well integrated into the broader system in DCPS and creates strong incentives for teachers. IMPACT is connected to numerous changes to human capital systems, from teacher recruitment and selection (Jacob et al., 2018), to feedback and support for struggling teachers (including professional development), to a focus on teacher retention (Toch, 2018, 2020). Compared to IPI, the IMPACT system resulted in far more differentiated evaluation ratings. It also offered far larger bonuses to teachers with highly effective ratings (these bonuses meant that DC's top teacher salaries went from $87,000 to $132,000) (Toch, 2017).[70] The size of these bonuses is important given evidence that teachers may need large increases in average compensation to offset the increased risk associated with high-stakes evaluation reform (Rothstein, 2009).

Beyond systems alignment, differentiated ratings, and strong incentives, IMPACT also highlights how politics and expectations can shape implementation success. IMPACT received sustained support under DC's mayoral control governance system that ultimately sustained the reform for multiple years. This longevity matters: IMPACT's positive effects only started to appear in its second year. Dee and Wyckoff (2017) speculate that the reform's delayed effects may have stemmed, in part, from the expectation among teachers that the system would abandon the controversial reform. But with continued support from the mayor and subsequent district leaders, the reform remained in place long enough to have an effect and, importantly, improve and change over time.

The fact that the national results (Bleiberg et al., 2021) are more like IPI is unsurprising. Because teacher evaluation reform addresses contentious issues such as pay and job security, it is controversial and requires strong political support and leadership to succeed.[71] As in the IPI sites, most states and districts that implemented evaluation reforms under RTTT did not use new teacher ratings to differentiate teachers by performance (Kraft & Gilmour, 2017; Walsh et al., 2017). As Bleiberg et al. (2021) summarize, "Despite the widespread adoption of teacher evaluation reforms, many states designed evaluation systems that only vaguely resembled the systems most reformers envisioned" (p. 25).

The contrasting results of IPI and IMPACT suggest that the failure of evaluation reform was a failure of implementation rather than a failure of theory. When making this distinction, we should ask whether the reforms were adequately implemented; whether there was enough engagement, uptake, and adherence to the design; whether intermediate outcomes were achieved; and whether final outcomes were achieved (Funnell & Rogers, 2011). In the case of IMPACT, the affirmative answers to these questions suggest an outcome consistent with the reform's underlying theory. In the case of IPI, shortcomings related to the first two questions suggest an implementation breakdown: The reforms were partly implemented (e.g., evaluations happened but were not linked to other systems), and adherence was uneven (e.g., evaluators did not use ratings to differentiate teachers by performance).

6.2 The Challenge of Turning Around Low Performing Schools

Research on school turnarounds unsurprisingly also includes stories of failure and success. As already noted, the national evaluation of the SIG program (Dragoset et al., 2017) concluded that the turnaround program nationwide did not improve student achievement. Other studies have reached similar conclusions (Heissel & Ladd, 2016), including some that argue turnaround efforts are not only ineffective but that they harm students (Trujillo, 2012). Other studies, however, point to more positive results (Schueler et al, 2017; Strunk et al 2016; Zimmer et al, 2016).[72] A rigorous look at turnaround efforts in two states illustrates the range of outcomes and offers some speculation about the implementation dynamics behind them (Dee, 2012; Doughery & Weiner, 2019).

In 2010, California awarded SIG grants to around 90 low performing schools to adopt one of the SIG program's turnaround strategies described in Section 4.3 (transformation, turnaround, restart, and closure). On average, schools received $1.5 million dollars. Most grant recipients (60%) adopted the transformation model. As we described earlier, this model involved replacing the school leader and introducing a host of reforms, including teacher evaluations, data-driven instructional strategies, and other initiatives. Around a third of the schools chose a more aggressive turnaround strategy, which required them to replace at least half of the school's teaching staff.[73] These interventions assumed that major changes to personnel and programming were necessary to improve low performing schools.

[70] In the first 2 years of IMPACT (2009–2010 and 2010–2011), the distribution of ratings was as follows: 14% of teachers were "highly effective," 69% were "effective," 14% were "minimally effective," and 2% were "ineffective." And teachers receiving one "highly effective" rating could receive a bonus of up to $25,000 (bonuses were larger for being effective and serving in a high-poverty school); those receiving consecutive "highly effective" ratings could increase their base pay by up to $27,000.

[71] Others have argued that state policy could have done more to ensure successful implementation. Walsh et al. (2017), for example, argue that the results of evaluation reforms in general may have been different if states had provided more guidance and/or rules about the components that feed into evaluation systems; states could have required, for example, that teachers deemed to have contributed very little to student learning (based on value-added measures) could not earn ratings that put them in effective or higher categories of performance.

[72] Redding and Nguyen (2020) conduct a meta-analysis of school turnaround studies and found turnarounds to be associated with improved attendance, standardized test scores, and graduation rates. They found little evidence, however, to suggest one approach to turnaround (e.g., replacing the school leader) produced better results than another (e.g., replacing a portion of the teaching staff).

[73] California is an interesting case here, in part because it had more SIG-eligible schools and SIG-awarded schools than any other state.

In 2012, Rhode Island was awarded funds through RTTT to support school improvement (see Section 4 for an overview of RTTT). Using the NCLB waiver that accompanied its RTTT award, Rhode Island created a system for ranking schools by performance and then required schools in the lowest three tiers of performance to implement a range of interventions. Per RTTT, all these schools were required to adopt college and career-ready standards, teacher evaluation reforms, and data-driven decision-making. Rhode Island also curated a menu of other approved interventions to support the schools' turnaround efforts. Unlike the SIG grants in California, Rhode Island required the lowest of its lower-performing schools to implement more interventions from the menu than other low-performing schools; the worse a school performed, the stronger the dose of interventions it received.

6.2.1 What Were the Results?

When Dee (2012) examined the impact of SIG awards in 82 low-performing schools in California, he found positive effects on performance as measured by California's Academic Performance Index (API), a school-level measure based on state testing. Other studies of the SIG program in individual California cities reach similar conclusions. Sun, Penner, and Loeb (2017), for example, also find positive effects on student achievement after 3 years of turnaround efforts in San Francisco's public schools. Strunk et al. (2016) also find positive student achievement results in English Language Arts (ELA) in SIG schools in Los Angeles (although not in math). Most recently, Sun et al. (2021) examined the longer run results for two cohorts of SIG schools in four different locations, including California (Washington State, North Carolina, San Francisco, and an anonymous school district). Their results suggest that turnaround schools saw gradual increases in math and reading achievement that sustained after the grants ended.[74] Rhode Island's results were less encouraging. Dougherty and Weiner (2019) find that the state's lowest performing schools that adopted the most interventions were worse off than comparable schools that had adopted fewer interventions; schools that adopted fewer interventions were also no better off than comparable schools that were not part of the initiative.[75]

[74] Positive findings about SIG and student achievement also emerged in studies of Ohio (Carlson & Lavertu, 2018) and Massachusetts (LiCalsi et al., 2015).

[75] It is worth noting that in both states, the turnaround efforts may have had other, unintended, consequences (e.g., impacts on staffing in other schools that are not receiving turnaround interventions in response to teachers leaving or joining a school receiving turnaround interventions).

6.2.2 Why Did the Results Differ?

Given the nature of their research questions and design, neither the California nor Rhode Island study provides direct evidence on how the implementation of turnarounds might explain their different results. Since the turnaround efforts in both states encompassed a range of interventions, it could be that the results were driven by different turnaround *models*. Interestingly, the positive effects in Dee's study of California (2012) appear to be driven by schools that adopted the turnaround model (which involved reconstituting part of the school's teaching staff). In their study of Los Angeles, Strunk et al. (2016) also find that positive results are largely driven by schools that reconstituted their teaching staff (as opposed to replacing only their school leader as in the transformation model). However, Kyse et al. (2014) offer a counter example in New Jersey, where less forceful turnaround models saw better results than those that called for new teaching staff.

In a useful and extended discussion, Doughtery and Weiner (2019) speculate that Rhode Island's disappointing results may reflect the state's lack of history with (and capacity for) supporting and overseeing turnaround efforts. Given the complexity of implementing school turnarounds, state and district capacity (and perhaps stability) may be critical supports for implementation. Echoing our earlier comments about IMPACT, Doughtery and Weiner suggest that the alignment and coherence of intervention strategies (or lack thereof) in Rhode Island schools may have affected the odds of improvement. Given that Rhode Island closely followed federal guidance on interventions, the authors speculate that,

In the end, these brief examples underscore the point that the implementation of the complex policies set in motion by the federal government in the 2000s and 2010s depended on a host of important decisions at the local level. Some of these decisions were technical—where to set cut scores, for example. But others were managerial (Are different elements aligned and mutually reinforcing? Do oversight agencies have the capacity to support schools?). The examples of teacher evaluation and school turnaround highlight the importance of coherence and beg questions about how the people, policies, and contexts in which reforms unfold support or inhibit coherence (Honig, 2006). Although the federal government's reach over local coherence is limited, some have argued that more consistent accounting rules and transparency could help states improve technical decisions with meaningful consequences for the application of rewards and sanctions for schools and teachers (Davidson et al., 2015).

7. Where to From Here?

The studies we reviewed in this paper capture rigorous evidence about the impact of federal initiatives on student achievement. But as our review suggests, clearly *causal* evidence on large-scale initiatives is hard to come by. And while the research we reviewed may be rigorous, it also carries its own assumptions and limitations; the types of evidence we reviewed are also only one factor among many that leaders consider as they chart a path forward. With all these caveats in mind, we want to close by underscoring four takeaways from our review and offering some thoughts about the future.

7.1 Four Takeaways

First, when it comes to student achievement, our review suggests that NCLB improved math outcomes, especially for younger students and students from historically marginalized groups. At the same time, we found that test-based accountability fell short of its lofty (rhetorical) ambitions to improve outcomes for all students in *all* schools. This is true of efforts to scale up teacher evaluation, standards, and school turnarounds as well. At the same time, it is clear that the outcomes-based policies of the 2000s and 2010s reverberated throughout the system in ways that do not always show up in test-scores; many of the era's impacts are hard, if not impossible, to measure quantitatively. Ironically, the era's focus on outcomes may have fueled push-back against the federal government's influence and use of testing by bringing the shortcomings of the system into high relief.

Second, nationwide judgments about federal education policy necessarily gloss over a tremendous amount of variation. Education policy and practice remain dominated by states and local school districts, even under more muscular federal initiatives. Federal policymakers can set priorities and requirements but translating them into reality depends on a wide range of actors working in situations that belie easy generalization. That some reforms succeeded in some places suggests we need to learn more about how the context and characteristics of implementing organizations moderate success, rather than make summary judgments about whether the reforms were a "good" or "bad" idea. This is especially important given that educational progress is, in our judgment, likely to depend on incremental and steady progress rather than on a moonshot reform or silver bullet policy.

Third, as we conducted our review, we were struck by how federal policy in the 2000s and 2010s tended to overlook some important components of the education system, especially school districts and teacher preparation programs. Both institutions clearly play mediating roles when it comes to school improvement, but both were largely untouched (at least directly) by federal K-12 initiatives.[76] This omission is puzzling. School districts play a critical role holding schools accountable for performance; these local institutions are where voters can directly act on information about what schools are doing and how students are learning. Meanwhile, the importance of effective teacher preparation seems axiomatic. That both these important institutions were largely left off the federal agenda looks, in retrospect, to have been an important oversight. They deserve more attention in policy and rigorous research going forward.

Fourth, NCLB drew attention to the performance of the system and of disadvantaged students in new and powerful ways. Indeed, as we conducted this review, we were reminded that the lack of systematic information about the academic performance of students prior to the NCLB era is one of the reasons that assessing the impact of reform in the 2000s and 2010s is so hard. As Fusarelli and Ayscue (2019) note,

[76] Note that under the Obama administration there was, briefly, an effort to hold teacher preparation institutions accountable for the placements and effectiveness of program graduates. But proposed regulations were rolled back before being implemented by the Trump administration. For more details, see Goldhaber and Brown (2016).

Thanks to the reforms of the 2000s and 2010s, we now live in an information rich education system and know far more about the impact—positive and negative—of schools and teachers on student achievement than ever before (As we noted earlier, economists using such data won a Nobel Prize in 2021 for their rigorous empirical research). But, as we wrote in the introduction, the backlash against testing and the lack of a clear policy agenda makes the path forward murky. As decision makers navigate that path, what do these findings suggest about what comes next? Below, we offer some speculative thoughts, beginning with some things that we wish we knew but do not.

7.2 A Few Thoughts on What We Need to Know Going Forward

Our review provides evidence about the impact of federal policies from NCLB to ESSA, but we also encountered numerous issues where the evidence about the impact of federally initiated reforms is limited. These knowledge gaps suggest several areas where we need to know more to prepare for the next wave of reform and improvement.

As we have emphasized earlier, test scores are important predictors of future academic and labor market success and can provide a gauge of student and school progress in the short run. But the field also needs a much better understanding of how schools influence student outcomes that go beyond math and reading test achievement. A new body of evidence is showing how schools contribute to non-test outcomes and how those outcomes relate to later life success (Gilraine & Pope, 2021; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2021; Petek & Pope, 2021), but this literature is still nascent and begs many questions about how different policy interventions influence non-test outcomes.

Our review also suggests an ongoing need to complement outcome-focused research with studies that deepen our understanding of how local organizations and contexts moderate the prospects of improvement and reform. There are countless opportunities to build knowledge about effective implementation practice. Researchers might, for example, learn about effective implementation by leveraging the school-level spending data now required under ESSA in combination with deep dives into district decision making to better understand how (or whether) additional federal resources reach the intended students and with what effects (Rosa & Anderson, 2020, April 32). And researchers might combine assessments of performance with deep dives into how schools use instructional time during the school day (Kraft & Novicoff, 2022). No matter the focus, understanding policy effects requires that we assess not only the changes that occur for students (e.g., increased learning or wellbeing) but also the changes that occur in how schools operate (Sandfort & Moulton, 2015).

Finally, our review also suggests that research should do more to situate schools in the broader contexts in which they operate. As we noted earlier in Section 2, the ultimate measure of K12 is what happens to students after they leave high school. So, we also need to continue to build our understanding of how students' experiences in school *and elsewhere* shape students' longer-run prospects. Thanks to the federal Statewide Longitudinal Data Systems Grant Program, states have made serious progress on this front and increased the amount and richness of information we have teachers and students and their pathways through the educational system and beyond.

In short, after reflecting on what is known and not known about the impact of accountability policies from the early 2000s, as well as the state of the field in the wake of ESSA, we see a need for a stronger evidence base that encompasses the full chain of logic associated with outcomes-based improvement efforts, capturing an expanded set of outcomes, the resources, contexts, and actors who contribute to those outcomes, and the longer-run results that arise from their efforts.

7.3 Some Thoughts on the Federal Role

Even as the future of education policy remains unclear, our takeaways from the review and the kinds of knowledge gaps we described in the prior section suggest that the federal government has an important role to play going forward regarding information and ideas.

First, the federal government should continue to use its influence to push states and districts to collect information that helps families and policymakers identify when schools are not helping students learn. ESSA's continuation of NCLB's testing requirement was important, but the government should also insist that states collect better performance information. At a minimum, the experience of the last 20 plus years suggests that the federal government should provide guidance and support to states on collecting and using measures of achievement growth and value-added measures, rather than proficiency alone, to gauge how effective schools are at supporting student learning. Growth-based test measures do a better job reflecting a school's contributions to its students' success and can help families and policymakers better understand how well schools are performing (Polikoff, 2017).

The federal government should also continue to insist that states collect common performance measures statewide. Having common measures helps policymakers and families understand school performance in context and assess the impact of different interventions and approaches (shared measures may also help schools meet the needs of students who move between school systems and help researchers assess the impacts of various policies across contexts). Indeed, the existence of shared measures is one of the main reasons we can gauge the impact of all the disruptions in schooling that occurred during the COVID-19 pandemic.

As we hinted at earlier, the federal government should also continue to encourage the use of measures that go beyond reading and math achievement. At the end of Section 4 we noted several opportunities to learn about and measure important, non-test outcomes. For example, the federal government could encourage and fund assessments of higher-order learning outcomes or achievement in subject areas besides math and reading. Other measures might assess a school's impact on long-range outcomes, from high school graduation to college enrollment (e.g., see Gross et al., 2021); other non-academic outcomes, such as student social and emotional wellbeing and civic participation, are already receiving increased attention that should continue. Indicators of educational equity are yet another set of measures worth exploring (National Academies of Sciences, Engineering, and Medicine, 2019). As Olson and Toch (2021) argue (also see Gill et al. 2016 and Bruno and Goldhaber, 2021), governments do not need to incorporate an expanded set of non-academic measures into formal accountability systems to create change (nor should they without strong claims to validity and reliability). Regardless of the measures, the federal government could provide guidance to states and districts on presenting information to ensure the intended audiences can use it—reporting performance in standard deviation units is unhelpful for most people, not just families.

Beyond providing guidance on the presentation of information, the federal government could also do more to clarify how different performance measures should (and should not) be used. For example, Scott Marion and Derek Briggs (2022, July 13) recently argued that the federal government could improve how states and districts use tests with a small change to the law. As they explain, ESEA currently requires state tests to provide accountability data and diagnostic data. But they write, this requirement feeds "the misconception that accountability tests can serve instructional purposes." As a result, tests often end up being longer than necessary for accountability purposes and teachers (and families) often receive the results of tests too late to be of instructional use. Marion and Briggs argue that if federal testing requirements focused on accountability, tests could be streamlined and make room for more useful assessments to inform instruction. They explain,

> Lifting the requirement for individual "diagnoses" could make space for other test designs that might be more effective for program and curriculum evaluation and inspire more ambitious teaching practices…States could fill the apparent gap…by supporting the development, selection, and use of resources such as modular interim assessments and formative tools that can more directly inform teachers and leaders about student performance in these specific domains.

Finally, the federal government could assess and/or encourage testing innovations, from efforts to reduce testing burdens (e.g., alternating year-over-year testing between grade spans and subject areas) to assessing younger students to guide earlier interventions. The federal government might, for example, incentivize policy innovations that allow schools that reach a performance threshold to opt-out of year-over-year testing and instead use less burdensome approaches (e.g., testing only a sample of students year-over-year).

Regardless of how one reacts to these specific ideas, a clear lesson of the NCLB-era is that we need to continue to insist that schools monitor how well they are teaching all students, but arguably to do it differently. We need new approaches to assessing schools that are both workable and productive. Indeed, we worry that if tests are not made more useful to teachers, families, and education leaders, it is likely that support for tests will continue to erode and, eventually, may collapse.

But what about accountability? Given the wide variation in local contexts and current political environment, we believe that near-term decisions about accountability will lean more heavily on the judgments and actions of local decision makers. Localism is a reasonable stance given that interventions that work in one context may not work elsewhere (See Section 6). For the foreseeable future, the federal government arguably needs to couple an insistence on measuring performance with flexibility around how states and districts hold schools accountable for performance.

In the same way the federal government could incentivize policy innovations in testing, however, it might incentivize state and local innovations in accountability policy and help the field identify what is working and how (Darling-Hammond & Hill, 2015). What would it take, for example, to incentivize, design, and assess an accountability system focused on a performance floor, rather than ceiling, as described by Deming (Barnum, 2017, May 10)? Deming argues,

> '...instead of having an accountability system that says these are the best schools and...rate their rank, we should think about it like the way FDA does drug approvals... certify this school is safe for children, and certify a minimum standard and set that minimum standard so that most schools are going to meet it, that are functioning, but then when schools start to really fail, and fail students in a way that's obvious—if just 20 percent of the kids are passing a basic test and they're way below grade level and you keep telling the school to improve and they don't—that's when the state needs to step in.

Again, regardless of how one reacts to such an approach, the federal government should support accountability innovations and help states and districts learn from them and improve.

Beyond supporting the collection and dissemination of information, the federal government should also continue to use its influence to help public education build cumulative knowledge about critical policy and practice issues that support effective schooling. Over the last two decades, the federal government's Institute of Education Sciences (IES) has provided nearly $4 billion in research funds (Klager and Tipton, 2021). This support has resulted in hundreds of rigorous studies that can be used to help make decisions about policies and practices. Many of the findings from these studies are reported in the What Works Clearinghouse (WWC), a database of intervention effects and related Practice Guides. These efforts have centralized a vast store of knowledge; but most practitioners report that they do not use it (Penuel et al., 2017). IES can and should continue to build on the WWC and Practice Guides to serve the field by exploring new and (even) more accessible reporting formats for different target audience (resources that work for district leaders, for example, may not work for teachers).

In addition to continuing to improve the dissemination of its current stock of knowledge, the federal government could also create new, problem-focused syntheses that draw on a wider array of evidence. For example, the WWC currently provides useful and rigorous evidence about the impact of individual training interventions; but none of the WCC Practice Guides summarize what we know about effective professional development *in general*.[77] Such syntheses would be useful not only for school leaders and policymakers. Colleges of education—important producers and consumers of knowledge—could also benefit from state-of-the-art summaries of empirical research (e.g., consider the uptake of the "science of reading" instruction in teacher preparation programs [Drake & Walsh, 2020]).

Finally, IES could help to push the field by supporting research in important areas that are understudied or lack a high-quality evidence base. We are not advocating against field-initiated studies. But IES could play a more assertive role identifying and filling knowledge gaps by revisiting the topic areas of research competitions on a regular basis, for example (National Academies of Sciences, Engineering, and Medicine, 2022). Such agenda-setting could be done through more targeted requests for applications (RFAs) or more competitions like IES's Learning Acceleration Challenges.

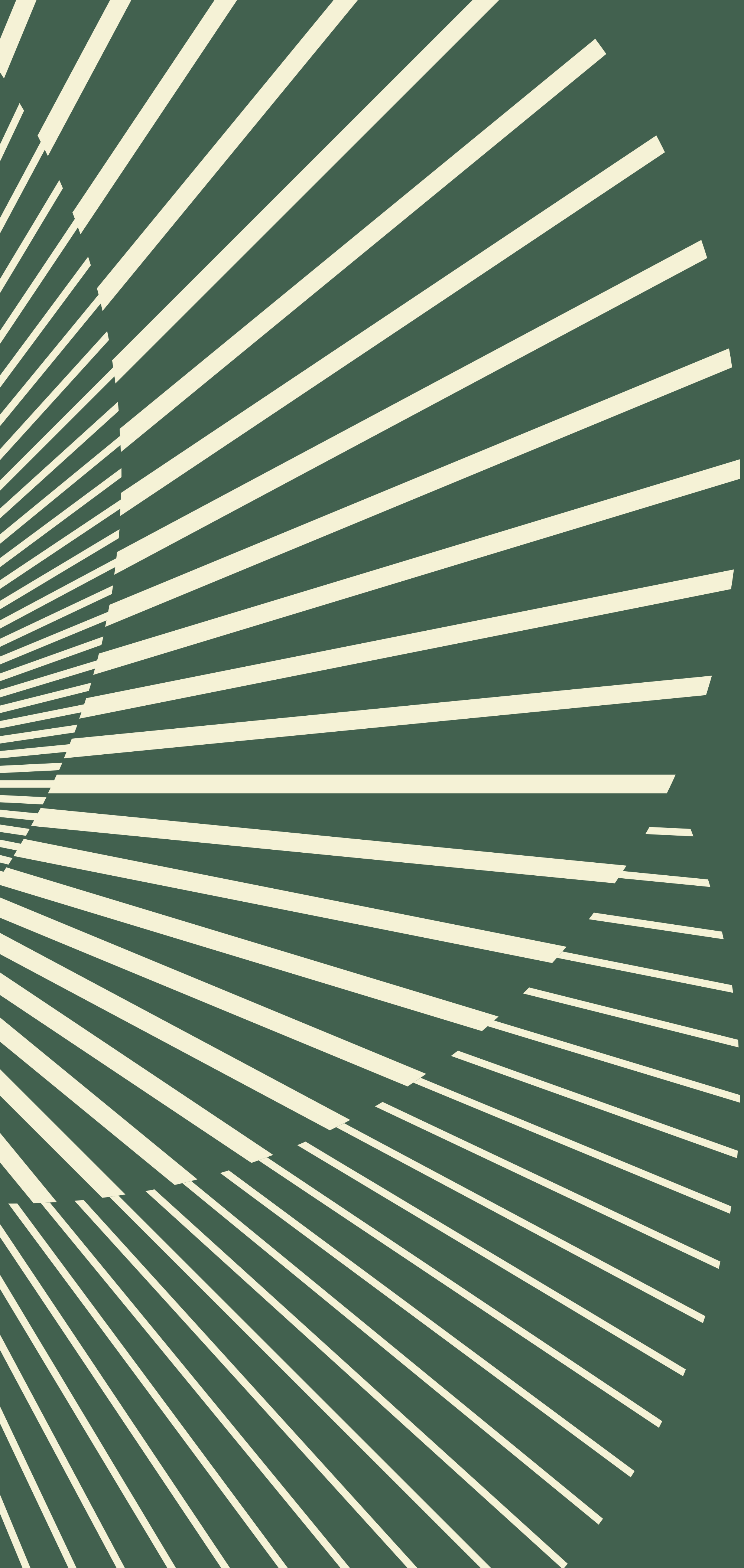
As the federal government seeds research agendas and syntheses, it will need to continue to ensure that decisions about research are driven by critical thinking and evidence-based thinking—especially in the current political climate. The somewhat underutilized National Board of Education Sciences could be a useful additional tool here for identifying critical gaps in knowledge and setting priorities for guide research competitions and synthesizing research.

[77] An example of this kind of summary can be found in Hill and Papay's (2022) recent review on professional learning.

8.4 Concluding Thoughts

In some sense, the reforms of the era were a coherent set of improvement levers grounded in compelling logic. But their rollout was at times disjointed, sequencing problems created unintended dissonance within the system, and key aspects of the system (districts and teacher preparation) were overlooked. Under NCLB, for example, school districts were largely bypassed as states were expected to hold schools and later, under RTTT, hold teachers accountable. Accountability sanctions were implemented before the development of more nuanced and fair measures of performance. And in too many places, state-driven teacher evaluations were launched before districts had thought through how they would evaluate principals. These sequencing problems could create dissonance in the system. Without more clarity and better alignment, unsystematic and partial policy roll outs sometimes sent mixed signals about school performance, with schools identified as failing even as most or all their teachers received positive performance evaluations.

Such issues notwithstanding, the era's focus on outcomes is still with us. In September and October 2022, the release of the NAEP results once again drew attention to the performance of the nation's schools. The results highlighted how academic outcomes have declined during the pandemic, both in general and especially for students who are struggling (Barnum, 2022, March 31; Willen, 2022). Thanks to $190 billion from the American Rescue Plan's Elementary and Secondary School Relief Fund (ESSER), schools have significant federal resources to help students who lost ground catch up. How will communities, families, and schools understand where the nation's schools are on the road to recovery? Whatever the path policymakers forge in the years ahead, the era of education reform from NCLB to RTTT suggest that the federal government has an important role to play in answering this question by drawing the attention of local decisionmakers and the public to outcomes, supporting knowledge generation and connecting it to practice and policy, and calling on states and districts to intervene when students are not learning. As policymakers fill the current vacuum in education policy, they need to build on and improve the federal government's influence in these areas, not abandon it.



Stakeholder Perceptions of the Legacy of 20 Years of Education Data and Accountability Efforts

Brightbeam

Introduction

The U.S. Chamber of Commerce Foundation (USCCF) Future of Data Working Group partnered with brightbeam to produce a qualitative analysis of perspectives from diverse communities and stakeholders on the use of data, assessments, and accountability in education. Guided by our belief in the original intent and promises of No Child Left Behind (NCLB) and other accountability reforms—that all students should have the chance to learn, excel, and live out their dreams—we sought to understand the extent to which this legislation has delivered on their aims. This report is a companion piece to the quantitative analysis of impacts produced by Dan Goldhaber and Michael DeArmond.

In speaking with diverse stakeholders, we uncovered varying perspectives on the lasting legacies of federal data and accountability reforms as well as several unintended consequences. We aim to contribute to an honest conversation about the real high stakes—our students' futures—by taking a closer look at public perceptions of the interventions, assumptions, and directives embedded within our accountability system and how these have played out in terms of student learning. Finally, we aim to distill these findings into takeaways that could inform USCCF and its partners' future efforts to engage the public and shape federal education legislation.

Methodology

During spring and summer 2022, brightbeam conducted interviews and focus groups with approximately 50 stakeholders from diverse backgrounds. In recruiting participants, we sought to balance the voices we heard from to include people on the political right and on the left, education practitioners and policymakers, grassroots and grasstops stakeholders, and advocates for diverse sub-populations of students including Black, Brown, and Indigenous children, children with disabilities, English learners, and low-income children.

We began by hosting focus groups with educators and parents. In order to encourage honesty and vulnerability in sharing their stories, these participants spoke anonymously. They also received a gift card for their time. We shifted to conducting interviews in order to gather more in-depth perspectives, particularly from grasstops stakeholders who had historical knowledge of the progression of education reform policies. To the right is a list of all interviewees organized alphabetically by last name.

In this report, at times we identify perspectives shared across demographic groups, as well as noting unique perspectives where relevant. However, we make no claims that our work constitutes a representative sample, nor have we asked participants to "speak for" a specific demographic group. Quantitative polling could be valuable to determine the extent to which these perspectives are truly representative of diverse education stakeholders in the country more broadly.

Brightbeam focuses on developing robust public engagement strategies to inspire Americans to fight for safe, affirming,, and liberating educational options where every child learns and thrives. We present these findings through that lens, with the aim of increasing public proficiency and identifying long-term communications strategies to bolster public support for data and accountability mechanisms within the education system, starting at the federal level.

Jason B. Allen, Atlanta educator

Kenya Bradshaw, Chief Program Officer of Reconstruction US

Dr. Jennifer Brown, Executive Director of KIPP Jacksonville

Anna East, Montana educator

Yvonne Field, Montana educator

Lindsay Fryer, former senior education policy advisor to Chairman Lamar Alexander (R-TN) on the Health, Education, Labor, and Pensions Committee and principal negotiator for Senator Alexander on ESSA

Dr. Howard Fuller, Distinguished Professor of Education, and Founder/Director of the Institute for the Transformation of Learning at Marquette University, former Milwaukee Public Schools Superintendent

Dr. Errick Greene, Superintendent of Jackson, Mississippi schools

Kati Haycock, founder and retired CEO of the Education Trust

Lindsay Jones, CEO of CAST

Rep. John Kline (R—MN), former chairman of the House education committee and key architect of the Every Student Succeeds Act (ESSA)

Sandy Kress, former Senior Adviser on Education in the White House for President George W. Bush, lead negotiator for NCLB

Delia Pompa, Senior Fellow for Education Policy at Migration Policy Institute's National Center on Immigrant Integration Policy

Dr. Michael Russell, professor in the Department of Measurement, Evaluation, Statistics & Assessment at Boston College's Lynch School of Education and Human Development

Dr. Sonja Santelises, CEO of Baltimore City Public Schools

Chris Stewart, CEO of brightbeam

Laura Waters, founder and Managing Editor of New Jersey Education Report

Executive Summary

No Child Left Behind (NCLB) was the first major federal education legislation dedicated to ensuring that all students should have the chance to learn, excel and live out their dreams. While the righteous goals it aimed to achieve have yet to materialize, there is no question that the reforms NCLB brought about, which were later built upon in subsequent initiatives such as Race to the Top (RTT) and the roll out of Common Core State Standards, ushered in a sea change in education that endures today.

Speaking with diverse education stakeholders about NCLB's lasting legacy, they lauded one lasting positive impact more than any other—the emergence of widespread availability and usage of comparable student data (and specifically disaggregated data), both for accountability purposes and as a pedagogical tool. Even some of the sharper critics of testing and accountability believe the movement towards data-based decision-making in education can be valuable for improving student outcomes, if done well. Very few folks truly want to "go back to how things were."

Without federally mandated testing, accountability systems, and the data revolution NCLB ushered in:

- Students of color, students with disabilities, English learners, and other subgroups of students with diverse needs would remain largely invisible in aggregate performance data, hiding the breadth and depth of opportunity gaps
- Parents and school leaders wouldn't have evidence to prove that their children, particularly for Black and Latine students and students from low income families, aren't getting what they need out of school
- Charter schools may still be an obscure blip in the education landscape, and those that serve targeted populations of students who are underserved by traditional schools would not be as prevalent, nor would they have the data to show they are beating the odds
- Teachers would still rely on their assumptions or intuition about their students rather than using data to identify trends and gaps that today inform their instruction
- School leaders wouldn't have data to differentiate the performance of teachers to inform their professional learning and growth opportunities, as well as personnel decisions
- Teacher preparation programs wouldn't face pressure to improve their practices to align with the science of learning
- Standards, curricula, and assessments would still be localized, with variations of quality and alignment among them
- Schools would still be operating in the "one room schoolhouse" mentality that limited sharing of resources and evidence-based practices across schools and districts
- The government would have limited evidence on practices that are worth investing federal dollars in and replicating
- Taxpayers would not have clear, reliable data to hold their elected officials accountable for delivering on the promise of public education

Despite these positive legacies, as well as the widespread support for the objectives it sought to achieve, when asked about it today, most stakeholders have an overall negative impression of NCLB and its associated reforms. According to most stakeholders, including ardent supporters and authors of the policies, the unintended negative consequences it wrought are significant and far-reaching within the education system. Any retrospective evaluation of the legacy of NCLB must acknowledge the lasting negative impacts upon students, families, and educators. Some of these consequences include:

- Narrowing of the curriculum, crowding out all other subjects besides reading and math, as well as extracurriculars, social emotional learning, and other elements of a well-rounded school
- Propagating and rewarding shallow definitions of school success, with a focus on proficiency on the test over everything else
- Labeling and tracking students based on their test performance or demographic factors
- Creating a culture of shame, blame, and urgency that raised stress levels for students and teachers alike, leading to burnout and negative associations with school
- Instigating cheating scandals
- Standardizing teaching methods, reducing teacher professional judgment, and perpetuating "teaching to the test"
- Directing large sums of public money away from students to for-profit enterprises and greater bureaucracy in schools

Some of these consequences are demonstrable, while others are perceptions held by various stakeholders, which have been shaped considerably by political battles and wider public narratives. Advocates and policymakers should learn from both perceptions and reality to inform future policy decisions and messaging efforts; however we will never be able to mitigate all risks and downsides. All policy decisions have implications—intended and unintended.

Now is a particularly important time for analysis because we are at a turning point in the public debate around education that will determine the survivability of the data and accountability movement. In light of recent testing waivers due to Covid, there has been a backsliding of accountability and a perception from many teachers, families, and administrators that testing is not as important today. Coupled with persistently negative public perceptions of testing even before Covid, the political landscape is ripe for removal of the accountability mechanisms we have fought for up until today. The extensive list of benefits above could be lost if we backslide. In turn, we run the risk of losing even more student progress than what has already been demonstrated in the recently released third grade National Assessment of Educational Progress (NAEP) results.

A solutions-oriented coalition of leaders from across the political spectrum enabled NCLB to pass. But they lacked the focus, funding, and strategy to sustain progress in the face of increased pushback. Opponents of data and accountability in education both on the right and the left have developed a sophisticated communications and ground strategy. They are in it for the long haul. Given the current polarized political climate and disintegration of the rational middle coalition, we need leaders who are willing to make sizable investments of time, energy, and money to bring the right people back to the table and keep them at the table for the long term fight for our students' futures. From tomorrow through the next reauthorization of the Elementary and Secondary Education Act (ESEA), USCCF must focus on cultivating that coalition.

How Are the Children?

For most stakeholders, the passage and subsequent implementation of NCLB marked a significant turning point in American educational practice. In the broadest sense, it clarified the expectation that schools (and ultimately educators) were, in fact, responsible for all students meeting basic academic standards. It gave us data about whether or not that expectation was being met. For the first time, educators, students, their families, and the general public had consistent access to information about how individual students as well as schools were performing relative to standards. When we spoke with stakeholders about this, there was widespread agreement that even beginning to answer the question "how are the children?" is important in and of itself.

The Old Days

Some stakeholders we interviewed have been involved in education long enough to remember the pre-NCLB education landscape. It's worth contrasting that period of time with today in order to highlight a few elements that we may take for granted because of how embedded they are in the current education system.

Limited Data

Pre-NCLB, stakeholders recalled an education system filled with assumptions as opposed to evidence. They described broad characterizations of student success based largely on averages, lacking nuance of student demographics. They shared how students of color, students receiving special education services, English Learners, poor kids, and students with other learning and thinking differences were rendered invisible at best, or at worst, deliberately left behind.

"Districts did not pay attention to how individual groups of students were doing. So you may have a predominantly white district with say 10% African American kids, and they would bury those test scores somewhere, and not educate the kids very well. NCLB shined a spotlight on how individual subgroups of students were doing."

—Chicago educator

"For as much as we critique it, and there are things that I think are worthy of critique, what it did do for the First time is force everybody to start paying attention to the academic achievement of different groups of children. Before that, we weren't talking about the academic achievement for Black children and Latin children and non-English speakers and students with disabilities or whether there was one year's worth of growth for those children. We weren't tracking data in that way. It's been a game changer in education to have been forced to look at the data disaggregated in that way."

—Dr. Jennifer Brown

The assessment landscape was marked by limited, unaligned, and sometimes invalid assessments created by educators or local districts. State-required assessments that did exist were not given annually, were norm-referenced as opposed to criterion-referenced, and often administered without clear goals or actions to follow.

"We have always assessed children. But the assessments were kind of just out there. Prior to NCLB, there was no sense of urgency to do anything differently because of the assessment data. The results were just the results. And in some places, unfortunately, more often than not, the results were thought of as the kid's fault."

—Dr. Jennifer Brown

Teachers may have used standards and curricula to drive their instruction holistically, but they often planned topical units and activities based on tradition or student interest. They relied on their own tests or simply their intuition about where students were performing and whether or not they had learned the intended objectives of a lesson. This subjective decision-making provided the breeding ground for bias and the famed "bigotry of low expectations," particularly for students of color, poor kids, and special education students.

"I can recall early on in my career in Baltimore getting this big old curriculum guide for middle school social studies. It was just a ton of information, and then 'go forth and teach.' I know at that time, and this was the early nineties, that there was no reference to the state standards or to what was required across the state. No sense of what would be on a state assessment or any of that. I actually don't even recall the state assessment."

—Dr. Errick Greene

"Prior to NCLB, data just wasn't even something that was on my radar. It wasn't something that you necessarily talked about or used as a classroom teacher."

—Minnesota educator

"I actually said this to somebody who was boasting about the recent move away from SAT in California, and I said, 'I understand why you did it. My only request is that you reflect on what life was before that.' And before that it was all about which high school you went to, or did your dean know somebody there who could write you a letter?"

—Dr. Sonja Santelises

Limited Accountability

Federal accountability was extremely limited. Local control ruled the day, with a tremendous amount of variation in state and district approaches to accountability. Limited testing meant limited data points to guide the basic accountability approaches that did exist.

"Before that, to the extent there was any accountability, you would look at whatever fragmentary test data there was to see whether the school was doing. You'd look at whether the school was doing a little better and you'd pat them on the back. Or maybe they did a little worse than before, but many states had no consequences."

—Sandy Kress

To the extent that there was a commitment to all students learning, there was limited action to back it up.

"Back before No Child Left Behind came out, we were using the phrase 'Every student can learn.' And it seemed at that point like such an empty phrase. Now instead of empty promises, we have focused the conversation on how you actually teach all students. It forced us to say, "Okay, we have to do this and there is accountability. What do we have to do to get there?"

—Delia Pompa

Local Variation

On the one hand, local control meant that student experience was remarkably disparate depending on local funding, priorities, and leadership. On the other hand, it did enable a few pioneering states to birth significant innovations, demonstrating the efficacy of data and accountability efforts early on.

"Some states, like Texas, accumulated a lot of data early on before all of this. That ended up making a lot of this reform possible. Because without data this is not as signifficant a movement in any respect. And so more data was being generated, more data was being required. Then it became important to look at data at a more granular level to see how each student was doing really."

—Sandy Kress

With initial success of data-driven decisions and accountability in Texas and a few other forward-thinking states, NCLB gained its contours. An emerging bipartisan coalition of student advocates rallied around an increased role for state and federal government to ensure "all students will have a better chance to learn, to excel, and to live out their dreams."

They paired requirements for annual state standards-aligned assessments with stronger accountability mechanisms, increased funding, and new options for students and their families if their school failed to meet its goals. A fairly immediate sea change began to occur in the nation's schools.

"I remember visiting schools a lot during the early years, especially when teachers thought the sanctions were going to be draconian. Of course, there was never the intention that the sanctions be tough, but that was what school folks thought. They were literally pouring through their data saying, 'here are the kids we can move to standard.' It was a very focused effort. That played out not just in getting kids over proficiency bars, but getting them into more AP classes, more honors classes. I mean, data was vastly more important. Looking at disaggregated data was vastly more important than it had been before. People felt under more pressure to create more opportunities and produce better results."

—Kati Haycock

Disaggregated Data

Not only did NCLB change the game in terms of accelerating the nationwide shift towards education data and accountability overall, stakeholders uplifted the ability to identify and examine outcomes disaggregated by student identity as the most notable benefit. Advocates pointed to the fact that pre-NCLB, there was no clear mechanism to account for how students of color, low-income students, English Learners, and students receiving special education services were actually served by schools. Families, policymakers, and advocates for these underserved student populations had ample qualitative evidence of disparities in schools, but limited quantitative basis to prove the educational injustices occurring. The theme of visibility and "shining a light" was prominent.

"There was a time where you could just laud and celebrate the relative performance of the district or of the school, but now at some point, we had the requirements to go deeper, disaggregate that data. We had to talk about the subgroups, how they're being supported, and how they're performing based on the supports that were there. So you've got a district that's high performing. Are there certain schools that are lower performing? You've got a school that's high performing. Are there certain children within that school that are lower performing based on that assessment? What does that tell you?"

—Dr. Errick Greene

"If we can't quantify that American schools are failing to educate huge numbers of Black boys, not teaching Black boys to read ... If we don't have that number, how can we do anything about it?"

—Laura Waters

"Unlike in the past, when data on kids who look like your kid, or shared other characteristics with your kid, had always been swept under the rug, finally, parents of color in particular, were going to be able to see data on how their kids were doing."

—Kati Haycock

"Never before did a lot of principals or superintendents care how their Hispanic students did, how their Special Education students did, how their English learners did. And now they were being forced to. I think a lot of that, unfortunately, got confounded with adult feelings of inadequacy. It became an easy scapegoat to say, 'Oh, it's because we have English learners that we're not doing well,' or 'It's because these kids are poor,' or 'It's because they don't have this or they don't have that."

—Delia Pompa

Special Education

Many advocates for students receiving special education services pointed to the transformative impacts of disaggregating data. In particular, they shared about how students receiving special education services can fall victim to low expectations. Having reliable access to unbiased assessment data has helped some students overcome those expectations. Parents also shared about how having an accountability mechanism forced their children's teachers into offering more support.

"Before No Child Left Behind, I was directly told by one of my principals when I was teaching, not to worry about a certain child because he has an IEP. One of the strengths of No Child Left Behind was that there was for the first time some urgency and pressure, whatever the reasons were, for individuals to care about these kids. There was a focus on different young people and school buildings that before had slipped below the radar for many reasons."

—Chicago educator

"My son was helped by high stakes tests because it would've been really easy to just lower standards for him and just say, 'Oh, he's never going to learn how to read. Let's not worry about it.' But because there was a test he had to take, guess what? He learned how to read. And I think that's largely due to higher standards for kids with intellectual disabilities. Because of No Child Left Behind, they had to disaggregate the groups by subgroups."

—Laura Waters

"One teacher told my son, 'I don't care if you all get this or not, because I'm going to get my paycheck regardless.' It's those types of things that just make me cringe. As another example, I remember probably around my son's junior year, the principal made comments on his performance on the ACT and some other things that had a direct correlation with the overall school standing. 'Well, Eric did well,' he said. And in other words, I think he just had some preconceptions about my son initially, and then his performance academically later proved him wrong."

—Memphis parent

"Prior to No Child Left Behind, kids with disabilities were not included in statewide standardized assessments. And what that really meant is that they were not included in the gen ed curriculum and they weren't taught standards that are appropriate for their age level. And in many cases, they were put on school buses and taken on field trips the days of the test. So our evidence indicates that for kids with disabilities, including them in the accountability system and testing them created a profound shift in the mindsets of educators who all of a sudden had to include them, had to teach them the content that might be on that test."

—Lindsay Jones

"For students with disabilities and particularly severe cognitive disabilities, research suggests that their educational opportunities 20 plus years ago were pretty minimal compared to the supports that they receive today. So I think for that subpopulation, there's certainly significant differences."

—Dr. Michael Russell

Additionally, one special education teacher described the importance of data to prove whether a student is receiving a free, appropriate public education (FAPE)—the hallmark educational requirement of the Individuals with Disabilities Education Act. Referencing the recent Supreme Court case Endrew F. v. Douglas County School District, she shared the perspective that consistent test data is necessary to determine whether or not students with IEPs are making progress and accessing the general education curriculum.

English Learners

A similar story can be told about NCLB's illuminating effect for English learners. Through disaggregated data, increased accountability mechanisms, and more widespread language acquisition training for general education teachers, advocates for English learners saw greater efforts to serve a population of students who had been largely ignored previously. Furthermore, a focus on English learners in the NCLB authorization of ESEA paved the way for even greater accountability for these students in the Every Student Succeeds Act (ESSA).

"English learners are often seen as the 'other.' And there is a false understanding that all these students are immigrants and if they would just go back, you wouldn't have this issue. When really, something like 75 percent of English learners are born in the United States.

With NCLB, we got the first mention of English learners and states being held accountable for them, but only in Title VII, which was then the bilingual education title. But it was a start. This was going to shine a spotlight on the kids that nobody had ever looked at before. As a result of having test scores and accountability, you started to see much more training for all teachers about English learners. Over the years, No Child Left Behind fell into disfavor, but those of us in the advocacy community who wanted the light shined on these kids could not have been further from that position. That's why, when I was at Unidos we worked with Ed Trust and other organizations to say, 'We've got to hang on to these principles. It's the only time anybody's paid any attention to these students.'"

—Delia Pompa

New Subgroups

Additionally, stakeholders described how access to data enabled teachers and policymakers to examine outcomes for student subpopulations that they hadn't previously considered as a group, such as foster students, students in military families, homeless and highly mobile students. These students with diverse needs could be better served when their outcomes are identified collectively.

"I was able to take data from our school system to look at students in single family households because they don't have two parents. And then we realized the data didn't really highlight that we do have a significant number of single dads that we never hear about in the public education system. We rarely hear about foster families and how we're supporting foster parents. We rarely, if ever, hear data about grandparents raising school-aged children. With this data we were able to ask, 'what are we doing to help these families?'"

—Jason B. Allen

Overall Impact

Thanks to NCLB and the accountability legislation that followed, we know where students are on a variety of measures. But has that meant improved outcomes for students? There was less agreement on how we answer that question.

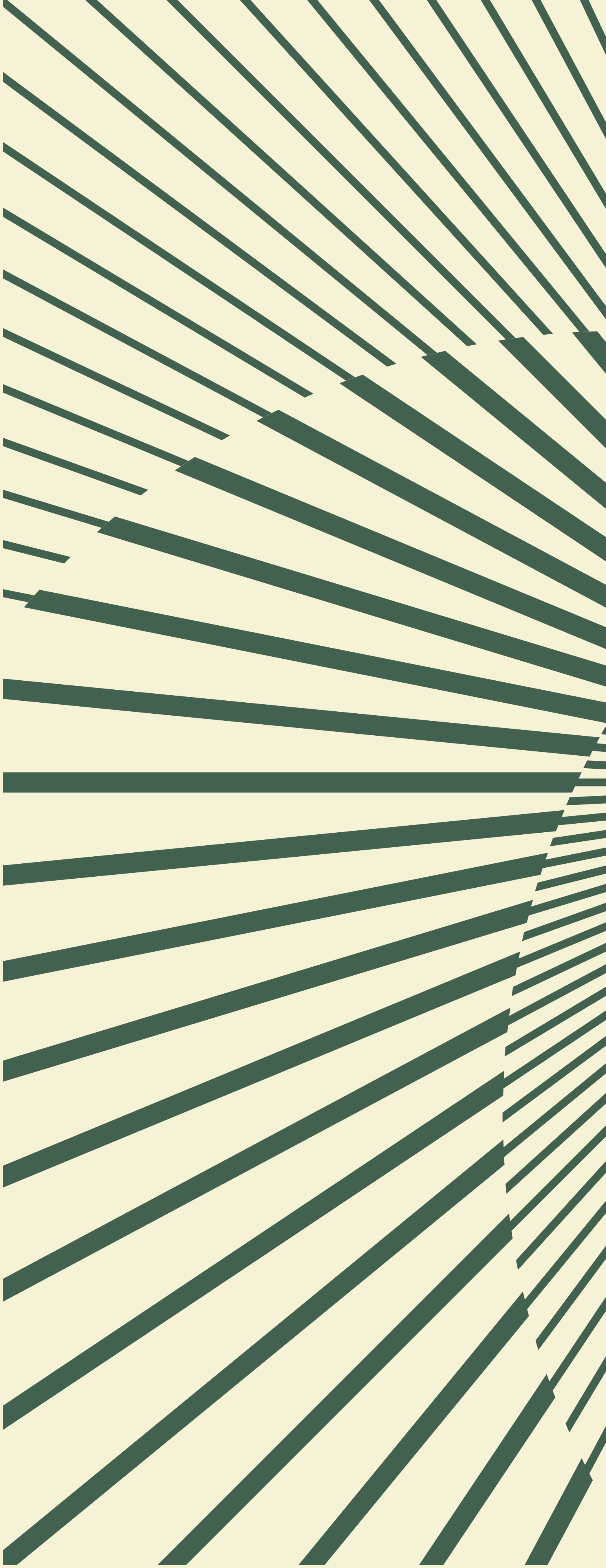
Looking at growth trends, we know NAEP scores and the state tests themselves show a sharp increase in overall student achievement in the early 2000s and then a slow plateau in the 2010s, gains that have been tragically erased in the post-pandemic era. But big picture, many stakeholders would answer that question recognizing that lagging overall proficiency rates mean we are still failing our students. And especially when asking this question for traditionally underserved student populations, advocates would suggest that increased attention has mattered, but that there is still significant room for growth.

"I think No Child Left Behind has benefited a number of children because it put a spotlight on schools in a different way. It was no longer possible to just hide kids who were not doing well in aggregate data. If No Child Left Behind did anything, it was to begin to talk about and spotlight the inequities that showed up based on race and class. To that extent, I think a significant number of kids have benefited. But just using the experience at our school and looking at the experience of the city of Milwaukee, there's no way that I can say that in the aggregate poor children, particularly poor children of color, are that much better off than they were before No Child Left Behind. Have some kids benefited from it? The answer to that of course is 'yes,' but I don't think that those kids represent the majority of the children in the categories that I'm talking about."

—Dr. Howard Fuller

"We have more data now than we ever had before. We know more about our kids than we've ever known before. Some of that makes it look worse than it was before because we know more. But if you were to say 'are things better or worse?,' I think for lots of kids, nothing's changed."

—California educator



Other Lasting Positive Legacies of NCLB

Beyond the notable impact of spurring a movement towards disaggregated data and accountability for student outcomes, NCLB also enabled a plethora of subsequent positive changes that have made a difference for students.

Without the consistent, reliable, and comparable testing data and focus on accountability for student outcomes, we would not have a host of reforms ranging from teacher evaluations to data-driven pedagogy to a growing school choice movement.

Advocates named the following effects as the most important to educational equity and student achievement more broadly.

Strengthened Accountability

Without testing data, we would not have outcomes-based accountability mechanisms at the state and federal levels pushing student achievement forward. Advocates discussed the informal and formal accountability mechanisms it enabled. For some, the greatest accountability mechanism was the "kinetic pressure" on educators brought about by raw data showing poor overall performance and sizable achievement gaps. For others, it was the ability to look at other schools and districts and see what was possible for kids of similar backgrounds. It forced honesty into the conversation about how we are serving our students, including for educators who might have preferred to look the other way.

"The data was bleak. And I remember people writing me back, because I would send it as a PDF or something in an email, and just being like, "Oh my God, really? This is where we are?" Yes. This is where we are. So I think maybe it kept everybody in 'real life land.'"

—Anna East

"If a Black kid just scored low, people would say 'oh, it was the Black kid's fault.' There was no sense of urgency to dig more deeply and to do something to be held accountable to those results. No Child Left Behind comes and says, 'We're now going to start looking at performance data. There's going to be some incentives and some accountability tied to it.' So if kids did not make a year's worth of growth in a year's worth of time, then there might be some consequences for schools and school districts. And then you had the onset of school improvement plans, where you had to say, 'This is how we are going to improve our progress as a school.' That kind of urgency and that culture of accountability did not exist previously."

—Dr. Jennifer Brown

"During a meeting with our principal association, we were having a really hard conversation. They were pushing on 'Oh the high school principal evaluations…' and I said to them, 'Wait, wait, wait, wait, one second. I don't want to pick on you all. We have a lot of work to do. Let's calibrate around reality. For a Black young man in Baltimore, I can improve the likelihood that they will graduate if I simply move them to DC and Chicago. Not Bel Air, not Bethesda, not Brookline. DC or Chicago. I want you to sit with that because those are also two cities who are experiencing increases in violence. Those are also two cities … ' And let me tell you, the call was quiet. But it shifted, for me, the excuse-making. So, I use data to celebrate, to instruct, and to say, 'Let's do reality checks on what really is possible.'"

—Dr. Sonja Santelises

Most stakeholders welcomed the change in accountability mechanisms over time as they evolved to include a host of factors beyond proficiency. Looking ahead, advocates see an opportunity to build on the changes in ESSA to create more comprehensive and locally-determined school report cards and accountability mechanisms.

"So, we have data now, student-specific data. Now we can begin to develop growth measures so that we can see whether these students are improving from year to year, which is really a fairer means of accountability for teachers in schools. Not to say that it wasn't important ultimately to get to where the standard was, but if you could at least advance the student by a grade level, then that has to be taken into account in accountability systems."

—Sandy Kress

"In ESSA, the flexibility provided to states to set their own interim and long-term goals for student achievement is the right policy move. The 100% proficiency math and reading proficiency standard in NCLB was set up to fail. NCLB was never meant to remain in place until 2014. ESSA creates more flexibility and accountability by saying states still have to have these student achievement goals but allows them to set the cadence and timeline for meeting them."

—Lindsay Fryer

Still, advocates point to the importance of proficiency and warn not to water down accountability systems too far away from the original goal of all students meeting basic standards.

"I appreciate the points that you get for growth in the Mississippi state accountability system. But with growth, you have to grow to a place. It's not just, 'oh, we moved from we suck' to 'we suck less.' You've got to get to a place of high performance. I've even had some former educational leaders in this district say to me, 'oh, well, we really only focus on growth.' That's not enough. When you know that many of your babies show up behind the starting line, getting them to the starting line is not enough."

—Dr. Errick Greene

"Saying every student can learn and 100 percent of them will learn was an incentive that we needed. Just the fact of saying we hold the same standards for all children, it was a huge move forward. And the fact that the disability community and advocates for English learners both support that, speaks to the importance of it."

—Delia Pompa

Alignment and Instructional Rigor Grounded in Higher Standards

Many educators described how the movement towards greater student achievement encouraged more frequent discussion of instructional rigor and alignment between curricula, assessments, and instruction. Thanks in large part to widespread adoption of and associated professional development on the Common Core standards, many educators developed a deeper understanding of the standards and clarified "the bar" for themselves and their students.

"With the SmarterBalanced assessment and the alternate assessment, we now have a definition of what it looks like to be proficient with the Common Core state standards. We know what rigorous, high expectations look like for kids. A lot of the professional development that I do with my teachers is actually around using the SmarterBalanced interim and formative resources, and their instructional resources, so they can tie thatdata into what they're doing every day. Some of those resources show really rigorous instruction targeted to specific needs of students."

—Yvonne Field

"I think the intentionality of the Common Core standards was very good. It gave educators a common language. It gave us a shared set of expectations. I think sometimes an educator's empathy can keep us from having the higher expectations that kids really want and can accomplish."

—Minnesota educator

Because many educators were bought into the content and depth of thinking required in the Common Core standards, they were more committed to working towards them. They employed backwards planning starting with the standards and then mapping out benchmarks along the way. They designed lessons and leveraged shared resources that were vetted with specific objectives in mind. They leveraged data cycles to monitor student learning, identify gaps, and re-teach where necessary. They pushed their students forward to meet higher bars.

The move towards instructional alignment with state standards also allowed for greater sharing of high-quality resources.

"With the new standards, I was more apt to have students do more analytical work. Because I believed in those standards, I was also more apt to analyze what I was doing and to make sure what I was doing actually made sense."

—Anna East

"We tend to get bogged down with the statewide assessment results that we report out publicly on a large scale. But a lot of what needs to happen to enable change is that teachers need to know how their kids are doing with the instruction that they're doing right at that moment. They need to have shared common assessments that they use to benchmark kids more on a six to eight week schedule. A lot of that is training them on how to identify assessments that are aligned to grade level, rigorous instruction. Then, once they have those assessment results, how to dig into those and really use them to know how each of their kids are doing. And then how to find some lessons and things to help with gaps, or figure out if they need to re-teach or do something differently."

—Yvonne Field

There are opportunities to continue strengthening these standards to ensure they are representative and accessible to the diverse communities we serve.

"There's this part of me that recognizes and honors that standards and rigor are important in academic settings in a lot of ways. I think more recent conversations about anti-racism and anti-oppression have helped me recognize that standards and definitions of rigor are often loaded with implicit or explicit biases. I think a tangential value of No Child Left Behind and Common Core is forcing a stronger recognition of the way some of these biases function. Whose standards are they, and whom do they benefit? Are the people of privilege or resources going to be the ones that continue to benefit from these because they're the ones who are developing them?"

—California educator

Better Assessments

Paired with this greater focus on aligning instruction and assessment with higher standards was a need to ensure we have valid, reliable, unbiased assessments that truly measure student learning. With PARCC and SmarterBalanced leading the charge for Common Core-aligned assessments, many states adopted new, more rigorous assessments that embed higher order thinking skills and can be administered online. With ESSA's encouragement of testing innovation, some advocates see great opportunity ahead to continue strengthening assessments and their utility for both formative and summative purposes.

"I think that we are getting better at evaluating learning than when I was taking the number two and bubble tests. Now we are able to use digital platforms that are designed to minimize frustration for kids and where we can really identify different strands that they need help in, or that they've mastered. I think technology has been a game changer. We're so much better off than 20 years ago in being able to gather data and use it and fine tune it and go back and get more without it being traumatic for kids or teachers. I only see that field exploding in different ways. Maybe in the future there'll come a point at which every day a kid will sit down for five minutes and from 10:00 to 10:05 in the morning at their school or in their house and answer a few questions. And that will point the teacher in the direction of what the kid should be focusing on."

—Laura Waters

"I think you could upend education in many ways if you have a criterion test that measures critical thinking—an assessment tool that takes out some of the bias and some of the cultural knowledge that gets rewarded in the current system of testing. It's going to highlight some of the inequities in teacher practice, showing that it's across a system. Some of the most atrocious teaching you're going to see is actually in some of our highest scoring schools. And that is crazy because you'd think that shouldn't happen because those kids are doing well because of the great teaching they're getting. When in fact, we know that that's not necessarily true. These kids actually are succeeding in many ways because of a system that is set up to support them."

—California educator

"Ten percent of students with disabilities should be on an alternative assessment. Those assessments are actually far better examples of assessments and far more robust in what they're measuring. That's the thing, when you design for the margins, you actually end up helping all people. We have Siri, text to speech, that wasn't created for you and me to use in our car to text our kids or whomever. As another example, we see massive, off the charts use of captions. The Wall Street Journal had an article about how Gen Z is watching multiple screens at once and they're turning captions on all the time as they watch, not just for language differences, they're using captions as they follow. We saw Zoom through the pandemic, and Microsoft, reported massive numbers of people, way more than statistically we would think have a disability, using captions during meetings. They're small examples but that's why those alternative assessments are stronger. When you're designing for your audience, you're capturing a lot more."

—Lindsay Jones

Focus on Teacher Quality

After NCLB paved the way for increased focus on data and accountability at the systemic level, RTT cemented a focus on teacher quality as the most crucial lever for change. States and districts increased their focus on all points of the teacher pipeline, seeking to strengthen the quality of new recruits, evaluate teacher performance, tie employment decisions to performance, offer tailored professional development, and more.

"Teacher evaluation had been very controversial. The unions were against it, but when it came down to the fact that the state might win billions of dollars if we had a teacher evaluation system in place, everybody was more eager to figure it out. Rather than just fight it, maybe say, "Well, how could we do it?" That was a sea change from, 'Absolutely not. You don't rate teachers; we're professionals,' and 'Test scores don't determine what we do.' It changed from that to, 'Okay, what would be a fair teacher evaluation system? If we had to construct one, as teachers, what would we do?'"

—Chris Stewart

"I have seen the value of teacher evaluation. Sometimes it doesn't always work, because it depends on the teacher. We've had some teachers for whom it was very, very useful. We saw them grow and implement things that they hadn't even thought of, or maybe were too lazy to even try before. We've seen other teachers that just stay stagnant and could care less about a teacher evaluation unless something happened to them, whether they got fired, or whatever the case may be."

—Chicago educator

Before widespread access to student performance data, teacher evaluations were even more subject to the biases and varying leadership abilities of school administrators than they are today. We now have more granular and outcomes-based data on what makes a high-quality teacher—what experiences, practices, mindsets, identity factors, etc. that we can use to improve upon teacher placement and ensure that every child has access to a high-quality educator.

"What are the teacher indicators we should look at? At that point, we mostly had just certification status and years of experience. Everybody thought it was important to prove outcomes and those things mattered, but not hugely. Now, obviously, a teacher's race is turning out to be vastly more significant than some people expected. I don't think folks assumed it was as critical to improving outcomes as we now believe it is."

—Kati Haycock

"As a union leader and grievance chairperson, one of the worst grievances that I ever had to negotiate was for a math teacher who had her bachelor's degree in math, her master's degree in math, her doctoral degree in math, and had taught math for her entire career. She was called three days before school started that year and was told that she would not be teaching math that year, she would be teaching basic skills reading. Just let that sink in. She knows math backwards, forwards, inside out. She knows how to teach math, she's very effective at it. She knows absolutely nothing about teaching reading. And yet, our principal could do that because at that time, if you held a license for grades K-8, you could be placed in any K-8 subject area. No Child Left Behind changed that. And I think that that was an incredibly positive change because you didn't want this teacher teaching reading, you wanted her teaching math."

—New Jersey educator

Increased Transparency and Public Engagement

NCLB promised that parents would have more information about the schools, and more say in how their children are educated. Transparency around student and school performance enabled greater democratic participation and student agency within school systems. Public dialogue about school quality has gotten richer, emphasizing the many facets of what makes a quality school. For better or worse, we now have many tools such as GreatSchools.com, niche.com, and rankings of schools on Zillow, providing the public with information about school performance.

"I think with the accountability systems, we saw greater public engagement in education. It began to demystify to some degree what happens in schools and in classrooms, as well as what relative and absolute success looks like. All of those things I think have helped to empower parents and community members in making smart choices and demanding more and different schools for their children. We can now differentiate between the mentality of 'I feel safe. I feel my kid is safe. I like the teachers and the staff at the school. I generally feel good about the school,' versus 'My kid is being prepared for success in future grades or in higher ed or wherever they go.' I think all of that has been a byproduct of the various reforms and the ways we've engaged in those reforms."

—Dr. Errick Greene

Greater transparency around student performance can also be a tool for increasing student ownership of their educational trajectory. It can invite families into participating in their children's education in meaningful ways.

"What I saw evolve with this is, I think it did drive both educators and students to talk more about data, to talk more about how students were doing. Before standardized tests came in, I would talk to my students about progress a little bit, but we wouldn't have the level and the depth and degree of goal-setting conversations that we have now. We wouldn't have student-led conferences. I think in some ways it did empower some students to feel 'I do have some autonomy. I am in control here somewhat for my learning.' I think it did to some degree make learning a little bit more tangible for everyone."

—Minnesota educator

"At Unidos, their parent training program had a module on accountability, assessment, and data. Once parents were able to look at the data at their school and compare it to the data in the school across the way that was doing much better, they were able to ask the principal, "Why is this? What are they doing that we're not?" Giving parents a simple but real look at how this data is used, empowers parents in a very different way."

—Delia Pompa

"Our inability as a profession to yield change fast enough for kids is now being used against kids. Parents don't want their schools labeled as underperforming, so then we close it and then we're not giving them anything better. That's what they're ticked off about. It's not that they don't want to know the data."

—Dr. Sonja Santelises

Additionally, having objective, comparable data allows for difficult conversations about rigor and expectations. In particular, it illuminated gaps for students who received good grades, but weren't in fact performing at grade level. Unfortunately, these students often did not find out they were actually behind until after graduation when they struggled with college placement or were placed in remedial courses.

"I remember the test results didn't match the grades that my child was getting. And I'm like, 'If my child got a B in your class, shouldn't their test results show that they're proficient in this area?' The teachers really didn't have any explanation for what the difference was. I didn't feel empowered to resolve that in any sort of way."

—California parent

Still, there is plentiful room for growth in engaging families and communities about what data matters most, as well as in educating the public about what to do with all of the information we have about schools.

"I would get the scores, I would look at them, and I'd file them. I didn't do anything with them because to me, they were meaningless. Unless the teacher or the school wanted to follow up and say, 'I see your child is not proficient in math. Here's what we want to do about it,' then I didn't have the time and the space to investigate. I was already trying to help them get through the grade and get all of their projects in. So the test scores, I didn't really pay too much attention to them."

—Michigan parent

"Most parents weren't getting information from States, districts, or schools that was helpful to them before Covid. State, district, and school report cards—if you can find them, you might not know what they mean. For example, what's an A school? What's a C school? We do such a poor job of communicating what we're collecting. That's been an issue since NCLB. Maybe the feds can play a deeper role in helping states, districts, and schools format and structure how information is made available. Additionally, we need more understanding of what parents want to know and make that information available in a helpful manner. For example, is it average test scores? Teacher credentials? Where high school graduates are going?"

—Lindsay Fryer

Increased School Choice and Proliferation of Charter Schools

Certainly prior to NCLB, charter schools and other forms of school choice existed across the country; but NCLB enabled school choice to become much more widespread for a few reasons.

First, it offered school choice as an accountability mechanism for low-performing schools. In his NCLB signing speech, President George W. Bush explained that for any school that doesn't perform satisfactorily, "any school that cannot catch up and do its job, a parent will have these options—a better public school, a tutor, or a charter school."

"Charter schools have been a resounding success. Go to Harlem, NY and see tremendous progress. You had no children going or preparing to go to college and now you can see a lot of kids are going to college. I visited a North Minneapolis charter school where the vast majority were kids of color. I was talking to the principal and he shared there were 400 kids there and they were trying to purchase a building. There were over 1,000 kids on the waiting list to get out of the nearby failing schools."

—Rep. John Kline

Second, increased availability of comparable test data allowed growing charter school networks to showcase their strong academic outcomes. Schools like KIPP and Success Academy were able to provide an apples-to-apples comparison of their schools' achievement with neighboring schools as a recruitment tool.

"I spend a lot of time looking at Newark and Camden where the traditional districts are really bad and there is a healthy public charter sector. Because of the reliability of the data we get, those charter sectors have been able to grow. There are still very long waiting lists, but more parents are putting their kids in higher achieving schools or better schools, more culturally responsive schools, because data is proving that these schools work, especially for low income kids of color."

—Laura Waters

"You now have examples of success. A major segment of Karin Chenowerth's career is using the data to point out schools that succeed with the same profile of student. When you read this, you know what's possible. You can not like how a charter school does things. You may not like it, and I say this as a traditional public school superintendent, but you can't deny the data."

—Dr. Sonja Santelises

We see an emergence of schools specifically designed with a population in mind. Most notably, we see a direct correlation with the growth of high-performing urban charter schools better serving Black and Brown students who were left behind in traditional district schools.

"I think it has birthed these very specific types of schools, particularly in the charter sector, where you have founders stepping up and saying, 'We have some historical data, some longitudinal data that shows Black kids are not thriving in these traditional district settings. We can start a school that is designed for Black children and Black families to thrive that's identity affirming and will demonstrate that they'll do better in these settings.' Even schools that are specifically designed for native children, where those charter schools are being started by Native Americans. And they want to educate native children in a specific way that is very much identity-affirming for them."

—Dr. Jennifer Brown

"It brought into the public system a new dynamism. Younger people came in with ideas about creating new schools, and how to serve kids differently. When you look at the data, it's clear that in aggregate, charter schools in certain areas outperform the traditional public school system for certain kids and in certain communities."

—Dr. Howard Fuller

With greater access to information and stronger public conversations about what makes a quality school, competition increased amongst schools aiming to meet the needs of students who were being failed. It established the concept of students and parents as active consumers of a public good, not just passive recipients. In turn, the competition encouraged traditional district schools to enhance their efforts to keep students (and associated per pupil funding).

"It has been good for all of us in traditional public schools to wake up and not take kids and families for granted, understanding that families have options. They will vote with their feet. I think it challenged us to focus more on customer service. Yes, we want to focus on academic outcomes for kids, but also how do we greet them? How are we communicating? How are we engaging? How are we keeping up the facility, all those things that matter? We're having to tend more to that now because parents have choices."

—Dr. Errick Greene

"I did threaten to take him out of the public school and put him in a charter school in our county. So I met with the school's head of the music department, explained what was happening to my son. I said I had decided to pull him out and enroll him in the charter school because of the music program. And she said, 'Please don't do that until you give us a chance to improve.' And I agreed to give them a chance. And they really did work with my son and with me to turn things around for him."

—New Jersey parent

"Race to the Top had a very direct impact, in my mind, on growing the number of charter schools. Initially, charter schools were not serving kids with disabilities. They were overwhelmingly counseling them out for many reasons. Today, charter schools know they have to serve kids with disabilities. They counsel them out far less because of years of advocacy. And there are many, many people in the charter school movement who care about kids with disabilities and are focused on that. I think that's great. I also think there are some really awesome laboratories of innovation in those schools. They were able to play with things like scheduling. Race to the Top, by and large, didn't actually address kids with disabilities in a robust way, but the massive push for charter schools and the disability advocacy movement, actually led us to some really good models that I'd love to see brought into traditional public schools."

—Lindsay Jones

Stronger Pedagogy and Best Practice Sharing

While some educators had long used formative assessment data to inform their instruction, NCLB made data-based decision-making much more widespread. For some educators, this meant paying attention to data as a pedagogical tool for the first time.

"I would say Race to the Top and No Child Left Behind helped to the extent that you have a generation of teachers now who are not afraid to look at data, who believe that looking at data is important in terms of their work."

—Dr. Howard Fuller

"In the high school English department during those Race To the Top and Common Core years, we got very specific data in terms of student performance on the literature competencies, the nonfiction competencies, and the different types of writing required in the assessments. We aligned it to our curriculum and said, 'Okay, where were there huge gaps in this grade level?' So for 500 ninth graders, where were those large gaps that we saw in the test results? We then reflected on our experience in teaching those parts."

—Indiana educator

Starting in teacher preparation, educators more actively engage in designing learning experiences to meet student needs based on data. They are more apt to see their job as ensuring their students actually learn, as opposed to just delivering content.

"Sometimes when you teach stuff, you think that you taught it, but it doesn't mean that the kid got it. It's really easy to just keep doing the same thing until something really demonstrates to you that your kids did not get the thing that you were teaching. I think that happens a lot. I know it happened in my district. The data reports that I was sending caused educators to have the moment where they go, 'God, I was teaching this, but they weren't learning it.'"

—Anna East

Teaching can often feel like an isolated endeavor—a teacher in her classroom all day with her students and limited collaboration with other adults. With a renewed focus on student achievement and teacher quality, and facing external pressures for improvement, many teachers turned to their peers to help strengthen their teaching practices. More teachers reported collaborating with their peers to identify what works, what doesn't, and how they could change their practices. Administrators also shared about how the spread of data and accountability mechanisms influenced their collaboration with other administrators.

"There's been this opportunity for us to step out of ourselves. No longer are we the one room schoolhouse. It is expected and assumed that best practices within a school are being shared in collaborative ways. It's the same across schools and the same across districts and states."

—Dr. Errick Greene

Targeted Interventions and School Improvement Efforts

In addition to the enhanced collaboration and scaling of best practices that happened organically within the teaching profession, NCLB, RTT, and Common Core equipped the state and federal governments with plentiful opportunities to drive school improvement efforts forward. A trove of new data enabled quantitative analysis of programs and strategies at a scale never seen before in American public education. With states adopting better data systems to provide schools, teachers and parents with information about student progress and subsequent massive investment in the School Improvement Grant program, teachers and school leaders would have support to become more effective.

"My building was considered underperforming, and so we were part of the Reading First program, which meant that state gurus would come in and provide extra professional development for the teaching staff. Some of it was good, but it just always felt a little bit forced upon us. We really had to remind ourselves that we weren't bad teachers because sometimes it felt like we were educators that weren't doing well by our children and that's why we were getting this extra help. And so as the building leader, I just really had to work with my staff and say to them, 'Nope, they do not see all of the amazing things you do every single day. Let's just take what they give us and embrace what we can. And when they walk out the door, we'll weed out what we really want to use and go forward from there in terms of how do we use this information to make us better along with the great skills that you already have.'"

—Minnesota educator

Shifting away from NCLB's focus on specific turnaround strategies, both RTT and ESSA called for using both innovative and effective approaches to improve struggling schools, doubling down on the intentionality of turnaround efforts. ESSA defined specific requirements around the term evidence-based in order to ensure federal funds are spent on practices that we know work.

"ESSA right-sized what to do related to school improvement and what to do when schools are failing. NCLB did not address the different reasons or needs for why schools were labeled as failing and what they needed to do to improve. Regardless of the reason for improvement identification, whether it was a whole school issue or one sub-group of students that needed attention, specific, prescriptive, predetermined school improvement models needed to be implemented. ESSA put in place a process for schools to do a needs assessment to understand their reasons for failing and then to develop a plan for how to improve by addressing specific needs, and this was the right policy move. NCLB had requirements that certain school improvement activities had to meet scientifically-based research standards, which are pretty high standards. Not too many things in education meet this gold star. So it was the right idea for ESSA to allow tiers of evidence to apply to interventions."

—Lindsay Fryer

Despite the promise, advocates are dubious about the effects of turnaround efforts at scale.

"Looking from a strictly high-level data standpoint at summative assessments and whether the schools that were identified for comprehensive or turnaround support have exited or gotten better or things have changed, they haven't. You can look across the board and pretty much see that across the state, the schools that were struggling 20 years ago are the schools that are struggling now."

—Yvonne Field

Unintended Consequences

Despite the extensive benefits detailed above, it should come as no surprise to education advocates that NCLB has a bad reputation amongst most stakeholders. Teachers, policymakers, parents, and students alike can point to a host of negative consequences of the legislation. In our conversations with stakeholders, we sought to understand the extent of negative perceptions and determine any differences across audiences. While few stakeholders gave the legislation five-star reviews, we found educators in particular were the most likely to speak about its negative impacts. Parents and policy wonks, too, shared in some critiques, but not nearly to the same level of vigor.

We also sought to differentiate between myths and realities. When we pressed people for evidence about their impressions, we found some claims had ample evidence of the negative impacts and some claims had little evidence or personal experience to back them up.

This differentiation is helpful in understanding the extent to which future policy and messaging efforts will be successful. It will be easier to win a narrative battle based on general impressions; whereas it is much more difficult to change people's perceptions grounded in their lived experiences.

Narrowing Focus

By far the most common critique of accountability efforts was the belief that everything within schools became focused on test outcomes, to the exclusion of so many other facets of what makes a quality school. Many participants, including prominent advocates of data and accountability, criticized the shallow definitions of school success that were driven by administrative pressure and public reporting.

"We were like, 'Wait, are you saying that we should have a school system that's supposed to teach people, but we should never find out if it is actually doing that? And that we should have no responsibility for whether or not it is actually producing results on the academic side?' But I think what happened is that those of us in the 'reform movement' pushed so hard on the test scores that we're now in a situation where it sounds like the value of a school could be determined solely by test scores. Value-added makes it at least a better argument, but it's still problematic in many instances."

—Dr. Howard Fuller

"I think the accountability structure here with letter grades is probably too complex, but I wish there were a way for us to tell a more robust, bigger, more comprehensive story about what schools are doing. There are some things that I think we do beautifully that will never get captured in our school report card grade."

—Dr. Jennifer Brown

Educators spoke extensively about the "culture of urgency" they attributed to federal testing and accountability efforts. They described pernicious impacts of this drive towards test scores above all else. They spoke about how they felt stymied from meeting students' needs more holistically or that they couldn't take the time to assist students with social emotional skills.

"There felt like a lack of awareness around the different systemic challenges and the poverty that impacts our community. There was constantly a sense of urgency. I felt like all of our classrooms were in a high pressure situation. We had to perform and produce. It was very dehumanizing. We felt like machines. Only in retrospect did I realize that we over-privileged the students' ability to perform on tests. We centered on scores versus each of the children's ability to perform in different ways. It really took up a lot of the pedagogical oxygen in the air. It really suffocated the teachers and the student leaders."

—Los Angeles educator

"We had a suicide cluster. We had a ton of trauma, lots of neglect, lots of child problems that were much bigger than what was in front of us just in the schools. So that is what teachers really focus on a lot in that district—trying to help overcome those issues. Sometimes the academic part isn't necessarily the top priority."

—Anna East

They commonly critiqued NCLB's expectation of 100 percent proficiency, seeing that as a goal that disregarded the herculean task educators often face when their students are coming into class multiple years behind.

"The problems start when you tie absolute performance to your pay, when you tie it to job security. Kids aren't widgets. But the problem is we sort of almost held teachers accountable for this idea that they were. So I'll give you an example. Are you going to hold teachers accountable that every kid reads at third grade? Well, we all want kids to read by third grade. Right? But the question is, what if the kid came in not knowing the alphabet? They came in the beginning of third grade and by the end of third grade, they might not be reading at a third grade level, but they might be reading at a second grade level. But that wouldn't count necessarily in some of the measures that teachers were held accountable for. I think the inflexibility in the system makes it really hard. The question then becomes 'How do we help? How do we move things forward and help teachers use data effectively, but not make it dogmatic?'"

—California educator

Others were critical of the implication that a narrow focus on test scores made schools compete in unhealthy ways.

"It seems really counterintuitive, like a fundamental flaw in the accountability matrix, that learners are pitted against each other. Schools are pitted against each other. A standards based approach assumes that each of us is capable of achieving a certain goal, but we're going to go about it in a different way. If you look at how the accountability framework is, certainly in Chicago, schools were ranked against each other. That's a complete disconnect from what I think we believe as educators."

—Chicago educator

"Extreme pressure for results was a very unhealthy thing for teachers, for students and for families. It definitely came from the changes that we were seeing in federal policy at the time and in state policy at the time. I think that the creation of the public school report cards also added to that pressure. Schools were competing now against one another. Competition can be healthy, but competition can also be harmful. I think that in this instance, competition between schools had a definitely harmful aspect to it."

—New Jersey parent

Educators, particularly elementary teachers, shared their experiences of narrowing the curriculum to focus their time squarely on mastery of reading and math standards. They shared their experience of "trimming the fat" from everything from social studies to arts to social emotional learning to field trips.

"When you're the core teacher for a classroom full of students and you're supposed to cover all those standards and you have this pressure of the state tests, and mandates, and this and that, and the other thing, science and social studies just got pushed out of the way. Students have so much to give, especially at that young tender age. We should be building on that culture of curiosity and risk taking mindset. But so often, as elementary teachers, we've let that opportunity go by the wayside. So that's what's been so sad to me is that tunnel vision of math and ELA, math and ELA."

—New York educator

"What endures in our system, and we're still trying to recover from, is we moved to hyper focus on teaching reading and math so much so that if it wasn't reading or math, it definitely fell by the wayside. Things like CTE became very secondary. In some schools, it got taken out completely. Now, what you hear from our community is we need plumbers, carpenters, and welders. We shifted so far to where I remember as a third grade teacher, my science had to be taught through literature and reading so that we were making sure we were focusing on those nonfiction reading strategies."

—Minnesota educator

"We narrowed our focus and reduced it to just math and ELA. As somebody who teaches social studies and civics, those worlds were pushed to the side in the younger grades. When students arrive in high school history and or civic spaces, there's a baseline of learning that isn't there as it was before. I think some of our national civic breakdown right now is probably somewhat related to the lack of understanding a lot of people have with history and civics because they were so strongly deemphasized in a No Child Left Behind environment. And so we're left in this polarized space that we're in right now."

—Chicago educator

Both teachers and parents described rote test prep becoming a time-consuming focus. They talked about how weeks of instruction were replaced by practice tests and test-taking strategies.

"As soon as the students got off the bus and were getting ready to enter the school, we would ask, 'Did you study these words?' Everything was test prep. Everything was about the test. Your lunch is silent and instead of having 25 or 30 minutes to eat, it's 15 minutes because we need you to eat real quick, and then we're going to do these math fractions during the time where you're supposed to be socializing. So, we took socialization away from students. We took teacher planning periods away from teachers, so teachers couldn't even really be prepared. You couldn't share best practices. I remember, especially during those times, for at least three years straight, we were teaching during our lunch period and teaching during our planning period. Every minute, the system was driving us to overkill, trying to teach test strategies and test examples to students. And we missed out on really teaching students."

—Jason B. Allen

"There was also a lot of time I felt like the classrooms had to prepare for that test, so they had to practice taking tests or they had to practice the material that was on the tests. And I remember when I was younger and taking the test, it was just a week. We never talked about it really before or after other than 'Make sure you eat a good breakfast' and 'Come to school prepared to take this test.' So I feel like there's more pressure to perform on the generation that my children are part of. I feel like it's harder for them because they're maybe less able to pursue the areas that they want because there are all these areas that they're sort of forced to perform in."

—Michigan parent

On top of the weeks they sometimes spent on test preparation, they also shared frustrations of the time it would take to conduct the tests themselves. Because of logistical challenges such as limited computers, student absences, and varying student accommodations, teachers shared that it would take up to six weeks to complete one round of testing. Additionally, because of various state and local decisions to administer other assessments, this time-draining task could be repeated multiple times per year.

"They just are excessively time consuming. One year, I remember counting how many days the 11th graders were missing for tests in the month of April, it was like 80%. It was unbelievable how much time they missed from class. How are you supposed to teach? You can't even teach to the test if you want to, because they're not in the classroom."

—Anna East

"When it comes to state assessments, I tend to cringe because it's literally a month out of normal class. During those assessments, schools just shut down and they focus on teaching children to take the test. Even now, I know here in our school district, this week starts the state assessment. And probably the two weeks leading up to it it's just all TCAP focus, TCAP focus, teaching them to take the test. I just don't think that's an effective way to do it. I just feel like it is not an accurate gauge of student performance."

—Memphis parent

Teachers also emphasized how drilling on the mechanisms of testing was hyper contextualized, that these skills were not transferable outside of a test. They shared that often, the skills were lower-order thinking skills and did not build critical thinkers. They shared that test scores might tell you something about the academic accomplishments of young people, but that they were limited in the quality and depth of learning.

"Learning time was taken away from other content areas dedicated to helping them do well in a test, which may or may not actually give them the skills they need to thrive in school. It seems counterintuitive, but some of the skills you might need to just get a few percentage points higher on a test or to raise certain subgroups of kids on a test made all kinds of bad things happen in schools that adversely impacted kids."

—California educator

“I think sometimes our focus was just so very, very, very specific about raising this particular strand on the test that skills weren’t taught in context or in necessarily meaningful ways. Some skills students couldn’t necessarily apply when they picked up a book because of how they were taught.”

—Minnesota educator

Labeling and Sorting Students

While the vast majority of stakeholders lauded efforts to disaggregate student data by various identities, some educators highlighted a pernicious flipside to this effect. They described how instead of diagnosing and supporting student needs in a more tailored way, disaggregation was used to label students and sort them into tracks.

“I think for the development of children, I feel like the data has done nothing but support the school to prison pipeline. I feel like it has done nothing but continue to support segregation in our educational system. When you get to school, school is supposed to help prepare you to be the best citizen that you can be, not the best Black citizen that you can be, or not the best Asian American citizen you can be. But that’s how we take data and we create models of learning.”

—Jason B. Allen

“We would sit and analyze testing data, formative and summative. What it led to was sorting lower-achieving students into classes that provided less robust instruction. And at the schools I was working at and high school level, we had regular honors, AP and IB. Unfortunately those lower level classes became all about housing students and managing them. Instruction was very poor and students could tell. They put less experienced teachers in there and it was viewed as hazing for those teachers. They weren’t well equipped to handle the students. They didn’t have the same robust lessons. That’s what happened when we put such a strong emphasis on data. We started sorting kids and all the Black kids were in the remedial class. And after eight years of doing that, we had a more progressive admin come in. They were like, ‘This is terrible. We have to stop doing this. It’s racist.’”

—Chicago educator

Troublingly, they described how the narrowing of focus at times also meant they focused on serving a specific subset of students who were most likely to achieve proficiency on the tests.

Some teachers shared how they felt pressure to resort to punitive disciplinary measures, kicking students out of the classroom instead of taking the time to de-escalate and address their needs. They described efforts to push kids out of the classroom or out of the school entirely to not have their test scores counted.

“With that increase in sense of urgency, we lost the space in classrooms to deal with social emotional needs of kids. Especially if you’re in a neighborhood that experiences trauma at a higher rate than you would in other neighborhoods, I really believe in the importance of having space in classrooms. When a lot of these policies were instituted, that space got taken away by our administrators because they needed us to hit these marks. I remember my admin coming in one day when I was trying to de-escalate a kid, basically telling me that we don’t have time for this and that they would just take that child outside of the classroom. Really it was a situation that if you had given me the five minutes that I needed to get them back on track, that student could have stayed for the lesson. But now the child was out of the classroom and continued to be in and out of the classroom for the rest of the year.”

—Chicago educator

“If we talk about state grades, most of the schools in our state who have A’s are schools that control their population. They’re either private or charter, which means they can expel a certain demographic of kid, or they can remove a certain type of kid. Public schools don’t have that luxury. As another example, Carmel has building codes, so they restrict a certain type of housing being built in the area and that’s how they keep certain kids out.”

—Indiana educator

“I have heard many, many stories of local special education directors and other leaders in districts talking to me about data manipulation and stigma, with educators being scared to have kids with disabilities in their classrooms. Teachers may want the students in their class, but when it comes to their scores they say, ‘I don’t own that kid. I don’t want to own their results. That’s somebody else’s kid.’”

—Lindsay Jones

Emotional Toll

What came through most viscerally in these interviews, particularly with educators, was a sense of deep anxiety and angst related to the perceived consequences of being labeled as "failing," particularly in such public fashion. There was also a perception that labeling schools as "failing" or labeling students as "not proficient" or using the term "achievement gap" wrongfully placed blame on students and communities.

"In 2005, I feel like when I started out my career having that whole 100% proficiency in math and English was just kind of hanging over my head. As a newer teacher, it was an impossible metric. To start out your career and have these people be like, 'Okay, we're going to publicize these results. And you have to get 100% ... ' You knew right from the gate that wasn't going to happen, and especially because with experience comes learning and growth."

—Michigan educator

"The kind of mass branding of a community's schools as under performing added another layer to the narrative of inadequacy. There's yet another narrative that the same communities—those for whom the work was designed to uplift, give access to, and increase the chances of—are the ones who are at fault. It's just what our country does. It has turned back to a 'blame the victim' kind of narrative."

—Dr. Sonja Santelises

Though none of the stakeholders we spoke with personally experienced negative consequences of accountability (such as losing a job, having a school closed, etc.), the looming threat of consequences and the underlying psychological impact of not doing a "good enough" job seemed ever present in their minds. They most frequently attributed the pressure to ineffective administrators who used threats and shame to try to motivate teachers to improve.

"I just remember hosting these meetings with teachers who were afraid of getting fired because they weren't meeting the standards. They were really trying to fill those gaps of where their kids were really struggling. They had a major fear of losing their job and not having a plan B."

—Chicago educator

"There was some aggressiveness on the part of school leaders that was caused by the very toxic environment that was being created by policies. Because I've heard this from several principals, 'If I lose the best teachers in my building, I'm going to have a failing school. How are my students going to be able to get what they need if the system takes away all of my good teachers because we didn't have a good rate on our test scores?' And so it went from, 'We want to help students' to 'We have to game the system because we need to have a better reflection in our test results.'"

—Jason B. Allen

"Everyone in my grade level got called down because we had just done our reading tests, and our admin basically yelled at everyone because for the most part, none of our kids made it to their proficiency scores. There was no celebration of, 'did they actually grow?' For me, that's the most important question."

—Chicago educator

It's worth reiterating that none of the folks we spoke with had actually personally experienced any sort of "high stakes" consequences of school failures. It's not to say that this did not happen at all, but the rampant fear seems to outpace the reality. Despite this gap between myth and reality, we do caution efforts to talk people out of their fears. These well-intentioned efforts can sometimes backfire and come across as tone-deaf.

"So to blame NCLB for high stakes teacher decisions, isn't grounded in reality. When I have said this before, there have been teachers who brought to my attention that there were places where teachers lost their jobs because their kids weren't performing or whatever. But I have been across this country and I have not seen a major situation where teachers literally lost their jobs because kids' test scores were low. As a matter of fact, if that were the case, America would have no teachers in whole cities. If teachers lost their jobs because of test scores being bad, there would be entire cities without teachers, because the test scores were bad back in the day and they're still bad today, right?"

—Chris Stewart

"Even when they took away tenure in Michigan and said like, 'Oh, we can fire people.' They never fired people because it was too much paperwork. They didn't have time to deal with it. So you just kept them in because it was easier unless they resigned, or quit, or whatever."

—Michigan educator

Some educators also hypothesized that these feelings of being shamed, the perceived standardization of teaching, and the removal of joyful elements of their jobs led to increased job dissatisfaction, burnout, and teacher turnover—particularly in low performing schools.

"It has an enduring negative value. I think we've seen over the last 15, 20 years a lot of teachers leaving the profession. Money is not really the main factor since that's not the main motivation for people to go into teaching. From personal experience and talking with all my friends who are teachers, I think it's the effect of these programs which took a lot of the power away from teachers."

—Chicago educator

"Being a parent now, I've learned to put more value on separating my work from my home life. Before I had kids and I was a teacher, I remember going to bed every night sick with anxiety that I wasn't doing a good enough job. I do believe that high accountability contributed to that. And I just don't have the energy for that."

—Chicago educator

Both educators and parents shared stories of students experiencing emotional tolls from testing. They described unfair burdens being placed on children, as well as examples of how increased pressure to perform on tests led to dissatisfaction with school.

"I was tested in school, but I don't feel like they carried the weight that they did for my children. My children knew that school funding was tied to their ability to do well on the test. That was something that was shared with them, so they knew that that carried a weight for their teacher and for their school."

—Michigan parent

"When you're teaching third grade and you've got kids that don't want to show their parents a grade of 89 because they're going to get punished, I mean it's awful. I see it all the time. Parents have this incredible expectation for performance."

—New York educator

"We had a remediation class specifically designed for students who had failed the GRE. These students had to keep taking the test for three years. It really impacted their ability to believe in themselves and to think that education had value because it all became about their low scores on that test."

—Indiana educator

"For kids who didn't have a track record of success with school, if they're coming into a space that has metal detectors and not the friendliest of people greeting them at the door and then add on the culture around testing, there's an impact on their thinking and feelings towards school. They don't want to be there."

—Chicago educator

Gaming the System

Faced with enormous pressure from administrators and potential job consequences, a few educators took unethical actions. Multiple stakeholders brought up the 2009 Atlanta cheating scandal as the most high-profile example of the lengths some people would go to in order to succeed on the tests. Some folks shared the perspective that high pressure accountability systems will always lead some people to cheat.

"Campbell's Law says basically, when you say 'this is the measure by which you will be judged or held accountable to something,' because people are people, the measure can become compromised. People will figure out a way to try to be successful on the measure, even if it means doing something that is unscrupulous. And then the very thing gets corrupted. So here, there's money that's at stake. Nobody wants to run a school or be in a school system where you might lose a quarter of your budget based on test performance. It's pretty high stakes."

—Dr. Jennifer Brown

"But as soon as we start to tie accountability to it, to tie funding to it, to tie job retention to it, you automatically make it high stakes. You'd be a ridiculous teacher not to want to help your kids thrive on it. And you've seen evidence of teachers being so stressed out about it, that they cheat. We've had whole systems actually caught in cheating scandals, and that's simply because of the accountability. We didn't have that 30 years ago. You have that today because of the accountability you put into a system."

—California educator

Standardization of Teaching

While some educators celebrated the increased resources and proliferation of best practice sharing that came with shared standards and assessments, many educators bemoaned the idea that teaching became more rote. Many educators talked about how as they were provided curricula and planning guides from their districts, they felt less autonomy in designing lessons to meet their students' needs. Some even described expectations from their administrators to be on the same page of the same textbook each day. They discussed how challenging it was to differentiate to meet students' varying needs while following a regimented curriculum calendar.

"When teachers are forced to say, 'We have to stay on this standard, we have to stay on this timeline,' one, it's irresponsible because that's not how children learn. Two, it is unrealistic to think that in 45 to 50 minutes a teacher who has 27 to 35 students in a classroom is going to be able to successfully model one lesson five different times and meet the needs for all of those students who some are on grade level or above, some are one grade level behind, and then others are two or three grade levels behind. We would not do that to a chef that is making a pound cake and saying, 'Now, we want you to bake a pound cake. But at the same time, we want you to also put in the oven cupcakes, and we want you to throw in an apple pie.'"

—Jason B. Allen

"I vividly remember what happened in school that day when Common Core was rolled out. It was towards the end of the school year, so teachers were throwing all the materials that they had created, all their hands-on materials that they had purchased for the classroom, and they were just throwing them in the hall. And they were just saying, 'We have to read from a script. We are now robots. We are required now that we can only follow the Common Core curriculum.'"

—New York educator

Some teachers critiqued the very idea of having common standards and assessments for all students in a state. They pointed out how biased assessments can be depending on the identities and biases of the test creators themselves, the cultural context of the questions, and the language used. They questioned how inclusive the standards are of their students' diverse identities and histories. They wondered about using multiple ways of showing learning as opposed to a multiple choice assessment.

"There have been children left behind for 400 years. It's not just since 2002. Children of color, communities of color have been left behind for hundreds of years. So when we think about Common Core, who is it for? Is it Western Eurocentric bodies or are we thinking about people of color?"

—Los Angeles educator

"So, if we're going to differentiate learning, why aren't we differentiating assessment? I do really believe each child has their unique skills and abilities, and all of that. And we're supposed to teach to that. And teaching encompasses assessment."

—New York educator

Other advocates remarked that the valid critique of bias in assessments doesn't mean that assessments aren't necessary. Instead, they would like to see efforts to address the biases, but keep the tests.

"The tests are biased. That doesn't mean that we don't need tests. We should address some of the biases that are in tests and we still need to test children."

—Kenya Bradshaw

"There's evidence that there is some bias in the tests themselves, in the measures themselves. But from a systems approach, even if we eliminated all the bias in the tests, we would still have this reproduction problem. Because even if, just make up a number, let's say that 10% of a score is a product of bias. And if we could eliminate that 10% of systematic error that is a product of bias, the tests are still an indication of what students are able to do, which are a product of the education that they've received. So even if you get rid of the bias, you're still going to see differences in scores across groups because we have differences in opportunities across groups."

—Dr. Michael Russell

Purpose of Assessments

Of the debates about the impacts of testing, an underlying question remains about the purpose of annual state tests. Some stakeholders shared their belief that these assessments should be relevant to educators for the purposes of informing instruction. Others indicated that these assessments should not strive to fulfill that purpose, and should instead focus on the purpose of accountability—enabling systems-level comparisons of outcomes across student populations.

"If tests are to assist teachers to improve instruction we need better formative assessments and we need to shorten the turnaround time of summative assessments so that they guide school based decisions more efficiently."

—Kenya Bradshaw

"If you want to use any of these large scale summative tests to inform classroom practice, none of them are going to be particularly useful because that's not what they're designed to do. They're not designed to provide student level information at a level of specificity that's going to be useful for classroom instruction. And there's this tension between educators wanting that level of information and yet not wanting to give up time for testing. And if you want that level of specificity, it's going to take a lot more testing time, or we're going to have to do testing much more frequently throughout the year, and the content of that test needs to be more aligned with instruction. But in order to do that, then you need what different schools are doing to be similar, so that when you're administering this sub-test, it's aligned with what kids have been learning over the last three or four weeks or whatever the period of time is. And if you allow for variation in curriculum and pacing, then either you have to give the teachers flexibility to pick which test they give and when, which is fine, but it's a very different approach ... So it's complicated when people want tests to not take time yet, provide a lot of information. You can't have both."

—Dr. Michael Russell

Money Not Reaching Students

A lesser known element of NCLB and its predecessor legislation was the influx of funding that came from the federal government to implement said reforms. Most people simply did not know that there was a massive investment to accompany NCLB. Among those who did recognize this and lauded the scale of investment, many wondered where the money ended up going. There was a widespread impression that if there was more funding flowing into the education system, it wasn't reaching students.

"When the unions characterized the law as all hammer and no help, they were actually wrong. There was a lot of money in the law that was intended to help building-level people to meet the goals of the law. The big problem was that the first years of law coincided with a massive decline in state revenues for education. And so, when you looked at education budgets by state, what you realized is the dollars were close to flat. They were not as improved as they should have been with the extra No Child dollars. And so, as we know, what school districts do when they get money like that from the feds is they backfill into their existing budget. They don't add the extras that they were supposed to be adding, either for kids or for educators. And so, the lived experience of educators was, 'I had this hard thing to do. They had hammers over my head and labels to put on my face, but no help.'"

—Kati Haycock

"I think that money was spent in so many foolish ways that didn't improve instruction, and not always maliciously. People genuinely didn't know what to do. That said, (thanks in part to NCLB) we have been blessed with a good amount of research the last 10 years that tells us what does work. And we have a better system for getting that information out. Ideally, the money would follow and be spent on what works, but it doesn't mean everybody's responding that way."

—Delia Pompa

An associated frequent critique of federal testing is that it spawned a multi-billion dollar testing and curriculum industry, as well as for-profit charter schools. Many stakeholders questioned the role these entities should play in public education and rebuked the perverse incentives that can come with a profit motive.

"All of these people have a stake in the game. Look at the testing industry. The standardized testing industry went from a million dollar industry to a billion dollar industry. Teachers are still out here having to ask for resources. Children are still not learning the standards the way that they need to, but everyone who had a stake in the game and education, they profited off these students who have been failed."

—Jason B. Allen

"The for-profit model of charter education has had a huge destabilizing effect in Michigan. They come into urban areas, they churn through teachers, and students, and buildings, and funding. There's a real glut that hasn't been good."

—Michigan educator

Others pointed to the growth of non-instructional administrative and consulting roles that they felt grew disproportionately as a result of the funding. They saw more coaching positions and district office roles, perceived as layers of administrative bureaucracy as opposed to authentic support for teachers. They experienced a carousel of consultants coming in and out of schools to offer strategic planning or professional development, which were characterized across the board as ineffective.

"One of the things that we've been challenging is the idea of consultants working with schools. They're very expensive. It can be $70,000 to hire a consultant who comes once a month for one or two days. And I think that that's a waste of money. And I think it's part of something that came about during No Child Left Behind, that's a holdover, that really is unhelpful to schools. Having somebody there one day a month, it really has not impacted or improved things. And I really see that as actually getting in the way of progress, because that money should go towards people that are actually in the schools locally all the time."

—Yvonne Field

Political Backlash and Subsequent Regression

Unsurprisingly, the deeply-felt negative consequences led to significant political backlash, organized predominantly by proponents of decentralization and later joined by teachers and their unions. This uneasy political alliance between the status quo left and the radical right emerged as a formidable force that pro-accountability advocates still face today.

"And for a variety of reasons, there began to be a strong resistance to it, an opposition to it. And the promoters, defenders of it, the people who had been involved in making it happen, many of them simply left the scene for a variety of reasons. They began to do other things. They tired out. They were beaten down. Those who support the status quo, those who don't want the pressure, don't want to be measured, don't want to have to be accountable to anybody, they began to be ascendant. Their views began to be dominant. We moved into a period where they were able to take over the policy reins. The landscape changed and it was not good for students."

—Sandy Kress

"There was the hopeful vision that eventually had to be implemented in the real world. As that happened, complexities arose, which made it easier to start changing the narrative from the hopeful one to a problematic one. There was a real effort to problematize everything having to do with standards, accountability, testing, outcome data—basically NCLB in total. And if your only goal is to make this look like a problem, boy, do you have so much fuel, because you're talking about big complex systems and there is plenty of stuff to make an issue. It didn't work right away because we had the narrative of 'Beat the Odds,' which included the hopeful schools, the brilliant charter schools getting the job done with kids, the 'Waiting for Superman' type of ethos. But slowly but surely, the hope community that was rising above the fray and doing a great job or whatnot, they waned. So then came, 'Charter schools are problematic. Testing is problematic. Measurements are problematic. Reform is problematic. Reformers are problematic. Funders are problematic,' blah, blah, blah, blah. That eventually won. A cancerous narrative metastasized while NCLB-ers were still trying to just push forward with their moonshot."

—Chris Stewart

"NCLB was destined to crash because every student was supposed to be reading at grade level by the 2014 deadline. When almost no schools could meet that goal, the law said those schools were failing. That built a huge movement to finally get rid of NCLB. The old testing regime became controversial. Teachers didn't like it. Parents, administrators, Republicans, and Democrats did not like it. No one liked it. Then Common Core became its own boogeyman."

—Rep. John Kline

Some stakeholders pointed to the pushback that began or crescendoed when privileged classes of students began experiencing ill impacts.

"What I found so ironic in the development of the school evaluation tool is that when they first rolled out the SQRP, there were a number of schools on the Northwest side of Chicago that had poor ratings because of growth measures. There was pushback from affluent parents and administrators of those schools. And CPS ultimately changed the policy so that if the school had a certain number of children performing at grade level, they no longer were held accountable to the growth metric."

—Chicago principal

Others described parochialism and indignation from local stakeholders based on their perceptions of the top-down nature of accountability.

"I would sit at State Board of Education meetings and listen to the board members rail against Common Core. I would think, 'Well, if we just called them the New Jersey State Standards instead, then it would be okay.' They really didn't have issues with the content of the standards. They had issues of the standards being perceived as national standards."

—New Jersey educator

On the opt-out movement, stakeholders shared about the dynamic relationships between teachers, students, and families driving local rates.

"A lot of kids do not take the test seriously. Teachers have been largely dismissive and students see that. There's a pretty high opt out percentage in urban cities like Chicago, where in the last eight years there's been such a backlash to measuring students by testing. Then during the pandemic, they just haven't done it at all."

—Chicago educator

Now, in our new post-Covid educational reality, we have seen even greater landscape shifts that again call our data and accountability systems into question. Faced with growing evidence of the monumental learning losses and deepened mental health crises, many people would rather tune out any talk about data and academic outcomes. Coupled with waivers from federal accountability requirements and a more polarized political climate, there are great risks to sustaining the progress we have fought for over two decades.

"There's a mentality with some parents and teachers who feel like, 'Well, we sent everybody home and some people were supposed to be remote and they never came, but they still went on to the next grade. So how important is school? How important are these tests?' So I think they're giving the kids this idea too. They're like, 'Don't worry about it. It doesn't matter. These days, just do whatever you can.' Nobody cares. Kids are finishing tests in 10 minutes and not really doing much the rest of the day. I don't think they're reliable or valid really anymore, if they ever were."

—New York educator

"I do feel a lot of watering down has happened. Expectations have become mushy. I don't want to say there's no accountability for standards anymore, but it's almost like everybody gets a participation ribbon. It's like, 'Oh, we addressed that standard on Tuesday, April 14th, so check. On to the next thing.' It's kind of unfortunate."

—Minnesota educator

"From 2009 up until 2017, I remember reading a lot about data, data, data … hearing about Michelle Rhee and the KIPP schools. When anyone walked into the classroom, they had better see the agenda and the objective on the white board. Students should be able to articulate that, if an observer inquired. There was a lot more emphasis placed on teachers meeting and discussing data. Now, and I don't even think it was just the pandemic, I'm just simply not having those conversations about data anymore. We're having conversations now dominated by social emotional learning and talking about feelings. Like Mondays are supposed to be dedicated to a social emotional lesson. That's fine, but there is no meaningful training. It turns into a throwaway lesson or just a worksheet. I don't think it is instructional time well spent. You can't collect any real data from that either. So it's not like you are discovering these profound truths about your students and how to help them."

—Chicago educator

The Work Ahead

With the political landscape evolving away from the bipartisan agreement that enabled NCLB, there is challenging work ahead for advocates of strong data and accountability systems in public education to regain momentum lost over the past few years.

Certainly there is no shortage of think pieces reflecting on what lessons we can draw from implementation over the past two decades, so we won't attempt to rehash all of the lessons here. We did ask stakeholders for their thoughts on the work ahead, so we share a few reflections below.

We hope these takeaways inform multiple dimensions of a winning issue advocacy campaign: a research-based, practical policy agenda; an electoral strategy to cultivate champions in positions of power; and robust public engagement efforts to build and sustain a constituency that will take action. Given brightbeam's focus on building public proficiency in and demand for quality educational opportunities, we will hone in on a few reflections stakeholders shared that have greatest relevance to that end. These include intentional stakeholder buy-in, a compelling and multi-faceted communications strategy, and long-term leadership.

Stakeholder Buy-In

When it comes to effective implementation, there's little that can match the impact of having goodwill of key players, which in this case includes teachers, students, and their families.

For students, families, and community members, the proposition of buying into the promise of greater educational opportunities seems valuable on its face. They are often an untapped resource, rarely asked to meaningfully participate in education conversations. Here, building the requisite trust that their input will be taken seriously is the main challenge. This takes strong relationships on the ground level, as well as consistently designed engagement strategies over time.

"Kids will tell you exactly what needs to happen in schools for them to learn better. I've had kids talk about, 'Well, I just didn't learn the way that teacher was teaching. I needed a different way to learn.' I had another student say, 'I could really tell that my teacher didn't love teaching. And if you don't love teaching, don't be a teacher.' And so those kids, and the parents, and those people actually in schools, if we really listen to them at the beginning of our initiatives when we're trying to do something new, I think that we would get to a better place."

—Yvonne Field

"Every four years, when it's a governor's election, then we hear about the state of education. Then there's a big political infighting about who knows best for our children, or how our children rank with students from across the United States. But other than that, we don't hear a whole lot about education."

—Michigan parent

A harder question is—how do you increase teacher buy-in when they are the ones who are ultimately being held accountable? Teacher advocates would suggest authentically engaging teachers in the design process, working collaboratively with unions, ensuring valid critiques are recognized and remedied. But, execution and political dynamics make it challenging to get and keep educators committed to and through implementation. With teachers oft-cited as the most trusted source of education information for parents, students, and general community members, it's worth the effort even if ultimately, this may not be a winnable audience at scale.

"One of the things that happens when you're trying to move at scale is if you don't do it right, you can fundamentally disempower teachers. I think that there is a tension there because when you're trying to move things forward, sometimes you mandate change in ways that take voice out of that equation. Sometimes that's really necessary if you're going to move things at scale. But it also raises the question—is that moving at that scale an actual way to make progress? I think that progress needs to come from communities and driven by communities and needs of communities."

—California educator

"Usually you hear theory dense, and practice thin. And I think K-12 is actually something the opposite, right? It's very theory thin and practice dense. And I think you need both of these pieces. You need to both have practice and theory. Marry the two, so that teachers, teacher educators, policymakers, families, community members, they all understand there's a political historical context that is occurring while all of these policies are in play."

—Los Angeles educator

"What we might be learning from all of this is that the keepers of the system can't be trusted to operate in good faith. I'm not talking about all teachers or all principals or all leaders. There are many, of course, who are great and who do a good job. We were ready. We were ready to modify the law to the extent that it needed to be improved and to deal with these unintended consequences, some of which were real and many of which were not. Margaret Spellings, the supporters of reform, Democrats and Republicans in Congress—we were all ready. I'm talking about a segment who, for the most part, are in charge of making policy and leading the way for the system. There's a sentiment among many of them that is resistant to standards and accountability. They'd rather not have it or much of it. So, with the support of various partners, they've pushed policies that have been adopted by policymakers that have eviscerated accountability. And, largely, as a result, children's achievement has suffered."

—Sandy Kress

Communications Strategy

Advocates, educators, and families all shared the belief that one of the biggest challenges to data and accountability efforts has been clear, consistent, compelling communications delivered by the right messengers.

We believe that a strong communications strategy is one with a few top-line messages that are repeated early and often by all messengers, across a variety of platforms. This can then be followed by tailored messages delivered by credible messengers to key audiences needed to move specific targets.

"Before implementation even began, the unions and administrators associations got out in front to characterize it as this horrible effort. All hammer, all stick, no carrot. It had these draconian sanctions. All schools would be failing. The administration was quite flat-footed without either the relationships, or they just didn't spend the money to try to communicate the positive story to parents, in particular, of all this new data or attention that was coming their way for them. There was nothing that was effectively messaged to either educators or parents on the positives of all this. So, what happened is stories started popping up of children who were crying because the tests were too hard, educators who were cheating, and educators who were just focusing on kids right under the bubble."

—Kati Haycock

"I would freaking hire a communications firm for some messaging help. I don't know how to keep it out of the political arena. The same thing's happening right now with CRT, it's just stupid. It's unfortunate how some of that really good effort was sacrificed because it became politicized."

—Anna East

Leadership

Another clear takeaway is the need for persistent leadership from respected people who are willing to work through challenges together. Leadership must come in many forms. It must include behind the scenes leadership in the form of a coalition of advocates working collaboratively to design and implement various elements of strategy. It can also come from prominent, trusted messengers who will carry the right messages forward to reach intended targets. It must come from political leaders who will negotiate compromises and keep diverse stakeholders at the table.

"Keeping the coalition together was a problem that we all had. Kennedy and Bush went onto other things. The people who put NCLB together, went on to other things. Of course, President Bush had a pretty good excuse. We were in a war. We were involved in a fight against terrorism. Could he have done more to have kept a working agreement, a working relationship going? I think that's a fair point made by Democrats. Senator Kennedy went on to other things. And so they weren't around to keep the cats in the same room, which I think hurt in all of this. Then some of the business groups, including the Chamber and the Business Roundtable, moved on to other issues. I think there was a feeling among some business leaders that we did it. Time to move on to the next issue. And what was the next issue? The next issue at the time was supporting STEM initiatives. Now, STEM is darn important. But what was happening was people were beginning to focus on inputs and move away from a focus on outputs on this reform strategy."

—Sandy Kress

"This idea of what happens when the centrists leave, what happens when the reasonable, technocratic, scientific thinking, bipartisan group, when they exit the room, what happens, I think, is the most ominous question for us to answer, and I think it's the one that has the most kind of critical input on our future. Like right now, we're already starting, I think, to not talk about it, but we do know that there's an unraveling of outcomes, like something bad is going on that is going to look way worse when we look back on it in retrospect in the future. And that is going to be the bipartisan dummies replacing the bipartisan smart people, the centrists."

—Chris Stewart

Conclusion

From this report, we hope student advocates take heed to not “throw the baby out with the bath water.” We should learn from NCLB’s unintended consequences and the political narratives that led to its demonization. We should also not take its impact for granted. While opponents may have gained traction by demonizing assessments and accountability efforts broadly, their case is a shallow one. Too many students and their families, particularly those from traditionally underserved communities, have experienced important advances that are worth protecting and building upon. We hope advocates take heart that their cause is not only a righteous one, but with the right policy choices and public engagement efforts, it is also a winnable one.

We have come away from this project sensing both urgency and inspiration about the work ahead to remedy the backsliding of progress we have seen after two decades of incremental gains for students. We see opportunity ahead for USCCF and other partners to shape a robust public engagement strategy to inspire Americans to fight for safe, affirming, and liberating educational options where every child learns and thrives.

References

What Does Empirical Research Say About Federal Policy From NCLB to ESSA

Arold, B., & Shakeel, M. D. (2021) The Unintended Effects of the Common Core State Standards on Non-Targeted Subjects. Available at SSRN: https://ssrn.com/abstract=3868271 or http://dx.doi.org/10.2139/ssrn.3868271

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135. https://doi.org/10.1086/508733

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76. https://doi.org/10.3102/0162373716663646

Ahn, T., & Vigdor, J. (2014). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina* (Working Paper No. 20511). National Bureau of Economic Research. https://doi.org/10.3386/w20511

Aldeman, C. (2017, January 18). The teacher evaluation revamp, in hindsight. *Education Next*, *17*(2). https://www.educationnext.org/the-teacher-evaluation-revamp-in-hindsight-obama-administration-reform/

Aldeman, C., & Chuong, C. (2014). *Teacher evaluations in an era of rapid change: From "unsatisfactory" to "needs improvement."* Bellwether Education Partners. https://eric.ed.gov/?id=ED553852

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*(2), 275–318. https://doi.org/10.1086/683665

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2010). Inputs and impacts in charter schools: KIPP Lynn. *American Economic Review*, *100*(2), 239–243. https://doi.org/10.1257/aer.100.2.239

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP? *Journal of Policy Analysis and Management*, *31*(4), 837–860. https://doi.org/10.1002/pam.21647

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27. https://doi.org/10.1257/app.5.4.1

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely* (Working Paper No. 26463). National Bureau of Economic Research. https://doi.org/10.3386/w26463

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 233285841560783. https://doi.org/10.1177/2332858415607834

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (Working Paper No. 20657). National Bureau of Economic Research. https://doi.org/10.3386/w20657

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The Common Core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, *62*, 48–65. https://doi.org/10.1016/j.econedurev.2017.10.004

Backes, B., & Hansen, M. (2018). The impact of Teach for America on non-test academic outcomes. *Education Finance and Policy*, *13*(2), 168–193. https://doi.org/10.1162/edfp_a_00231

Ballou, D., & Springer, M. G. (2017). Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains. *Education Finance and Policy*, *12*(1), 77–106.

Barnum, M. (2017, May 10). 74 Interview: Harvard researcher David Deming takes the long View on Head Start, Integration. *The 74*. https://www.the74million.org/article/74-interview-harvard-researcher-david-deming-takes-the-long-view-on-head-start-integration/

Barnum, M. (2021, October 13). *The latest Nobel Prize winner: Researcher who helped show money matters for schools.* Chalkbeat National. https://www.chalkbeat.org/2021/10/13/22724766/economics-nobel-prize-education-research-school-spending

Barnum, M. (2022, March 31). *Has the federal government underestimated the progress of high school students for decades?* Chalkbeat National. https://www.chalkbeat.org/2022/10/24/23417139/naep-test-scores-pandemic-school-reopening

Barnum, M. (2022, October 23)). *Nation's report card: Massive drop in math scores, slide in reading linked to COVID.* Chalkbeat National. https://www.chalkbeat.org/2022/3/31/23005371/high-school-test-scores-underestimate-naep-dropout-nces

Bay-Williams, J. M., Duffett, A., & Griffith, D. (2016). *Common Core math in the K–8 classroom: Results from a national teacher survey*. Thomas B. Fordham Institute. https://fordhaminstitute.org/national/research/common-core-math-k-8-classroom-results-national-teacher-survey

Berliner, D. C. (2013). Problems with value-added evaluations of teachers? Let me count the ways! *The Teacher Educator*, *48*(4), 235–243. https://doi.org/10.1080/08878730.2013.827496

Bidwell, A. (2014a, March 6). The politics of Common Core. *U.S. News & World Report*. https://www.usnews.com/news/special-reports/a-guide-to-common-core/articles/2014/03/06/the-politics-of-common-core

Bidwell, A. (2014b, January 31). More states seek to repeal Common Core. *U.S. News & World Report*. https://www.usnews.com/news/articles/2014/01/31/more-states-seek-to-repeal-common-core

Birman, B. F., Boyle, A., Le Floch, K. C., Elledge, A., Holtzman, D., Song, M., & Yoon, K. S. (2009). *State and local implementation of the No Child Left Behind Act* (Volume VIII— Teacher quality under NCLB: Final report). U.S. Department of Education. https://www2.ed.gov/rschstat/eval/teaching/nclb-final/report.pdf

Black, S. E., & Machin, S. (2011). Housing valuations of school performance. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 485–519). Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00010-7

Bleiberg, J. (2021). Does the Common Core have a common effect? An exploration of effects on academically vulnerable students. *AERA Open*, *7*, 233285842110107. https://doi.org/10.1177/23328584211010727

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2021). *The effect of teacher evaluation on achievement and attainment: Evidence from statewide reforms* (EdWorkingPaper No. 21-496). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-496

Bloom, H., Ham, S., Melton, L., & O'Brien, J. (2001). *Evaluating the accelerated schools approach: A look at early implementation and impacts on student achievement in eight elementary schools* [Report]. Manpower Demonstration Research Corporation. https://www.mdrc.org/publication/evaluating-accelerated-schools-approach

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *The Future of Children*, *17*(1), 45–68. https://www.jstor.org/stable/4150019

Boyer, E.L. (1983) *High School: A Report of the Carnegie Foundation for the Advancement of Teaching*. New York: Harper & Row.

Bruno, P., & Strunk, K. O. (2019). Making the cut: The effectiveness of teacher screening and hiring in the Los Angeles Unified School District. *Educational Evaluation and Policy Analysis*, *41*(4), 426–460. https://doi.org/10.3102/0162373719865561

Burns, J., & Strunk, K. O. (2021). School accountability. In Brian P. McCall (Ed.), *The Routledge Handbook of the Economics of Education*. Routledge.

Burtless, G. (1996). Does money matter? *Brookings*. https://www.brookings.edu/book/does-money-matter/

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Camera, L. (2021, October 13). America's kids earn disappointing grades on nation's report card. U.S. News and World Report. https://www.usnews.com/news/education-news/articles/2021-10-14/americas-kids-earn-disappointing-grades-on-nations-report-card

Candelaria, C., & Shores, K. (2017). *Court-ordered finance reforms in the adequacy era: Heterogeneous causal effects and sensitivity* (SSRN Scholarly Paper ID 3009784). Social Science Research Network. https://papers.ssrn.com/abstract=3009784

Card, D., & Payne, A. A. (2002). School finance reform, the distribution of school spending, and the distribution of student test scores. *Journal of Public Economics*, *83*(1), 49–82. https://doi.org/10.1016/S0047-2727(00)00177-8

Carlson, D., & Lavertu, S. (2018). School Improvement Grants in Ohio: Effects on student achievement and school administration. *Educational Evaluation and Policy Analysis*, *40*(3), 287–315. https://doi.org/10.3102/0162373718760218

Carmichael, S. B., Wilson, W. S., Porter-Magee, K., & Martino, G. (2010). *The state of state standards—and the Common Core—in 2010*. Thomas B. Fordham Institute. https://fordhaminstitute.org/national/research/state-state-standards-and-common-core-2010

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331. https://doi.org/10.3102/01623737024004305

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(43), 17176–17182. https://doi.org/10.1073/pnas.1315746110

Chambers, J. G., Lam, I., Mahitivanichcha, K., Esra, P., Shambaugh, L., & Stullich, S. (2009). State and Local Implementation of the No Child Left Behind Act. Volume VI-—Targeting and Uses of Federal Education Funds. *US Department of Education. https://eric.ed.gov/?id=ED504207*

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Chingos, M. M. (2013). Class size and student outcomes: Research and policy implication. *Journal of Policy Analysis and Management*, *32*(2), 411–438. https://doi.org/https://www.jstor.org/stable/42001539

Chingos, M. M. (2018). *What matters most for college completion. Academic preparation is a key predictor of success.* American Enterprise Institute; The Third Way Institute. https://www.aei.org/wp-content/uploads/2018/05/What-Matters-Most-for-College-Completion.pdf?x91208

Cohen, D. K., & Moffitt, S. L. (2009). *The ordeal of equality. Did federal regulation fix the schools?* Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674035461

Cohodes, S. R., & Parham, K. S. (2021). *Charter schools' effectiveness, mechanisms, and competitive influence* (Working Paper No. 28477). National Bureau of Economic Research. https://doi.org/10.3386/w28477

Collins, C. A., & Kaplan, E. K. (2022). Demand for school quality and local district administration. *Economics of Education Review*, *88*, 102252. https://doi.org/10.1016/j.econedurev.2022.102252

Cowan, J., Goldhaber, D., & Theobald, R. (2022). Performance evaluations as a measure of teacher effectiveness when implementation differs: Accounting for variation across classrooms, schools, and districts. *Journal of Research on Educational Effectiveness*, *15*(3), 510–531. https://doi.org/10.1080/19345747.2021.2018747

CREDO. (2009). *Multiple choice: Charter school performance in 16 states*. Center for Research on Education Outcomes. https://credo.stanford.edu/wp-content/uploads/2021/08/multiple_choice_credo.pdf

Cremata, E., Davis, D., Dickey, K., Lawyer, K., Negassi, Y., Raymond, M. E., & Woodworth, J. L. (2013). *National charter school study 2013*. Center for Research on Education Outcomes. https://credo.stanford.edu/wp-content/uploads/2021/08/ncss_2013_final_draft.pdf

Cullen, J. B., Koedel, C., & Parsons, E. (2021). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, *16*(1), 7–41. https://doi.org/10.1162/edfp_a_00292

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In T. Gronberg & D. Jansen (Eds.), *Advances in Applied Microeconomics* (Vol. 14). Elsevier Science.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011). *Getting teacher evaluation right: A background paper for policy makers*. National Academy of Education. https://eric.ed.gov/?id=ED533702

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15. https://doi.org/10.1177/003172171209300603

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher*, *44*(6), 347–358. https://doi.org/10.3102/0013189X15601426

D.C. Board of Education, (2021, March 17). Results from the 2021 D.C. All-Teacher Survey. https://sboe.dc.gov/sites/default/files/dc/sites/sboe/publication/attachments/2021-03-17-FINAL-DC%20State%20Board%20All-Teacher%20Survey%20Report%20%28March%202021%29.pdf

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996. https://doi.org/10.1257/aer.20181169

Dee, T. (2012). *School turnarounds: Evidence from the 2009 stimulus* (Working Paper No. 17990). National Bureau of Economic Research. https://doi.org/10.3386/w17990

Dee, T. S., & Jacob, B., (2010). The impact of No Child Left Behind on students, teachers, and schools. *Brookings papers on economic activity*, 149-207.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418–446. https://www.jstor.org/stable/23018959

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252–279. https://doi.org/10.3102/0162373712467080

Dee, T. S., James, J., & Wyckoff, J. (2021). Is effective teacher evaluation sustainable? Evidence from District of Columbia Public Schools. *Education Finance and Policy*, *16*(2), 313–346.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Dee, T. S., & Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. *Education Next*, *17*(4), 58–66. https://www.educationnext.org/a-lasting-impact-high-stakes-teacher-evaluations-student-success-washington-dc/

Denice, P., & Gross, B. (2016). Choice, preferences, and constraints: Evidence from public school applications in Denver. *Sociology of Education*, *89*(4), 300–320. https://doi.org/10.1177/0038040716664395

Dillon, S. (2009, October 28). After complaints, Gates Foundation opens education aid offer to all states. *The New York Times*. https://www.nytimes.com/2009/10/28/education/28educ.html

Dobbie, W., & Fryer, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3*(3), 158–187. https://doi.org/10.1257/app.3.3.158

Dobbie, W., & Fryer, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, *5*(4), 28–60. https://doi.org/10.1257/app.5.4.28

Dobbie, W., & Fryer, R. G. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, *123*(5), 985–1037. https://doi.org/10.1086/682718

Donaldson, M., & Papay, J. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed.). Routledge.

Dotter, D., Chaplin, D., & Bartlett, M. (2021). *Impacts of school reforms in Washington, DC on student achievement*. Mathematica. https://www.mathematica.org/publications/impacts-of-school-reforms-in-washington-dc-on-student-achievement

Drake, G. and Walsh, K. (2020). 2020 Teacher Prep Review: Program Performance in Early Reading Instruction. Washington, D.C.: National Council on Teacher Quality. Retrieved from: www.nctq.org/publications/2020-Teacher-Prep-Review:-Program-Performance-in-Early-Reading-Instruction.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., Boyle, A., Upton, R., Tanenbaum, C., & Giffin, J. (2017). *School Improvement Grants: Implementation and effectiveness* (NCEE 2017-4013). U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED572215.pdf

Draper (1987, April 22) Many graduates can't read their high school diplomas. *Chicago Tribune.* https://www.chicagotribune.com/news/ct-xpm-1987-04-23-8701310386-story.html

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, *32*(4), 692–717. https://doi.org/10.1002/pam.21715

Education Commission of the States. Task Force on Education for Economic Growth. (1983). *Action for Excellence: A comprehensive plan to improve our nation's schools*. The Education Commission of the States.

Egalite, A. J. (2018). Federal support for charter schooling. In F. Hess & M. McShane (Eds.), *Bush-Obama school reform: Lessons learned* (pp. 125–144). Harvard Education Press.

Elpus, K. (2014). Evaluating the effect of No Child Left Behind on U.S. music course enrollments. *Journal of Research in Music Education*, *62*(3), 215–233. https://doi.org/10.1177/0022429414530759

Fantz, A. (2015, April 14). *Prison time for some Atlanta school educators in cheating scandal*. CNN. https://www.cnn.com/2015/04/14/us/georgia-atlanta-public-schools-cheating-scandal-verdicts/index.html

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics*, *90*(4–5), 837–851. https://doi.org/10.1016/j.jpubeco.2005.01.003

Figlio, D., & Getzler, L. (2006). Accountability, ability, and disability: Gaming the system? In *Advances in Applied Microeconomics* (Vol. 14). Elsevier Science.

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*(9), 1069–1077. https://doi.org/10.1016/j.jpubeco.2009.07.003

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review*, *94*(3), 591–604. https://doi.org/10.1257/0002828041464489

Finn, C. E., Jr., & Hess, F. M. (2022, summer). The end of school reform? *National Affairs, 52*. https://www.nationalaffairs.com/publications/detail/the-end-of-school-reform

Forte, D. (2021, November 11). NAEP is telling us again that it's past time to close long-standing resource gaps. *Eduwonk*. http://www.eduwonk.com/2021/11/why-is-naep-flat-or-falling-with-denise-forte.html

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. John Wiley & Sons.

Fuhrman, S. H., Goertz, M. E., & Weinbaum, E. H. (2007). Educational governance in the United States: Where are we? How did we get here? Why should we care?. In *The State of Education Policy Research* (pp. 41-61). Routledge.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher, 36(5)*, 268-278.

Fusarelli, L. D., & Ayscue, J. B. (2019). Is ESSA a retreat from equity?. *Phi Delta Kappan*, *101*(2), 32-36. https://doi.org/10.1177/0031721719879152

GAO. (2012). *School Improvement Grants: Education should take additional steps to enhance accountability for schools and contractors* (GAO-12-373). United States Government Accountability Office. https://www.gao.gov/assets/gao-12-373.pdf

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). Reading First impact study. Final Report. NCEE 2009-4038. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., & Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED519922

Gill, B. (2021, September 21). Making sure school performance measures provide the right diagnosis to improve student outcomes. *REL Mid-Atlantic*. https://ies.ed.gov/ncee/rel/Products/Blog/30243

Gilraine, M., & Pope, N. G. (2021). Making teaching last: Long-run value-added (Working paper no. 29555). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w29555/w29555.pdf

Ginsburg, A., & Smith, M. S. (2018). *Revisiting SIG: Why critics were wrong to write off the federal School Improvement Grants Program*. FutureEd. https://www.future-ed.org/wp-content/uploads/2018/12/REPORT_Revisiting-SIG.pdf

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study* (NCEE 2010-4027). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED565837

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years: Final report*. Mathematica Policy Research, Inc. https://eric.ed.gov/?id=ED530098

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). *The evaluation of charter school impacts* (NCEE 2010-4030). U.S. Department of Education. https://ies.ed.gov/ncee/pubs/20104029/pdf/20104030.pdf

Goertz, M. E., & Duffy, M. C. (2001). *Assessment and accountability systems in the 50 states, 1999-2000*. CPRE Research Report Series. https://eric.ed.gov/?id=ED450639

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. Educational Researcher, 44(2), 87-95. https://doi.org/10.3102/0013189X15574905

Goldhaber, D. (2019). Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education*, *70*(2), 90–101. https://doi.org/10.1177/0022487118800712

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources*, *32*(3), 505–523. https://doi.org/10.2307/146181

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145. https://doi.org/10.3102/01623737022002129

Goldhaber, D., & Brown, N. (2016). Teacher policy under the ESEA and the HEA: A convergent trajectory with an unclear future. In C. Loss & P. McGuinn (Eds.), *The convergence of K-12 and higher education: Policies and programs in a changing era* (pp. 87-102). Cambridge, MA: Harvard Education Press.

Goldhaber, D., & Dee, T. S. (2017, April 26). Understanding and addressing teacher shortages in the United States. *Brookings*. https://www.brookings.edu/research/understanding-and-addressing-teacher-shortages-in-the-united-states/

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2017). Screen twice, cut once: Assessing the predictive validity of applicant selection tools. *Education Finance and Policy*, *12*(2), 197–223. https://doi.org/10.1162/EDFP_a_00200

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, *100*(2), 250–255. https://doi.org/10.1257/aer.100.2.250

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612. https://doi.org/10.1111/ecca.12002

Goldhaber, D., Krieg, J., Naito, N., & Theobald, R. (2020). Making the most of student teaching: The importance of mentors and scope for change. *Education Finance and Policy* 15 (3): 581–591. https://doi.org/10.1162/edfp_a_00305

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*(5), 293–307. https://doi.org/10.3102/0013189X15592622

Goldhaber, D., & Özek, U. (2019). How much should we rely on student test achievement as a measure of success? *Educational Researcher*, *48*(7), 479–483. https://doi.org/10.3102/0013189X19874061

Goldhaber, D., & Startz, R. (2017). On the distribution of worker productivity: The case of teacher effectiveness and student achievement. *Statistics and Public Policy*, *4*(1), 1–12. https://doi.org/10.1080/2330443X.2016.1271733

Goldhaber, D., Theobald, R., & Fumia, D. (2022). The role of teachers and schools in explaining STEM outcome gaps. *Social Science Research*, *105*, 102709. https://doi.org/10.1016/j.ssresearch.2022.102709

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in US public schools. American Educational Research Journal, 55(1), 171-201. https://doi.org/10.3102/0002831217733445

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Government Accountability Office. (2000). *Stronger accountability needed for performance of disadvantaged students*. (GAO/HEHS-00-89). Washington, D.C.: U.S. Government Printing Office. https://www.gao.gov/assets/hehs-00-89.pdf

Government Accountability Office (2007). *Education should clarify guidance and address potential compliance issues for schools in corrective action and restructuring status* (GAO-07-1035). Washington D.C.: U.S. Government Printing Office. https://www.gao.gov/assets/gao-07-1035.pdf

Greene, J. P. (2016, November 5). Evidence for the disconnect between changing test scores and changing later life outcomes. *Jay P. Greene's Blog*. https://jaypgreene.com/2016/11/05/evidence-for-the-disconnect-between-changing-test-scores-and-changing-later-life-outcomes/

Griffith, D. (2021). *NAEP shows recent achievement flat but longer-term progress made*. ASCD. https://www.ascd.org/blogs/naep-shows-recent-achievement-flat-but-longer-term-progress-made

Gross, M., Shiferaw, M., Deutsch, J., & Gill, B. (2021). *Using promotion power to identify the effectiveness of public high schools in the District of Columbia* (REL 2021-098). U.S. Department of Education, Institute of Education Sciences, Regional Educational Laboratory Program. https://ies.ed.gov/ncee/rel/Products/Publication/30249

Hanushek, E. (2009). The economic value of education and cognitive skills. In G. Sykes, T. Ford, D. Plank, & B. Schneider (Eds.), *Handbook of Education Policy and Research* (pp. 39–56). Routledge.

Hanushek, E. A. (2003). The failure of input based schooling policies. *The Economic Journal*, *113*(485), F64–F98. https://doi.org/10.1111/1468-0297.00099

Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2020). *Long-run trends in the U.S. SES-achievement gap* (Working Paper No. 26764). National Bureau of Economic Research. https://doi.org/10.3386/w26764

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327. https://doi.org/10.1002/pam.20091

Hashim, S. A., Kane, T. J., Kelley-Kemple, T., Laski, M. E., & Staiger, D. O. (2020). *Have income-based achievement gaps widened or narrowed?* (Working Paper No. 27714). National Bureau of Economic Research. https://doi.org/10.3386/w27714

Hastings, J. S., & Weinstein, J. M. (2007). *Information, school choice, and academic achievement: Evidence from two experiments* (Working Paper No. 13623). National Bureau of Economic Research. https://doi.org/10.3386/w13623

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482. https://doi.org/10.1086/504455

Heissel, J. A., & Ladd, H. F. (2018). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, *62*, 302–320. https://doi.org/10.1016/j.econedurev.2017.08.001

Henry, G. T., McNeill, S. M., & Harbatkin, E. (2022). Accountability-driven school reform: are there unintended effects on younger children in untested grades?. *Early Childhood Research Quarterly*, 61, 190-208. https://doi.org/10.1016/j.ecresq.2022.07.005

Hess, F. M., & McShane, M. Q. (Eds.). (2018). Bush-Obama school reform: Lessons learned. Harvard Education Press. https://www.hepg.org/hep-home/books/bush-obama-school-reform

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. Educational Researcher, *42*(9), 476–487. https://doi.org/10.3102/0013189X13512674

Hill, H. C., & Papay, J. P., (2022). *Building better PL: How to strengthen teacher learning*. Research Partnership for Professional Learning. https://annenberg.brown.edu/rppl/what-works

Hill, P. T., & Celio, M. B. (1998). *Fixing urban schools*. Brookings Institution Press.

Hitt, C., McShane, M. Q., & Wolf, P. J. (2018). *Do impacts on test scores even matter? Lessons from long-run outcomes in school choice research*. American Enterprise Institute. https://www.aei.org/wp-content/uploads/2018/04/Do-Impacts-on-Test-Scores-Even-Matter.pdf

Holbein, J. B., & Hassell, H. J. G. (2018). When your group fails: The effect of race-based performance signals on citizen voice and exit. *Journal of Public Administration Research and Theory*, *29*(2), 268–286. https://doi.org/10.1093/jopart/muy075

Honig, M. I. (2006). *New directions in education policy implementation: Confronting complexity*. Suny Press.

Hoxby, C. M., Murarka, S., & Kang, J. (2009). *How New York City's charter schools affect achievement*. New York City Charter School Evaluation Project. http://users.nber.org/~schools/charterschoolseval/how_NYC_charter_schools_affect_achievement_sept2009.pdf

Hyman, J. (2017). Does money matter in the long run? Effects of school spending on educational attainment. *American Economic Journal: Economic Policy*, *9*(4), 256–280. https://doi.org/10.1257/pol.20150249

Jackson, C., & Cowan, J. (2018). *Assessing the evidence on teacher evaluation reforms* (CALDER Policy Brief No. 13-1218). CALDER. https://caldercenter.org/publications/assessing-evidence-teacher-evaluation-reforms

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. https://doi.org/10.1086/699018

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108. https://doi.org/10.1257/app.1.4.85

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, *131*(1), 157–218. https://doi.org/10.1093/qje/qjv036

Jackson, C. K., & Mackevicius, C. (2021). *The distribution of school spending impacts* (Working Paper No. 28517). National Bureau of Economic Research. https://doi.org/10.3386/w28517

Jackson, C. K., Wigger, C., & Xiong, H. (2021). Do school spending cuts matter? Evidence from the Great Recession. *American Economic Journal: Economic Policy*, *13*(2), 304–335. https://doi.org/10.1257/pol.20180674

Jacob, B. (2017). The changing federal role in school accountability: Point/counterpoint. *Journal of Policy Analysis and Management*, *36*(2), 469–477. https://doi.org/10.1002/pam.21975

Jacob, B. A., & Dee, T. S. (2010, September 1). The impact of No Child Left Behind on students, teachers, and schools [with Comments and Discussion]. *Brookings*. https://www.brookings.edu/bpea-articles/the-impact-of-no-child-left-behind-on-students-teachers-and-schools-with-comments-and-discussion/

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics*, *166*, 81–97. https://doi.org/10.1016/j.jpubeco.2018.08.011

Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379–407. https://doi.org/10.1080/19345747.2016.1273411

Jacobsen, R., Saultz, A., & Snyder, J. W. (2013). When accountability strategies collide: Do policy changes that raise accountability standards also erode public satisfaction? *Educational Policy*, *27*(2), 360–389. https://doi.org/10.1177/0895904813475712

James, J., Kraft, M. A., & Papay, J. (2022). *Local supply, temporal dynamics, and unrealized potential in teacher hiring* (EdWorkingPaper No. 22-518). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai22-518

James, J., & Wyckoff, J. H. (2020). Teacher evaluation and teacher turnover in equilibrium: Evidence from DC public schools. AERA Open, 6(2). https://doi.org/10.1177/2332858420932235

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250. https://doi.org/10.3368/jhr.44.1.223

Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments: Why states are quitting the PARCC and Smarter Balanced testing consortia. *Education Next*, *16*(4), 44–52. https://www.educationnext.org/the-politics-of-common-core-assessments-parcc-smarter-balanced/

Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies: Accountability and time allocated for science. *Science Education*, *97*(4), 621–636. https://doi.org/10.1002/sce.21058

Rice, J. K. (2013). Learning from experience? Evidence on the impact and distribution of teacher experience and the implications for teacher policy. *Education Finance and Policy*, *8*(3), 332-348. https://doi.org/10.1162/EDFP_a_00099

Kendi, I. X. (2016, October 20). *Why the academic achievement gap is a racist idea.* African American Intellectual History Society. https://www.aaihs.org/why-the-academic-achievement-gap-is-a-racist-idea/

Klager, C.R., and Tipton, E.L. (2021). Commissioned Paper on the Summary of IES Funded Topics. Paper prepared for the National Academies of Sciences, Engineering, and Medicine, Committee on the Future of Education Research at the Institute of Education Sciences in the U.S. Department of Education. https://nap.nationalacademies.org/resource/26428/READY-KlagerTipton_IES_Topic_Analysis_Jan2022v4.pdf.

National Academies of Sciences, Engineering, and Medicine (2022). The Future of Education Research at IES: Advancing an Equity-Oriented Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/26428.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. The University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo24695545.html

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36. https://doi.org/10.3368/jhr.54.1.0916.8265R3

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Kraft, M. A., and Novicoff, S. (2022). Instructional Time in U.S. Public Schools: Wide Variation, Causal Effects, and Lost Hours. (EdWorkingPaper: 22-653). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/1xxp-9c79

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500. https://doi.org/10.3102/0162373713519496

Kress, S. (2021, October 21). Accountability works, until it's no longer accountability. *Eduwonk*. http://www.eduwonk.com/2021/10/why-is-naep-flat-or-falling-part-1.html

Krieg, J. M. (2011). Which students are left behind? The racial impacts of the No Child Left Behind Act. *Economics of Education Review*, *30*(4), 654–664. https://doi.org/10.1016/j.econedurev.2011.02.004

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532. https://doi.org/10.1162/003355399556052

Kyse, E. N., Swann-Jackson, R., Marini, J., Benton, J., Byrne, A., Sceppaguercio, A., & Wilson, K. (2014). *Final evaluation report for the School Improvement Grant (SIG) evaluation: Summary and recommendations*. Center for Research and Evaluation on Education and Human Services (CREEHS). https://nj.gov/education/title1/sig/EvaluationReportC1and2.pdf

LaFortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics*, *10*(2), 1–26. https://doi.org/10.1257/app.20160567

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62. https://doi.org/10.3102/01623737024001037

LaVenia, M., Cohen-Vogel, L., & Lang, L. B. (2015). The Common Core State Standards initiative: An event history analysis of state adoption. *American Journal of Education*, *121*(2), 145–182. https://doi.org/10.1086/679389

Layton, L. (2014, June 7). How Bill Gates pulled off the swift Common Core revolution. *The Washington Post*. https://www.washingtonpost.com/politics/how-bill-gates-pulled-off-the-swift-common-core-revolution/2014/06/07/a830e32e-ec34-11e3-9f5c-9075d5508f0a_story.html

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, *10*, 179–214.

Le Floch, K. C., Martinez, F., O'Day, J., Stecher, B., Taylor, J., & Cook, A. (2007). *State and local implementation of the No Child Left Behind Act* (Volume III—Accountability under NCLB). U.S. Department of Education.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*(5), 635–652. https://doi.org/10.1080/03054985.2012.734800

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255 \

LiCalsi, C., Citkowicz, M., Friedman, L. B., & Brown, M. (2015). *Evaluation of Massachusetts Office of District and School Turnaround assistance to commissioner's districts and schools*. American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/15-2687_SRG_Impact-Report_ed_FINAL.pdf

Lin, D., Lutter, R., & Ruhm, C. J. (2018). Cognitive performance and labour market outcomes. *Labour Economics*, *51*, 121–135. https://doi.org/10.1016/j.labeco.2017.12.008

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. Russell Sage Foundation.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*. https://ies.ed.gov/ncser/pubs/20133000/

Liu, J., & Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, 56(2), 343–379. http://jhr.uwpress.org/content/early/2019/07/02/jhr.56.2.1216-8430R3

Loeb, S., Miller, L. C., & Wyckoff, J. (2015). Performance screens for school improvement: The case of teacher tenure reform in New York City. *Educational Researcher*, *44*(4), 199–212. https://doi.org/10.3102/0013189X15584773

Loss, C. P., & McGuinn, P. J. (Eds.). (2016). *The convergence of K-12 and higher education: policies and programs in a changing era*. Harvard Education Press

Loveless, T. (2015). Measuring effects of the Common Core. *Brookings*. https://www.brookings.edu/research/measuring-effects-of-the-common-core/

Loveless, T. (2016, March 24). Reading and math in the Common Core era. *Brookings*. https://www.brookings.edu/research/reading-and-math-in-the-common-core-era/

Loveless, T. (2018). Why standards produce weak reform. In F. M. Hess & M. Q. McShane (Eds.), *Bush-Obama school reform: Lessons learned*. Harvard Education Press.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293. https://doi.org/10.3102/0162373719849044

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

Marchitello, M. (2014). *Politics threaten efforts to improve K-12 education*. Center for American Progress. https://www.americanprogress.org/article/politics-threaten-efforts-to-improve-k-12-education/

Marion, S., & Briggs, D. (2022, July 13). *Please Make One Small Change in Federal Testing Law to Yield Big Improvements*. Just Give Us a Little, Center for Assessment. https://www.nciea.org/blog/just-give-us-a-little/

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., & Epstein, S. (2011). *A big apple for educators: New York City's experiment with the schoolwide performance bonuses*. Rand Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606. https://doi.org/10.1162/edfp.2009.4.4.572

McGuinn, P. J. (2006). *No Child Left Behind and the transformation of federal education policy, 1965-2005*. University Press of Kansas.

Mulligan, C. B. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy*, *107*(S6), S184–S224. https://doi.org/10.1086/250108

Murnane, R. J. (2013). U.S. High school graduation rates: Patterns and explanations. *Journal of Economic Literature*, *51*(2), 370–422. https://doi.org/10.1257/jel.51.2.370

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, *19*(4), 547–568. https://doi.org/10.1002/1520-6688(200023)19:4<547::AID-PAM2>3.0.CO;2-#

NAEP scores rise; NCLB gets credit. (2007, September 25). *Education Week*. https://www.edweek.org/education/naep-scores-rise-nclb-gets-credit/2007/09

National Academies of Sciences, Engineering, and Medicine. (2020). Changing expectations for the K-12 teacher workforce: Policies, preservice education, professional development, and the workplace. National Academies Press. https://doi.org/10.17226/25603

National Council for Accreditation of Teacher Education. (2010). *Transforming teacher education through clinical practice: A national strategy to prepare effective teachers. Report of the Blue Ribbon Panel on Clinical Preparation and Partnerships for Improved Student Learning*. ERIC Clearinghouse.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. National Academies Press.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263–283.

Neill, M. (2007). *NAEP and NCLB*. Schools Matter. http://www.schoolsmatter.info/2007/10/naep-and-nclb.html

Padilla, C., Skolnik, H., Lopez-Torkos, A., Woodworth, K., Lash, A., Shields, P. M., Laguarda, K. G., & David, J. L. (2006). *Title I accountability and school improvement from 2001 to 2004*. U.S. Department of Education. https://www2.ed.gov/rschstat/eval/disadv/tassie3/tassie3.pdf

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359–388. https://doi.org/10.1257/pol.20170709

Penuel, W.R., Briggs, D.C., Davidson, K.L., Herlihy, C., Sherer, D., Hill, H.C., Farrell, C., and Allen, A.-R. (2017). How school and district leaders access, perceive, and use research. AERA Open. https://doi.org/10.1177/2332858417705370.

Petek, N., & Pope, N.G. (2021). The multidimensional impacts of teachers on students. Unpublished manuscript.

Peterson, P. E. (1983). *Making the grade. Report of the Twentieth Century Fund Task Force on federal elementary and secondary education policy.* The Twentieth Century Fund Task Force.

Petrilli, M. J. (2021, October 21). Declining NAEP scores are flashing red lights for the COVID generation. *Education Next*. https://www.educationnext.org/declining-naep-scores-are-flashing-red-lights-for-the-covid-generation/

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal*, *58*(3), 527–566. https://doi.org/10.3102/0002831220905580

Phillips, K. J. R. (2010). What does "highly qualified" mean for student achievement? Evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *The Elementary School Journal*, *110*(4), 464–493. https://doi.org/10.1086/651192

Polikoff, M. (2021). It's easy to blame Common Core for lagging NAEP scores. But the evidence doesn't really add up. *The 74*. https://www.the74million.org/article/polikoff-its-easy-to-blame-common-core-for-lagging-naep-scores-but-the-evidence-doesnt-really-add-up/

Polikoff, M. S. (2012). The redundancy of mathematics instruction in U.S. elementary and middle schools. *The Elementary School Journal*, *113*(2), 230–251. https://doi.org/10.1086/667727

Polikoff, M. S. (2014). *Common Core State Standards assessments*. Center for American Progress. https://www.americanprogress.org/article/common-core-state-standards-assessments/

Polikoff, M. S. (2017). Is Common Core "working"? And where does Common Core research go from here? *AERA Open*, *3*(1), 233285841769174. https://doi.org/10.1177/2332858417691749

Polikoff, M. S. (2017) *Proficiency vs. growth: Toward a better measure*. FutureEd. https://www.future-ed.org/work/proficiency-vs-growth-toward-a-better-measure/

Polikoff, M. S., Desimone, L.M., Porter, A.C., Garet, M.S., Stornaiuolo, A., Pak, K., Smith, T.M., Song, M., Flores, N., Fuchs, L.S., Fuchs, D., and Nichols, T.P. (2022). *The Enduring struggle of standards-based reform: Lessons from a national research center on college and career-ready standards*. (EdWorkingPaper: 22-622). https://doi.org/10.26300/00ev-gk28

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis*, *31*(3), 238–268. https://doi.org/10.3102/0162373709336465

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity* (pp. 91–116). Russell Sage.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5), 1394–1415. https://doi.org/10.1016/j.jpubeco.2007.05.003

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*(3), 207–241. https://doi.org/10.1257/pol.6.3.207

Redding, C., & Nguyen, T. D. (2020). The relationship between school turnaround and student outcomes: A meta-analysis. *Educational Evaluation and Policy Analysis*, *42*(4), 493-519. https://journals.sagepub.com/doi/abs/10.3102/0162373720949513

Rentner, D. S., & Kober, N. (2014). *Common Core State Standards in 2014: Curriculum and professional development at the district level*. Center on Education Policy. https://eric.ed.gov/?id=ED555414

Richmond, G. (2022, January 25). Choice, flexibility, accountability drive school improvement. *Education Next, 22*(2). https://www.educationnext.org/choice-flexibility-accountability-drive-school-improvement-what-explains-charter-success/

Rice, J. K. (2013). Learning from experience? Evidence on the impact and distribution of teacher experience and the implications for teacher policy. *Education Finance and Policy*, *8*(3), 332–348. https://doi.org/10.1162/EDFP_a_00099

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252. https://doi.org/10.1257/0002828041302244

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*(7), 3184–3213. https://doi.org/10.1257/aer.102.7.3184

Rodriguez, L. A., Swain, W. A., & Springer, M. G. (2020). Sorting through performance evaluations: The influence of performance evaluation reform on teacher attrition and mobility. *American Educational Research Journal*, *57*(6), 2339–2377. https://doi.org/10.3102/0002831220910989

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. American Educational Research Journal, 50, 4–36. https://doi.org/10.3102%2F0002831212463813

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571. https://doi.org/10.1162/edfp.2009.4.4.537

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Teachers College Press.

Roza, M. & Anderson, L. (2020, April 32). New financial data spotlight the district role in distributing dollars across schools. *Education Next*. https://www.educationnext.org/new-financial-data-spotlight-district-role-distributing-dollars-across-schools-opportunities-education-leaders/

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Sandfort, J., & Moulton, S. (2015). *Effective implementation in practice: Integrating public policy and management.* Jossey-Bass.

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago Public Schools. *Journal of Human Resources*, *51*(3), 615–655. https://doi.org/10.3368/jhr.51.3.0514-6390R1

Sartain, L., & Steinberg, M. P. (2021). *Can personnel policy improve teacher quality? The role of evaluation and the impact of exiting low-performing teachers* (EdWorkingPaper No. 21-486). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-486

Sass, T. R., Apperson, J., & Bueno, C. (2015). *The long-run effects of teacher cheating on student outcomes* [Technical report]. Georgia State University. https://www.atlantapublicschools.us/crctreport

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2–3), 104–122. https://doi.org/10.1016/j.jue.2012.04.004

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the Common Core State Standards for mathematics. *Educational Researcher*, *41*(8), 294–308. https://doi.org/10.3102/0013189X12464517

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* [Think Tank White Paper]. American Educational Research Association. https://www.aera.net/portals/38/docs/causal%20effects.pdf

Scholastic and the Bill & Melinda Gates Foundation. (2014). *Primary sources: America's teachers on teaching in an era of change*. https://www.scholastic.com/primarysources/teachers-on-the-common-core.htm

Schwartz, H. (2022, March 14). What is really polarizing schools right now? *Education Week*. https://www.edweek.org/leadership/opinion-what-is-really-polarizing-schools-right-now/2022/03

Shakeel, M. D., & Peterson, P. E. (2020). Charter schools show steeper upward trend in student achievement than district schools. *Education Next*, *21*(1). https://www.educationnext.org/charter-schools-show-steeper-upward-trend-student-achievement-first-nationwide-study/

Shakeel, M. D., & Peterson, P. E. (2022). A Half Century of Student Progress Nationwide. *Education Next*, *22*(3). https://www.educationnext.org/half-century-of-student-progress-nationwide-first-comprehensive-analysis-finds-gains-test-scores/

Smith, J. M., & Kovacs, P. E. (2011). The impact of standards based reform on teachers: The case of "No Child Left Behind." *Teachers and Teaching*, *17*(2), 201–225. https://doi.org/10.1080/13540602.2011.539802

Smith, T. M., Garet, M. S., Song, M., Atchison, D., & Porter, A. (2021). *The impact of a virtual coaching program to improve instructional alignment to state standards*. American Institutes for Research. https://www.c-sail.org/sites/default/files/uploads/12/FAST-Impact-Brief.pdf

Song, M., Garet, M. S., Yang, R., & Atchison, D. (2022). Did states' adoption of more rigorous standards lead to improved student achievement? Evidence from a comparative interrupted time series study of standards-based reform. *American Educational Research Journal*, *59*(3), 610–647. https://doi.org/10.3102/00028312211058460

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, *27*(5), 556–563. https://doi.org/10.1016/j.econedurev.2007.06.004

Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V.-N., Pepper, M., Lockwood, J. R., & Stecher, B. M. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. National Center on Performance Incentives.

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Robyn, A., Baird, M. D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G. P., & Chambers, J. (2018). *Improving teaching effectiveness: Final report: The intensive partnerships for effective teaching through 2015–2016*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2242.html

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340–359. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Strauss, V. (2020, June 21). It looks like the beginning of the end of America's obsession with student standardized tests. *The Washington Post*. https://www.washingtonpost.com/education/2020/06/21/it-looks-like-beginning-end-americas-obsession-with-student-standardized-tests/

Strunk, K. O., Marsh, J. A., Hashim, A. K., Bush-Mecenas, S., & Weinstein, T. (2016). The impact of turnaround reform on student outcomes: Evidence and insights from the Los Angeles Unified School District. *Education Finance and Policy*, *11*(3), 251–282. https://doi.org/10.1162/EDFP_a_00188

Stullich, S., Eisner, E., & McCrary, J. (2007). *National assessment of Title I final report* (NCEE 2008-4012). U.S. Department of Education. https://ies.ed.gov/ncee/pdf/20084012_rev.pdf

Sun, M., Kennedy, A. I., & Loeb, S. (2021). The longitudinal effects of School Improvement Grants. *Educational Evaluation and Policy Analysis*, *43*(4), 647–667. https://doi.org/10.3102/01623737211012440

Sun, M., Penner, E. K., & Loeb, S. (2017). Resource- and approach-driven multidimensional change: Three-year effects of School Improvement Grants. *American Educational Research Journal*, *54*(4), 607–643. https://doi.org/10.3102/0002831217695790

Swanson, C. B., & Chaplin, D. (2003). Counting high school graduates when graduates count: Measuring graduation rates under the high stakes of NCLB. https://eric.ed.gov/?id=ED474605

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Taylor, K., & Rich, M. (2015, April 21). Teachers' unions fight standardized testing, and find diverse allies. *The New York Times*. https://www.nytimes.com/2015/04/21/education/teachers-unions-reasserting-themselves-with-push-against-standardized-testing.html

Toch, T. (2017, June 11). Hot for Teachers. Washington Monthly. https://washingtonmonthly.com/2017/06/11/hot-for-teachers/

Toch, T. (2018). *A policymaker's playbook: Transforming public school teaching in the nation's capital*. Future Ed, Georgetown University. https://www.future-ed.org/wp-content/uploads/2018/06/APOLICYMAKERSPLAYBOOK.pdf

Toch, T. (2020). Disrupted: Public-education reform in the nation's capital. *Education Next*. https://www.educationnext.org/wp-content/uploads/2022/01/ednext_XX_3_toch.pdf

U.S. Department of Education (2003). *Standards and Assessments Non-Regulatory Draft Guidance* U.S. Department of Education. https://www2.ed.gov/policy/speced/guid/nclb/standassguidance03.pdf

U.S. Department of Education (2007). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001*. U.S. Department of Education. https://www2.ed.gov/policy/elsec/guid/saaprguidance.pdf

U.S. Department of Education. (2009). *Race to the Top program executive summary*. U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED557422.pdf

U.S. Department of Education. (2012). *ESEA flexibility*. U.S. Department of Education. https://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc

U.S. Department of Education. (2021). *Awards*. Office of Elementary and Secondary Education. https://oese.ed.gov/offices/office-of-discretionary-grants-support-services/charter-school-programs/credit-enhancement-for-charter-school-facilities-program/awards/

U.S. Senate Committee on Health, Education, Labor & Pensions. (2015, November 18). *Alexander: House, Senate conference committee begins process to fix No Child Left Behind, a "law that everyone wants to fix"* [Press release]. https://www.help.senate.gov/chair/newsroom/press/alexander-house-senate-conference-committee-begins-process-to-fix-no-child-left-behind-a-law-that-everyone-wants-to-fix

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in place: How new teacher evaluations fail to live up to promises.* National Council on Teacher Quality. https://www.nctq.org/dmsView/Final_Evaluation_Paper

Watts, T. W. (2020). Academic achievement and economic attainment: Reexamining associations between test scores and long-run earnings. *AERA Open*, *6*(2), 233285842092898. https://doi.org/10.1177/2332858420928985

Wei, X. (2010). Are more stringent NCLB state accountability systems associated with better student outcomes? An analysis of NAEP results across states. *Educational Policy*, *26*(2), 268–308. https://doi.org/10.1177/0895904810386588

Wei, X. (2012). Does NCLB improve the achievement of students with disabilities? A regression discontinuity design. *Journal of Research on Educational Effectiveness*, *5*(1), 18–42. https://doi.org/10.1080/19345747.2011.604900

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project. https://files.eric.ed.gov/fulltext/ED515656.pdf

West, M. R. (2007). Testing, learning and teaching: The effects of test-based accountability on student achievement and instructional time in core academic subjects. In C. E. Finn Jr. & D. Ravitch (Eds.), *Beyond the basics: Achieving a liberal education for all children*. Thomas B. Fordham Institute.

West, M. R., & Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal*, *116*(510), C46-C62. https://doi.org/10.1111/j.1468-0297.2006.01075.x

Willen, L. (2022). Plunging NAEP scores make clear the long and difficult road ahead to pandemic recovery. The Hechinger Report. https://hechingerreport.org/plunging-naep-scores-make-clear-the-long-and-difficult-road-ahead-to-pandemic-recovery/

Why are the Common Core State Standards important? (n.d.). Common Core State Standards Initiative. Retrieved June 7, 2022, from http://www.corestandards.org/faq/why-are-the-common-core-state-standards-important/

Wong, K. K., Anagnostopoulos, D., Rutledge, S., & Edwards, C. (2003). The challenge of improving instruction in urban high schools: Case studies of the implementation of the Chicago academic standards. *Peabody Journal of Education*, *78*(3), 39–87. https://doi.org/10.1207/S15327930PJE7803_04

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, *8*(2), 245–279. https://doi.org/10.1080/19345747.2013.878011

Xu, Z., & Cepa, K. (2015). *Getting college and career ready during state transition toward the Common Core State Standards* (Working Paper No. 127). National Center for Analysis of Longitudinal Data in Education Research. https://caldercenter.org/publications/getting-college-and-career-ready-during-state-transition-toward-common-core-state

U.S. Chamber of Commerce
Foundation