PHOTOGRAPH/WHITE HOUSE, GETTY IMAGES

THE TEACHER EVALUATION REVAMP, IN HINDSIGHT

WHAT THE OBAMA ADMINISTRATION'S SIGNATURE REFORM GOT WRONG

WHEN PRESIDENT OBAMA TOOK OFFICE IN 2009, his administration quickly seized on teacher evaluations as an important public-policy problem. Today, much of his legacy on K–12 education rests on efforts to revamp evaluations in the hopes of improving teaching across the country, which his administration pursued via a series of incentives for states. In response, many states adopted new systems in which teachers' performance would be judged, in significant part, on their contributions to growth in student achievement.

Those moves have paid off in some ways, but in others, they backfired. Teacher evaluations today are more nuanced than they were eight years ago, and have contributed to better decisionmaking and enhanced student achievement in some districts. But progress was uneven, hampered by both design flaws and capacity challenges. And the changes were unpopular, helping generate a backlash against much of the reform playbook for the last few decades—as well as a strong federal role in education policy writ large. As we look ahead into the next four or eight years, an honest reflection can yield useful lessons about the potential, and limits, of federally led reform.

In this piece, I attempt to assess what went right, what went wrong, and what we can learn from the Obama administration's efforts to improve teacher evaluation systems. I do this as someone who played a role in the events that I describe: in 2011 and 2012, I was part of the policy team working on the No Child Left Behind (NCLB) waiver initiative and grant

by CHAD ALDEMAN

programs like the Teacher Incentive Fund (TIF), and played a role in spreading the Obama administration's teacher evaluation policies across the country.

That work rested on a particular theory of action—that more rigorous evaluation systems would empower districts to make better decisions about which teachers to put at the front of the room and thereby improve student outcomes. How well considered was that theory? How effective were the administration's efforts? These are my opinions, and mine alone.

A Focus on "Widgets"

The Obama administration's interest in teacher evaluations was spurred by two uncontroversial facts that, together, seemed to demand policymakers' attention. First, compelling new data confirmed that teacher quality was the most important in-school factor affecting growth in student achievement. Second, most school districts ignored these important differences in performance by treating all teachers as interchangeable parts, a phenomenon dubbed the "widget effect" in a timely 2009 report by TNTP (formerly The New Teacher Project). Districts rated 99 percent of teachers as "satisfactory" and ignored performance altogether when making decisions about recruitment, professional development, promotion, pay, or dismissal.

RATHER THAN KEEPING ITS FOCUS ON COMPETITIVE GRANT PROGRAMS LIKE RACE TO THE TOP, under the NCLB waiver program, the administration asked all states, regardless of interest or capacity, to tackle teacher evaluation systems—whether they wanted to or not.

The administration's ability to act quickly was spurred by a third factor: the massive, $787 billion economic stimulus bill passed by Congress in response to the 2008 financial crisis. The American Recovery and Reinvestment Act (ARRA) included roughly $115 billion in education spending, of which $4.35 billion was allotted for competitive state grants through a new program, Race to the Top (RTT). While the bulk of the stimulus funding went out with few strings attached to help cash-strapped states avoid layoffs, the grant program provided a unique opportunity to encourage innovation.

The Race to the Top program, announced in 2009, allotted $4.35 billion for competitive state grants and encouraged states and districts to revamp their teacher evaluations.

PHOTOGRAPH / AP PHOTO,HARAZ N. GHANBARI

RTT encouraged states and districts not only to revamp their teacher and principal evaluation policies but also to use evaluation results to make personnel decisions. It embedded "improving teacher and principal effectiveness based on performance" into its rubric for scoring applications and awarded the category more than 10 percent of the total available points. States and participating districts were to evaluate teachers and principals using multiple measures, including, "in significant part," student growth. The term "student growth" was further defined to mean the change in student achievement as measured on statewide assessments and other measures that were "rigorous and comparable across classrooms." The administration also embedded these requirements and definitions in subsequent grant competitions, its proposal to reauthorize NCLB, and,

starting in 2011, conditions for states seeking NCLB waivers.

In addition, the administration greatly expanded the TIF program, which awards grants to high-need districts to fund performance-based compensation systems, and established a new rule for winning applications: proposals would need to differentiate teacher and principal effectiveness, based in significant part on student growth, and create compensation systems that reflected those results. In the administration's first year, some 62 districts and schools across the country shared $437 million in TIF grants.

The reaction was swift, as state legislators and policymakers across the country made sweeping changes in areas that had long been dormant. According to the National Council on Teacher Quality (NCTQ), the number of states requiring objective measures of student achievement to be included in teacher evaluations nearly tripled from 2009 to 2015, from 15 to 43 states nationwide (See Figure 1). Even more striking, the number of states requiring districts to consider teacher evaluations in tenure decisions grew from 0 to 23 over that same period.

What Went Right

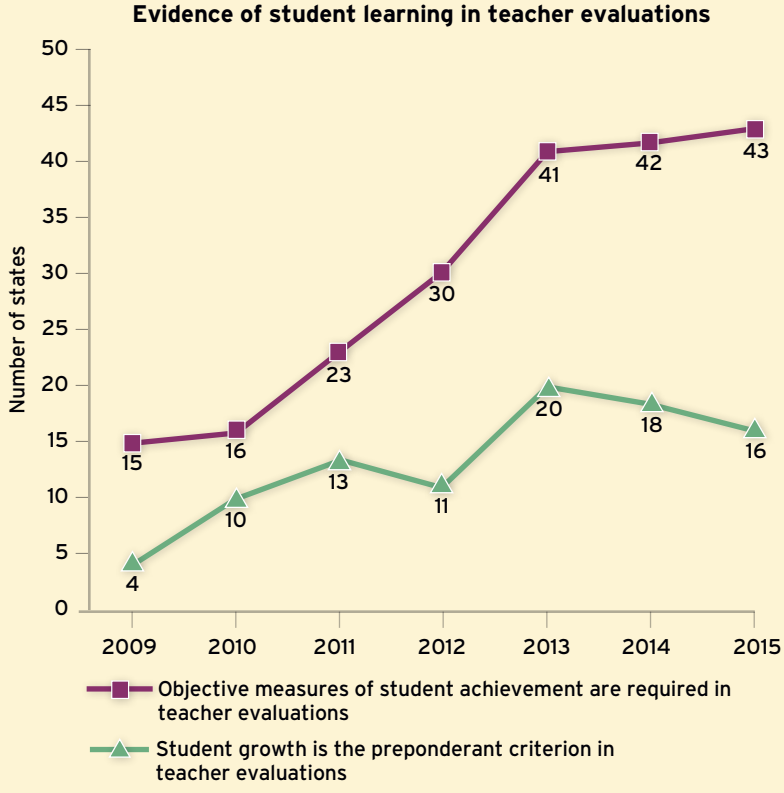
As a result of these initiatives and investments, teacher evaluation systems today are much stronger than they were before Obama took office. Teachers are evaluated more frequently, evaluators use higher-quality observation rubrics to assess their performance, and teachers receive more detailed feedback on their performance. More states and districts now factor teacher effectiveness into decisions regarding promotion and compensation.

There is evidence that these changes have had a positive effect on student learning—even if, in some cases, the changes were not implemented well. That was the finding of an independent evaluation by the Institute of Education Sciences (IES) of the administration's investments in educator evaluation and compensation through the TIF program. The review unearthed numerous problems: Districts chose to base teacher performance awards on measures that don't reflect individual performance, such as raw, unadjusted student achievement scores or school-wide average growth rates. They shared smaller incentives among large numbers of teachers and principals

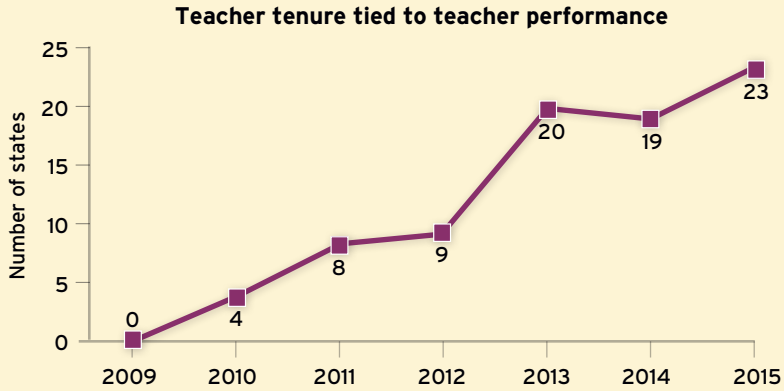
Sweeping Changes to Teacher Evaluation

(Figure 1)

(1a) By 2015, 43 states required that objective measures of student achievement be included in teacher evaluations—up from 15 states in 2009. The number of states making student growth the preponderant criterion in teacher evaluations jumped from four to 16 over this time period.



(1b) Although no state required districts to consider student learning when making decisions on teacher tenure in 2009, 23 states did by 2015.



SOURCE: National Council on Teacher Quality State Teacher Policy Yearbook 2015

rather than giving larger awards to the highest-achieving staff. And they failed to communicate the program well to teachers and principals, leading to mass confusion about who was eligible for awards and how large the awards would be. Yet, despite those flaws, a randomized controlled trial found that the program led to gains equivalent to 10 percent of a year's worth of learning in math and 11 percent in reading.

We don't know *why* the TIF program succeeded, but it does provide one piece of evidence in support of the Obama administration's underlying theory of action. It suggests that performance-based evaluation and compensation systems can drive improvements in student outcomes. Yet, we shouldn't apply that lesson universally, as evidenced by similar efforts spurred by RTT and the NCLB waiver initiative.

What Went Wrong

There were successes, but we can also learn from the weak spots. In my view, the Obama administration's efforts to revamp teacher and principal evaluation systems got at least four major things wrong.

A Universal Approach. Let's look again at the TIF districts, compared to districts compelled to undertake reforms under RTT or NCLB waivers. Here we find Lesson No. 1: the dangers of a universal approach.

TIF districts voluntarily applied for additional support to pursue reform. They chose to participate, put together winning plans, and won five-year grants worth millions of dollars to implement them. And while the TIF competition had multiple components, it was entirely focused on evaluation and compensation systems.

By contrast, under RTT, states were competing for a share of billions of dollars, and districts could opt to sign on, or not, to plans they did not themselves create. Teacher and principal evaluation systems were just one component of those plans, and nearly 90 percent of a plan's score in the grant competition was awarded for elements other than evaluation and compensation systems. In fact, a 2016 study commissioned by IES concluded that "across all states, use of policies and practices promoted by RTT was…lowest for teacher and principal certification and evaluation."

The NCLB waiver initiative was even less targeted. That effort was similar to RTT in that it involved states making reforms across a number of areas, with evaluation systems being only one component. But it offered no new money to implement the changes. Perhaps most importantly, the waiver initiative put sometimes-reluctant state departments of education in the role of enforcing evaluation systems in local school districts that never agreed to them. The prospect of flexibility from NCLB requirements may have succeeded in making states adopt evaluation reforms, but this approach left the federal government with few levers to make states implement them well.

This reveals the central design flaw in the administration's universal strategy. Rather than keeping its focus on competitive grant programs like RTT or TIF, the Obama administration sought to apply its ideas everywhere. In the NCLB waiver program, all states, regardless of interest or capacity, were asked to tackle teacher evaluation systems—and to do so in all of their districts. Places that didn't really want to tackle this particular challenge were forced to anyway. While bold initiatives can be admirable, it's important to get the scope right.

A Narrow Definition of Reform. In all its grant competitions and funding programs, the administration included language that pushed states and districts to create multi-tiered evaluation systems to "differentiate" among educators based "in significant part" on their contributions to "student growth." Lesson No. 2: this definition of reform was too rigid.

It was an understandable response after decades of evaluations divorced from student outcomes in which virtually all teachers and principals received cursory "satisfactory" ratings. But by requiring that all teachers be evaluated in this way, we forced states and districts to come up with a suite of new pre- and post-test measures to track changes in student achievement over time. We also left ourselves open to grossly misleading claims about our policies, such as the myth that we advocated evaluating teacher performance based on test scores alone.

States and districts should have been focusing on the real end goal—differentiating the best teachers from those who are merely satisfactory and those who continue to struggle—a task that would not have required complicated mathematical formulas designed to measure each teacher's "value-added" to student achievement. It would have been better to allow or even encourage states and districts to use any set of measures

WHEN THE ADMINISTRATION MATCHED THE TIMELINE FOR IMPLEMENTING EVALUATION SYSTEMS TO THAT FOR COMMON CORE, it quickly became a liability to hold teachers accountable for results on tests they had never seen before.

that came to broadly similar results. This approach also would have addressed concerns that the state-created teacher evaluation systems locked in existing one-teacher-one-classroom staffing arrangements rather than allowing more flexible staffing models. While I don't believe that the new systems truly stifled innovation in the field—schools and districts operating under one set of state rules still report widely varying results—it did present a potential obstacle, and once again left us open to easy criticism.

Focusing on end results would also have allowed districts to spend their time developing and implementing high-quality observation and rating tools instead of developing new assessments to measure student growth. Principals can be effective at identifying high- and low-performing teachers (see "When Principals Rate Teachers," *research,* Spring 2006), and while all observation rubrics may not be perfectly aligned with student growth, they can be applied to all teachers—not just those in tested grades and subjects. A back-end check that the evaluation results corresponded with evidence of impact on student achievement, where available, could have accomplished our purposes more effectively. And it could have helped avoid widespread conflict about the precise weighting of student growth in teacher evaluation systems and the adoption of additional tests to measure student performance. Focusing on the systems as a whole also would have encouraged districts to be more honest in their observation ratings rather than creating the incentive for subjective observation ratings to compensate for value-added results that, by definition, grade teachers on a curve. Thanks in part to those incentives, Brown University researcher Matthew Kraft found that the share of teachers receiving a less-than-satisfactory rating hardly budged in most states as the new systems were implemented (See Figure 2).

The notion of value added was itself both a strength and a liability. The models themselves are what allowed policymakers, and district officials, to operationalize and put to use the concept of a teacher's contribution to student learning. In addition, research showing that value-added measures outperform other teacher characteristics at predicting a teacher's impact on student growth in future years—and that they also capture information on teachers' impacts on longer-term life outcomes like teen pregnancy, college going, and adult earnings—served as an important justification for differentiating teacher effectiveness. But value-added scores can be complicated to interpret and, on their own, do not provide teachers with guidance on how to improve. Moreover, some teachers may fundamentally disagree with the notion that their skill can be fairly evaluated based on their students' outcomes, and no empirical evidence can persuade them otherwise.

An elementary school teacher burns his evaluation during a protest in front of the Albuquerque Public Schools headquarters in October 2016.

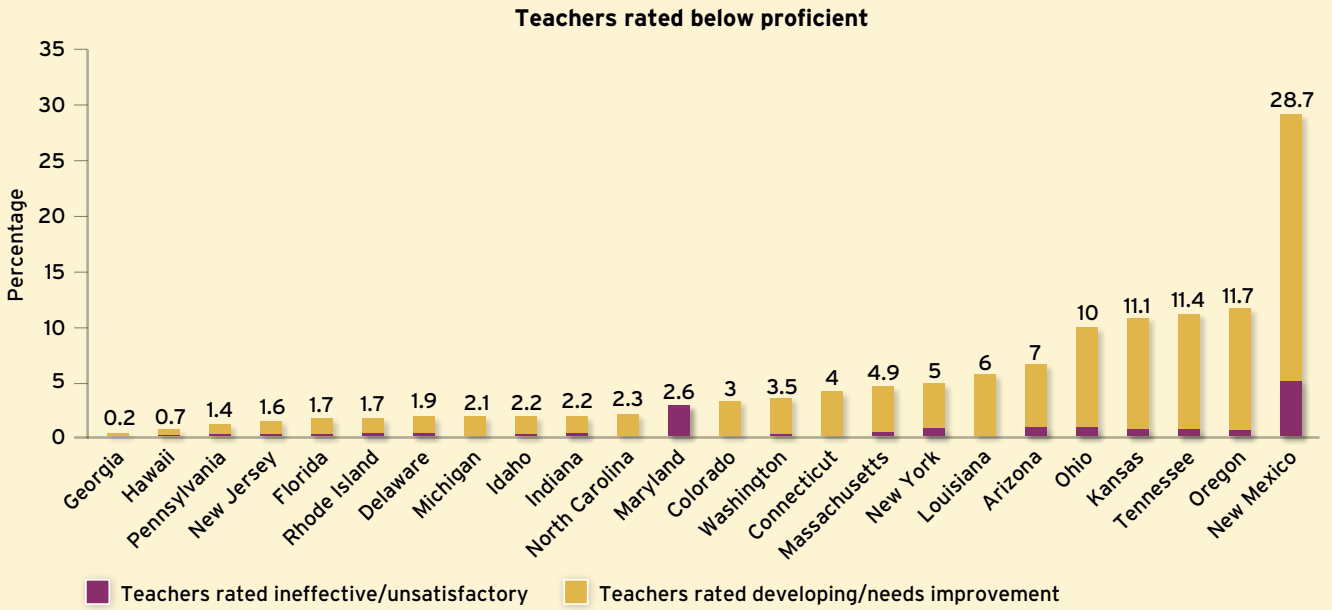
PHOTOGRAPH / JIM THOMPSON, *ALBUQUERQUE JOURNAL*

Process over Purpose. This relates to Lesson No. 3: the perils of prioritizing a process over its end result. The perceived complexities of evaluating teaching and, in particular, the mysterious-sounding nature of value-added models, captured much of the public conversation—and the time and efforts of state and district officials. The push to revamp evaluation systems ended up focusing too much on the evaluation systems themselves, and never actually got around to *using* those systems to make decisions.

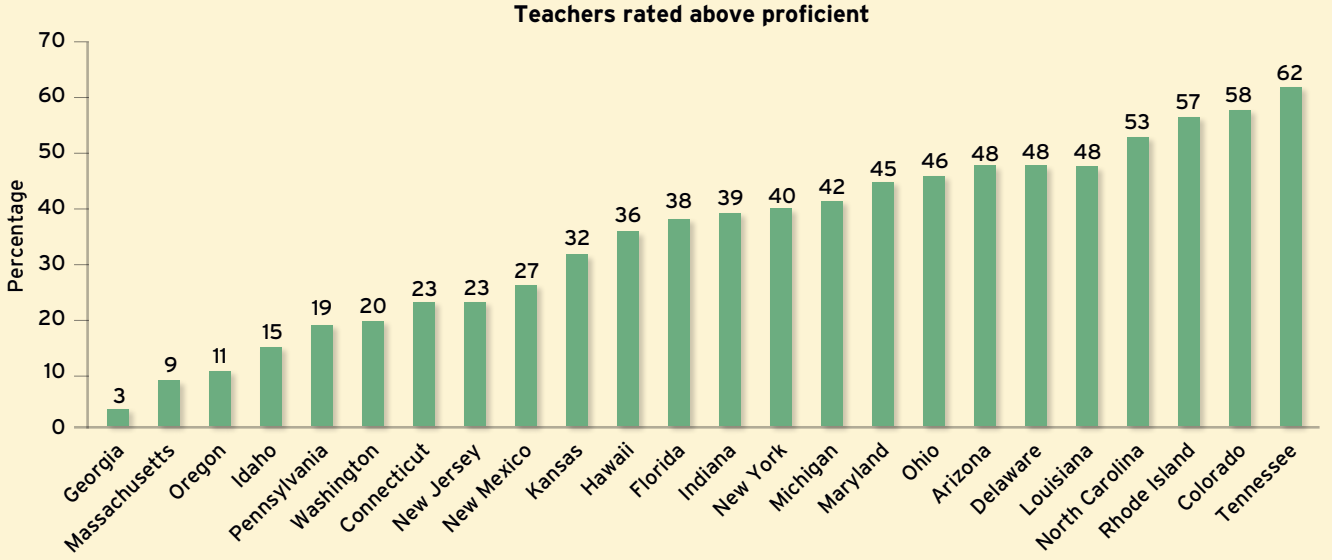
When Obama took office, unhelpful evaluation systems were just one of several barriers preventing districts from effectively managing their teaching staffs, such as tenure rules and lockstep compensation systems based solely on seniority and graduate degrees. The administration made a conscious choice to focus on evaluations first, believing that employees need to be evaluated honestly before their performance can be used for consequential decisions. That premise may have been

New Evaluations Identify Few Teachers as Ineffective (Figure 2)

(2a) In 22 of the 24 states that had implemented a new teacher evaluation system by the 2014-15 school year, one percent or less of teachers received an unsatisfactory rating. The total share of teachers rated below proficient exceeded five percent in only seven states.



(2b) The same evaluation systems have led to more teachers being recognized for their effectiveness, with substantial shares being rated above proficient in most states.



NOTE: Figure 2b combines the shares of teachers receiving one of the top two ratings in states with five categories.

SOURCE: Kraft, M.A. and Gilmour A. (2016). "Revisiting the Widget Effect: Teacher evaluation reforms and distribution of teacher effectiveness ratings." Brown University Working Paper

flawed from the outset. It is possible, for example, that creating incentives to expand principals' decisionmaking authority within district rules and collective-bargaining agreements could also have been a feasible approach. It is also possible that no amount of investment in new evaluation systems would ever make teachers comfortable with consequential decisions flowing from those systems.

In addition, the administration wildly overestimated the field's capacity to improve teacher evaluation systems. In 2009, most states lacked basic data systems linking teachers to their students over time, and few had growth models in place that could be used to measure teacher performance. Most districts were using simple checklists to assess teacher performance rather than the more sophisticated rubrics that can successfully differentiate instructional skills.

Moreover, these systems had to be implemented by people, and there were very few people who had been trained in evaluating teachers and providing them actionable feedback on how to improve. As an administration, we asked the country to quickly move from evaluating almost no teachers seriously to suddenly evaluating all teachers in a brand-new, much more comprehensive way. It was a case of too fast, too soon.

Common Core Collision. As these new systems were coming online in 2013 and 2014, many states and districts were also starting to implement the Common Core State Standards and related assessments. Lesson No. 4: proper pacing is critical when pursuing multiple reforms.

In its NCLB waiver initiative, the administration matched the timeline for teacher and principal evaluation systems to that for Common Core: pilot in 2013–14 and implement in 2014–15. This alignment made sense logistically as a way for states to have time to plan implementation efforts and give them a test run. But the two reforms amounted to a one-two punch in the public eye and gave critics an easy-to-understand argument against reform: too many uncertainties, all at once.

Despite critics' claims that simultaneously implementing new standards and new evaluation systems would undermine the latter's validity, subsequent research demonstrated that teacher value-added scores remained generally stable, even as states made changes to their standards and assessments and the percentages of students passing the new tests plunged. For example, when New York administered new, tougher assessments in 2013, the percentage of students deemed "proficient" was virtually cut in half, to 31 percent from 55 percent in reading and to 31 percent from 65 percent in math. However, despite lower scores for students, teacher value-added rankings on the official state-provided student growth measure stayed nearly identical. In other words, it was *technically* feasible to implement multiple reforms at once.

But technical feasibility does not translate perfectly into the political realm. It quickly became a liability to hold teachers accountable for results on tests they had never seen before, and much of the Republican establishment seized on Common Core as the embodiment of federal overreach. The twin controversies threw flames on one another.

Many teachers viewed the new evaluations as a threat—which, in retrospect, appears overblown, given how few teachers were dismissed as a result of new systems. Numerous states did change their laws and regulations so that districts had new discretion to dismiss teachers who received poor ratings. But potential dismissals are not the same as actual firings, and few teachers were ever removed from the classroom. In New York State, for example, as of late 2015 only one tenured teacher had been fired through its revamped evaluation and dismissal process. Between 2012 and 2014, the entire state of New Jersey dismissed just 23 teachers for poor performance out of more than 100,000 statewide. In these and many places, the threat of dismissal caused widespread political conflict without districts getting much upside in terms of removing low performers.

THE THREAT OF DISMISSAL CAUSED WIDESPREAD POLITICAL CONFLICT without districts getting much upside in terms of removing low performers.

Despite the minute share of teachers unfairly affected by these policies, they helped fuel a backlash against testing in general. The percentage of parents choosing to "opt out" of statewide testing grew sharply; for example, about 15 percent of high-school juniors in New Jersey and sophomores in Colorado skipped testing in 2015, and 20 percent of all students did so in New York State. Common Core was and remains a political concern, and the number of states planning to use the Common Core–aligned PARCC and Smarter Balanced assessments dropped from 45 in 2011 to just 20 that actually used one of the two tests in 2016 (see "The Politics of the Common Core Assessments," *features,* Fall 2016). But testing opposition appears to be more closely linked to concerns about teacher evaluation policies: the top two reasons chosen among a national survey of parents who opted out were, "I oppose using students' performance on standardized tests to evaluate teachers" and "standardized tests force teachers to teach to the test."

The education policy community, particularly those who lean toward a stronger federal role, has not fully grappled with the consequences of the failures of the federal push for improved teacher evaluation policy, but they are severe.

Missed Signals

Some of these problems were foreseeable, and many critics of the Obama administration's policies pointed out their flaws as they were being implemented. I can't fully explain why those critiques never forced a course correction, at least not until late in Obama's second term, but I can offer a few reasons why the critics didn't resonate, at least not with me.

For starters, we faced inaccurate criticisms of our policies, like the assertion that we were forcing teachers to be fired based on a single test score. The actual policy was much more nuanced, but that didn't matter in the public debate. Other critics claimed we were "mandating" that states adopt these policies, but states always had an option of whether or not to pursue grant funding or NCLB waivers (five states chose not to apply for a waiver, and others did so only after watching many other states earn approval). I recognized at the time that our timelines may have been too rapid, but I also saw some states and districts moving much more quickly than others, which suggested that political will was as much an issue as capacity constraints. And presidential administrations think in four-year cycles, so any policy that would have ended after Obama left office (and thus rely on another administration to complete) would have been a nonstarter.

There was also probably some inertia that came into play. After the Obama administration formalized its signature policies in 2009 and 2010, those who followed (including me) were tasked with implementing those policies through various grant programs and initiatives. After that process was underway, it would have taken a massive effort—not to mention an unlikely mea culpa—to change course. All of our budget documents, grant competitions spanning the entire U.S. Department of Education, and official administration talking points, goals, and objectives would have had to have been rewritten, ideally according to a well-considered alternative theory of action. It took years of work from hundreds, if not thousands, of people to devise the original policies; it would have taken a similarly massive effort to undo them.

Looking Ahead

With the 2015 passage of the Every Student Succeeds Act eliminating all federal oversight of teacher evaluation systems, states now have full discretion to chart their own course. There's much to be learned from the federal government's efforts over the past eight years.

States and districts would be wise to focus on the goals of their evaluation systems, including differentiating teachers based on their observed practice, providing actionable feedback on how to improve, and using the results to make consequential personnel decisions. They should gather data on results, keeping in mind that there is no "correct" distribution of teacher effectiveness. An honest system should identify some portion of educators as excellent, some as solid but with areas for improvement, and some who need significant support or who may not be a good fit for the profession.

WITH THE 2015 PASSAGE OF THE EVERY STUDENT SUCCEEDS ACT ELIMINATING ALL FEDERAL OVERSIGHT OF TEACHER EVALUATION SYSTEMS, states now have full discretion to chart their own course, and they would be wise to focus on the goals of their evaluation systems.

Districts should have the flexibility to use the results of their evaluation systems to make decisions regarding compensation, professional development and advancement, and dismissals. But states should not attempt to define the specific components of evaluation systems or to mandate one system for all of its districts.

And what of future federal efforts, in this policy area and others? Federal bureaucrats would do well to focus on targeted competitive grant programs to encourage the adoption of their desired policies in places that really want to pursue them. They should think deeply about end goals rather than getting bogged down in specific design choices. They should take to heart the adage that the federal government can make states and districts do something, but it can't make them do those things well. And they should be ambitious in their aspirations for educational improvement in the United States but humble about the potential unintended consequences of their work.

Chad Aldeman is a principal at Bellwether Education Partners. Previously, he was a policy adviser at the U.S. Department of Education, where he worked on ESEA waivers, teacher preparation, and the Teacher Incentive Fund.