



Why Does Education Policy so Frequently Disappoint? Lessons from Teacher Evaluation

James H. Wyckoff

University of Virginia

Decades of K–12 education reform — much of it focused on teaching — have produced a persistent paradox: policies grounded in strong evidence disappoint when scaled, even as some implementations achieve striking success. Teacher evaluation is the clearest case. Cincinnati, Dallas, Chicago, and Washington, DC produced meaningful achievement gains, while forty-four state systems, Houston, and a heavily funded, pre-screened seven-site demonstration produced little or none. The weight of the evidence indicates that teacher evaluation policy failed in implementation rather than design. Less clear is why. I explore three complementary reasons: limits of the mechanisms, failure to share responsibility, and failure to adapt; each made predictable by the political economy of education reform. Recent NAEP results have renewed calls for reform. This paper summarizes what design and implementation guidance the record supports as necessary, which remain unresolved, and what research would resolve them.

VERSION: October 2026

Suggested citation: Wyckoff, James H. (2026). Why Does Education Policy so Frequently Disappoint? Lessons from Teacher Evaluation. (EdWorkingPaper: 26-1601). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/e015-5789>

Why Does Education Policy so Frequently Disappoint? Lessons from Teacher Evaluation

Jim Wyckoff *
September 2026

Abstract

Decades of K–12 education reform — much of it focused on teaching — have produced a persistent paradox: policies grounded in strong evidence disappoint when scaled, even as some implementations achieve striking success. Teacher evaluation is the clearest case. Cincinnati, Dallas, Chicago, and Washington, DC produced meaningful achievement gains, while forty-four state systems, Houston, and a heavily funded, pre-screened seven-site demonstration produced little or none. The weight of the evidence indicates that teacher evaluation policy failed in implementation rather than design. Less clear is why. I explore three complementary reasons: limits of the mechanisms, failure to share responsibility, and failure to adapt; each made predictable by the political economy of education reform. Recent NAEP results have renewed calls for reform. This paper summarizes what design and implementation guidance the record supports as necessary, which remain unresolved, and what research would resolve them.

The Achievement Imperative and the Role of Teachers

The case for raising student achievement in the United States is social, economic and moral. Academic achievement among most American students has experienced substantial decline over the last 15 years (Hanushek, 2025; Wyckoff, 2026). Learning losses have been greatest among already low-performing students—the lowest decile declined 0.51 SD in reading and 0.47 SD in math since 2013 (Wyckoff, 2026). The consequences extend well beyond the individual: a growing body of research links educational attainment and cognitive skill development to earnings, health, civic participation, and long-run economic growth (Doty et al., 2025). Three decades of research establish that teachers are central to these outcomes: differences in teacher effectiveness are large and consequential both in school (Aaronson et al., 2007; Chetty et al., 2014a; Kane & Staiger, 2008; Rivkin et al., 2005) and in adulthood (Chetty et al., 2014b).

Decades of widely heralded policy reform, much of it aimed at teaching quality, resulted in disappointing student outcomes—despite instances of meaningful improvement. By 2015, enthusiasm had waned, and key features of these efforts were effectively neutered. Policymakers appear to have concluded that the designs don't work, moving on to other interventions.

Other policies hold promise and should be pursued. Yet, it is difficult to identify a meaningful school intervention whose theory of change does not depend on the quality of teaching. Improved school leadership, stronger instructional materials, reduced class sizes, rigorous

* Thanks to Brendan Bartanen, Daphna Bassok, Julie Cohen, Laura Rogers, and participants at EdPolicyWorks Policy Lab for useful feedback. Claude was employed to provide feedback on some of the ideas in the paper, suggestions on references, and feedback on some of the writing.

accountability, in-school tutoring, school choice—each may be the lever through which a policy acts, but each depends for its effectiveness on the quality of teaching it produces, sustains, or assumes. This does not diminish the importance of out-of-school factors. Income security, housing stability, and access to healthcare all shape children's readiness to learn, and reforms in those domains could produce larger educational gains than anything schools alone can achieve. The narrower claim is that most in-school policies either target teaching quality or depend on it.

Effective teaching is a highly practiced, cognitively demanding craft. Experienced teachers hold well-established instructional routines, anchored in internalized habits and knowledge about how students learn (Leinhardt & Greeno, 1986). Changing those routines is not a matter of informing teachers of a new approach or handing them new materials. It requires altering routines, building knowledge and skill, and sustaining the conditions under which deliberate practice produces durable change (Cohen & Ball, 1999; Cohen & Mehta, 2017) — a process that implementation researchers estimate takes two to four years of supported effort, and longer still to become embedded in a school's professional culture (Fixsen et al., 2005).

That challenge is sharpened by the current condition of the profession, where declining teacher preparation enrollment, elevated attrition, and competition from other sectors have left many districts unable to staff classrooms with fully credentialed teachers (Cowen & Goldhaber, 2024; Kraft & Lyon, 2024). The challenge of improving teaching quality is not simply a matter of developing the current teachers, but of rebuilding a profession that is losing ground.

Periodically, a direct focus on teaching quality has animated reform efforts. Yet the historical record suggests these efforts have more often reshaped policy than transformed practice or improved student outcomes at scale (Cohen & Mehta, 2017; Cuban, 2022; Hess, 1998; Tyack & Cuban, 1995) and the results have been strikingly inconsistent. Teacher evaluation reform produced significant gains in student achievement in Washington, DC, Cincinnati, Chicago and Dallas, but null results in many states and districts adopting ostensibly similar systems. The science of reading is credited with dramatic improvements in Mississippi but modest effects in states that adopted legislation with similar features. Instructional coaching programs have shown large effects in randomized trials of boutique programs but much smaller effects when scaled.

The conventional explanations for this inconsistency—political resistance, insufficient funding, differences in context—have merit but are incomplete. This paper argues that education policy has systematically underestimated the difficulty of improving teaching quality and has treated the adoption of a reform as functionally equivalent to implementing the mechanisms that change what teachers do in diverse classrooms (Cohen & Ball, 1999; Elmore, 1996; Goertz, 2012; Tyack & Cuban, 1995). That conflation also explains why policy churn is so destructive (Hess, 1998; 2017; 2026): the initiatives being churned rest on theories of change requiring years of supported practice to activate, and they are abandoned precisely when the organizational learning needed to make them work would begin to accumulate. Rarely do researchers do the autopsy to better understand whether the policy is fatally flawed or merely in need of adjustment.

Policymakers risk repeating the pattern of promising policies leading to disappointing outcomes if they fail to understand and learn the lessons of the last three decades of education reform.

The Reform-as-Policy Problem

Education policy depends on a well-grounded, promising design *and* a detailed, informed implementation plan. Policy design and policy implementation are distinct, and much of the

discussion of education policy underappreciates the distinction, either assuming the conditions for effective implementation will follow policy adoption or by ignoring it. Although there is a large literature on implementation, its scope and importance for policy success is less fully appreciated by policy researchers (for exceptions see Coburn, 2003; Elmore, 1996; Hess, 1998; McLaughlin, 1987).

Design rests on a *change model*: an account of why the policy should lead to changes for the target population. Implementation rests on an *action model*: an account of why that change should reliably occur in practice (Chen, 2005). A sound change model identifies realistic goals, specifies credible causal mechanisms by which the policy can be expected to reach them, and prioritizes the features necessary for those mechanisms to operate (Sabatier & Mazmanian, 1979; Chen, 2005; Capano & Howlett, 2021). A sound action model specifies the resources, tools, and supports required to deliver those features with fidelity across varying contexts (Pressman & Wildavsky, 1973; Lipsky, 1980). In assessing policy failure, it is helpful to locate where in the causal chains the policy went awry, and more useful still to diagnose why.

Teacher evaluation illustrates the distinction. Suppose the goal is to improve student outcomes through teacher development. This design mechanism runs through four linked components: valid and reliable assessment of the instructional skills that matter for student learning; clear, timely, specific feedback on that assessment; coaching or targeted professional development that helps teachers translate feedback into altered classroom practice; and sustained, cumulative practice on these skills to entrench them in teachers' routines, behaviors and practices. These are the policy's claims about *why* teaching should improve. If any of them is wrong — if observation cannot capture the skills that matter, or if feedback and coaching do not improve practice, the policy fails on its design, and no amount of implementation effort will rescue it. If strong implementation across the crucial elements of the change model failed to produce meaningful outcomes, that would be evidence the change model failed.

The action model addresses a different question: whether those components reached teachers. If they don't, the policy failed due to implementation. Did policymakers and practitioners attend to the components of teacher evaluation to ensure their effectiveness? Doing so requires selecting an observation protocol, specifying the frequency and duration of observations, training observers and establishing that they rate reliably, protecting time for feedback conversations, staffing coaching with people who have the expertise to deliver it, and allowing sufficient time for these changes to embed themselves in the practices of teachers and the culture of the school.

The implementation literature has long recognized that policies are mediated and reinterpreted by the administrators and practitioners who enact them—the street level bureaucrats (Pressman & Wildavsky, 1973; Fixsen et al., 2005; Lipsky, 1980; McLaughlin, 1987). In education, that mediating layer ultimately runs through teachers, because every meaningful pedagogical reform asks teachers to do something differently. Policy success depends on the alignment of teachers' classroom goals with the policy mechanisms.

Scale compounds the problem, as does the federal nature of U.S. education policy. The 3.2 million public K-12 teachers in the United States bring varying content knowledge, entrenched instructional habits, and different capacities for change, and they work in settings — students, communities, resources — that vary enormously (McLaughlin, 1987). An action model adequate to a workforce this large and this heterogeneous cannot assume a uniform pace or a single mode of support. It must differentiate and support teacher cumulative development (Elmore, 1996).

When policy is determined in Washington or state capitals but enacted in states, districts and classrooms, the risk of inadequate implementation substantially increases. Federal and state policymakers need to anticipate this misalignment and build systems to reduce these risks.

Reform that does not build this complexity into both models will produce uneven results. And when results disappoint, the appropriate response depends entirely on the diagnosis. A faulty change model calls for a different policy; a faulty action model calls for modifying implementation. Conflating the two leads policymakers to abandon sound ideas and to persist with unsound ones. Teacher evaluation reforms of the 2000s are a case in point. Widely promoted at adoption, they were judged failures by 2015, and policymakers have since been reluctant to revisit these policies. I argue below that reading the disappointing outcomes of teacher evaluation as a design failure is a mistake. Rather the evidence suggests that implementation is at fault, but the exact source of this failure remains unclear. This paper produces a graded inventory of which design features the record supports, which are contingent, and which remain open.

Teacher Evaluation Reform: Uneven Impacts

The reform of teacher evaluation became a national priority in the years surrounding *Race to the Top* (2009), driven by the research consensus that teacher quality varies enormously and that traditional evaluation systems—in which the vast majority of teachers received satisfactory ratings regardless of student outcomes—were incapable of identifying, differentiating, or developing teachers based on effectiveness (Gordon et al., 2006; Weisberg et al., 2009). This research consensus coupled with a compelling theory of change provided the foundation for including teacher evaluation in policies such as the federal *Race to the Top* and the Teacher Incentive Fund. The policy response was nearly universal: forty-four states adopted new evaluation systems with multiple ratings tiers, observation rubrics, and, in many cases, value-added measures.

The results were, by most accounts, disappointing. Most state evaluation reforms produced little evidence of improvement in teaching quality or student achievement, (Bleiberg, et al., 2025; Stecher, et al., 2018). Researchers, pundits and the popular press were quick to conclude that teacher evaluation was a failed reform and that policymakers should move on (e.g., Dynarski, 2016; Will, 2021). Against this background, the experience of four school districts— Dallas, Cincinnati, Chicago, and Washington DC—stand out.

The design of teacher evaluation policy typically allows for three mechanisms that could improve teacher quality—teacher recruitment, differential retention and development. A well-conceived and executed teacher evaluation system that includes incentives for strong performance could attract prospective teachers who value such a system. This mechanism relies on measures of performance that are valid and reliable. How teaching effectiveness is measured and the incentives associated with performance will dictate its effect on teacher supply. These are implementation components and were largely left to school districts to determine. The second mechanism is differential selection or retention—requiring or incenting low performing teachers to exit and using incentives to retain more effective teachers. Differential retention also relies on a valid and reliable measure of teaching effectiveness paired with meaningful incentives. Teacher development is the third mechanism. In addition to a measure that accurately identifies teacher performance, development requires the ability to identify deficient skills and the difficult, time-consuming, and iterative work to alter teachers’ beliefs, behaviors and practices so they

improve. Taken together, these three mechanisms represent the change model by which policy is hypothesized to improve teaching quality and student outcomes.

The implementation of teacher evaluation policy, the action model, involves all the systems, tools, and processes that ensure the altered teacher routines, behaviors, and practices that flow from the change model are reaching classrooms. As an example, for a classroom observation rubric to be an accurate assessment of teaching skills requires that it can measure many of the skills that result in improved student learning and do so reliably across teachers and school settings. The classroom observers need to be calibrated to consistently judge teaching skills across settings and over time. These measured skills, in combination with other processes and systems, then can be employed as part of each of the three mechanisms to alter classroom behaviors of teachers.

Evidence for teacher evaluation comes from five school districts (Cincinnati, Dallas, Chicago, Houston, and Washington DC), three districts which were part of an intensive research partnership, a performance feedback experiment involving 127 schools, and a national quasi-experimental summary to explore the features that may have caused teacher evaluation policy to produce disappointing outcomes. That evidence is summarized in Table 1.

Cincinnati PAR. Results from Cincinnati's Peer Assistance and Review (PAR) program demonstrate that evaluation can generate meaningful improvement in teachers' skills (0.11 SD in math). PAR, which was designed through a partnership of district administrators and the teachers' union, pairs teachers with an expert peer evaluator who is external to the school. More than 90 percent of teachers receive summative scores in the highest two categories, but the scores on individual rubric items and individual observations are much lower providing teachers with the impetus to improve (Taylor & Tyler, 2012).

Dallas PEI & TEI. Dallas Independent School District implemented the Principal Excellence Initiative (PEI) in 2013 as the first step in a comprehensive effort to improve teaching quality and student outcomes. PEI established strong financial incentives and job security consequences for principals tied to their effectiveness in improving teaching effectiveness. Principals received training in teacher evaluation and ongoing support to align classroom observations with the other measures of teaching effectiveness that would be embedded in the Teacher Excellence Initiative (TEI), adopted two years later in 2015. Teachers and administrators have access to a shared library of videos and modules designed to support a common understanding of effective teaching and to guide individual development (Hanushek et al., 2026).

Evaluation scores are used to assign teachers to one of nine performance categories, which directly determines salaries. The district sets a target distribution for each category: the top three performance levels are capped at a combined 20 percent of teachers, and the lowest level is targeted at 3 percent. For most teachers, TEI salaries compare quite favorably to those of the single salary schedule employed in Dallas in 2014-15.

Hanushek and colleagues attribute the math gains primarily to compositional change in the teacher workforce: the evaluation system identified lower-performing teachers and created incentives for their exit, while attracting and retaining more effective teachers. However, differential retention explains roughly half of the total student math achievement improvement, leaving the remainder to be explained by a variety of other factors, including teacher development, school leadership, supports for learning, etc.

Table 1. Design and implementation features of teacher evaluation systems and student achievement effects

System	Rater independence	Observations per year	Sustained rater calibration	Feedback protocol	Link to targeted development	Stakes	Differentiation	Achievement/teacher retention
Cincinnati PAR (Taylor & Tyler, 2012)	Yes — external peer evaluator (3 of 4 observations)	4 (3 peer, 1 principal)	Yes — intensive training, scoring calibrated to rubric	Written feedback after each observation; ≥1 conference	Peer assistance program for low scorers	Low — tenure and some promotion decisions	>90% in top two categories; item-level scores far lower	+0.11 SD math VA, years following evaluation (teachers with 8–17 yrs experience)
Dallas PEI + TEI (Hanushek et al., 2026)	No — administrators, but principals trained and incentivized under PEI two years first	11 (10 unannounced short, 1 × 45 min)	n.r.	Written feedback after every observation; conference after 45-min observation	Shared video and module library (not individualized)	High — nine levels determine salary (\$45k–\$90k)	Lowest level targeted at 3% by district-set distribution	+0.10 SD math by 2019 vs. synthetic control; reading null. ~half attributed to differential retention
Chicago EITP, Cohort 1 (Steinberg & Sartain, 2015)	No — principals	2	Yes — calibration reports vs. normed experts	Rubric-aligned structured feedback within one week	Support for principals coaching teachers	None	n.r.	+0.10 SD reading in year 1; math positive, n.s. Gains in lower-poverty, higher-achieving schools
Chicago EITP, Cohort 2 (Steinberg & Sartain, 2015)	No — principals	2	Reduced central-office support	Same design, less support	Same design, less support	None	n.r.	None
Chicago REACH (Sartain & Steinberg, 2025)	No — administrators	n.r.	n.r.	n.r.	n.r.	High — consequences for low ratings	1% of tenured teachers rated Unsatisfactory	Unsatisfactory rating → exits +50%; replacements +0.60 SD on observation rating, no sig. VA difference
AIR RCT (Garet et al., 2017)	Yes—1 observation by principal, 3 by research team	4	n.r., but experiment lasted 2 years	Written score reports with oral briefing	no	None	98% of CLASS teachers, 87% of FFT in top 2 groups	+0.05 math in year 1, not significant in year 2, reading null
Houston ETI (Cullen et al., 2021)	No—principals	4	n.r.	n.r.	n.r.	Modest financial incentives	At full phase in 64% of teachers effective or better	Exit of top quintile -4.0 pp; Exit of bottom quintile +6.2 pp. No effect on student achievement.
DCPS IMPACT (Dee & Wyckoff, 2015; Adnot et al., 2017; Dee et al., 2021)	Partly — Master Educators (subject specialists) alongside administrators	n.r.	Yes — calibrated annually; cross-observer scores reconciled	Rubric-based standards with ongoing feedback	Weak initially; LEAP coaching from 2016	Very high — dismissal, large bonuses and base-pay increases	15% (2% Ineffective, 13% Minimally Effective), first three years	exit of low performers → +0.14 SD reading, +0.21 SD math; low performers improved by 0.27 SD teaching effectiveness
Typical state system (Bleiberg et al., 2025; Kraft & Gilmour, 2017; NCTQ, 2013, 2015)	No — principals and assistant principals	Varies; observers: insufficient time	Rare	~2/3 of states required feedback linked to PD as adopted	Nominal	States requiring growth measure fell 43 (2015) to 30 (2022)	Very few rated below effective	Little evidence of improvement
Intensive Partnership districts (Stecher et al., 2018)	No — administrators	2–7	No — calibrated at outset, no sustained effort	Principals recommended PD based on evaluation	Available but not required, not tracked; districts struggled to tailor	Low — bonuses of \$2,000–\$3,000 to ~25% of eligible teachers; <1% dismissed	Collapsed once stakes attached: 17%→4%, 13%→8%; 2% in third district	None after six years

n.r. = not reported in the sources cited here.

Chicago EITP, REACH. Chicago's implementation of teacher evaluation provides useful insight across several relevant margins of evaluation. It proceeded in two phases. In a 2008–09 pilot program (EITP), forty-four Chicago public schools adopted a new evaluation system that combined structured classroom observation with intensive implementation support for principals (Cohort 1). Principals received detailed training on how to apply the standardized observation rubric, calibration reports comparing their scoring to normed experts, and support for coaching teachers. Teachers received two observations per year and within a week of each observation principals provided rubric aligned structured feedback to teachers. In short, EITP was designed to differentiate teacher performance and to promote teacher development with no stakes associated with evaluation ratings (Steinberg & Sartain, 2015). The following year, forty-eight additional schools adopted the same evaluation design but with substantially reduced central-office support (Cohort 2). By 2009–10, even Cohort 1 principals received little of the implementation support they had in the pilot year (Steinberg & Sartain, 2015).

The achievement gains realized for Cohort 1 differentially accrued to lower poverty schools and to schools that had already been higher achieving pre-pilot. When Cohort 2 schools adopted the program the following year without comparable support, they saw no equivalent gain, and the advantage of Cohort 1 schools remained but did not grow (Steinberg & Sartain, 2015). These effects reflect the initial response to the first year of the pilot program, so can't reflect differential retention. The authors attribute this directly to implementation: "The fact that cohort 2 schools never saw the benefit that the initial implementers did is likely explained by changes in district leadership and reduced support for principals implementing the pilot system."

Following that first year, Sartain and Steinberg (2016) find EITP also produced improvements resulting from differential retention of low performing and nontenured teachers in Cohort 1 schools relative to the comparison schools. The authors estimate that EITP increased the exit rate of low performing teachers by 30 percent (This result should be interpreted cautiously as there were relatively few low rated teachers); exit rates for untenured teachers increased by about 40 percent (Sartain & Steinberg, 2016). Exit rates for higher performing and tenured teachers were not statistically different. On average exiting teachers were replaced by teachers whose teacher observation ratings by principals were meaningfully higher, suggesting that exit decisions on average improved teaching quality.

By 2012, Chicago had implemented teacher evaluation across all schools under the Recognizing Educators Advancing Chicago Students (REACH) program. That program was similar to EITP but included high-stakes consequences for teachers with low ratings. This resulted in large improvements in teaching attributable to differential retention. So, while evaluation had the intended effect on differential retention, it affected relatively few classrooms.

Chicago teacher evaluation illustrates in miniature the reform as policy problem: in EITP the same policy design, implemented with and without serious support for the adults responsible for carrying it out, produces systematically different results for both the teacher improvement and the differential retention mechanisms; REACH in a broader context shows that evaluation can produce intended outcomes and suggests that stakes may motivate behavior change.

AIR RCT. In 2012 AIR began a two-year randomized controlled trial in 127 elementary and middle schools to better understand the effects of performance feedback on teachers and principals. Half of the treatment teachers received feedback four times a year from the Danielson Framework for Teaching rubric and half received feedback from the CLASS rubric; teachers in

grades and subjects that allowed, received value added scores in math and/or reading (Garet et al., 2017). Teachers did not receive coaching beyond the feedback sessions, and no incentives were tied to ratings' performance. Treatment teachers were more likely to receive feedback than control teachers (80 percent v. 50 percent) and received feedback more than four times as often.

The results were modest. Treatment teachers outperformed control teachers on classroom practice when measured by CLASS but not when scored by FFT. The CLASS measured gains moved teachers from the 50th to the 57th percentile of the distribution. In year 1 student math achievement was 0.05 SD higher for treatment than control teachers, year 2 differences were about the same but not statistically significant. Reading achievement was smaller and not statistically significant (Song et al., 2021). The authors indicate that the scores on specific aspects of classroom observation were not sufficiently reliable to differentiate areas of focus for individual teachers (Garet et al., 2017). As a result, the scores provided limited information to help teachers know where to focus their efforts to improve or specific practices to change. In addition, teacher ratings were unconnected to any incentives to improve.

Houston ETI. By 2012 Houston Independent School District had fully phased in the Effective Teacher Initiative (ETI). The goal of ETI was to inform more rigorous reviews of teacher performance, which could be used for personnel decisions, such as involuntary exit. In 2012-13 64 percent of teachers received ratings in the top two evaluation categories. Teachers in these evaluation categories are eligible for modest financial rewards. The policy increased the attrition of bottom quintile teachers by 6.2 percentage points and reduced attrition of top quintile teachers by 4.0 percentage points. A back of the envelope calculation showed the net effect of the bundle of reforms pre-policy to post-policy on student achievement was effectively zero. That resulted from an overall increase in attrition across all school achievement quintiles and that replacement teachers were roughly as effective as bottom quintile teachers. The authors can't discern whether it was only the differential retention policy that led to the overall increase in attrition or whether other contemporary changes contributed.

The effects of the policy were consistent with the differential retention theory of change. However, the composition change in the workforce was too small and the policy was not sufficiently attractive to many higher performing teachers to remain in the district, when combined with a relatively weak replacement pool led to disappointing achievement outcomes.

Washington DC IMPACT. IMPACT is perhaps the most robust example of teacher evaluation improving student outcomes. District of Columbia Public Schools (DCPS) launched IMPACT in the 2009–10 school year, a system that set clear standards for effective instruction and provided supports and incentives for teachers to meet those standards. Research following three years of IMPACT implementation found evidence of both differential retention—the lowest performing teachers involuntarily exited, lower-performing teachers voluntarily exited, replaced by higher performing teachers, and higher performing teachers received large financial incentives; teachers facing incentives to improve did so (Dee & Wyckoff, 2015).

More recent work demonstrated that these effects persisted through IMPACT's first decade, a notable finding given that implementation fatigue and leadership transitions had caused other evaluation systems to weaken over time (Dee et al., 2021). In sum, research on IMPACT demonstrates that it had large gains associated with both differential retention and teacher development.

What might account for these gains in teaching effectiveness and student achievement? The answer lies less in the specifics of how evaluation was designed than in the attention to implementation including the infrastructure built around the system and the institutional commitment to improving the system over time (Toch, 2018).

From the outset, IMPACT was designed both to help teachers develop and to hold teachers accountable—and it took both functions seriously. On the accountability side, the system provided powerful incentives: very large increases in base pay and substantial bonuses for teachers rated Highly Effective, and dismissal for those rated Ineffective or the credible threat of dismissal for those rated Minimally Effective. These incentives were large enough to meaningfully alter the calculus of staying, leaving, or improving.

The developmental infrastructure was equally serious. Teachers received clear, rubric-based standards that defined what effective instruction looked like, along with ongoing feedback to help them move toward those standards. Classroom observations were conducted not only by school administrators but by subject-matter specialists—designated Master Educators—who brought deep pedagogical content knowledge to the observation process. All observers were calibrated annually against the standardized rubric, and within-teacher scores across observer types were compared and reconciled to ensure consistency.

These design choices reflect a coherent and consistent focus on the goal of improving teaching quality. By investing in measurement accuracy and consistency, IMPACT could differentiate meaningfully among teachers rather than sorting them into broad, undifferentiated categories. Over the first three years 2 percent of teachers were rated as Ineffective, 13 percent Minimally Effective, 69 percent Effective, and 16 percent Highly Effective. That differentiation, in turn, made individualized feedback possible, giving each teacher a picture of where their practice fell short. Most DCPS teachers believed that expectations for effective teaching were clearly defined (80 percent), that IMPACT ratings were an accurate reflection of teacher effectiveness (65 percent), and that the evaluation process helped identify teaching strengths and weakness (69 percent) (James & Wyckoff, 2020).

DCPS operated with an advantage that some districts lack: a strong external supply of teachers that made the replacement of Ineffective teachers both feasible and productive. When low-performing teachers exited under IMPACT, DCPS could draw on a deep applicant pool to recruit more effective replacements—a condition that made credible dismissal incentives not merely administratively possible but educationally meaningful. Dismissing an ineffective teacher is of limited value if the replacement is no more effective, and strong accountability provisions may therefore carry higher risks and lower returns in lower-supply labor markets. In these contexts, a rebalanced evaluation design—one that places greater weight on teacher development and professional support, and less on high-stakes accountability—may be better suited to local conditions.

Over time IMPACT policymakers realized the original design had flaws. DCPS surveyed teachers annually on their experience with IMPACT and used that feedback to drive meaningful revisions. When teachers reported that value-added scores, which initially composed 50 percent of the evaluation for teachers for whom they could be calculated — felt opaque and that school value added was dependent on the performance of colleagues, policymakers responded. Three years after launch, the weight of individual value-added was reduced to 35 percent, and the school-wide value-added component was eliminated entirely. Policymakers recalibrated the

thresholds tied to high-stakes personnel decisions and restructured the bonus system to concentrate the largest financial incentives in the highest-poverty schools where the need for effective teaching was greatest. While that original system may have been necessary to garner initial buy in, policymakers tightened the focus on schools with the lowest student performance. Realizing that early efforts to foster teacher development were inadequate, in 2016 policymakers eliminated the external coaches and introduced the Learning Together to Advance our Practice (LEAP) initiative, an intensive coaching program designed to provide teachers with timely coaching and professional development tailored to their specific subjects and developmental needs (Toch, 2018).

This iterative quality — the willingness to treat the evaluation system as a work in progress rather than a finished product — is as instructive as any of IMPACT's design features. Most state evaluation reforms were mandated, implemented similarly across districts, and then largely left unchanged as political attention moved on.

Most states and districts. Many states subscribed to the idea of teacher evaluation but were less effective in ensuring the implementation of its critical features. By 2013 virtually all states had some form of teacher evaluation, but just 12 states required a specific system for all of their districts and of the states where districts can choose different approaches only 15 required state approval for the evaluation system. How actual implementation may have varied from policy is unclear (NCTQ, 2013).

Relatively few states included what design experts considered to be important components of teacher evaluation systems. For example, only about a third of the 44 states adopting teacher evaluation policies included several rigorous features of performance and just two-thirds required evaluation-based feedback that links to coaching or professional development (Bleiberg et al., 2025). Based on these individual components of an idealized teacher evaluation system, Bleiberg and colleagues classified fewer than half the states as “high quality”, signaling some design weaknesses. Even so, the measure is based on whether states include these features in their adopted systems, and not whether, or the extent to which, districts implemented these requirements with fidelity. Perhaps not surprisingly, Bleiberg and colleagues found no link between the features included in evaluation systems and their effect on student outcomes, suggesting that design did not differentiate among more and less successful systems.

A key mechanism of teacher evaluation is the accurate differentiation of teacher effectiveness and constructive feedback aligned with the evaluation. As of 2013 44 states and DC were required by policy to perform classroom observations as part of that evaluation, but only 35 required evaluators to be trained, 13 states and DC required evaluators to be certified, only 4 states and DC required multiple evaluators, 21 states and DC explicitly required teachers receive feedback on observations, and 19 states and DC require in policy that evaluation results inform professional development (NCTQ, 2013). In practice, principals often suffered from leniency bias (Kane & Staiger, 2012; Kraft & Gilmour, 2017), typically had insufficient time to conduct rigorous evaluation and provide feedback to all teachers (Neumerski et al., 2018; Sartain et al., 2011) and these administrators were often not well trained to provide constructive feedback (Hunter & Springer, 2022; Kraft & Christian, 2022; Steinberg & Sartain, 2015).

These and other differences between design and practice led Bleiberg and colleagues to conclude that states had limited control over district adaptations. “Consequently, when brought to scale,

new evaluation systems tended to differ in meaningful ways from the stylized ideal envisioned by reformers.” (Bleiberg et al. 2025, p. 576)

States created accountability mechanisms without ensuring the instructional improvement infrastructure necessary to make accountability productive. These systems did not build rigorous measures of teacher evaluation, thwarting both the ability to effectively execute differential retention and the ability to accurately inform the process of teacher development. In addition, they did not build the infrastructure necessary to provide the feedback and coaching needed for meaningful teacher development. In doing so, they separated the lever—evaluation—from the mechanism by which it was supposed to operate—improved teacher practice. An evaluation system that does not tell teachers what to do differently is, for most teachers, professionally useless.

Intensive Partnerships for Effective Teaching (IP). The Gates Foundation designed and funded an intensive examination of teacher evaluation in three school districts and four charter CMOs. IP had a clear design theory of change that identified the mechanisms commonly hypothesized to improve teaching quality and student outcomes. The Foundation chose sites based on their interest in transforming their HR practices to align with the Foundation teacher evaluation theory of change and their commitment to improving teaching effectiveness. The Foundation provided substantial funding and guidance about several of the design elements and a program officer who could assist in planning implementation. Yet, after six years of operation the RAND and AIR team assembled to evaluate the outcomes found no evidence of improvement on either teaching effectiveness or student achievement (Stecher et al., 2018). IP required participants to address each of the primary mechanisms required for effective teacher evaluation—rigorous recruitment, differential retention and teacher development. It provided generous support and guidance on many aspects of the theory of change. In short, it seems to have attempted to head off the reasons why many districts failed to successfully implement teacher evaluation. A review of their very thorough analysis adds useful nuance to our assessment.

A potentially important lesson can be drawn from the pattern of how teacher evaluation ratings evolved over the project. In two of the three districts a reasonable degree of rating differentiation existed in the first-year of implementation—13 percent in the lowest two of four categories in one district, 17 percent in the other (the third district had only 2 percent of teachers in the lowest two categories). By 2014, when stakes began to be associated with ratings, many fewer teachers were assigned to those lowest ratings categories—(17 percent dropped to 4 percent, 13 percent to 8 percent) (Stecher et al., 2018, p. 76). None of the districts made sustained efforts at rater calibration to guard against leniency bias, a well-known problem in the personnel literature (Murphy & Cleveland, 1995). It is also true that in 2014 one district included the use of teacher designed SLOs in the evaluation and required that VAMs have the same distribution across ratings categories as the SLOs. While it is possible that teaching quality made substantial improvements coinciding with these changes, that would be quite coincidental.

Because nearly all teachers received ratings of at least Effective, very few teachers were subject to dismissal or required improvement plans. Less than 1 percent of teachers in the two districts for which data are available were dismissed (Stecher et al., 2018, p. 171). Because evaluations were largely undifferentiated, few teachers were given the signal they needed to improve, unlike the culture in Chicago and DCPS. As a result, neither differential retention nor development had the key ingredient to be effective.

Principals in all three districts used evaluations to guide specific recommendations to teachers for professional development and all three districts had combinations of mentors, PD workshops, coaches, videos, and other supports. Surveys and interviews made clear that districts struggled to tailor individual professional development to a teacher's evaluation (Stecher et al., 2018, p. 216) citing the burden on already overworked principals.

Under IP the district policies of additional compensation based on teacher evaluation scores varied widely but generally drove modest bonuses, available in only some schools and for some years. Additional compensation at this level is well below that typically found to be effective in altering teacher behavior and the RAND/AIR surveys and interviews found that few teachers believed that additional compensation altered their behavior (Stecher et al., 2018, p. 277).

The IP analysis shows that the three districts designed the measurement of teacher effectiveness consistent with best practice at the time and quite similarly to what occurred in the four districts described above as successes. Design across other elements of teacher evaluation—the use of meaningful stakes attached evaluation ratings to motivate behavior and the link between ratings and individualized professional development conformed less well. Importantly, implementation across most dimensions of teacher evaluation was insufficient. Most telling is the differentiation of teacher ratings. Once stakes were attached to the evaluation, very few teachers were assigned ratings that would signal a need for improvement, suggesting that leniency bias may have crept into principals' evaluations. Moreover, despite good intentions from principals, districts had not established professional development systems that linked to a teacher's evaluation. Most teachers are motivated to improve if they trust the system assessing their skills and behaviors. For these teachers, IP failed to connect these teachers to resources targeted to their weaknesses. Even though the IP districts had benefited from substantial additional resources and from guidance around best practice, these districts failed to implement key elements of the teacher evaluation theory of change successfully.

Beyond Design Versus Implementation

The evidence shows that teacher evaluation policy largely failed on implementation, but the reasons for implementation failure matter. Cincinnati, Dallas, Chicago, the AIR/IES experiment, and Washington implemented the change model somewhat differently, but all achieved at least some success when that same model failed nearly everywhere else, which is evidence that the change model is not inherently broken. Five cases cannot establish that its mechanisms would work anywhere, a limit Houston and the Intensive Partnership results make impossible to ignore. Three reasons for implementation failure stand out—failure to foresee the limits of some mechanisms, failure to share responsibility for implementation, and failure to monitor and adapt.

Limits of mechanisms. The natural inference from concluding that implementation is at fault is that districts simply need to focus harder or spend more. That is misleading because some of the model's requirements are not matters of local effort. Differential retention that requires or encourages the exit of ineffective teachers works only where the labor market can supply teachers who are more effective. Schools in many districts have an insufficient supply of teachers to withstand the short run effects of high-stakes induced exits. Houston is a good example. The mechanism works when states and districts address this supply condition, either by expanding the labor pool or by modifying how differential retention is implemented. A weak labor supply should have been accounted for as teacher evaluation was rolled out. This doesn't eliminate the need for all districts to implement the other components of evaluation—the

rigorous measurement of teaching skills that are the basis for high-stakes incentives to high performing teachers and development for all teachers.

Shared responsibility. Some of the conditions necessary for teacher evaluation to function may be beyond the capacity of many districts. Differential retention and teacher development only work if evaluations are a valid measure of effectiveness. In practice that means that classroom observations must reliably measure performance accurately. Whether most districts could have met that requirement depends on the conditions necessary for its fulfillment. External evaluators with subject and grade-level expertise, calibrated annually, are expensive and require enough scale to justify the specialization. DCPS could afford Master Educators and ongoing calibration only with generous external support, the AIR schools did so with outside funding. Cincinnati and Washington are both large systems. Districts with perennial teacher shortages or who are quite small are unlikely to have the capacity to accurately assess teaching or to provide the expertise for effective coaching. The state, which could have, did not. Most states mandated the components of evaluation while retaining almost none of the functions that made them work, leaving each district to build accurate measurement, calibration and coaching capacity alone.

Adaptation. Policymakers at all levels must adapt to emerging evidence. Rarely is policy enacted where all the conditions for its success can be anticipated and built into the initial policy design and implementation. By 2013 the ratings differentiation problem was documented and more to the point should have been recognized by states and districts performing their implementation due diligence. They would have also observed schools where fewer than half the students passed statewide achievement tests even though nearly all its teachers were rated as effective or better. While not conclusive, this should have at least raised concerns about leniency bias in evaluation. Yet nothing changed in those states and districts. The lack of adaptation by states and districts to these signals is telling. Well documented adaption in Washington, DC is the exception.

The Intensive Partnership sites fit less comfortably, which is why they matter. Each opted into the IP partnership with the advantages it provided in advance of state designed teaching evaluation systems. Each decided to employ principals as classroom observation raters and experienced leniency bias once stakes were tied to ratings. That weakened differential retention and teacher development. IP districts made professional development available, but did not require its use except under an improvement plan (which covered very few teachers), did not track it, and reported that tailoring development to a teacher's evaluation exceeded what already-overburdened principals could do (Stecher et al., 2018, p. 216). Two explanations seem plausible. Although the Gates Foundation surely understood the importance of accurate ratings (Kane & Staiger, 2012), they may have underestimated the potential for leniency bias to compromise ratings accuracy. Alternatively, the Foundation may have been unwilling to provide the level of support necessary for external evaluators who are regularly calibrated, or if they did, may not have mandated to the IP districts that they use funds for that purpose. The evidence doesn't distinguish. Regardless, the experience of the IP districts implies important lessons.

Implementation of teacher evaluation likely suffered in different ways across varying circumstances. Taken together they suggest that systemwide reform is challenging, especially when its goal is to change the routines, knowledge and skills that teachers rely on to produce durable changes in classroom practices. Similar implementation challenges have weakened other promising policies.

Other Policy Disappointments, Similar Implementation Failures

Teacher evaluation is not the only policy targeting teaching effectiveness whose broader policy rollout was preceded by a compelling design with empirical evidence that raised expectations for meaningful impacts only to disappoint. Below a summary of two other policies provide additional evidence that improving teaching quality is demanding in expertise, resources and time; factors that are often lacking when policies are implemented at scale.

The Science of Reading: The right law, weak investment in implementation. The scientific evidence on early reading instruction is unusually robust. Decades of cognitive, linguistic, and neuroscientific research have established that systematic, explicit phonics instruction is essential for most children to develop foundational reading skills, and that the dominant approach of the last generation — balanced literacy, which includes phonics but balances that with whole language strategies— is not supported by evidence and produces worse outcomes for a substantial portion of students (Moats, 2020; National Reading Panel, 2000). This consensus has given rise to a wave of state legislation mandating science-of-reading approaches, widely hailed as an overdue reckoning with decades of failed practice.

Yet the translation of this evidence into improved student outcomes has been uneven (O'Donnell, 2025), and the variation is instructive. Mississippi's experience — widely described as the "Mississippi Miracle" — represents the most striking success. In 2013, Mississippi enacted the Literacy-Based Promotion Act, which required a multi-component implementation process including teacher training, early screening, intensive intervention, and third-grade retention, with the goal that all children would read at grade level by the end of third grade (Mississippi SB2347, 2013; RMC Research Corporation, 2019; Spencer, 2024). Mississippi's fourth-grade NAEP reading scores rose dramatically; by 2024, the state ranked 9th nationally in fourth-grade reading, having risen from 49th in 2013 (NAEP, 2026b). Research found that students exposed to the LBPA from kindergarten through third grade showed gains of approximately 0.23 standard deviations in reading and 0.29 standard deviations in math on the grade-4 NAEP — effects equivalent to roughly 75 percent of a year of academic progress; effects for all students, some of whom with partial exposure were 0.14 SD in reading (Spencer, 2024).

What Mississippi did was categorically different from what most states have done, and the difference lies in the investment in implementation. In its first year, the state committed \$9.5 million to provide all K-3 educators with intensive LETRS (Language Essentials for Teachers of Reading and Spelling) training, a program requiring approximately 160 hours of engagement over two years, designed to change not merely teachers' instructional behaviors but their foundational understanding of how children learn to read. The state also deployed literacy coaches to its lowest-performing schools — trained professionals who spent two to three days per week in a school for an entire year, modeling instruction, observing, and providing feedback (Wright, 2022). The coaches provided the sustained, contextualized feedback required for new instructional behaviors to take root. Mississippi's gains also reflected a broader system-level commitment — encompassing third-grade retention, a 2015 state test realignment with NAEP, and strengthened standards-based accountability, making it difficult to attribute gains to any single component (Barnum, 2023).

By contrast, most states have operationalized reform primarily through curriculum adoption mandates, requiring districts to purchase materials aligned with structured literacy without

adequate attention to the instructional routines and knowledge teachers bring to those materials (CCSSO, 2024). Most reading laws passed between 2019 and 2022 provide no roadmap for changing instructional practice or developing instructional leadership (Albert Shanker Institute, as cited in Schwartz, 2023). This matters because the obstacle to better reading outcomes is the teacher. Most current elementary teachers were trained under balanced literacy frameworks whose theory of reading development is fundamentally incompatible with systematic, explicit phonics instruction (Ellis et al., 2023).

Asking these teachers to deliver explicit, systematic decoding instruction requires a fundamental reconceptualization of reading development, one that takes sustained professional learning to produce. Teacher knowledge can be increased and practice can shift, but these shifts do not reliably translate into higher student achievement without aligned curriculum, expert coaching, and sustained accountability for implementation fidelity (Schwartz, 2022). Adopting a different approach depends on whether that policy supports and sustains altered classroom practices aligned with the science of reading. Whether policymakers facilitate that process with complementary policies and rigorous implementation seems to differentiate successful and unsuccessful adoptions.

Instructional Coaching: The promise and the paradox of scale. Instructional coaching has become one of the most widely adopted approaches to teacher development in American education, and for good reason: the research base is encouraging. A meta-analysis of 60 causal studies found pooled effect sizes of 0.49 standard deviations on instructional practice and 0.18 standard deviations on student achievement—effects that, if realized at scale, would represent large improvements in the quality of teaching (Kraft et al., 2018).

Yet embedded in the same meta-analysis is a finding that policy designers have too often overlooked: average effects from large-scale effectiveness trials (0.10 SD student achievement) are less than half of those from smaller efficacy trials of more tightly controlled programs (0.28 SD). The paradox of coaching at scale is that it appears to change teacher practice when it is done well but doing it well at scale is genuinely difficult. The research consistently finds that coach quality varies substantially—and that the attributes most predictive of impact, including deep content knowledge, preparation in adult learning, small number of teachers per coach, and the ability to identify and prioritize specific instructional behaviors from classroom observation, are frequently absent in large-scale implementations (Desimone & Pak, 2017; Marsh, McCombs & Martorell, 2012). Yet these are precisely the dimensions of effective coaching that get compromised at scale (Kraft et al., 2018).

The mechanism by which coaching changes instructional practice is well established: a coach observes specific classroom behaviors, provides feedback that surfaces the gap between current and targeted practice, and supports the teacher in rehearsing the alternative behavior before enacting it independently (Joyce & Showers, 2002; Kraft et al., 2018). Coaching that substitutes general encouragement, logistical support, or relationship maintenance for this cycle does not activate that mechanism and does not produce instructional change. Kraft and colleagues' meta-analysis finds that the number of coaching contact hours is not associated with larger impacts, which leads them to conclude that “the quality and focus of coaching may be more important than the actual number of contact hours” (Kraft et al., 2018, p. 565). What drives improvement is whether the interaction actually shapes what that teacher does in the classroom.

This implies that scaling coaching requires screening candidates for coaching effectiveness, providing rigorous preparation to those who need it, and monitoring coaches' impact on teacher practice rather than merely their activity levels. When coaching is scaled, the features of pilot programs that produce focused professional dialogue that result in changed instructional behavior are often neglected leading to limited improvements in teaching practice. Visible, countable features of a reform are treated as proxies for the harder-to-observe work of observing changing instructional behavior.

The Political Economy of Underinvestment

Understanding why education reforms so consistently underinvest in features necessary for teacher behavioral change requires more than pointing to the technical failures of implementation planning. There are structural features of the political economy of education reform that systematically bias toward shallow, rapid adoption of policies and against the sustained, expensive, and slow-moving work of changing instructional practice at scale. These pressures are predictable features of the democratic governance of public education.

The most fundamental problem is the mismatch between political and implementation time horizons. Elected officials—governors, legislators, and school board members—operate on electoral cycles of two to four years. The benefits of genuine instructional reform, by contrast, accumulate over five to 10 years or more, as implementation matures, teacher knowledge deepens, and the full cohort of students moves through a changed instructional environment. This creates what political economists call a “diffuse and delayed gains” problem: the costs of reform (political conflict, union resistance, budget expenditure, disruption of teaching) are concentrated and immediate, while many of the benefits are spread across numerous stakeholders and will not be visible until long after the next election. Rational political actors have incentives to adopt noticeable reform quickly—enacting a law, announcing a new system—and weaker incentives to invest in the unglamorous, years-long work of implementation that produces results (Hess, 2017; Moe & Wiborg, 2017; Tyack & Cuban, 1995).

This dynamic is compounded by the structure of education governance in the United States, which is unusually decentralized by international standards. Federal and state governments can mandate evaluation reform or literacy legislation, but the actual work of implementation falls to approximately 13,000 local school districts, each with its own governance structure, political dynamics, and capacity constraints. State mandates typically arrive with limited funding and technical assistance, creating a “capacity constraints” problem: districts are required to implement complex systems—observation rubrics, feedback protocols, curriculum, coaching programs—that require significant professional capacity to execute well, without the resources to build that expertise (Bleiberg et al., 2025).

The interest group dynamics of education reform introduce further distortions. Teacher unions represent the most organized and politically active constituency in K-12 education. They have strong incentives to resist reforms that increase accountability for teacher performance, and they are well-positioned to negotiate away the provisions most directly tied to changing teacher behavior—high-stakes consequences for low performance or differentiated performance-based compensation, frequent observation cycles, coaching and supervision structures that make evaluation meaningful, and school accountability that provides incentives for improvement. In many states, the most consequential elements of evaluation reform were progressively weakened through union negotiation or legislative revision before they could take effect. The National

Council on Teacher Quality documented that the number of states requiring student-growth data in teacher evaluations dropped from 43 in 2015 to 30 by 2022, as states retreated from the accountability provisions that were generating political opposition (NCTQ, 2022).

This is not simply a story of interest group obstruction; unions also expressed genuine concerns about the fairness and validity of evaluation systems implemented without adequate teacher input or professional development support. But the political resolution of these conflicts tended to hollow out the reforms rather than address their underlying implementation deficiencies.

The current teacher shortage has sharpened this dynamic. As districts struggle to fill classrooms, the political case for rigorous accountability becomes harder to sustain, and both unions and administrators face stronger incentives to ease pressure on teachers. Yet the evidence points in the opposite direction: teachers who work in schools with strong instructional leadership, useful feedback, and genuine professional development opportunities are more likely to stay in the profession (Kraft & Papay, 2014; Simon & Johnson, 2015). The shortage does not make investment in teaching quality less necessary, it makes it more urgent. Reducing expectations may ease short-term recruitment and retention pressures, but it does so by making teaching less professionally rewarding, which compounds attrition over time. The more productive response is to make investments that address the working conditions driving teachers out of the profession — in compensation, planning time, coaching, and manageable caseloads — without sacrificing commitments to instructional improvement.

Finally, the funding structures of American education create a persistent bias against the recurrent investment in human capacity that sustained implementation requires. State and federal revenues, which constitute about 55 percent of district budgets, fluctuate with economic cycles and shifting political priorities. When revenues decline, districts have consistently reduced the most discretionary expenditures, including professional development, coaching, and capacity-building investments, before cutting core instructional or contractually obligated costs (Jackson et al., 2021). Sustained multi-year coaching and teacher evaluation that produce genuine changes in instructional practice are recurrent expenditures whose benefits are hard to quantify, which are often viewed as discretionary, and whose beneficiaries, future students, have no political voice.

Designing Reform to Reach Teachers

The 2024 NAEP results — showing continued declines in most grades and subjects, achievement gaps widening, and the share of scoring below basic in reading at the highest level over the last 20 years (NAEP, 2026a)— have reinvigorated calls for education reform among policymakers, researchers, and the press (Hess, 2026; Mahnken, 2026; Smith, 2025).

As the policy community considers how best to address these challenges, two broad lessons are worth considering. First, it is difficult to imagine sufficient progress on student achievement without meaningful improvements in teaching quality, for which differential retention and teacher development are fundamental. Policymakers should recognize that the teacher quality agenda and the teacher supply crisis are not separate problems requiring separate solutions. Investing in teacher development is not simply a strategy for improving instruction, it is a strategy for building the sustainable profession that the achievement imperative requires. A better informed, revitalized, teacher evaluation system is not just a policy option, it is an approach policymakers concerned about student outcomes should embrace.

Second, across the three examples discussed here, the variable that most reliably separates successful from unsuccessful adoptions is whether the reform invested seriously in implementation procedures to change teachers' skills, routines and classroom practices. Each domain rests on a well-grounded body of research; in each, some implementations produced substantial improvements in student outcomes; in each, broader rollouts produced weaker, more inconsistent, or null results. Scaling requires a clear understanding of the influential mechanisms and a bit of humility. Demonstration programs and early pilots almost always operate with resources, leadership attention, and talent pipelines that have not been replicated at scale. DCPS had concentrated philanthropic funding, mayoral control that simplified governance, and the recruitment advantage of operating in a high-profile reform environment that attracted unusually committed principals and coaches. Mississippi's literacy reform benefited from a sustained political commitment across multiple gubernatorial administrations and from significant state investment in training infrastructure. When reformers observe these successes and call for replication, they rarely account for the full bundle of conditions that produced them, nor do they appreciate how implementation needs differ across the quite varied contexts of school districts.

The review of teacher evaluation policy provides guidance around several more specific recommendations for policymakers and researchers. The evidence from this era shows that some features are required for the evaluation mechanisms to be effective; others are important, but their intensity, scale, or frequency are uncertain; while the usefulness of a third group of features is unresolved (column 1, Table 2). A summary of evidence of the details of the policy design and implementation are provided in columns 2 and 3; column 4 suggests a research agenda that could resolve remaining design and implementation questions.

The evidence identifies four features that are necessary if teacher evaluation mechanisms are to function properly: valid and differentiated classroom observation ratings of teacher performance for differential retention and item level differentiation for development, a sufficient reserve pool of effective teachers, specific, rubric-based feedback on teachers' own classroom performances, and sufficient time for the culture of the system to accept evaluation and for teachers' skills and practices to embrace feedback. Any system considering a teacher evaluation policy must ensure these features are successfully implemented. Otherwise, student outcomes will be disappointing. The open questions for these features pertain to their nature not whether they are necessary.

A second group of features—the contingent bucket—each have open implementation questions that vary by school context. For example, systems that want to make use of differential retention need to consider whether the teacher labor market can support incentives to exit for low performing teachers or whether financial incentives are consistent with current and future compensation capacity. Importantly, we have much to learn about how to implement successful teacher development at scale. Most of the achievement gains in the four district successes accrue to differential retention; Cincinnati and the AIR schools being the exception. Development holds the greatest promise for broad-based improvements in teaching quality and student outcomes. It remains to be seen if that promise can be broadly realized.

The final bucket includes features for which the evidence is too weak to provide much guidance. For example, whether principals can serve as effective raters who provide useful feedback depends on whether safeguards to avoid leniency bias can be implemented, and whether they can devote sufficient time to rating and feedback.

Table 2 What the evidence supports, what it doesn't, and research to fill in the gaps

Feature	What Evidence Shows	Confidence	What Resolves the Open Question
A. Default Requirements—presumed necessary absent evidence to the contrary			
Ratings that differentiate performance	Present in DCPS, Dallas, not summative ratings in Cincinnati. Collapsed in the IP districts when stakes were attached. Very few rated below effective in most states.	Moderate. Claim is analytic as well as empirical: neither differential retention nor targeted development can operate on a signal with no variance.	Does it matter if summative rating is not very differentiated if individual rating items are?
Supply of effective teachers to address induced increases in exits	Systems that rely on differential retention had sufficient supply, e.g., DCPS, Dallas. Systems with high levels of policy induced attrition may not, e.g., Houston	High. Challenge is determining the effectiveness of reserve labor supply, which is influenced by many factors in addition to the policy.	How can schools mitigate against this by making teaching more attractive, e.g., compensation, working conditions coincident with policy.
Specific, rubric anchored feedback on teachers' own practice	Present in all four successes and 21 states.	Moderate to high on nature; low on components and timing.	Direct comparison of cadence and format: written alone versus written plus conference; 1 week versus 1 month.
Sustained commitment	Four successes persisted for many years, with effects downstream; research finds durable practice change is 2 to 4 years;	High — analytic as well as empirical, entailed by any change model running through skill acquisition.	Open questions: how long and whether the clock starts at adoption or when implementation stabilizes.
B. Contingent—the conditioning variable is identified; the magnitude is unclear			
Stakes intensity for differential retention	DCPS's retention gains presuppose a strong replacement pool. Kraft et al. (2020) find high-stakes evaluation reduced teacher newly licensed 16–18% but increased quality.	Moderate on direction. None on threshold. That the returns to dismissal fall as applicants per opening fall is close to arithmetic. Where the crossover sits is unknown.	Cross-district variation in applicants per opening against measured returns to the retention mechanism. Largely answerable with administrative data.
Observation regime	Dallas used eleven short observations plus one long; Cincinnati used four long ones. Both produced effects. Several states only evaluated teachers with ratings below effective.	Moderate. To what extent does frequency v. length matter? Omitting evaluations for teachers rated as effective or better meaningfully limits development mechanism.	Within-system comparison of frequency-length combinations against a common reliability target.
Development access to coaching expertise	Development in IP districts available but not required or tailored, missing AIR/IES schools. Coaching effects attenuate substantially at scale.	Moderate/low	Whether automated feedback substitutes at the margin. Whether consortium or external-provider models deliver tailoring that districts cannot.
C. Unresolved—empirical evidence insufficient			
Rater insulation from supervisor	Cincinnati and DCPS produced greater ratings variation. Four states required multiple evaluators. Dallas produced effects with administrators as raters.	Largely untested.	Within-district random assignment of teachers to external-evaluator versus principal observation.
Sustained calibration	Most states trained; 17 states, IP, and Cincinnati, calibrated initially; Chicago, DCPS calibration annually.	Low. Every success calibrated somehow; no two did it alike. Most states calibrated initially but not ongoing.	Within district design, varying calibration regime.
Stakes for development	Cincinnati produced development at modest stakes; DCPS produced both mechanisms at very high stakes.	None.	Within-system variation in stakes holding observation and feedback fixed. By necessity different districts.
Student growth measure	All successes included some student growth component. IP's value-added distributions forced to match SLO distribution.	None. Field has never tested observation-only against composite systems.	Direct comparison of observation-only and composite systems, and validation of whatever measure is proposed for non-tested grades and subjects.

Note: Evidence presented in table is taken from earlier discussion.

Better understanding and elucidating the mechanisms, including implementation, that cause a policy to improve teaching will enable better policy (Coburn & Penuel, 2016). For example, teacher evaluation with these features, not merely teacher evaluation. It is unlikely that a single study or even a single research group could embrace each of the suggested research dimensions. Rather, research funders working with a network of researchers could coordinate a structured research program that would include the features of a policy's implementation, as well as its important dimensions of the treatment design (Bryk et al., 2015).

As noted at the outset of the paper, student achievement has fallen sharply. Since 2013 NAEP 8th scores for the students at the lowest decile of test takers have declined by 0.51 SD in reading and math scores have declined by 0.47 SD—each more than a year of education. These declines are illustrative of much broader declines in student outcomes. Teacher quality is among the few school inputs with a large and well-documented effect on student outcomes which is why policies aimed at improving it warrant serious attention.

The logic of evaluation is straightforward: identify what a teacher does well and poorly, then supply the feedback and coaching that close the gap. The record examined here indicates that this logic holds where the necessary features are present and fails where they are not. Cincinnati, Dallas, and Washington produced math gains of 0.10 to 0.21 SD through differential retention or development; Chicago and Washington produced reading gains of 0.10 to 0.14 SD. Though seemingly modest when compared to the achievement declines since 2013, gains of this magnitude are among the largest 10-20 percent of published interventions implemented with large student populations (Kraft, 2020). Whether the gains grow when both mechanisms are used well, and whether such systems can be implemented and sustained at scale, remain open questions. What the record does not support is the conclusion drawn at the close of the last reform cycle — that teacher evaluation was a failed intervention. For the most part, it was not implemented as designed, and the evidence from the places that did implement it deserves a more careful read.

References

- 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.
- 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.
- Albert Shanker Institute, Reading Reform Across America,
<https://www.shankerinstitute.org/literacy>
- Barnum, M. (2023) “Mississippi made big test score gains. Here is what to make of them” Chalkbeat, July 18, 2023. <https://www.chalkbeat.org/2023/7/18/23799124/mississippi-miracle-test-scores-naep-early-literacy-grade-retention-reading-phonics/>
- Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2025). Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment. *Journal of Political Economy Microeconomics*, 3(3), 568-610.
- Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.
- Capano, G., & Howlett, M. (2021). Causal logics and mechanisms in policy design: How and why adopting a mechanistic perspective can improve policy design. *Public policy and administration*, 36(2), 141-162.
- Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Sage.
- 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.
- 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.
- Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational researcher*, 32(6), 3-12.
- Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, 45(1), 48–54.
- Cohen, D. K., & Ball, D. L. (1999). Instruction, capacity, and improvement. CPRE RR-43.

Cohen, D. K., & Mehta, J. D. (2017). Why reform sometimes succeeds: Understanding the conditions that produce reforms that last. *American educational research journal*, 54(4), 644-690.

Council of Chief State School Officers. (2024). *Science of reading legislation and implementation state scan*. <https://learning.ccsso.org/science-of-reading-legislation-and-implementation-state-scan>

Cowen, J. M., & Goldhaber, D. (2024). *The declining status of the teaching profession*. NBER Digest, August 2024. <https://www.nber.org/digest/202408/declining-status-teaching-profession>

Cuban, L. (2022). *Confessions of a school reformer*. Harvard Education Press.

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.

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

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.

Desimone, L. M., & Pak, K. (2017). Instructional coaching as high-quality professional development. *Theory into practice*, 56(1), 3-12.

Doty, E., Kane, T. J., Patterson, T., & Staiger, D. O. (2025). What do changes in state NAEP scores imply for birth cohorts' later life outcomes? *Journal of Policy Analysis and Management*. <https://doi.org/10.1002/pam.70018>

Dynarski, M. (2016) “[Teacher observations have been a waste of time and money](#)” Brookings Institution.

Ellis, C., Holston, S., Drake, G., Putman, H., Swisher, A., & Peske, H. (2023). *Teacher Prep Review: Strengthening Elementary Reading Instruction*. National Council on Teacher Quality. <https://www.nctq.org/publications/Teacher-Prep-Review:-Strengthening-Elementary-Reading-Instruction>

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard educational review*, 66(1), 1-27.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. University of South Florida, Louis de la Parte Florida Mental Health Institute.

- Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). The Impact of Providing Performance Feedback to Teachers and Principals. NCEE 2018-4000. *National Center for Education Evaluation and Regional Assistance*.
- Goertz, M. E. (2012). Policy analysis: Studying policy implementation. In *Handbook of complementary methods in education research* (pp. 701-710). Routledge.
- Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. Brookings Institution, Hamilton Project.
- Hanushek, E.A. (2025) The pandemic in perspective: US learning losses in the twenty-first century. <https://hanushek.stanford.edu/index.php/publications/pandemic-perspective-us-learning-losses-twenty-first-century>
- Hanushek, E., Luo, J., Morgan, A., Nguyen, M., Ost, B., Rivkin, S., & Shakeel, A. (2026). The effects of comprehensive pay reform on achievement in urban schools. *Journal of Urban Economics*, 153, 103857.
- Hess, F. M. (1998). *Spinning wheels: The politics of urban school reform*. Washington, DC: Brookings Institution.
- Hess, F. M. (2017). *Letters to a young education reformer*. Harvard Education Press.
- Hess, F.M. (2026) “Five Lessons for School Reformers: 2026 Edition” Education Next, Downloaded 5/2/26 from <https://www.educationnext.org/five-lessons-for-school-reformers-2026-edition/>
- Hunter, S. B., & Springer, M. G. (2022). Critical feedback characteristics, teacher human capital, and early-career teacher performance: A mixed-methods analysis. *Educational Evaluation and Policy Analysis*, 44(3), 380-403.
- 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.
- James, J., & Wyckoff, J. H. (2020). Teacher evaluation and teacher turnover in equilibrium: Evidence from DC public schools. *AERA Open*, 6(2), 2332858420932235.
- Joyce, B. R., & Showers, B. (2002). *Student achievement through staff development* (Vol. 3). Alexandria, VA: Association for Supervision and Curriculum Development.
- Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.
- Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, 49(4), 241-253.
- 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.
- Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics*, 188, 104212.
- Kraft, M. A., & Christian, A. (2022). Can teacher evaluation systems produce high-quality feedback? An administrator training field experiment. *American Educational Research Journal*, 59(3), 500-537.
- 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.
- Kraft, M. A., & Lyon, M. A. (2024). The rise and fall of the teaching profession: Prestige, interest, preparation, and satisfaction over the last half century. *American Educational Research Journal*, 61(6), 1192-1236.
- 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.
- Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of educational psychology*, 78(2), 75-95.
- Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell sage foundation.
- Mahnken, K. (2026) “‘We’re Adrift’: Arne Duncan on Democrats’ Education Agenda” The74, April 30, 2026 found at <https://www.the74million.org/article/were-adrift-arne-duncan-on-democrats-education-agenda/>
- Marsh, J. A., McCombs, J. S., & Martorell, F. (2012). Reading coach quality: Findings from Florida middle schools. *Literacy Research and Instruction*, 51(1), 1-26.
- McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational evaluation and policy analysis*, 9(2), 171-178.
- Mississippi SB2347 (2013). *Senate Bill 2347: Literacy-Based Promotion Act*. State of Mississippi. <https://legiscan.com/MS/text/SB2347/id/818137>

Moats, L. C. (2020). *Teaching reading is rocket science: What expert teachers of reading should know and be able to do*. American Federation of Teachers.

Moe, T. M., & Wiborg, S. (Eds.). (2017). *The comparative politics of education*. Cambridge University Press.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.

National Assessment of Educational Progress, Achievement Gap Dashboard, downloaded from https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx June 2026a.

National Assessment of Educational Progress, calculated using the NAEP Data Explorer <https://www.nationsreportcard.gov/ndecore/xplore/NDE> September 1, 2026b.

National Council on Teacher Quality. (2015). *State of the states 2015: Evaluating Teaching, Leading and Learning*. https://www.nctq.org/wp-content/uploads/2025/03/StateoftheStates2015_lores.pdf

National Council on Teacher Quality. (2022). *State of the states 2022: Teacher and principal evaluation policies*. <https://www.nctq.org/research-insights/state-of-the-states-2022-teacher-and-principal-evaluation-policies/>.

National Council on Teacher Quality. (2024). *State of the states 2024: Five policy actions to strengthen implementation of the science of reading*. <https://www.nctq.org/research-insights/state-of-the-states-2024-five-policy-actions-to-strengthen-implementation-of-the-science-of-reading/>

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development.

Neumerski, C. M., Grissom, J. A., Goldring, E., Rubin, M., Cannata, M., Schuermann, P., & Drake, T. A. (2018). Restructuring instructional leadership: How multiple-measure teacher evaluation systems are redefining the role of the school principal. *The Elementary School Journal*, 119(2), 270-297.

O'Donnell, P. (2025, October 22). 'Disappointing': Ohio's science of reading switch not yet bringing results. *The 74*.

Pressman, J. L., & Wildavsky, A. (1973). *Implementation: How great expectations in Washington are dashed in Oakland; Or, why it's amazing that federal programs work at all, this being a saga of the Economic Development Administration as told by two sympathetic observers who seek to build morals on a foundation* (Vol. 708). Univ of California Press.

- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- RMC Research Corporation. (2019). *Mississippi's Literacy-Based Promotion Act: An inside look*. U.S. Department of Education. <https://eric.ed.gov/?id=ED613758>
- Sabatier, P., & Mazmanian, D. (1979). The conditions of effective implementation: A guide to accomplishing policy objectives. *Policy analysis*, 481-504.
- 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.
- Sartain, L., & Steinberg, M. P. (2025). Can personnel policy improve teacher quality? The role of evaluation and the impact of exiting low-performing teachers. *Education Finance and Policy*, 20(1), 137-163.
- Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago*. Consortium on Chicago School Research.
- Schwartz, S. (2022). What is LETRS? Why one training is dominating 'science of reading' efforts. *Education Week*. <https://www.edweek.org/teaching-learning/letrs-program-teacher-training>
- Schwartz, S. (2023, December). The science of reading in 2023: 4 important developments. *Education Week*. <https://www.edweek.org/teaching-learning/the-science-of-reading-in-2023-4-important-developments/2023/12>
- Simon, N. S., & Johnson, S. M. (2015). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record*, 117(3), 1–36.
- Smith, C. (2025) “Are Red States Leading the Way on Education Reform?” Discussion with Mike Petrilli, April 29, 2025 found at <https://www.governing.com/policy/are-red-states-leading-the-way-on-education-reform>
- Song, M., Wayne, A. J., Garet, M. S., Brown, S., & Rickles, J. (2021). Impact of providing teachers and principals with performance feedback on their practice and student achievement: Evidence from a large-scale randomized experiment. *Journal of Research on Educational Effectiveness*, 14(2), 353-378.
- Spencer, G. (2024). Comprehensive early literacy policy and the "Mississippi Miracle." *Economics of Education Review*. <https://doi.org/10.1016/j.econedurev.2024.102556>
- Stecher, B., Holtzman, D., Garet, M., Hamilton, C., Engberg, J., Steiner, E., Robyn, A. (2018). *Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching through 2015–2016*. Santa Monica, CA: RAND.

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.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628-3651.

Toch, T. (2018). A Policymaker's Playbook. Future Ed <https://www.future-ed.org/wp-content/uploads/2018/06/APOLICYMAKERSPLAYBOOK.pdf>

Tyack, D. B., & Cuban, L. (1995). *Tinkering Toward Utopia: A Century of Public School Reform*. Harvard University Press.

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.

Will, M. (2021) "[Efforts to Toughen Teacher Evaluation Show No Positive Impact on Students](#)" *Education Week*, November 29, 2021.

Wright, C. (2022). How Mississippi reformed reading instruction. *ExcellInEd*. <https://excelined.org/2022/12/02/how-mississippi-reformed-reading-instruction/>

Wyckoff, J. (2026). Puzzling over declining academic achievement. *Journal of Policy Analysis and Management*, 45(2), e70052.