frontiers | Frontiers in Education

TYPE Hypothesis and Theory
PUBLISHED 23 August 2023
DOI 10.3389/feduc.2023.1221569

OPEN ACCESS

EDITED BY
Stefinee Pinnegar,
Brigham Young University, United States

REVIEWED BY
Mary Frances Rice,
University of New Mexico, United States
Cheryl J. Craig,
Texas A&M University, United States

*CORRESPONDENCE
Drew H. Gitomer
✉ drew.gitomer@gse.rutgers.edu

RECEIVED 12 May 2023
ACCEPTED 02 August 2023
PUBLISHED 23 August 2023

CITATION
Gitomer DH and Marshall BL (2023) The sizzle and fizzle of teacher evaluation in the United States and the selective use of research evidence.
Front. Educ. 8:1221569.
doi: 10.3389/feduc.2023.1221569

The sizzle and fizzle of teacher evaluation in the United States and the selective use of research evidence

Drew H. Gitomer* and Brittany L. Marshall

Graduate School of Education, Rutgers University, New Brunswick, NJ, United States

In 2009, the United States funded the largest federal educational reform effort in the nation's history. Referred to as *Race to the Top* (RTTT), a cornerstone of this effort was the high-stakes evaluation of all teachers, with a significant emphasis on the use of highly researched statistical methods that ascribed changes in student test scores to a teacher's quality. The widespread endorsement of these policies across a broad range of the political spectrum was based on a theory of action that faced technical, organizational, and political challenges. Enthusiasm for these evaluation efforts was substantially muted in a mere 5 years. Among a number of factors, we argue that the framing of the problem together with privileging particular lines of research and voices, as well as the lack of consideration of other frames and attention to other research and voices, resulted in an evidence base that was wholly insufficient to justify the large-scale policy changes that were enacted.

KEYWORDS

teacher evaluation, assessment, evidence use, teacher quality, policy formation

1. Introduction

Teacher evaluation in the United States has been an important K-12 education policy issue for the past 25 years. In this article, we will describe the evolution and design of national in-service teacher evaluation policies as part of a major educational reform initiative, how policies were implemented, and why many of them failed. We argue that these policies were doomed from the start for many reasons, including weak theories of action as a result of inadequate attention to research and critical stakeholders, weak measures to explain causal attribution, organizational issues, and lack of consideration to how teacher evaluation systems affect schools in marginalized communities.

As part of the federal response to an economic crisis, the U. S. Congress enacted the American Recovery and Reinvestment Act of 2009, a massive and unprecedented stimulus package of over $800B (Congressional Budget Office, 2012). Included in this package was an equally unprecedented $4.35B for educational reform, known as *Race to the Top* (RTTT). The most important consideration in states' applications was their plan for implementing the evaluation of educators, including both teachers and principals.

These evaluation systems represented a change in how teacher evaluations in the United States were to be conducted, as they focused, in large part, on how individual teachers contributed to student learning as measured by standardized test scores and other types of assessment measures. While evaluation systems also included measures like classroom observations, this focus on using student learning measures to evaluate teachers was an effort relatively unique to the United States (Williams and Engel, 2012). The push for these systems was strongly bipartisan, motivated by concerns about student learning as well as very pointed critiques of teachers and, particularly, teacher unions (see Katz and Rose, 2013; Maranto et al., 2016). This bipartisan agreement also led to the charter school boom of the 2000s.

The enthusiasm for teacher evaluation was fully shared by policy leaders across the country, as they argued that evaluation would be a powerful tool to aid teachers in their ability to support their students. The two largest funders of these efforts were the U. S. Department of Education and the Bill and Melinda Gates Foundation. Arne Duncan, U. S. Secretary of Education at the time, said, "Teachers support evaluations based on multiple measures: student growth, classroom observation and feedback from peers and parents" (Duncan, 2009). Bill Gates, speaking for his Foundation, stated, "Students deserve great teachers. And teachers deserve the support they need to become great" (Gates and Gates, 2018).

Though RTTT marked a major policy shift in American education, its genesis was long in the making. For some 40 years, policymakers had consistently focused on the comparatively poor academic performance of U. S. students as measured by national and international assessments. The most recent policy iteration was based on a broad body of research evidence that was used to justify the need to improve teaching quality, generally, and the need to reform teacher evaluation practices, specifically. Indeed, it was virtually certain that research papers and policy statements alike would begin their arguments by pointing out that teachers were the most important school-based factor in determining students' academic outcomes. This research was used to support the implementation of teacher evaluation policies in 40+ states by 2013. The fervor for these policies represented the confluence of the promise that teachers were the single most important factor in determining student outcomes (the qualifier of *school-based* was often lost in policy discussions) and the promise of measurement technologies that could identify teacher quality with appropriate precision. The sizzle was palpable.

The enthusiasm for teacher evaluation and its related policies was short-lived. By 2015, the federal government had abandoned teacher evaluation as a requirement for federal funding. Foundations that had been major supporters of these initiatives shifted their attention elsewhere. While teacher evaluation did not disappear completely, many states abandoned the use of student growth scores as a required component of teacher evaluations.

Research over the last number of years has revealed the many ways in which the policies did not live up to their promise. For the most part, the goal of improving student achievement was not realized. Constituent measures were shown to be unreliable and biased. Inadequate attention was given to implementation and organizational issues and their impact on students, teachers, and schools in marginalized communities. Educators, in general, soon became vocal opponents of the policies.

In this paper, we argue that a critical reason for the failure of RTTT to realize its promise was that the research base that was used to support the theory of action for teacher evaluation was, from its inception, inadequate to support ambitious policy goals. We consider the arc of history that led to teacher evaluation as a core educational reform policy, the research that motivated the policy, the limits of that research, and the resulting outcomes of the policy. We use this to highlight that using research evidence to create policy is limited to the extent that the research is not sufficient to address the complexity of the problem it is trying to address.

2. Setting the stage for RTTT – the role of federal policy in educational reform

Historically, educational policy in the United States was a responsibility of individual states and local districts. The establishment of a cabinet-level Department of Education did not occur until 1980 and was politically contested as usurping states' responsibilities (Stallings, 2002). During the 1980s, several landmark reports that laid the groundwork for RTTT (National Commission on Excellence in Education, 1983; Carnegie Forum on Education and the Economy, 1986) were issued. These reports were authored by commissions that consisted of leaders in education, government, and business and came to a set of conclusions, largely based on test score performance and international comparisons, that were at the core of reform efforts for the next 40 years:

- Public schools are bastions of mediocrity, and students are underachieving.
- This mediocrity has direct implications for the nation's economic well-being.
- The federal government has a role in improving our nation's education.

These reports led to two generations of educational reform efforts characterized by various initiatives to: specify what both students and teachers needed to know and be able to do in the form of standards; increase testing of student achievement; increase testing of teachers for licensure and certification; and implement a range of accountability efforts to hold states and schools accountable for educational performance. These policies were embodied in landmark legislation such as the *Improving America's Schools Act* (IASA) of 1994 and the *No Child Left Behind Act of 2001* (NCLB; officially, the *Elementary and Secondary Education Act* [ESEA]).

NCLB was particularly interesting in that it called for schools to make adequate yearly progress (AYP) on achievement scores in such a way that all students would be 100% proficient 13 years later (2013–14). It became clear that states were trying to navigate the policy by setting lower standards for proficiency, setting minimal growth targets early in the AYP trajectory, and seeking exceptions. All of this had significant implications for how schools were judged and for which schools were labeled as "failing" (Polikoff et al., 2014; Davidson et al., 2015). By most metrics, NCLB did not lead to meaningful gains for students, and international comparisons remained troubling for policymakers (e.g., Dee and Jacob, 2011; Lee and Reeves, 2012). The ineffectiveness of school-based accountability led policymakers to shift their focus to teachers as the target of educational reform. Several lines of research laid the foundation for what was to become the most far-reaching policy initiative focused on teacher evaluation, both globally and historically.

3. The research basis and process for teacher evaluation

Gitomer and Marshall (in press) reviewed key research efforts that provided the justification for the teacher evaluation policies embedded in the RTTT program. The first line of research focused on *teacher effects*, a statistical determination in which the outcome was changes in student year-to-year achievement on annual standardized achievement scores, and the target input(s) were the teachers who taught each student. Using a range of regression-based approaches (Nye et al., 2004), researchers identified teachers as the single most important school-based factor associated with student outcomes. These studies attempted to control for student and school characteristics in order to obtain unconfounded estimates of teacher effects, although such efforts are imperfect in controlling for all non-teacher effects (Lockwood and Castellano, 2017).

For many years, researchers had tried to identify teacher characteristics that were associated with teacher effects on student learning. Looking at metrics commonly used for teacher compensation, such as years of service, degree attainment, and academic credits, researchers consistently found limited associations with student achievement (e.g., Kane et al., 2008; Harris and Sass, 2011). Though student experience was initially related to student outcomes, that relationship disappeared after the first 5 years of practice (Clotfelter et al., 2010). Similarly, professional certification status and domain-specific coursework had minimal relationships with student achievement growth (Wayne and Youngs, 2003; Goe, 2007).

If policymakers could not rely on teacher inputs as a measure of teacher quality, research also makes clear that traditional teacher evaluation practices did not lead to very credible or informative reports about teacher practice. Though teacher evaluation was long embedded in educational systems, Weisberg et al. (2009) reported that teacher evaluation systems did not identify or remove weak teachers and provided inflated and non-differentiated reports of teacher quality.

The inability to find consistent relationships of teacher inputs to student outcomes and the limited utility of evaluations led policymakers and researchers to turn their attention to other directions. Specifically, they were intrigued with the statistical approaches being promoted by prominent statistician, William Sanders, who had developed an approach known as *Value-Added Modeling* (VAM; Sanders and Horn, 1994). VAM used multiple years of prior test scores for each student to estimate the contribution of a specific teacher to the annual growth of all the students in that teacher's classroom. Aggregate VAM scores are standardized so that all teachers in a particular cohort (e.g., a school district or state) are compared in terms of a standardized score relative to the mean score (0) of the cohort. The promise and allure of Sanders' VAM was that it was designed to address potential issues of fairness by using prior student achievement as a control to encompass all potential factors that might influence student achievement. Other VAM models that largely followed Sanders' approach also emerged, but these models varied on how they treated covariates and other model specifics (see Braun, 2005; Harris, 2011, for basic introductions to VAM).

Policymakers also became interested in whether compensation systems could be used to improve the quality of teaching. Pay-for-performance systems were developed in a number of states and districts. The Tennessee system, using Sanders' VAM models, provided additional compensation to teachers with high VAM scores (Sanders and Horn, 1994). Denver public schools developed a more comprehensive compensation model that included annual evaluations and working in high-needs schools.

Finally, research that examined the relationship of teacher practice to student outcomes had also been conducted. Studies examined the effects of particular pedagogical strategies (e.g., Murnane and Phillips, 1981) as well as the relationship of teachers' scores on classroom observation protocols to the achievement growth of their students (Milanowski, 2004; Kane et al., 2010).

3.1. The interplay of research and teacher evaluation policy

The convergence of the aforementioned research, and the evidence it produced, was used to shape the teacher evaluation policy that was central to RTTT. To understand why and how these particular lines of research were used, we borrow from two theoretical perspectives—one that considers policy formation in general terms (McDonnell and Weatherford, 2020) and one that considers the sociopolitical context of teaching from a critical race perspective (Nasir et al., 2016). Together, these perspectives help us better understand why certain research evidence was so salient in policy formation, why other research was not attended to, and, ultimately, why the research that guided policy was insufficient to adequately satisfy the ambitious policy goals of RTTT.

McDonnell and Weatherford (2020) described the strategic use of evidence by policymakers to achieve political objectives given a set of goals and beliefs about how best to achieve those goals. In that context, they argued that it was important to understand *what* evidence is given attention as well as *who* is engaged in the production and use of evidence. The *who* includes:

- *researchers*: those who produce original research;
- *policy entrepreneurs*: those who have a strong policy position and marshal research and other evidence to support that position;
- *translators and disseminators*: those people and organizations that have a goal of identifying and communicating high-quality research to policymakers;
- *advocates*: those who represent particular policy positions and put priority on the ends they are trying to achieve; and
- *hybrids*: those who have an advocacy position and also try to operate as translators and disseminators.

Nasir et al. (2016) argued that, in order to have a comprehensive understanding of teaching, one must take into account the multi-level context in which teaching is situated. Yet, research on teaching and the resulting policies have often ignored such complexity. They further contended that the research and policies over the recent past have been guided by particular kinds of framing of the problems to be addressed.

In Nasir et al.'s (2016) framework, there are three levels of context that need to be addressed in any full analysis of teaching. First, there are broad economic and policy macro-trends that include: significant and growing economic inequality; the paradox of increasing racial and ethnic diversity in American schools combined with increasing social class segregation in society and schools; and marketized neoliberalism (bringing free-market principles to social issues). The second level includes ways that schools and districts adapt to these broader economic and policy macro-trends. The third level focuses on how these other levels influence the nature of instruction and learning environments that students, and particularly marginalized students, encounter. The focus on accountability testing, for example, often results in low-skill test preparation teaching for marginalized students.

Nasir et al. (2016) also adopted Hand et al.'s (2012) conception of operating frames "as a way to examine and reorganize race and power within learning environments. Power plays out in everyday social interaction as individuals become attuned to, coordinate and mobilize around *frames* they engage in during moments of interaction" (Hand et al., 2012, p. 251). The first frame they identify is one of *colorblindness*, a view that "minimizes the existence or consequentiality of race and views policy solutions as best when universal in nature" (Nasir et al., 2016, p. 354). A second frame is *meritocracy*, one that ascribes accomplishment as solely due to the actions of individuals and "allows policy makers to act without acknowledging the systemic nature of racial disparities and diverts attention to the choices of individual actors" (Nasir et al., 2016, p. 353). The final frame, also located in their multi-level hierarchy is *neoliberalism*, which has led to the marketization of schooling and "emanates from three decades of policy that positioned the private sector to be superior to the public sector in providing more efficient social services" (Nasir et al., 2016, p. 355).

Indeed, accountability efforts, particularly those involving the federal government, have engendered significant debate about their role in supporting education as a public good. While the dominant policy argument has long been that accountability efforts exist to improve education and support the enterprise as a public good, others have been far more critical. They have taken more critical stances, such as those embodied by Nasir et al. (2016), to argue for a much more nuanced understanding of how accountability efforts have also served to diminish education as a public good (see Anagnostopoulos et al., 2013).

The research that guided teacher evaluation policy, and how it was conducted and by whom, can help us make sense of how the policy took shape, why the policy was problematic in its uptake by states and districts, and, ultimately, why the initiatives were largely abandoned or dramatically reduced in scope. Using the frameworks provided by both Nasir et al. (2016) and McDonnell and Weatherford (2020), we highlight key aspects of research development and use.

The guiding principles of teacher evaluation grew out of the dominant framing noted by Nasir et al. (2016) that has directed policy perspectives on education for the last several decades. Embedded within this work was the meritocratic perspective that teachers are the primary agent associated with student growth and that their relative success is deserved and an outcome of choices and actions by individuals (teachers and administrators). The VAM models were proffered as ways of overcoming the influence of any contextual factors and, thus, were designed to be pure measures of a teacher's contribution. By controlling for factors such as race and socio-economic status, these models also subscribed to the framing of colorblindness—that evaluation scores are fair estimates of a teacher's quality regardless of a teacher's (or their students') background. Neo-liberal framing was evident throughout the system in hiring and retention policies as well as in the various pay-for-performance schemes that were linked to teacher evaluation.

These framings had important consequences for the kinds of research that was done, who did the research, and how the work was supported. Research on teacher effects was almost always guided by researchers (e.g., Kane et al., 2008; Rockoff et al., 2011) who adopted the three frames identified by Nasir et al. (2016). These researchers, often educational economists, were focused on identifying the "effects" of teaching by adopting methods that were designed to control for contextual effects rather than trying to understand their influence.

As McDonnell and Weatherford (2020) argued, there are multiple actors involved in how research shapes policy and vice versa. The emergence of teacher evaluation policy, including its central features, represented the confluence of a strategic use of evidence to achieve a particular set of objectives. By the time RTTT was developed, the lines between researchers, policy entrepreneurs, and translators and disseminators/advocates had become highly blurred (see DeBray and Houck, 2011). Reckhow and colleagues described how think tanks, foundations, government policymakers, and researchers set a research agenda and policy coordinated to elevate teacher evaluation (see Reckhow and Tompkins-Stange, 2018; Reckhow et al., 2021). All of these players, likewise, were guided by the three frames identified by Nasir et al. (2016). The Bill and Melinda Gates Foundation and the U. S. Department of Education were the two primary drivers of this work. The Gates Foundation funded research, supported intensive district-level reform efforts, provided advocacy, and worked with the U. S. Department of Education. The U. S. Department of Education, through the RTTT program as well as through funding from the Institute of Education Sciences (IES), not only led the policy initiative but was instrumental in leading advocacy efforts (e.g., Duncan, 2009) and funding programs of research that were supportive of the endorsed teacher evaluation efforts. RTTT was driven by a set of core beliefs about public schools, teachers and teacher unions, neo-liberal approaches to the marketization of education, and concerns about academic performance by students in marginalized communities, along with the emergence of research that offered potential solutions.

3.2. The theory of action and implementation plan guiding teacher evaluation policy

In 2009, the U. S. Department of Education announced the RTTT competition and invited states to compete for funds to support educational reform (U. S. Department of Education, 2009). The initiative was based on a theory of action that improved teacher quality would lead to improved student learning. Theories of action specify a cause-and-effect relationship between a policy intervention and a set of desired outcomes (e.g., McDonald, 2009). As articulated by Gitomer and Bell (2013), teacher evaluation was championed as improving teacher quality through four complementary drivers. First, teacher evaluation served an accountability purpose in which teachers (and principals) could be held accountable for student performance. Second, evaluation could support what came to be called *the strategic use of human capital*. In a market-based approach, evaluation results could be used to guide a system of incentives and disincentives to manage the supply of teachers by increasing the supply of effective teachers and removing less effective teachers (e.g., Gordon et al., 2006; Heneman et al., 2006). A third purpose was to improve individual teacher and institutional capacity by including direct measures of classroom instructional quality that could be used as a tool for providing feedback to teachers (e.g., Borko, 2004; Johnson et al., 2004;

Kardos and Johnson, 2007). Finally, teacher evaluation could be used to support evidence-based instructional policy by determining the efficacy of particular policies and interventions (e.g., Rowan et al., 2004, 2009; Rowan and Correnti, 2009).

There were two components related to teacher evaluation that all proposals needed to satisfy:

1. building data systems that measure student growth and success and inform teachers and principals about how they can improve instruction; and
2. recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most.

Specific criteria that had to be met included:

1. measuring student growth for every student;
2. creating evaluation systems that:
    a. differentiated effectiveness using multiple rating categories and treated student growth as a significant factor; and
    b. were designed and developed with educator involvement;
3. conducting annual teacher evaluations that provided feedback, including information on student growth from their students and classes; and
4. using teacher evaluations to inform decisions regarding:
    a. coaching and development;
    b. compensation, promotion, retention, and advancement;
    c. tenure; and
    d. removal.

By 2011, 19 states received RTTT funding. However, far more states and localities (42 in total) adopted these policies in order to obtain waivers from the NCLB mandates that were still in effect (Gitomer and Marshall, in press). While states, and often, districts within states, varied in how they developed the specifics of their systems, Gitomer and Marshall (in press) described the key features of all systems, their technical limitations, and how they varied across and within states.

One requirement for determining any teacher growth measure is the necessity of defining which students' growth scores should be used in determining a teacher's value-added. The realities of schooling made this a non-trivial problem, and the solutions varied greatly. For example, how should students with high levels of absenteeism be treated? If students move between schools multiple times across the year, how should they be treated in the VAM models? What about situations in which multiple teachers are responsible for the students in a particular classroom (e.g., special education)? Of course, less stable student populations are typically associated with schools with high proportions of minoritized and economically insecure students (see Everson, 2017).

A second issue concerned the inclusion of particular test score results for each teacher. In some models, teachers had evaluation scores that included test results for which they ostensibly had no teaching responsibility.

Third, states had to decide which measures contributed to an evaluation system. Almost all states included a student growth measure and a classroom observation score. But other measures, including student learning objectives (SLOs), principal rating, overall achievement levels of grades and/or schools, and student surveys, were also used in some systems.

Fourth, states needed to decide how the scores from different measures were aggregated for a final evaluation score. Aggregation methods could be compensatory, in which each component is weighted in a linear combination of scores, and a total score is used to determine the appropriate evaluation category for an individual. Another option was to use a conjunctive model, in which a minimum score is required for each of the constituent measures. How scores were weighted in the overall model depended on how highly particular measures were valued relative to others, as well as how much variation was associated with particular measures. Measures that have scores that vary more across individuals will have a greater influence on overall evaluative judgments than measures on which most individuals receive the same score, even if the latter measures are assigned nominal weights.

Fifth, states differed in both the consequences and supports given for particular evaluation scores. Typically, an ineffective rating was associated with some type of probationary status for the first year, which then would require some type of additional professional development and support.

Sixth, measures, particularly those associated with classroom observations, required some type of training of principals and other administrators. While researchers have given great attention to observer training, calibration, and overall quality control of scores (National Research Council, 2008; Bell et al., 2014), in practice, these procedures were often compromised as school districts did not have the time or resources to undertake the kinds of procedures that had been used to validate measures from research studies.

4. Measures and challenges

While states adopted a large number of measures for their teacher evaluation systems, the three measures that were most ubiquitous and most prominent across systems are discussed here. Each of these measures was used to support inferences about a teacher's quality. However, each of these measures had significant technical issues that challenged the validity of using them for such a consequential process.

4.1. Student growth measures

RTTT advocated the use of growth measures to overcome inherent problems associated with making any relative judgments of teachers based on their students' achievement status by separating the effects of teachers from other factors such as demographics, resources, and student prior achievement. The basic logic was that any attributions to teacher effectiveness must be made with respect to the relative year-to-year growth in student achievement. A broad range of growth models were used, including different versions of VAM, as well as a related method, *Student Growth Percentiles* (SGP; Betebenner, 2009). All growth models required multiple years of student test data linked to individual teachers.

Research on growth models made clear that precise, causal estimates of a teacher's contribution to student learning were very fragile. Rowan and Raudenbush (2016) provided a detailed overview of the challenges in using growth models to make high-stakes decisions about teachers. Reardon and Raudenbush (2009) explained how the fundamental statistical assumptions that are foundational to these models can never be satisfied. Studies have revealed how relative estimates of teacher quality can shift dramatically because of using different estimation techniques (Goldhaber and Theobold, 2013) or different achievement measures (Lockwood et al., 2007; Grossman et al., 2014). Multiple studies have shown that VAM estimates can be statistically biased toward classrooms that have students with higher levels of prior achievement, a situation that growth models were supposed to overcome (Rothstein, 2009, 2017; Raudenbush, 2013). Researchers also found that VAM scores in one testing domain (e.g., reading) could be influenced by the quality of teaching in another domain (e.g., mathematics) (Koedel, 2009).

Scholars in measurement and statistics issued several statements to caution about the use of these models for high-stakes decisions. Baker et al. (2010) produced a consensus statement of several leading educational scholars that cautioned the use of these measures and also highlighted potential unintended consequences, including discouraging teachers from wanting to work in schools with students who had the most academic needs. The American Statistical Association (2014) also released a statement, recognizing the value of VAM to help understand the relationship of different factors to student outcomes when results are aggregated across teachers but also cautioning against using these models to make strong causal statements about individual teachers. Other cautionary and critical statements were made by the National Association of Secondary School Principals (NASSP) in 2015 (see https://www.nassp.org/top-issues-in-education/position-statements/ for the most recent version, updated in 2019; National Association of Secondary School Principals, 2019) and the American Educational Research Association (American Educational Research Association Council, 2015). Several lawsuits challenging the consequences of teacher evaluation efforts were also instituted (see Paige and Amrein-Beardsley, 2020).

4.2. Student learning objectives

As much as growth measures based on student achievement scores were central to this evaluation movement, the fact is that a very large proportion of teachers did not have testing data that would be appropriate for estimating student growth. Testing was only federally mandated, for example, in grades 3–8 mathematics and reading, meaning that teachers in earlier and later grades, as well as those who taught other subjects, would not have students who had multiple years of testing data to analyze. Certain states did, however, impose more encompassing testing requirements.

Thus, in order to address the legislative mandate that teacher evaluation needed to include a "student growth measure," most states adopted SLOs for teachers in non-tested subjects, but many states also used them for all teachers as a complementary measure of student growth. SLOs are a locally determined evaluation of teacher effectiveness by which measurable targets for student achievement are set following an analysis of baseline data. Essentially, SLOs include some prior to instruction measure of student understanding (pre-test) and a post-instruction measure or assessment. The extent to which those targets are met is then used to evaluate the teacher. Within this common definition, specific features of the SLO process have varied substantially (see Crouse et al., 2016).

An SLO consists of three components. The first is the population of students it covers—is the teacher evaluated on the basis of performance by all students in all classrooms and subjects taught by the teacher or just a subset (e.g., only mathematics or reading for an elementary teacher, only one section of a course for a secondary teacher)? The second component is the target of the SLO—do all teachers with the same teaching assignment in a school or district have the same target, or is greater variability part of the design? In addition, the meaningfulness of SLO-based scores is largely a function of the quality control procedures used in the implementation of the SLO process (Crouse et al., 2016).

The third component is the assessment to measure student learning. SLO assessments include locally generated measures as well as standardized, externally developed assessments. Often, classroom assessments such as portfolios or some type of performance assessment are used.

While little research about the quality of SLO measures was done, Crouse et al. (2016) described the inherent problems of using SLOs as a measure of student growth in teacher evaluations. They argued that the validity of such measures for evaluating and comparing teachers could not be justified because of the idiosyncratic nature of their design and implementation. They also pointed out that making causal attributions to a teacher was problematic in light of external factors such as district curriculum, outside tutoring, and student background characteristics that can influence student outcomes. Finally, the use of SLOs was highly variable across states and districts. In some cases, all teachers needed to have an SLO as part of their evaluation. In other instances, only those teachers who did not have standardized test-based growth scores were required to have SLOs. Because the distribution of scores for test-based growth models and SLOs tends to be different, the net effect is that overall evaluation scores could be lower for teachers who have growth estimates based entirely or, in part, on standardized tests as compared with those only having SLOs as their growth measure component.

4.3. Classroom observations

Structured observation protocols, originally designed as tools for professional development (e.g., Danielson, 2007; Pianta et al., 2008), soon became the object of study in research and a key component of teacher evaluation systems under RTTT. These protocols were created around particular views of teaching that drew on research and were organized along sets of cognitive, social, emotional, and classroom management dimensions of instructional quality.

The protocols adopted for teacher evaluation systems were designed to be used across grades and subject areas. Each protocol provided guidelines for how to observe a period of classroom instruction, how to code what was observed, and how to score instruction for the set of criteria that were described in the protocol's scoring rubric (see Bell et al., 2012).

Scores typically involved some form of aggregation of dimensional scores into a total lesson score as well as aggregation of scores across multiple lessons. The management of observations and recording and maintaining of data within school systems was often done with the assistance of commercial observation tools that were designed specifically to support teacher evaluation processes.

Research has shown the limitations of observation protocols in assuring precise and valid estimates of teacher quality. For one, many factors, other than the quality of the instruction itself, can influence the scores for a particular observation, most especially the observers themselves. Research efforts have tried to moderate these sources of error through careful training and monitoring of observers, using multiple and different observers across multiple observations, and ensuring that there were no conflicts of interest between the observer and the observed that might bias scoring (see Bell et al., 2012, 2014).

As observation measures were used in evaluation systems, it became clear that findings from research studies did not generalize to practice settings. Observation scores in practice are uniformly higher than scores from research studies, for example. Scores in research studies that typically fell in the 2–3 range on 4-point scales fell between 3 and 4 when used in practice (see Sartain et al., 2010; Briggs et al., 2014).

Of course, conditions within practice settings were quite different as observers were not disinterested parties. They knew the teachers and worked with them as part of a professional staff (Harris et al., 2014; Kraft and Gilmour, 2017; Donaldson and Woulfin, 2018), and they gave higher scores for teachers they worked with than for teachers with whom they were not familiar (Ho and Kane, 2013). School administrators must conduct observations by statute, regardless of how well qualified they are to score. Typically, fewer observations were conducted in school evaluations than in research studies, and it was very rare for any system to include multiple observers.

It also became clear, in both research studies and studies of observation in practice, that personal characteristics of the teacher, and especially the students, affected observation scores. There have been consistent findings that teachers of students with weaker academic profiles are assigned lower observation scores (Gitomer et al., 2014; Campbell and Ronfeldt, 2018). In addition, Steinberg and Sartain (2021) found that observation scores of Black teachers were substantially lower than scores for White teachers and that those differences could be accounted for by the achievement levels of their students. Campbell and Ronfeldt (2018) found that male teachers tended to have lower than expected scores than female teachers and that scores were also lower than expected in classrooms with higher concentrations of Black, Latin*[1], male, and low-performing students, a result also found by Garrett and Steinberg (2015).

5. The fizzle of teacher evaluation policy – promises not kept

Despite the tremendous amount of resources, attention, and effort given to teacher evaluation, teacher evaluation had a very short shelf-life as a major educational reform policy. By 2015, the core idea of linking teacher evaluation to student outcomes was abandoned when the ESEA was reauthorized in the form of the *Every Student Succeeds Act* (ESSA, 2015):

> Nothing in this Act shall be construed to authorize or permit the Secretary … as a condition of approval of the State plan, or revisions or amendments to, the State plan, or approval of a waiver request submitted under section 8401, to … prescribe … 'any aspect or parameter of a teacher, principal, or other school leader evaluation system within a State or local educational agency; … indicators or specific measures of teacher, principal, or other school leader effectiveness or quality; (pp. 42–43)

ESSA reflected a change in the entire policy landscape, as teacher evaluation was no longer perceived as the key to improving America's schools. Actors like the Gates Foundation, which had played a major role in advocating for and influencing teacher evaluation policy, also relatively quickly moved in other directions. By 2018, the Foundation publicly acknowledged the modest impact their efforts had made (Gates and Gates, 2018). Other foundations that had been players in the teacher evaluation movement also switched priorities.

There certainly was a great deal of political pushback to the increasing federal role in public education, most especially with the Common Core curricular standards and associated assessments (Loveless, 2021). By 2015, the two cornerstones of education reform—ambitious standards and teacher evaluation—had gone from broad endorsement to policies that were increasingly shunned. Indeed, as the federal mandate disappeared, large numbers of states abandoned, or gave great flexibility to, the use of growth models built on student test scores (Close et al., 2020). Many states, however, continued to mandate some type of classroom observation.

Gitomer and Marshall (in press) reviewed evidence addressing the extent to which teacher evaluation policy efforts met the ambitious goals that were promised upon the launch of RTTT. While the results summarized in this section are representative of what happened across the country, there was variation in how systems were implemented and the kinds of results that were observed. The most notable exception to general findings was found in Washington D. C., which implemented a very well-resourced, comprehensive reform effort that resulted in significant changes to the district's schools (National Research Council, 2015; James and Wyckoff, 2020). The intensive, multi-faceted systemic approach of Washington D. C. stands in contrast to how teacher evaluation was conceptualized and operationalized in most settings.

5.1. The promise of identifying weak teachers

One goal of teacher evaluation was to differentiate teachers based on their effectiveness. However, Kraft and Gilmour (2017) and Stecher et al. (2018) found that evaluation score distributions were largely unchanged from the findings of Weisberg et al. (2009). Grissom and Loeb (2017) noted that principals would give higher scores in an accountability context than they would for professional development. Not surprisingly (see Rowan and Raudenbush, 2016), evaluators in professional contexts consider many factors aside from the performance itself in making ratings (Harris et al., 2014; Donaldson and Woulfin, 2018). Two explanations for the failure to identify weak teachers are (1) inconsistent training of evaluators (i.e.,

[1] Latin* is a term that encompasses fluidity of social identities. The asterisk considers variation in self-identification among people of the Latin American diaspora and origin (Salinas, 2020). Latin* responds to (mis)use of *Latinx*, a term reserved for gender-nonconforming peoples of Latin American origin and descent (Salinas and Lozano, 2019).

school leaders); and (2) the difficulty of negatively evaluating colleagues.

5.2. The promise of improving student performance

If the end goal for improving teacher quality through teacher evaluation is that students would benefit, results were disappointing. Stecher et al. (2018) observed null effects in terms of mathematics and English language arts (ELA) achievement in three large school districts across the 6 years of an intensive push to embed teacher evaluation systems. Bleiberg et al. (2021) conducted a cross-state analysis of student achievement by examining test scores before and after each state implemented their evaluation system. They also found null effects that did not vary over time since implementation.

5.3. The promise of changing the composition of the teaching force

Critical to the theory of action underlying this policy was the idea that weaker teachers could be replaced with more effective teachers. Substantial effects were found in Washington D. C. (Dee and Wyckoff, 2017; James and Wyckoff, 2020). While studies with other samples found changes in the teaching force, although to a lesser degree (e.g., Grissom and Bartanen, 2019; Nguyen et al., 2019; Cullen et al., 2021), Stecher et al. (2018) found null effects.

5.4. The promise of supporting more effective professional development

One of the key policy mechanisms for improving teaching quality was to provide better and more targeted professional development. Kraft and Gilmour (2017) and Stecher et al. (2018) did not find any evidence of such improvement and attributed this to the inherent tension between evaluations being used for accountability and high-stakes evaluations on the one hand and then being used for professional development on the other. In such cases, the accountability uses typically dominated and crowded out the professional development messages.

5.5. The promise of contributing to equity

Arguably the most important goal of this educational reform initiative was to improve the quality of teaching in schools that had histories of poor academic performance. Schools in urban and impoverished communities were of particular interest as those areas were the face of the U. S. education crisis (Cuban, 1989). These districts typically had high proportions of Black, Latin*, and Indigenous students, English language learners (ELLs), students considered in need of special education services, as well as the highest proportion of minoritized teachers (Boyd et al., 2010; Ronfeldt et al., 2016; D'Amico et al., 2017).

Again, Washington D. C. was relatively unique in making progress toward these goals, but this was an exception. In most targeted districts, teachers were disincentivized from working in low-performing schools. As we have discussed, teachers of students with weaker academic profiles fare more poorly on teacher evaluations (Drake et al., 2019). The bias that has been observed in these systems across measures is alarming.

There are multiple reasons why teachers of students with weaker academic profiles fare more poorly in these evaluation systems. As previously mentioned, there appears to be some statistical biases in the growth model estimates. Additionally, some low-achieving students have high levels of absenteeism, yet their test scores contribute as much to a teacher's estimate as those of students who rarely miss school. Cowen (2017) found that unhoused students, more likely to be Black and Latin*, are much more transient, almost always impoverished, and have lower achievement levels (i.e., classroom assessments and standardized test scores). Yet, teachers of these students are unfairly treated identically in the growth estimate models.

Classroom observations raise a number of additional issues with respect to equity. Jacob and Walsh (2011), Gitomer et al. (2014), Garrett and Steinberg (2015), Campbell and Ronfeldt (2018), and Steinberg and Sartain (2021) have all found that observation scores are systematically lower for teachers who teach students with weaker academic profiles.

Additionally, Black teachers are more likely to receive lower observation scores (Campbell and Ronfeldt, 2018; Steinberg and Sartain, 2021), and Black teachers who work in schools with predominantly White staff are more likely to receive lower evaluation ratings than those who work at schools with mostly Black colleagues (Drake et al., 2019). Campbell (2020) found that Black women received lower observation scores than White women, even when accounting for other measures of teaching quality, especially in schools where the race of the evaluator differed from that of the teacher.

Unfortunately, there is a paucity of research on the effects of teacher evaluation policies on teachers of students who represent the full range of students in American schools. However, there have been several studies that have discussed the complexity of conducting evaluations of teachers of special education students (Jones et al., 2022) and of English language learners (Turkan and Buzick, 2016).

6. Post-mortem

By almost any definition, the exuberant adoption and endorsement of teacher evaluation as a panacea for the educational problems facing the United States in the 2000s was hardly justified. None of the ambitious goals were satisfied. One of the primary reasons for this disappointment, we argue, is that the research foundations upon which all of this was built were myopic and insufficient to effectively implement and produce results that were technically valid and substantively robust enough to address the complex issues of teaching and learning in American schools.

We can return to the three operating frames that Nasir et al. (2016) identified to highlight the gaps in the research base and policy interpretation and also to demonstrate that the limitations of these frames also have consequences for the technical quality of the evaluation measures. We do not claim that these operating frames are the only reason for the technical problems that surfaced, but we do claim that they played a major role.

The first frame is *colorblindness,* which minimizes the consequences of race and argues that all policy prescriptions should be the same, independent of racial considerations. Yet, the failure to address race and racism in our educational system had profound negative influence on the utility of the evaluation systems. We see that historical forces that have located minoritized students in lower-performing schools and inadequately resourced neighborhoods actually have direct effects on all measures, independent of the actual skills of a particular teacher. We see evidence of significant bias in observation systems that produces lower scores for Black teachers and lower scores for teachers of Black children. And we see all of this as raising skepticism of the fairness of assessment-based systems (Gitomer and Iwatani, 2022).

The second frame is *meritocracy,* which would ascribe accomplishment as solely due to the actions of the individual teacher and their impact on the student, ignoring the systemic nature of racial disparities and also ignoring the interdependence of teachers with other educators, and the resources and constraints they are provided, the tests their students are given, what students experience in other classrooms and at home, and the complex interrelated web of other factors that all have an influence on what goes on in a classroom. Such an approach also ignores the complexity and messiness of conducting assessments and evaluations. Treating all of these factors as either measurement error or factors that can be statistically controlled is to ignore reality and trivialize the educational process. From a technical perspective, we see measures that have a tremendous amount of error associated with them and the fundamental problem that causal claims at an individual level cannot be supported. And, of course, much of the system was predicated on using student standardized achievement test scores as the primary, if not sole, marker of student progress.

The final frame is *neoliberalism*, the idea that market-based incentives and practices can be applied in the educational system. Such simple explanations do not help account for the range of motivations that institutions and individuals have in assigning evaluation scores. The fact that distributions of teacher ratings barely changed, despite being an explicit goal of the policy, points to the failures to understand complex organizational behaviors associated with performance judgments (Rowan and Raudenbush, 2016). The fact that pay-for performance systems have had modest to no effects (Springer et al., 2016) suggests that economic incentives are not sufficient enough to result in desired changes in teaching.

The idea that one could build evaluation systems based on the emergence of a set of attractive technologies and limited and limiting frames, without attending to social, cultural, organizational, political, and even measurement theory and research led to a system that was bound to fizzle.

Essential problems included, first, the failure to resolve the tension between the goals of accountability and professional development. Second, all constituent measures had very significant problems in supporting the kinds of inferences that were needed for a high-stakes evaluation system. While the measures were not without value, they were being asked to carry far more water than the system could support.

Finally, if it was not clear to some at inception, it should be abundantly clear after this grand social experiment that teacher evaluation was not the policy lever to challenge the ubiquitous inequities in our educational system. The systems tended to reify historical inequities rather than upend them. Had attention been paid to researchers who were considering the multi-level nature of educational influences as the system was designed, it is possible that certain missteps could have been avoided.

This experience also highlights the risks associated with conducting policy formation within an echo-chamber of researchers, funders, and intermediaries who all adopt a similar framing of the problem. Without challenge, one can continue to wind up with expensive and taxing policies that are ephemeral.

The fizzle of Race to the Top does not negate concerns about instructional quality, nor does it negate the need for thoughtful evaluation, hiring, and retention practices that are essential to any well-functioning institution.

There is no doubt that much was learned during the time preceding and concurrent with this policy. Classroom observation instruments and SLOs have the potential to be used as they were initially designed—to support professional development. VAM can be useful to understand educational issues at aggregate levels. But having measures alone, developed and researched in one context, is not a warrant for a massive policy initiative. In order to move forward on any kind of major educational reform policies in the future, much more sophisticated and nuanced theories of action will be required.

What might such more productive reforms look like? While it would be presumptuous to suggest a particular design, it is possible to outline certain principles that are critical to consider. We can draw on research that has studied effective schools and effective teaching across different contexts and countries to imagine policies to encourage, as well as those to avoid, in designing approaches to the evaluation of teaching within schools.

1. Teaching is contextually bound, and any attempt to understand and evaluate teaching as a reflection of the teacher alone is inherently misguided. Factors as far-ranging as curriculum, community, food and housing insecurity, school leadership, school and classroom resources, and students' language and culture all have profound effects on what transpires within a given classroom. A history of educational accountability policy in the United States has focused on particular entities in the system (students, schools, principals, teachers) apart from all these contextual issues, and each effort has failed. Any productive evaluation system needs to understand how teaching is influenced by, and influences, this larger context. Only then can more reasonable interpretations of particular actors and actions be made, and only then can more thoughtful decisions of follow-up actions be made.
2. Any system should pay explicit attention to issues of race, language, culture, and power in understanding and supporting classroom interactions. It is not sufficient to simply put forth standards that say all students' needs should be met. We know that there are specific challenges and approaches that engage and support students from different backgrounds (e.g., Ladson-Billings, 2009).
3. To the extent that teachers are held accountable for their teaching, measures should be transparent, actionable, and under teachers' control. A central critique of growth models used in teacher evaluation systems was that they did not meet any of these criteria. Measures that focus on teacher actions,

interactions, and decision-making are those that individuals and systems are more apt to be able to address.

4. The criteria against which teacher effectiveness is measured should reflect a full vision of teaching. The attractiveness of using growth measures was that these metrics were available for large numbers of teachers. They also led to mathematics and reading test scores receiving overwhelming attention, often to the exclusion of other subject areas and almost always to the exclusion of important outcomes of classroom instruction that were not measured by standardized achievement tests. Focusing on a small set of proxy measures for teacher evaluation will inevitably distort school practices (see Rowan and Raudenbush, 2016).
5. Systems should anticipate and try to avoid predictable reactions of how policies will be interpreted and acted upon. The inflation of observation scores and the far lower than anticipated classification of teachers as needing improvement should not have been surprising in light of what we know about how systems respond to performance appraisal systems (Rowan and Raudenbush, 2016). Actors will be less apt to shape responses to policy goals in unintended ways if they are invested in the goals and processes of the system. Any policy needs to be informed and have buy-in from practitioners in the field that is far greater than what was evident in Race to the Top.
6. Systems should have as a dominant goal the development of the educational system, which would include professional development for teachers and school leaders, curricular reform, community relationships, resource analysis, etc. While the Race to the Top system endorsed the rhetoric of professional development, effective efforts that built on the evaluations were not commonplace. Policy, resources, and attention were given to the mechanics of evaluation and human resource management far more than they were to system development. If future teacher evaluation efforts are to be successful, these priorities need to be inverted.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

American Educational Research Association Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educ. Res.* 44, 448–452. doi: 10.3102/0013189X15618385

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Available at: https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Anagnostopoulos, D., Rutledge, S. A., and Jacobsen, R. (Eds.). (2013). *The infrastructure of accountability: Data use and the transformation of American education*. Cambridge, MA: Harvard Education Press.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers (EPI briefing paper #278)*. Washington, DC: Economic Policy Institute.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., and Qi, Y. (2012). An argument approach to observation protocol validity. *Educ. Assess.* 17, 62–87. doi: 10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., et al. (2014). "Improving observational score quality: challenges in observer thinking" in *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. eds. T. J. Kane, K. A. Kerr, and R. C. Pianta (San Francisco, CA: Jossey-Bass), 50–97.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educ. Meas. Issues Pract.* 28, 42–51. doi: 10.1111/j.1745-3992.2009.00161.x

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., and Springer, M. (2021). *The effect of teacher evaluation on achievement and attainment: Evidence from statewide reforms*. Providence, RI: Annenberg Institute at Brown University.

Borko, H. (2004). Professional development and teacher learning: mapping the terrain. *Educ. Res.* 33, 3–15. doi: 10.3102/0013189X033008003

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., and Wyckoff, J. (2010). The role of teacher quality in retention and hiring: using applications to transfer to uncover preferences of teachers and schools. *J. Policy Anal. Manage.* 30, 88–110. doi: 10.1002/pam.20545

Braun, H. I. (2005). *Using student Progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Briggs, D. C., Dadey, N., and Kizil, R. C. (2014). *Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness*. Boulder, CO: Center for Assessment, Design, Research and Evaluation, University of Colorado.

Campbell, S. L. (2020). Ratings in black and white: a QuantCrit examination of race and gender in teacher evaluation reform. *Race Ethn. Educ.*, 1–19. doi: 10.1080/13613324.2020.1842345

Campbell, S. L., and Ronfeldt, M. (2018). Observational evaluation of teachers: measuring more than we bargained for? *Am. Educ. Res. J.* 55, 1233–1267. doi: 10.3102/0002831218776216

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: teachers for the 21st century*. New York: Carnegie Forum on Education and the Economy.

Close, K., Amrein-Beardsley, A., and Collins, C. (2020). Putting teacher evaluation systems on the map: an overview of state's teacher evaluation systems post-every student succeeds act. *Educ. Policy Analysis Archives* 28:58. doi: 10.14507/epaa.28.5252

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. (2010). Teacher credentials and student achievement in high school: a cross-subject analysis with student fixed effects. *J. Hum. Resour.* 45, 655–681. doi: 10.3368/jhr.45.3.655

Congressional Budget Office. (2012). Estimated impact of the American recovery and reinvestment act on employment and economic output from October 2011 through December 2011. Available at: http://www.cbo.gov/sites/default/files/cbofiles/attachments/02-22-ARRA.pdf

Cowen, J. M. (2017). Who are the homeless? Student mobility and achievement in Michigan 2010–2013. *Educ. Res.* 46, 33–43. doi: 10.3102/0013189X17694165

Crouse, K., Gitomer, D. H., and Joyce, J. (2016). "An analysis of the meaning and use of student learning objectives" in *Student growth measures in policy and practice: Intended and unintended consequences of high-stakes teacher evaluations*. eds. K. Kappler Hewitt and A. Amrein-Beardsley (New York: Palgrave Macmillan), 203–222.

Cuban, L. (1989). The 'at-risk' label and the problem of urban school reform. *Phi Delta Kappan* 70, 780–801.

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

D'Amico, D., Pawlewicz, R. J., Earley, P. M., and McGeehan, A. P. (2017). Where are all the black teachers? Discrimination in the teacher labor market. *Harv. Educ. Rev.* 87, 26–49. doi: 10.17763/1943-5045-87.1.26

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching 2nd ed.* Alexandria, VA: Association for Supervision and Curriculum Development.

Davidson, E., Reback, R., Rockoff, R., and Schwartz, H. L. (2015). Fifty ways to leave a child behind: idiosyncrasies and discrepancies in states' implementation of NCLB. *Educ. Res.* 44, 347–358. doi: 10.3102/0013189X15601426

DeBray, E., and Houck, E. A. (2011). A narrow path through the broad middle: mapping institutional considerations for ESEA reauthorization. *Peabody J. Educ.* 86, 319–337. doi: 10.1080/0161956X.2011.579009

Dee, T. S., and Jacob, B. (2011). The impact of no child left behind on student achievement. *J. Policy Anal. Manage.* 30, 418–446. doi: 10.1002/pam.20586

Dee, T., and Wyckoff, J. (2017). A lasting impact: high-stakes teacher evaluations drive student success in Washington, DC. *Educ. Next.* 17, 58–66.

Donaldson, M. L., and Woulfin, S. (2018). From tinkering to going "rogue": how principals use agency when enacting new teacher evaluation systems. *Educ. Eval. Policy Anal.* 40, 531–556. doi: 10.3102/0162373718784205

Drake, S., Auletto, A., and Cohen, J. M. (2019). Grading teachers: race and gender differences in low evaluation ratings and teacher employment outcomes. *Am. Educ. Res. J.* 56, 1800–1833. doi: 10.3102/0002831219835776

Duncan, A. (2009). "Robust data gives us the roadmap to reform" in *Address by the secretary of education to the fourth annual Institute of Education Sciences research conference* (Washington, DC). Available at: https://education44.org/speeches/robust-data-gives-us-the-roadmap-to-reform/

ESSA. (2015). Every student succeeds act, 20 U.S.C. § 6301. Available at: https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Everson, K. C. (2017). Value-added modeling and educational accountability: are we answering the real questions? *Rev. Educ. Res.* 87, 35–70. doi: 10.3102/0034654316637199

Garrett, R., and Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: evidence from the randomization of teachers to students. *Educ. Eval. Policy Anal.* 37, 224–242. doi: 10.3102/0162373714537551

Gates, B., and Gates, M. (2018). 10 tough questions we get asked (2018 annual letter). Available at: https://www.gatesnotes.com/2018-Annual-Letter

Gitomer, D. H., and Bell, C. A. (2013). "Evaluating teaching and teachers" in *APA handbook of testing and assessment in psychology*. ed. K. F. Geisinger, vol. *3* (Washington, DC: American Psychological Association), 415–444.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., and Pianta, R. C. (2014). The instructional challenge in improving teaching quality: lessons from a classroom observation protocol. *Teach. Coll. Rec.* 116, 1–32. doi: 10.1177/016146811411600607

Gitomer, D. H., and Iwatani, E. (2022). "Fairness and assessment: engaging psychometric and racial justice perspectives" in *Race and culturally responsive inquiry in education: Improving research, evaluation, and assessment*. eds. S. L. Hood, H. T. Frierson, R. K. Hopson, and K. N. Arbuthnot (Cambridge, MA: Harvard Education Press).

Gitomer, D. H., and Marshall, B. (in press). "The bold and unfulfilled promises of teacher evaluation as policy" in *Handbook of education policy research*. eds. L. Cohen-Vogel, J. Scott, and P. Youngs (Washington, DC: American Educational Research Association).

Goe, L. (2007). *The link between teacher quality and student outcomes: a research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., and Theobold, R. (2013). *Do different value-added models tell us the same things?* Stanford, CA: Carnegie Knowledge Network.

Gordon, R., Kane, T. J., and Staiger, D. O. (2006). *Identifying effective teachers using performance on the job (the Hamilton project discussion paper 2006–01)*. Washington, DC: The Brookings Institution.

Grissom, J. A., and Bartanen, B. (2019). Strategic retention: principal effectiveness and teacher turnover in multiple-measure teacher evaluation systems. *Am. Educ. Res. J.* 56, 514–555. doi: 10.3102/0002831218797931

Grissom, J. A., and Loeb, S. (2017). Assessing principals' assessments: subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Educ. Finance Policy* 12, 369–395. doi: 10.1162/EDFP_a_00210

Grossman, P., Cohen, J., Ronfeldt, M., and Brown, L. (2014). The test matters: the relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educ. Res.* 43, 293–303. doi: 10.3102/0013189X14544542

Hand, V., Penuel, W. R., and Gutiérrez, K. D. (2012). (Re)framing educational possibility: attending to power and equity in shaping access to and within learning opportunities. *Hum. Dev.* 55, 250–268. doi: 10.1159/000345313

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harris, D. N., Ingle, W. K., and Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: a comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *Am. Educ. Res. J.* 51, 73–112. doi: 10.3102/0002831213517130

Harris, D. N., and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *J. Public Econ.* 95, 798–812. doi: 10.1016/j.jpubeco.2010.11.009

Heneman, H. G., Milanowski, A., Kimball, S. M., and Odden, A. (2006). *Standards-based teacher evaluation as a Foundation for Knowledge- and Skill-based pay (CPRE policy brief RB-45)*. Philadelphia, PA: Consortium for Policy Research in Education.

Ho, A. D., and Kane, T. J. (2013). *The reliability of classroom observations by school personnel (MET project research paper)*. Seattle, WA: Bill and Melinda Gates Foundation.

Jacob, B. A., and Walsh, E. (2011). What's in a rating? *Econ. Educ. Rev.* 30, 434–448. doi: 10.1016/j.econedurev.2010.12.009

James, J., and Wyckoff, J. H. (2020). Teacher evaluation and teacher turnover in equilibrium: evidence from DC public schools. *AERA Open* 6, 1–21. doi: 10.1177/2332858420932235

Johnson, S. M., Kardos, S. M., Kauffman, D., Liu, E., and Donaldson, M. L. (2004). The support gap: new teachers' early experiences in high-income and low-income schools. *Educ. Policy Analysis Archives* 12:61. doi: 10.14507/epaa.v12n61.2004

Jones, N. D., Bell, C. A., Brownell, M., Qi, Y., Peyton, D., Pua, D., et al. (2022). Using classroom observations in the evaluation of special education teachers. *Educ. Eval. Policy Anal.* 44, 429–457. doi: 10.3102/01623737211068523

Kane, T. J., Rockoff, J. E., and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Econ. Educ. Rev.* 27, 615–631. doi: 10.1016/j.econedurev.2007.05.005

Kane, T. J., Taylor, E. S., Tyler, J. H., and Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data*. Cambridge, MA: National Bureau of Economic Research.

Kardos, S. M., and Johnson, S. M. (2007). On their own and presumed expert: new teachers' experience with their colleagues. *Teach. Coll. Rec.* 109, 2083–2106. doi: 10.1177/016146810710900903

Katz, M. B., and Rose, M. (Eds.) (2013). *Public education under siege*. Philadelphia, PA: University of Pennsylvania Press.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Econ. Educ. Rev.* 28, 682–692. doi: 10.1016/j.econedurev.2009.02.003

Kraft, M. A., and Gilmour, A. F. (2017). Revisiting *the widget effect*: teacher evaluation reforms and the distribution of teacher effectiveness. *Educ. Res.* 46, 234–249. doi: 10.3102/0013189X17718797

Ladson-Billings, G. (2009). *The Dreamkeepers: Successful teachers of African American children. 2nd edn.* San Francisco, CA: Jossey-Bass.

Lee, J., and Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: state NAEP 1990–2009 reading and math achievement gaps and trends. *Educ. Eval. Policy Anal.* 34, 209–231. doi: 10.3102/0162373711431604

Lockwood, J. R., and Castellano, K. E. (2017). Estimating true student growth percentile distributions using latent regression multidimensional IRT models. *Educ. Psychol. Meas.* 77, 917–944. doi: 10.1177/0013164416659686

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., and Martínez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *J. Educ. Meas.* 44, 47–67. doi: 10.1111/j.1745-3984.2007.00026.x

Loveless, T. (2021). *Between the state and the schoolhouse: understanding the failure of Common Core*. Cambridge, MA: Harvard Education Press.

Maranto, R., McShane, M. Q., and Rhinesmith, R. (2016). *Education reform in the Obama era: the second term and the 2016 election*. New York: Palgrave Macmillan.

McDonald, S.-K. (2009). "Scale-up as a framework for intervention, program, and policy evaluation research" in *Handbook of education policy research*. eds. G. Sykes, B. Schneider, and D. N. Plank (Washington, DC: American Educational Research Association), 191–208.

McDonnell, L. M., and Weatherford, M. S. (2020). *Evidence, politics, and education policy*. Cambridge, MA: Harvard University Press.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: evidence from Cincinnati. *Peabody J. Educ.* 79, 33–53. doi: 10.1207/s15327930pje7904_3

Murnane, R. J., and Phillips, B. R. (1981). What do effective teachers of inner-city children have in common? *Soc. Sci. Res.* 10, 83–100. doi: 10.1016/0049-089X(81)90007-7

Nasir, N. S., Scott, J., Trujillo, T., and Hernández, L. (2016). "The sociopolitical context of teaching" in *Handbook of research on teaching*. eds. D. H. Gitomer and C. A. Bell (Washington, DC: American Educational Research Association), 349–390.

National Association of Secondary School Principals. (2019). *Value-added measures in teacher evaluation (NASSP position statement)*. Available at: https://www.nassp.org/top-issues-in-education/position-statements/value-added-measures-in-teacher-evaluation/

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Available at: https://edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

National Research Council. (2008). *Assessing accomplished teaching: advanced-level certification programs*. Washington, DC: The National Academies Press.

National Research Council. (2015). *An evaluation of the public schools of the District of Columbia: Reform in a changing landscape*. Washington, DC: The National Academies Press.

Nguyen, T. D., Pham, L., Springer, M., and Crouch, M. (2019). *The factors of teacher attrition and retention: an updated and expanded meta-analysis of the literature*. (Ed Working Paper No. 19-149) Providence, RI: Annenberg Institute at Brown University.

Nye, B., Konstantopoulos, S., and Hedges, L. V. (2004). How large are teacher effects? *Educ. Eval. Policy Anal.* 26, 237–257. doi: 10.3102/01623737026003237

Paige, M. A., and Amrein-Beardsley, A. (2020). "Houston, we have a lawsuit": a cautionary tale for the implementation of value-added models for high-stakes employment decisions. *Educ. Res.* 49, 350–359. doi: 10.3102/0013189X20923046

Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS)*. Baltimore, MD: Paul H. Brookes.

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., and Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educ. Res.* 43, 45–54. doi: 10.3102/0013189X13517137

Raudenbush, S. W. (2013). *What do we know about using value-added to compare teachers who work in different schools?* Stanford, CA: Carnegie Knowledge Network.

Reardon, S. F., and Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Educ. Finance Policy* 4, 492–519. doi: 10.1162/edfp.2009.4.4.492

Reckhow, S., and Tompkins-Stange, M. (2018). Financing the education policy discourse: philanthropic funders as entrepreneurs in policy networks. *Interest Groups Advoc.* 7, 258–288. doi: 10.1057/s41309-018-0043-3

Reckhow, S., Tompkins-Stange, M., and Galey-Horn, S. (2021). How the political economy of knowledge production shapes education policy: the case of teacher evaluation in federal policy discourse. *Educ. Eval. Policy Anal.* 43, 472–494. doi: 10.3102/01623737211003906

Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Educ. Finance Policy.* 6, 43–74. doi: 10.1162/EDFP_a_00022

Ronfeldt, M., Kwok, A., and Reininger, M. (2016). Teachers' preferences to teach underserved students. *Urban Educ.* 51, 995–1030. doi: 10.1177/0042085914553676

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

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *Am. Econ. Rev.* 107, 1656–1684. doi: 10.1257/aer.20141440

Rowan, B., Camburn, E., and Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: a study of literacy teaching in third-grade classrooms. *Elem. Sch. J.* 105, 75–101. doi: 10.1086/428803

Rowan, B., and Correnti, R. (2009). Measuring reading instruction with teacher logs. *Educ. Res.* 38, 549–551. doi: 10.3102/0013189X09349313

Rowan, B., Jacob, R., and Correnti, R. (2009). Using instructional logs to identify quality in educational settings. *New Dir. Youth Dev.* 2009, 13–31. doi: 10.1002/yd.294

Rowan, B., and Raudenbush, S. W. (2016). "Teacher evaluation in American schools" in *Handbook of research on teaching*. eds. D. H. Gitomer and C. A. Bell (Washington, DC: American Educational Research Association), 1159–1216.

Salinas, C. Jr. (2020). The complexity of the "x" in *Latinx*: how Latinx/a/o students relate to, identify with, and understand the term *Latinx. J. Hisp. High. Educ.* 19, 149–168. doi: 10.1177/1538192719900382

Salinas, C. Jr., and Lozano, A. (2019). Mapping and recontextualizing the evolution of the term *Latinx*: an environmental scanning in higher education. *J. Latinos Educ.* 18, 302–315. doi: 10.1080/15348431.2017.1390464

Sanders, W. L., and Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. *J. Pers. Eval. Educ.* 8, 299–311. doi: 10.1007/BF00973726

Sartain, L., Stoelinga, S. R., and Krone, E. (2010). *Rethinking teacher evaluation: findings from the first year of the excellence in teaching project in Chicago public schools*. Chicago, IL: Consortium on Chicago School Research, University of Chicago.

Springer, M. G., Swain, W. A., and Rodriguez, L. A. (2016). Effective teacher retention bonuses: evidence from Tennessee. *Educ. Eval. Policy Anal.* 38, 199–221. doi: 10.3102/0162373715609687

Stallings, D. T. (2002). A brief history of the U. S. Department of Education, 1979–2002. *Phi Delta Kappan* 83, 677–683. doi: 10.1177/003172170208300910

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., et al. (2018). *Improving teacher effectiveness: Final report: The intensive partnerships for effective teaching through 2015–2016*. Santa Monica, CA: RAND Corporation.

Steinberg, M. P., and Sartain, L. (2021). What explains the race gap in teacher performance ratings? Evidence from Chicago public schools. *Educ. Eval. Policy Anal.* 43, 60–82. doi: 10.3102/0162373720970204

Turkan, S., and Buzick, H. M. (2016). Complexities and issues to consider in the evaluation of content teachers of English language learners. *Urban Educ.* 51, 221–248. doi: 10.1177/0042085914543111

U. S. Department of Education. (2009). *Race to the top program: executive summary*. Washington, DC: U. S. Department of Education.

Wayne, A. J., and Youngs, P. (2003). Teacher characteristics and student achievement gains: a review. *Rev. Educ. Res.* 73, 89–122. doi: 10.3102/00346543073001089

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York: The New Teacher Project.

Williams, J. H., and Engel, L. C. (2012). How do other countries evaluate teachers? *Phi Delta Kappan.* 94, 53–57. doi: 10.1177/003172171209400414