THE DESIGN AND PRACTICE OF U.S. TEACHER EVALUATION

Veronica Katz Luke C. Miller Jim Wyckoff

September 2019

EdPolicyWorks
University of Virginia

This paper has been prepared for the Virginia Department of Education under contract 7.7-PR8703018- 2019. We appreciate helpful comments on a prior draft from Jen Piver-Renna. All errors remain the sole responsibility of the authors.

THE DESIGN AND PRACTICE OF U.S. TEACHER EVALUATIONS

Introduction

Education has long been viewed as a crucial catalyst for economic and social development, and mounting evidence points to teacher quality as the most important school factor that influences student outcomes. As we discuss below, the past decade witnessed both growing evidence regarding the importance of teacher quality for student success and a fertile political landscape that enabled the proliferation of new teacher evaluation systems throughout the United States.

We summarize the research literature examining the design, measurement properties and effects of teacher evaluation systems in the U.S. with a particular focus on systems that have emerged over the last eight years. We ground our discussion of each of these evaluation systems with a discussion of the theoretical underpinnings that link evaluation with proximal and distal outcomes of interest. We then summarize how these systems have been implemented in several states and school districts. These evaluation systems have been chosen to illustrate differences across five features we consider influential in realizing improved teacher effectiveness and student outcomes. Taken together, this review is intended to provide a high-level orientation to the general logic, measurement properties and structure of evaluation as currently practiced in the United States.

We begin with a description of the historical and political context that facilitated the increasingly commonplace use of evaluation systems. We then examine teacher evaluation systems.

Background

Education is critical to economic and social development. A large conceptual and empirical literature documents these relationships and how they evolve as countries develop.[1] Educational outcomes are influenced by a variety of school and non-school factors suggesting that policymakers could pursue a variety of policies that would improve student outcomes. We now have strong evidence for what may be common sense to most: teachers are the single most important school-based factor in determining improvement in student outcomes.[2] Research also documents that large differences exist within and across the teachers in school districts. These differences have important effects on student learning.[3] Policies intended to improve the quality of teaching in public schools are many and varied.[4] These policies attempt to recruit potentially effective teachers, differentially retain more effective teachers, and develop teachers to become more effective. Systematic teacher evaluation offers the potential to rigorously identify the strengths and weaknesses of teachers to guide their development and to hold them accountable for their effectiveness.

Evolution of Teacher Evaluation in the U.S.

The evaluation of U.S. public school teachers dates to early in the 20th century. Throughout the ensuing century there were a variety of attempts to use assessments of teachers

[1] For a good summary of this from the U.S. perspective see Goldin and Katz (2009).
[2] See, for example, Aaronson, Barrow, and Sander, 2007.
[3] See Chetty, Friedman, and Rockoff, 2014a and 2014b.
[4] See, for example, Goldhaber (2015) or Katz & Wyckoff (2017).

to improve the quality of teaching and student outcomes. And while there were periods and pockets of success, for the most part teacher evaluations rarely had any implications for teachers. A combination of concerns regarding unbiased measures of teacher effectiveness and the competing views of important stakeholders contributed to most teacher evaluations being viewed as uninformative and inconsequential. However, about ten years ago a confluence of pressure to improve student outcomes, newly available data, research on teacher effectiveness and interest from policymakers brought on a new wave of teacher evaluation reform in the U.S.

Following the publication of *A Nation at Risk* in 1983, pressure mounted on states to improve student achievement which resulted in 2001 in bipartisan national support for No Child Left Behind (NCLB), federal legislation which imposed increasing sanctions on schools whose students did not meet increasing student achievement goals. At about the same time, employing newly available administrative data, researchers documented that teachers are the single most influential factor in student achievement (Aaronson, Barrow, & Sander, 2007) and that teachers differ substantially in their ability to improve student outcomes (Aaronson et al., 2007; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Taken together these findings created a sense of urgency around policies to improve teacher quality in the nation's schools, especially those in urban areas where student disadvantage is concentrated. Soon thereafter, *The Widget Effect* (Weisberg et al., 2009) documented the nature of teacher evaluation in the typical school district. That study found that most teacher evaluation processes were perfunctory, contained little actionable feedback for teachers, rated nearly all teachers Satisfactory, and had no consequences for teachers. As a result, many teachers viewed teacher evaluation at best as unhelpful or, more frequently, as a demoralizing process that did not value their skills or their development.

These findings informed several policy initiatives from the U.S. Department of Education. The Race to the Top (RttT), the Teacher Incentive Fund (TIF), and waivers under No Child Left Behind (NCLB) encouraged states and districts to focus more directly on rigorous teacher evaluation which included student achievement test growth as one of the measures of teachers' performance (Donaldson & Papay, 2015). As a result, most states and school districts modified teacher evaluation systems with the goal that these systems would improve teacher quality and student outcomes. The goal was to design valid, transparent and reliable teacher evaluation systems. That is, evaluation systems that accurately identified future teacher effectiveness in a way that teachers understood the basis for their evaluation and did so consistently across teachers. We describe these evaluation systems in much greater detail below, but they all included features which were intended to realize these goals and address the perceived limitations of earlier teacher evaluation systems.

Eligibility for federal support available through RttT and TIF required states and school districts to make substantial changes in teacher evaluation, which needed to be designed and implemented quickly. While the design of many of these systems employed the best available evidence from the Measures of Effective Teaching (MET) project (Kane, McCaffrey, Miller, & Staiger, 2013) and other research, little time for implementation and political constraints often influenced their design and implementation. For example, the compressed timeframe associated with RttT and TIF left many states with little time and insufficient resources for careful implementation. For example, most systems included some in-class observation of teachers by a trained expert employing a standards-based observation rubric. Implementing teacher observations rigorously calls for the use of a well-designed rubric, training of observers until they are reliable on the instrument, frequent observation of teachers and systematic feedback from the observer to teachers.[5] Implementing observation protocols to such a standard is expensive and time-consuming. While we are aware of no systematic evidence, there is anecdotal evidence that few states or districts invested sufficiently to create such systems and work with teachers to ensure successful implementation. As we describe in more detail below, what emerged are a collection of teacher evaluation systems with some superficial similarities but with important variation across and within states

U.S. teachers' unions have historically opposed evaluation systems when those systems meaningfully differentiate among teachers and when evaluation differences are associated with differing consequences, e.g., differential professional development, bonus pay or dismissal. This opposition was often expressed as concerns over the validity and reliability of evaluation measures or anxiety over the lack of adequate due process. The political economy of teachers' unions caused them to weigh the misclassification of teachers much more heavily than the benefits of improved teacher quality that an evaluation system could yield. This perspective, coupled with the rushed nature of the design and implementation in most states, caused unions to oppose the use of teacher evaluation systems. As a result, some states and districts chose not to apply for federal support, and others withdrew from TIF support when it became apparent they could not implement their teacher evaluation systems as proposed. Nonetheless, the momentum for improved teacher quality led most states and districts to design and implement new evaluation systems, at times overriding union opposition. However, union opposition would resurface as these systems began to function.

Theory of Change

Student outcomes are influenced by a variety of family, community and school-based factors. The research literature identifies the ways in which teacher evaluation can lead to improved quality of instruction and, in turn, improved student outcomes. We summarize these relationships in Figure 1. A very large literature documents the many ways in which students, their families and their community context influences student outcomes (denoted by dashed arrows in Figure 1). However, we know that teachers are the most influential school-based factor in determining student outcomes. Teacher evaluation can improve teacher quality through two primary mechanisms: the composition and development of teachers (denoted by the light green boxes in Figure 1). The process of evaluation, especially if teachers receive useful feedback and coaching and are motivated to improve, can help teachers develop.[6] This development will improve teacher quality and outcomes for students. Evaluation may also lead to the differential retention of more effective teachers, improving the composition of teachers, which in turn improves average teacher quality and student outcomes.

There is increasing evidence that the design and implementation of teacher evaluation can importantly influence teacher development, teacher composition, and student outcomes. Teacher evaluation embodies a number of design decisions—what performance measures to include, the nature of feedback that teachers receive, whether they receive coaching linked to this feedback and whether there are stakes associated with the evaluations. There is growing evidence that each of these design decisions can influence proximal outcomes of teacher development and the composition of the teacher workforce. For example, if the sole measure of teacher effectiveness

[5] See, for example, Kane et al. (2013)

[6] There is a growing literature that documents the effectiveness of teacher evaluation coupled with coaching or other targeted professional development. For a summary of the coaching literature see Kraft, Blazar & Hogan (2018).

is value-added, teachers will learn how much more or less effective they are than their peers, but they are given very limited guidance of how to improve. However, if value-added is a good predictor of student outcomes and teachers face meaningful consequences for their value-added performance, e.g., financial rewards or dismissal, depending on their performance, then evaluation based on value-added would likely improve the composition of teachers. Rigorous teacher observations have the potential to provide teachers with actionable feedback that could improve their teaching skills and, in turn, student outcomes. Observations also are much more transparent with the attendant ability to increase confidence in the evaluation system. As this simple example shows, and we develop in greater detail below, including multiple measures in has may meet several complementary goals of teacher evaluation systems.

Yet, no system of evaluation, including teacher evaluation, will operate without mistakes—at times two individuals with equivalent effects on student outcomes will have their performance measured differently. This may arise because measures don't adequately account for all aspects of teaching that are valued or because these differing outcomes are not measured well. While unfortunate, and an outcome that should lead to the development of more robust evaluation measures, such mistakes don't necessarily negate the usefulness of an evaluation system. Rather, conditional on the available evaluation measures, such misclassifications of teachers should be weighed against the gains in teacher quality and student achievement that evaluation systems may yield.[7]

We are beginning to learn that as important as the design of teacher evaluation may be, how it is implemented is also important. Here we highlight three aspects that appear relevant (denoted by the orange boxes in Figure 1). A variety of forces may mitigate against differentiating effectiveness among teachers, even when it exists. For example, if effectiveness is measured by the school leader's observation of teacher classroom performance, these observations may be biased due to interpersonal issues between teachers and the school leaders. Lack of differentiation limits the ability of teacher evaluation to improve the composition and development of teachers. Similarly, it is important that measures of teacher effectiveness are consistently applied across teachers. Lack of reliability again leads to misclassification of teachers for development or other meaningful interventions; it also reduces teachers' confidence that the system is credible. Finally, implementation is influenced by the sense of collegiality and trust that school leaders have built, especially in creating acceptance of negative evaluations and motivating development. A recent study of the effectiveness of teacher evaluation in three school districts and four charter management organizations concluded that evaluation had not led to improved student outcomes. While not definitive, the report also indicated that due to financial and human resource constraints teacher evaluation had not been well-implemented (Stecher et al., 2018).

As this brief overview suggests, the potential for teacher evaluation to improve teacher quality depends on several design and implementation factors. Next, we examine the teacher evaluation design elements for which evidence appears most promising to yield improved teacher quality and student outcomes. To this end, we explore three aspects of teacher evaluation:

- How does the design of teacher evaluation vary across states and school districts?

[7] Atteberry, Loeb and Wyckoff (2015) provide a more detailed discussion of these tradeoffs in the context of value-added measures of teacher effectiveness.

- What role does value-added play in these systems and what evidence do we have regarding its properties?
- What effect does teacher evaluation have on improving the quality of teaching and student outcomes?

Teacher evaluation design

Strictly speaking, a teacher evaluation system need only measure teacher effectiveness. However, as the theory of change suggests, measuring effectiveness is just the first step toward improving teacher quality and student outcomes. How this information is employed by employers and teachers determines whether there is improvement in teacher quality through the development of teaching skills and the composition of teachers. We examine both the measurement of teacher effectiveness and the mechanisms that employ this information to improve teacher quality and student outcomes.

Measuring Teacher Effectiveness

We identify and describe the measures of teacher effectiveness that may be included in teacher evaluation systems, how these measures can be combined to create an overall rating of teacher effectiveness, and the sanctions and rewards that might be tied to final ratings.

Test-based measures of teacher effectiveness. Test-based measures of teacher effectiveness are at the heart of our summary because one may reasonably expect effective teachers to improve student outcomes. Although test scores are not the only (nor necessarily the most important) student outcome education aims to improve, there are several advantages associated with test-based measures of teacher effectiveness. First, as a result of NCLB, all U.S. school systems administer standardized achievement tests annually; student-level achievement data is therefore readily available and there are few additional costs associated with analyzing these data to create measures of teacher effectiveness (Donaldson & Papay, 2015). Second, standardized achievement tests capture one student outcome that school systems aim to improve. Third, estimates of current teacher value-added have been shown to be the best predictor of future value-added—they are valid estimates (Chetty, Friedman, & Rockoff, 2014a; Kane et al., 2013). Finally, because student performance on standardized achievement tests is correlated with other student outcomes of interest (Chetty, Friedman, & Rockoff, 2014b), these data provide a useful proxy for other outcomes school systems aim to improve.

Despite these advantages, there are also distinct disadvantages associated with test-based measures of teacher performance. First, the validity of the measures depends heavily on the strength of the achievement test administered to students. Some student achievement tests have been shown to be more reliable than others, and this reliability (or lack thereof) is in turn reflected (indeed, amplified) in test-based measures of teacher effectiveness. Second, student achievement tests are typically administered to students in certain grades and subjects (typically math and reading in grades three through eight) thereby limiting the number of teachers who can be linked to students with achievement data. As a result, test-based measures of teacher effectiveness are typically generated for fewer than 20 percent of teachers. Third, test-based measures of teacher effectiveness typically cannot be estimated until the end of the school year, and therefore cannot be used to provide teachers with timely feedback. Fourth, because test-based measures of teacher effectiveness do not measure teacher performance on specific instructional practices, these measures provide teachers with little information about how to improve student outcomes. (Donaldson & Papay, 2015)

Nonetheless, there is growing evidence to support the inclusion of test-based measures of teacher effectiveness as one of multiple components in a robust teacher evaluation system. When combined with other measures of teacher effectiveness, test-based measures of teacher effectiveness not only link teacher performance to student outcomes but also provide an important benchmark against which to measure the reliability of other component measures (Cantrell & Kane, 2013; Harris, 2011; Tennessee Department of Education, 2016).

There are two primary test-based measures of teacher effectiveness: value-added measures (VAMs) and student growth percentiles (SGPs). SGPs compare one student's achievement growth against the growth of other students with similar prior-grade performance histories. Teacher effectiveness is then measured using the mean or median SGP, where the 50th percentile represents average student growth. Like SGPs, VAMs also control for prior student achievement but do not compare student growth to the growth of similar peers. Instead, VAMs use prior student achievement and other student characteristics to estimate predicted achievement for each student. Predicted achievement is then subtracted from actual achievement, yielding the difference in student performance relative to each student's expected performance. These differences are then averaged for all students in a teachers' classroom to produce teacher value-added. In this manner, VAMs capture a teacher's average contribution to student learning after controlling for each student's baseline performance and background characteristics.

The inclusion of background characteristics in the estimation of value-added is one of many points of contention in the ongoing dialogue regarding teacher evaluation (Donaldson & Papay, 2015; Harris, 2011). For instance, the inclusion of a student's socio-economic status (SES) as a control estimates different levels of predicted achievement for students with different economic profiles. On the one hand, including this demographic characteristic acknowledges that students from less affluent families face considerable hardships that likely affect their academic performance. From this perspective, controlling for student characteristics seeks to estimate teacher value-added in a manner that accounts for classroom composition that influences student achievement that is beyond the control of teachers. On the other hand, controlling for differences in performance due to student SES can be perceived as an implicit effort to set different learning standards for student subgroups.

How does value-added compare with principal perceptions of teacher effectiveness? The teachers identified by principals as their most and least effective teachers (without any value-added information) correlated very highly with those identified by value-added. (Harris & Sass, 2014; Jacob & Lefgren, 2008). While value-added has important limitations, it nonetheless typically passes a face-validity threshold with principals.

While there is no easy answer to this issue, ongoing debate regarding the estimation of value-added highlights one of the fundamental questions about teacher evaluation in general: how can school systems fairly and reliably evaluate teacher effectiveness? Emerging evidence from teacher evaluation systems points to the importance of implementing rigorous teacher evaluation systems that draw on multiple measures of teacher effectiveness. While test-based measures of teacher effectiveness play an important role in teacher evaluation, school systems may also opt to include other student growth outcomes, especially as a means of holding all teachers accountable for student learning.

Other student achievement measures. Test-based measures of teacher effectiveness can only be calculated for roughly 20 percent of teachers (i.e., those whose students participate in annual standardized examinations). In an effort to link all teachers to student learning outcomes, school systems have sought to develop student achievement-based measures of teacher effectiveness for teachers in grades and subjects not subject to state standardized tests. Two popular options include school value-added (SVA) and student learning objectives (SLOs). SVA is the measure that results when value-added estimates are aggregated to the school level. Because the measure has the same statistical root as VAMs, it is subject to most of the same commentary: the measure is only summative and therefore lacks timeliness; the measure does not identify specific practices to support teacher development; and there is the usual contention around the inclusion of student covariates. However, SVA is also often critiqued due to the high level of aggregation. That is, critics argue that it is unfair to hold teachers accountable for the performance of their colleagues (Steinberg & Donaldson, 2016). On the other hand, one can logically argue that teacher value-added fails to account for the contributions others (e.g., tutors, specialists) make to student academic progress (Donaldson & Papay, 2015).

In many ways, SLOs seek to strike a balance between teacher value-added and school value-added. The measure can be aggregated to the teacher level, but it does not rely on standardized achievement tests. Most commonly, SLOs are learning goals established by a teacher for each of her students at the beginning of the school year. These learning goals are reviewed by an administrator and approved. The teacher then tracks student progress toward the completion of the approved learning goals. One advantage of SLOs is that they can be established for all teachers, regardless of the subject or grade of instruction. However, it is also very hard to standardize SLOs and ensure faithful implementation. Indeed, what little evidence exists suggests SLOs are only modestly correlated with student achievement on standardized tests (Goldhaber & Walch, 2012; Slotnick & Smith, 2004).

Standards-based observations. Standards-based observations (SBO) are intended to provide teachers with timely feedback regarding their performance on specific instructional practices. The instructional practices are usually identified in an observation rubric that describes what each practice looks like when implemented at different levels of efficacy. These observation rubrics provide teachers with a transparent description of the practices they will be evaluated on during each classroom observation. The possibility of providing teachers with timely and specific feedback on measures which are readily accepted by teachers are the strongest merits of SBOs; these merits also stand in direct contrast with two of the most common criticisms leveled again test-based measures of teacher effectiveness. Nonetheless, SBOs are not without their unique limitations. In particular, there are significant time and personnel costs associated with the implementation of SBOs (Donaldson & Papay, 2015; Steinberg & Donaldson, 2016). Teacher observation measures of effectiveness are typically correlated with VAM between 0.20 and 0.35 (Kane, McCaffrey, Miller & Staiger, 2013), suggesting they measure different aspects of teacher effectiveness. In addition to designing an observation rubric that reflects essential instructional practices, school systems must make a plethora of decisions in order to implement SBOs. Specifically, school systems must decide:

- Who will conduct observations? If current employees (e.g., teachers, administrators, instructional coaches) are conducting observations, how will they find time to meaningfully engage in classroom observations? If the school system will employ external evaluators, how much will this cost?

- How will observers be trained and what systems will be implemented to gauge reliability?
- How often will teachers be evaluated?
- Are all observations formal or are some observations informal?
- Should observations be announced or unannounced?
- What are the post-observation conferencing requirements?
- Should the number and pacing of observations differ for teachers with different levels of experience? For instance, perhaps new teachers should be the first to receive informal observations, whereas teachers who consistently demonstrate strong instructional practices could receive fewer observations over the course of the school year?
- Should the observation rubric focus on general instructional practices, or should school systems develop different rubrics for teachers based on their subject(s) and grade(s) of instruction?

Despite the numerous intricacies associated with SBOs, this measure is the most commonly observed component in teacher evaluation systems in the U.S. This may in part be due to the fact that SBOs are a logical extension of the checklist style of observation that new evaluation systems sought to replace. That is, prior to the widespread push to reform teacher evaluation systems, most teachers received an annual classroom observation that largely focused on checking for the presence of certain artifacts: a visibly posted learning standard, an orderly classroom, teacher professional attire, and other readily observable features. In some ways, SBOs are a logical descendant of checklist observations; the checklist is replaced with a rubric and the observations occur at a more regular interval, but the general logic is similar and, indeed, familiar.

Other measures of teacher effectiveness. While SBOs and achievement-based measures of teacher effectiveness are well known, there are other less common measures of teacher effectiveness designed to evaluate different domains of teacher effectiveness. These measures include community involvement, teacher professional conduct, student surveys, and parent/caregiver surveys.

Measures of community involvement seek to capture the extent to which a teacher nurtures a positive learning environment for students, families, and colleagues. In some instances, administrators use a rubric to evaluate this construct. In other school systems, student or caregiver surveys may include items that speak to a teachers' community involvement. Regardless of the manner in which this information is gathered, the inclusion of this construct in a teacher evaluation system sets an expectation that teachers will contribute to the broader school community. By the same token, inclusion of the measure implies that teachers will be recognized for their service to the broader community.

Some school systems have also piloted the inclusion of student and/or caregiver surveys. Studies indicate that students are able to identify facets of teacher effectiveness that extend well-beyond the likeability of their teacher (Cantrell & Kane, 2013; Peterson, Wahlquist, Bone, Thompson, & Chatterton, 2001).

Finally, the inclusion of teacher professional conduct as a component measure sets clear expectations for teachers as professionals. In addition to improving student outcomes, demonstrating effective instructional practices, exhibiting commitment to the school community, teachers are expected to take their profession seriously. This construct may include domains such as professional appearance and demeanor, attendance and timeliness, evidence of preparation for instruction, participation in meetings and events, and more.

Rating Teacher Effectiveness

States and districts typically combine some subset of the previously discussed measures of teacher effectiveness to generate an overall rating of teacher effectiveness. This is consistent with what is currently considered best practice (Kane et al., 2013). Once again, school systems must make many decisions before arriving at an overall rating of teacher effectiveness. The following section provides more detail about deciding which component measures to include and how to combine them to obtain an overall rating.

Combining measures of teacher effectiveness. Growing evidence points to the importance of combining multiple measures of teacher effectiveness to assign final evaluation scores and ratings. Combining multiple measures of teacher effectiveness not only increases reliability, but also paints a more holistic measure of teacher effectiveness. The seminal MET project demonstrated that different measures of teacher effectiveness are positively, but not perfectly, correlated. Positive correlations between measures means that teachers who do well on one measure also do well on another measure. The fact that measures are not perfectly correlated is typically taken as an indication that the component measures capture distinct facets of effective teaching. For instance, teachers who receive high value-added scores also score higher on SBOs, suggesting that these teachers not only improve student outcomes but also demonstrate more effective instructional practices. (Cantrell & Kane, 2013; Kane et al., 2013)

While there seems to be growing consensus around the importance of combining multiple measures of teacher effectiveness, school systems have less of a research base to rely on when trying to determine which measure to include and how much weight to assign to each measure. Once again turning to the MET study, there is good evidence to support the inclusion of student achievement gains and classroom observations as key components of a rigorous teacher evaluation system. The study even goes so far as to suggest that student achievement gains should receive between 33 and 50 percent of the overall weight in order to ensure consistency and avoid a narrow focus that might promote strategic behavior (Cantrell & Kane, 2013). Although this leaves little guidance regarding the weighting of components for teachers without student achievement gains, several studies point to the inclusion of VAMs as an essential source of credibility for other component measures and therefore the evaluation system as a whole (Harris, 2011; Tennessee Department of Education, 2016). As noted previously, VAMs provide an important benchmark for other measures of teacher effectiveness. For example, if the correlation between VAMs and SBOs changes appreciably, this may point to a need for rater calibration, which would in turn improve the reliability of SBOs for all teachers (Harris, 2011; Tennessee Department of Education, 2016). In brief, the evidence available to date suggests that rigorous teacher evaluation systems would do well to include multiple measures of teacher effectiveness; among these measures, student achievement gains and classroom observations should be included as preponderant measures.

Assigning final ratings. After assigning weights to component measures to generate an overall index of teacher effectiveness, school systems must decide how many rating categories to assign to teachers and how these rating categories will be assigned to teachers. Rigorous teacher evaluation systems must provide rating systems that do more than simply distinguish unsatisfactory performance from satisfactory performance (Weisberg et al., 2009). In some sense, it is challenging to think of final ratings without considering the sanctions and rewards that might be linked to final ratings. For instance, if a school system establishes a four-level rating system (e.g., Ineffective, Developing, Effective, and Highly Effective), what consequences will distinguish Ineffective teachers from teachers rated Developing? What kind of recognition will communicate the benefit schools derive from teachers who were rated Highly Effective? What opportunities will be offered to help Effective teachers become Highly Effective teachers? In brief, it seems to serve little purpose to create labels that are not linked to policies that will further advance the ultimate goal of improving teacher effectiveness.

Linking Teacher Effectiveness with Supports, Sanctions and Rewards

As suggested by the theory of change in Figure 1, evaluations systems will be more effective if teacher ratings are linked to supports, rewards, and/or sanctions, as there are the mechanisms by which teacher quality is hypothesized to improve. In this section, we discuss these options.

Supports. One avenue through which a teacher evaluation system is theorized to improve student outcomes is by driving improvements in teacher practice. As a first step in this process, teacher evaluations must be communicated to teachers. We know from learning theory that when and how this communication occurs matters (Allen, Pianta, Gregory, Mikami, & Lun, 2011). In general, communicating results shortly following the evaluation component improves responsiveness, both because context is fresh and because it facilitates timely improvement. For example, in a system with SBOs that occur three times a year, it is beneficial for teachers to receive the results of each evaluation shortly following the evaluation. The ability to receive feedback and make adjustments during the school year is one of the advantages of SBOs.

It also matters how the communication occurs. There is increasing evidence that providing teachers with feedback, targeted professional development and/or coaching allows teachers to better understand how to translate results of their performance to improve teaching (Allen et al., 2011; Desimone & Garet, 2015; Kraft, Blazar, & Hogan, 2016). For instance, school systems could use teacher evaluation data to identify and prioritize the needs of low-performing teachers.[8]

Professional development. Professional development is intended to support teacher improvement, but few professional development programs that have been implemented at scale have been linked to improved student outcomes. Notwithstanding, there seems to be a logical connection between teacher evaluation and professional development: a rigorous teacher evaluation system should not only identify teachers in need of improvement but should also

[8] This is not to say that low-performing teachers should be the sole recipients of professional development. Indeed, high-performing teachers may very well be proactive in seeking out opportunities to continue honing their skills. One possibility would be to allow high-performing teachers more flexibility in choosing professional development activities that would enrich their instructional practices. On the other hand, low-performing teachers might receive intensive coaching and/or select from professional development opportunities designed to target their greatest areas of need, as determined using teacher evaluation data.

home in on specific instructional practices that need to be shored up. In this manner, the information gathered in the service of teacher evaluation can dovetail with efforts to support teacher improvement through the provision of targeted professional development opportunities.

Salary freeze. Most public school teachers in the United States are paid based on a strict salary schedule that provides for increases in salary based on years of experience teaching and degrees obtained, regardless of classroom performance. One possible sanction for low-performing teachers is to impose a salary hold, whereby a teacher does not receive the annual salary increase indicated on the school system's salary schedule. For instance, a teacher who receives an Ineffective rating in his or her first year of teaching would continue to be paid like a first-year teacher the following school year, rather than receiving the salary increase typically extended to a second-year teacher.

Revoke contract or tenure. Another type of sanction that may be brought against low-performing teachers affects the conditions of their teaching contract. Specifically, school systems may seek to dismiss low-performing teachers (i.e., revoke their teaching contract) or they might extend the pre-tenure probationary period for low-performing teachers. The latter option delays the provision of tenure for low-performing teachers until they demonstrate effectiveness. The vast majority of teachers in the U.S. receive tenure after two years of service, and it is typically very challenging to dismiss a tenured teacher. Extending a teacher's pre-tenure probationary period therefore delays the onset of contractual guarantees afforded to tenured teachers. For these reasons, dismissal of tenured teachers is rarely employed as an outcome in teacher evaluation systems and when it is would either require an unusual labor management contract or substantial negotiation.

Career advancement. School systems may also wish to implement policies to reward high performing teachers with opportunities for career advancement. For instance, high-performing teachers might be eligible for leadership roles or advanced training. Teachers could be promoted to the position of mentor teacher or instructional specialist or coach. High-performing teachers could also receive advanced training to prepare them for administrative roles (i.e. principal or assistant principal). These kinds of career advancement opportunity not only recognize high-performing teachers for their performance but can also serve to retain talent within the school system.

Financial rewards. Finally, school systems may also choose to offer high-performing teachers additional remuneration. Financial rewards can take the form of annual bonuses and/or permanent salary increases and can differ based on the nature of a teacher's assignment. This approach has been relatively common in various versions of teacher pay-for performance or merit pay. For example, recipients of Teacher Incentive Fund grants were required to provide financial incentives tied to some form of teacher evaluation. School systems may use financial rewards to target improvement among teachers in certain contexts. For example, some systems offer larger financial rewards to high-performing teachers who teach in tested grades/subjects and who serve a large share of low-income students. Ultimately, the school system must develop a consistent set of rules that determine eligibility for performance-based financial incentives.

Unintended consequences. In general, the intended goals of teacher evaluation are to provide teachers and school leaders with actionable information that can guide their improvement as well as with incentives that encourage those changes. Accountability raises concerns that while it may successfully encourage improvement in intended outcomes, such as teaching skills and student achievement, it may also cause some individuals to engage in inappropriate, or in some cases illegal, activities. These unintended outcomes raise important questions about the design and implementation of teacher evaluation. For example, pressure to perform may induce some teachers to game the observations, especially if they are announced, to present a more favorable view of their teaching than is typically the case. Principals may assign observation scores that don't reflect performance but are intended to encourage teachers. More troubling, teachers or administrators may manipulate student achievement scores to reduce sanctions or increase rewards. Indeed, there are examples of such activity associated with accountability in education.[9] This issue raises both philosophical and empirical questions. Should the use of incentives be constrained by the potential unethical or illegal acts that follow? Empirically, are their safeguards that mitigate the potential for unintended consequences? Do the positive effects of sanctions and rewards outweigh the cost of unintended consequences?

Teacher Evaluation in the United States

States and districts in the U.S. navigated the many decisions of redesigning teacher evaluation with many similarities but with some important differences. Their decisions were informed by a convergence of research and federal policy around 2009 that led many states to adopt systems with common features. As we describe in more detail below, differences arose in the way these systems were implemented which largely reflected local political and economic realities. Our synthesis draws heavily on an excellent review by Steinberg and Donaldson (2016). Steinberg and Donaldson examined primary sources from all 50 states, the 25 largest school districts in the United States, and Washington D.C. The authors then refine their sample to focus on school systems that made revisions to their teacher evaluation system in the last five years, thereby limiting their analysis to 46 states and 23 large districts, including Washington D.C. We complement Steinberg and Donaldson's review with a state level summary prepared by the National Council on Teacher Quality (NCTQ) and another paper (Kraft & Gilmour, 2017) that describes the distribution of teacher effectiveness in 24 states.

Measures of Teacher Effectiveness. School systems may choose to evaluate teachers using a variety of measures of teacher effectiveness. Here we describe how states resolved these competing measures in the design of their teacher evaluation systems by examining how frequently each measure is observed. Figure 2 provides a high-level summary of the prevalence of each component measure across 46 states and 23 large districts. As shown in the first set of bars in Figure 2, all states and large districts included in the analysis incorporate classroom observations in their teacher evaluation system. Test-based measures of teacher effectiveness (i.e., VAMs and/or SGPs) are the next most prevalent component. According to Steinberg and Donaldson, 80 percent of U.S. school systems incorporate at least one test-based measure of teacher effectiveness in their teacher evaluation system (2016, p. 347). This statistic is generally consistent with the figure reported by the NCTQ, in which 80 percent (N = 40) of states required evidence of student learning in their teacher evaluation system as of 2016 (see Figure 3).[10]

[9] Perhaps the most egregious example is the 2007-008 Atlanta public schools cheating scandal. Ten educators were convicted of a criminal conspiracy to alter students answers on state tests to improve school performance (Tagami, 2017).

[10] According to the NCTQ's review of state policy, "30 states require measures of student academic growth to be at least a significant factor within teacher evaluations; another 10 states require some student growth, and 11 states do not require any objective measures of student growth." (Walsh, Joseph, Ross, & Lubell, 2017, p. 5). NCTQ

Steinberg and Donaldson contrast the use of VAMs versus SGPs between states and large districts. As shown in Figure 2, large districts are more likely than states to rely on value-added measures (VAMs), whereas states tend to favor student growth percentiles (SGPs) over VAMs. Specifically, of 46 states with newly implemented teacher evaluation systems, close to half (N = 22) use only SGPs, 20 percent (N = 9) use only VAMs, and another 10 percent (N = 5) use both SGPs and VAMs. On the other hand, of 23 large districts with newly implemented teacher evaluation systems, more than half (N = 12) use only VAMs, one quarter (N = 6) use only SGPs, and 2 districts (less than 10 percent) use both VAMs and SGPs. (Steinberg & Donaldson, 2016, p. 347)

Although several states and large districts also incorporate SLOs in their teacher evaluation system, states are more likely than large districts to do so (52 percent versus 39 percent). On the other hand, large districts are more likely than states to include a measure of teacher professional conduct in their teacher evaluation system (52 percent versus 22 percent). The last two component measures (SVA and student surveys) are not widely observed in either states or large districts.

In brief, evidence suggests that student learning outcomes are an increasingly common feature of teacher evaluation systems in the U.S., with a majority of school systems incorporating student learning outcomes as a complement to classroom observations. In the following section, we will turn our attention to the manner in which school systems are combining measures of teacher effectiveness to assign teacher effectiveness ratings.

Combining Measures of Teacher Effectiveness

Steinberg and Donaldson's analysis reveals that classroom observations are not only the most common component measure of teacher evaluation systems in the U.S. but also the most heavily weighted component. On average, classroom observations account for at least half of a teacher's final rating (see Table 1 and Table 2). Notably, classroom observations are the preponderant measure regardless of analytic weights and the level of aggregation (i.e., state or large district). This provides clear evidence that teacher evaluation systems in the U.S. continue to rely heavily on classroom observations.

Tables 1 and 2 also demonstrate that test-based measures of teacher effectiveness (i.e., VAMs and/or SGPs) are often the next largest component for teachers assigned to tested grades and subjects, with VAMs accounting for roughly 15 - 20 percent of tested teachers' final ratings. Analyses aggregated at the state level also identify SGPs as a relatively salient measure for tested teachers, for whom SGPs contribute 10 - 12 percent of the final rating (see Table 1 Panel B and Table 2 Panel A). Steinberg and Donaldson privilege the average component weights among typical teachers (rather than tested teachers), most likely because measures based on standardized tests of teacher effectiveness cannot be generated for more than 30 percent of teachers (Donaldson & Papay, 2015). Thus, for the majority of teachers, test-based measures of teacher effectiveness do not tend to account for a substantial portion of teachers' final ratings. In fact, "no more than 13 percent of the evaluation rating of a typical teacher nationwide [depends] on VAMs and/or SGPs. For the typical teacher teaching in one of the nation's largest school

identifies student academic growth as a significant factor if it accounts for at least 30 percent of a teacher's summative rating.

districts, at most 10 percent of their evaluation rating will reflect student test score performance (see Table 2, Panel B)" (Steinberg & Donaldson, 2016, p. 349).

Despite significant debate regarding the inclusion of test-based measures of teacher effectiveness, student achievement measures play only a limited role in the evaluation of the typical teacher in the U.S. Indeed, as the NCTQ notes, of 30 states where student growth accounts for at least 30 percent of teachers' summative rating, only two states require teachers to meet student growth goals in order to receive a proficient rating (Walsh, Joseph, Ross, & Lubell, 2017, p. 6). Although teachers shouldn't necessarily have to demonstrate evidence of student growth in order to receive a proficient rating, the evidence available to date suggests that assigning 33 - 50 percent of the weight to student learning outcomes produces a teacher evaluation system that is more balanced and reliable (Cantrell & Kane, 2013). Steinberg and Donaldson's analysis suggests most school systems in the U.S. have not fully adopted this guiding principal.

Beyond classroom observations and test-based measures of teacher effectiveness, the one remaining component that at times receives a preponderance of weight is SLOs. In particular, SLOs often seem to be used as a student learning outcome for teachers who are not assigned to tested grades/subjects. Thus, for teachers in non-tested grades and subjects, SLOs may account for up to 25 percent of the teacher's final rating.

Final Ratings

As the NCTQ notes, most states (n = 38) require at least four rating categories, but the specific labels for rating categories vary from state to state (Walsh et al., 2017, p. 2). While the specific labels vary across states, the four rating categories that are commonly observed tend to identify teachers as Highly Effective, Effective, Developing and Ineffective. Notwithstanding, three of the 24 states included in Kraft and Gilmour's study opted to use a rating scale with five categories. The main feature distinguishing teacher evaluation systems with five rating categories from teacher evaluation systems with four rating categories appears to be the addition of an "Exemplary" rating, a distinction that few teachers attain (see Figure 4).

Despite the creation of multiple rating categories, Kraft and Gilmour's analysis points to limited differentiation in final ratings. For instance, only four of 24 states in their analytic sample rated more than 10 percent of teachers Ineffective (see Figure 5). Stated differently, the vast majority (i.e., generally more than 90 percent) of teachers in these 24 states receive a final rating of Effective or higher. New Mexico is a stark outlier in Figure 5, having rated almost 30 percent of teachers Ineffective in 2013-14. Other studies have also highlighted New Mexico for its high degree of differentiation (see Table 3), but it is not entirely clear what is driving this outcome. One hypothesis is that New Mexico experiences greater differentiation because the evaluation system assigns less weight to classroom observations than other school systems (Walsh et al., 2017, p. 11). Indeed, a forthcoming simulation study suggests that placing less weight on criterion-referenced measures, such as classroom observations, may indeed lead to greater differentiation of teacher effectiveness (Steinberg & Kraft, 2017).

Linking Teacher Effectiveness with Supports, Sanctions and Rewards

An important decision in designing evaluation systems is whether they attach meaningful stakes to the evaluation outcomes. Financial rewards or the potential of dismissal are intended to increase teachers' responses to evaluation outcomes. Table 4 illustrates the prevalence of different sanctions and rewards associated with teacher evaluation systems in the U.S. Most school systems adopt a developmental stance to teacher evaluation, as evidenced by the large proportion of states and large districts that link evaluation outcomes to professional development. States are more likely than large districts to attach high-stakes contractual consequences (i.e. termination and/or tenure revocation) to teacher evaluation outcomes. Although teacher evaluation systems may allow for contractual consequences, it is less clear how often these policies result in termination of tenured teachers. Kraft and Gilmour (2017) surveyed and interviewed school leaders in one large (anonymous) district and found that principals avoided assigning teachers the lowest rating due to the bureaucratic implications of "evaluating out" a low-performing teacher.

At the other end of the performance distribution, 20 percent of school systems in the U.S. offer financial incentives to high-performing teachers. Below we describe how seven school systems address the issue of financial incentives. Finally, given that only a handful of school systems have linked their teacher evaluation systems with career advancement opportunities (prevalence not shown in Table 4), we explore how specific school systems have managed to merge their teacher evaluation system with a career ladder.

Teacher Evaluation Design Summary

Our summary of teacher evaluation in the U.S. post 2009 reveals four important features. First, virtually every system allows for meaning differentiation in performance. In most states teacher ratings include at least four performance categories. This is in sharp contrast to the systems in place in nearly all states prior to 2009. This potential for differentiation addresses important concerns regarding the ability to recognize differences in teacher performance that nearly everyone agreed existed in most schools. More importantly, differentiation in measured performance allowed for improvement through several mechanisms.

Second, virtually every system included some version of SBOs. This is unsurprising given that observations have been a part of the informal evaluation process practiced in most schools for many decades. Importantly, SBOs have the potential to make this process much more rigorous, allowing it to successfully differentiate among teachers. This is crucial to the effectiveness and credibility of a formal evaluation system.

Third, although contentious, many states and districts included in their evaluation systems some rigorous measure of teacher performance based on standardized student achievement outcomes—either VAMs or SGPs. While these measures apply to only a fraction of teachers, they may be important to both the face validity of the system but also to maintaining its rigor. VAMs and SGPs measure an important outcome of teaching. SBOs are inputs. As such many observers believe the inclusion of VAMs or SGPs aligns performance to desired policy outcomes. Perhaps as importantly, VAMs and SGPs can be loosely employed to calibrate other measures, such as SBOs, when they diverge too dramatically.

Finally, despite the requisite tools, we find that few states meaningfully differentiate teacher performance. In most states, nearly all teachers are judged to fall in one of the top two performance rating categories. This may represent an improvement relative to the pre-2009 systems where nearly everyone was rated Satisfactory depending on how performance differences are viewed and treated. For example, if Effective teachers were provided with incentives and supports to become Highly Effective, this would create an environment conducive to teacher development. If, however, being labeled Effective or Highly Effective was viewed broadly as satisfactory, few teachers may feel the need to improve. Ideally, more dispersion would be employed, as most research suggests such dispersion exists in practice.

Any summary of teacher evaluation in the U.S. is by necessity broad and general. Our summary of the teacher evaluation in the U.S. led us to explore some of these themes in more detail. To do so we purposely chose seven examples of teacher evaluation systems to better understand how these systems work in practice.

U.S. Teacher Evaluation in Practice: Seven Illustrative Examples

The rapid redesign of teacher evaluation systems in most states and school districts beginning about 2010 provides an opportunity to better understand how different design and implementation decisions influence the functioning of teacher evaluation. We employ this variation to pursue three goals: a) to explore in more detail the choices that states and school districts made in the design and implementation of teacher evaluation systems, b) to identify some of the competing forces that shaped the differential development of teacher evaluation across states and school districts, and c) to identify some of the promising practices emerging from the operation of these systems in their first few years. Each of these goals is a substantial task taken individually. Because there is so little research on the forces that determine the design of teacher evaluation systems and their effects on teacher and student outcomes, our conclusions are posed as a series of conjectures and promising practices.

We see this summary as a first approximation at identifying design and implementation differences, forces that may have led to these differences and the effects of these differences. There is far too little research that explores the details of these systems to draw any firm conclusions. Our summary is premised on the belief that policymakers who are developing or revising their evaluation systems benefit from some evidence, even if it is limited and preliminary, rather than none. This is an area of research that is moving rapidly. We expect a summary of this sort to be much more clearly defined even three years from now.

The broad parameters of most teacher evaluation systems are defined by states. Most states afford school districts discretion in design details and in implementation. As a result, teacher evaluation systems have the potential to differ on dozens of dimensions. However, the variation in system designs is remarkably small. For example, most states have designed systems that evaluate teachers across multiple measures of teacher performance but allocate most of the weight to a standards-based observation protocol and at least one measure of student academic improvement. These systems classify teachers into one of four, or in some cases five, performance categories. Teachers rated in the top two categories are often identified as “highly effective” and “effective,” or similar names that connote proficiency. Teachers in the bottom two categories are often identified as “ineffective” and “developing,” or titles indicating improvement is expected.

Despite the enormous time and expense invested in revising teacher evaluation in most states and districts, these systems still classify the vast majority of teachers with labels connoting proficiency. In most states over 90 percent of all teachers fall into one of the top two (or three in the case of five category systems) performance categories. In many states 95 percent or more teachers are so classified (Kraft and Gilmour, 2017). If nearly all teachers are being given the signal that their performance is fine, there may be little incentive for principals or teachers to invest in the time-consuming process of improvement. This has led some observers to conclude that reforms to the teacher evaluation process have been a failure (Dynarski, 2016; Walsh et al., 2017).

When differences in the distribution of teachers across performance categories arise, they appear to do so as a result of differences in the nuance of design or in implementation rather than differences in more fundamental design approaches. For example, some states prevent teachers who score less than effective on the student achievement growth component of the teacher evaluation system to receive an overall effective rating, while others make that outcome quite likely (Walsh et al., 2017). More frequently, differences in the implementation of measures may lead to differences in outcomes. For example, some districts appear to invest heavily in the validity and reliability of standards-based observations, while others don't. Our examination of seven teacher evaluation systems leads us to conclude that more rigorous implementation of observations is correlated with greater differentiation of teacher evaluation ratings.[11]

We illustrate some of the differences among teacher evaluation systems by examining seven systems in more detail. We chose these systems to illustrate what we believe are key dimensions of design and implementation. Our choices were guided by five criteria around which we sought variation:

- The use of test-based measures of teacher effectiveness as an important component of the system;
- The use of high stakes associated with evaluation outcomes;
- The process that led to the design of the system;
- The rigor by which standards-based observations of teachers were designed and implemented; and
- The role of coaching and professional development.

Employing these criteria led us to examine the teacher evaluation systems in four districts—Denver Public Schools (DPS), District of Columbia Public Schools (DCPS), New Haven Public Schools (NHPS) and New York City Public Schools (NYCPS)—and three states—New Mexico, North Carolina, and Tennessee —in more detail. Some version of value-added is employed in New Mexico, North Carolina, Tennessee and the District of Columbia, but not in Denver, New Haven or New York City (NYC). Denver and New Haven arrived at their systems through a collaborative process of negotiation between the teachers' union and the school district; DCPS, NYC, New Mexico, North Carolina, and Tennessee had their systems completely or largely imposed by the district or state. While each system we studied connected teacher performance to stakes for teachers, in all but DCPS these performance-based consequences were not meaningfully differentiated across teachers.

Methodology

We examine the design of teacher evaluation in districts and states by employing publicly available information found on most state and district websites. To facilitate our comparisons, we designed a rubric that lists the key features of teacher evaluation system (Appendix A). For each of our seven evaluation systems we completed this rubric. In general, we were able to find most of the information from online sources. These sources are available in Appendix B.

[11] We suspect that a deeper analysis of the observation rubrics would yield other interesting differences.

Results

There are many aspects of these systems that may be of interest. We focus our discussion on the five elements described above: use of test-based measures of teacher effectiveness, use of stakes, implementation process, rigor of SBOs, and role of coaching and professional development.

Design: Use of test-based measures of teacher effectiveness. Four of the seven systems we focus on estimate value-added: DCPS, NC, TN, and NM. (See Figure 6 for a summary of the composition of teacher evaluation measures in our seven systems.) Despite the small number of systems, we nonetheless observe interesting variation across these four systems. For instance, DCPS estimates value-added for 4th-10th grade math and English Language Arts (ELA) teachers and includes student demographics in their estimation of value-added. On the other hand, NC, TN and NM estimate value-added for a wider-range of subjects[12] and grades[13] and do not include student demographics in their estimation of value-added. The variation observed across these four school systems serves as a reminder of the different approaches to estimating value-added. Ultimately, the decision of how to estimate value-added must be determined in consultation with statisticians and psychometricians who have a nuanced understanding of value-added measurement approaches and the standardized assessments available for analysis. Generally speaking, to the extent that school systems have reliable and rigorous assessments of student learning across a broad range of grades and subjects, value-added can in turn be estimated for a broader range of teachers.

Another key difference across these four systems is the inclusion of school or district-level value-added estimates, a practice that is currently observed only in TN.[14] For TN teachers in non-tested grades and subjects, group value-added accounts for 25 percent of the overall evaluation. This is a surprisingly large weight to assign to a measure over which teachers have little individual control. Paraphrasing Harris (2011), school systems must align measures of effectiveness with the type of accountability they are pursuing. Group measures have the benefit of reducing year to year noise and potentially encouraging collaboration but undoubtedly mis-measure any given teacher's performance. In this vein, 25 percent appears high and to under value the potential disadvantages of group measures for the purpose of teacher evaluation systems intended to discern individual performance.

[12] Depending on the assessment practices of a given district, TN can estimate value-added for 2nd and 3rd grade teachers as well as high-school teachers whose students take the following end-of-course (EOC) exams: Algebra I and II; English I, II, and III; Biology; Chemistry; Geometry; Integrated Math I, II, and III; U.S. History (SAS EVAAS, 2017). The state also estimates value-added using 3rd-8th grade science and social studies assessments (SAS EVAAS, 2015). Finally, the state also uses the ACT (11th grade), PLAN (10th grade), and EXPLORE (8th grade) assessment in English, math, reading, and science to estimate value-added for a broader range of teachers (Klafehn, 2015). NC can estimate value-added using the following assessments: English language arts and math in grades four through eight, science in grades five and eight, Math I, English II, and Biology; the ACT, SAT, PSAT, AP; many of the Career and Technical Education Post-Assessments (CTE), and many of the NC Final Exams (formerly known as Common Exams) (https://ncdpi.sas.com/welcome.html?as=b&aj=b).

[13] Depending on the assessment practices of a given district, NM can estimate value-added for 4th-12th grade math teachers and K-12 ELA teachers (New Mexico Public Education Department, 2017, pp. 15–16).

[14] The first version of DCPS' IMPACT evaluation system included a school value-added measure. The component accounted for 5 percent of a teacher's overall evaluation. Prior to the 2016-17 school year, NC also used to incorporate school-wide value-added in its teacher evaluation system (Walsh et al., 2017, p. 53). In fact, for NC teachers in non-tested grades and subjects, school-wide value-added accounted for 100% of their student-growth measure (Public Schools of North Carolina, 2013, p. 8).

With the exception of TN's use of group value-added, there is a great deal of similarity with respect to the weight each system assigns to value-added. In fact, three of these systems (i.e., DCPS, TN, NM) use individual value-added (when available) to account for 35-50 percent of teachers' overall evaluation. This approach aligns well with best practices from the MET study, which recommends using test-based measures of student achievement to account for one-third to one-half of the overall evaluation (Cantrell & Kane, 2013). Additionally, NM varies the amount of weight assigned to value-added based on the number of years of assessment data employed in estimation; value-added accounts for 35 percent of the overall evaluation for teachers who only have 1-2 years of student assessment data, whereas value-added accounts for 50 percent of the overall evaluation for teachers who have at least 3 years of student assessment data. In this manner, NM assigns more weight to value-added when data allow for more accurate estimation.

Although NC still estimates value-added, the state no longer incorporates student growth in its teacher evaluation system; as of the 2016-17 school year, value-added is used only as a tool for targeting professional development and for school, district, and state reporting purposes (Walsh et al., 2017, p. 53). However, the six other systems we examined were designed such that measures of student learning accounted for roughly half of teachers' overall evaluation; the remaining half of the overall evaluation can generally be described as measures of teacher practice (e.g., standards-based observations, parent or student surveys, measures of teacher professionalism and/or commitment to the school community). In the absence of value-added, the remaining systems we examined (i.e., Denver, NYC, New Haven) opted to measure student learning using student growth percentiles (SGPs; observed in Denver and NYC) and/or student learning objectives (SLOs; observed in Denver, DCPS, New Haven, NYC, and TN). Whereas SGPs are a rigorous, test-based measure of student learning, SLOs offer much flexibility but little standardization. In other words, only SGPs can be considered a viable alternative to value-added measures (VAMs). Research has found that when there is nonrandom sorting of teachers to students, as is often thought to occur, that SGP estimates are more biased than VAM estimates (Guarino, Reckase, Stacy, & Wooldridge, 2015).

Both Denver and NYC include SGPs in their teacher evaluation system, but they tend to assign less weight to this measure of student learning: Denver uses SGPs to account for only 10 percent of the overall rating for teachers in tested grades and subjects, and NYC allows school-based committees to select individual and/or group SGPs and/or SLOs to account for 50 percent a teacher's overall rating. Teachers assigned to tested grades and subjects in NYC could ostensibly have half of their overall rating determined by SGPs, but because the measures vary from school to school, it is very difficult to ascertain how much weight SGPs receive on average. Moreover, as the state is transitioning to new learning standards, NYC is currently in the midst of a three-year moratorium on the use of 3rd-8th grade math and ELA assessments for teacher evaluation, thereby limiting the potential use of SGPs until the moratorium ends.

Design: Use of stakes. Although there is some evidence to suggest that evaluation alone can improve student outcomes (Taylor & Tyler, 2012), theory suggest that tying teacher evaluation to rewards and/or sanctions could intensify the effects of teacher evaluation (Lazear, 1995). However, there is also evidence to suggest that performance-based consequences might decrease intrinsic motivation (Deci, 1971; Deci, Koestner, & Ryan, 2001) or encourage strategic behavior, i.e., cheating. On the latter point, teacher evaluation best practices (e.g., limiting the amount of weight given to any single measure) are intended to mitigate unintended consequences. Nonetheless, when confronted with a high-stakes teacher evaluation system, teachers may shift their priorities to protect their continued employment or to increase their chance of receiving a financial incentive.

We observed an array in the use of sanctions across the seven systems we examined. (See Table 5 for a summary of the use of sanctions and rewards.) For instance, low-performing teachers in NC may be demoted or dismissed, and low-performing teachers in DCPS and New Haven are subject to salary freezes as well as dismissal.[15] Rather than dismissing low-performing teachers, school systems may opt to extend the pre-tenure probationary period for low-performing teachers. This practice was adopted in Denver and TN.

Although New Haven and DCPS imposed similar sanctions, it is worth noting variation in how these policies were implemented. Specifically, New Haven teachers subject to termination are described as having "accepted resignation" (Bailey, 2011, 2012, 2014), suggesting that low-performing teachers can leave New Haven as though it were voluntary. Additionally, low-performing teachers in New Haven can avoid a salary freeze by completing five professional development sessions over the summer. In other words, New Haven appears to have adopted a more forgiving approach to salary freezes and dismissals.

NYC and NM do not appear to have adopted sanctions for low-performing teachers. Although low-performing teachers in NYC are placed on an improvement plan, this is explicitly non-disciplinary. Similarly, NM intended to place low-performing teachers on improvement plans, but the local teachers union won an injunction in 2015 that blocks any "consequential actions" based on teacher evaluations (Burgess, 2017b).

There was also variation in the use of rewards for high-performing teachers. We found that career advancement opportunities were a popular option in the seven systems we examined. Denver, DCPS, NYC, New Haven, and NC all implemented policies that allow high-performing teachers to apply for leadership positions. Although there is no guarantee that applicants will receive a leadership position, these roles are typically associated with a stipend, thereby offering high-performing teachers a financial incentive of sorts. DCPS and Denver are the only systems that guaranteed performance-based financial incentives for all high-performing teachers. In DCPS these rewards can be quite substantial for repeated evaluations as highly effective, allowing teachers to earn more than $130,000 annually. Although NM asks districts to submit annual requests for funding to award performance-based financial incentives, the availability of these funds varies from district-to-district and from year-to-year. In brief, of the seven systems we examined, DCPS and Denver appear to have the most transparent performance-based financial incentives.

Implementation process. The systems we examined also differed in their general approach to design and implementation, as well as the level of centralization of their teacher evaluation system. (See Table 6 for a summary of the implementation process.) Denver and New Haven are examples of two districts that engaged early and extensively with a variety of stakeholders, soliciting feedback intended to refine the design of the evaluation system and build

[15] Until the current school year (2017-18), Denver also froze the salaries of low-performing teachers, but this provision was removed from the most recently negotiated teacher contract.

positive momentum around the planned revisions. Although both Denver and New Haven received a lot of positive press for their collaborative approach to teacher evaluation design and implementation (Jerald, 2013; “The New Haven Model for Teacher Evaluations,” 2010), both district’s ultimately adopted evaluation systems that rely heavily on some of the more subjective measures of teacher effectiveness (i.e., standards-based observations, SBOs, and student learning objectives, SLOs). While we do not intend to suggest that there is a direct relationship between collaboration and evaluation design, it is not entirely surprising that evaluation systems designed in collaboration with teachers and their unions would privilege SBOs and SLOs; these are, after all, measures that teachers may believe they have more control over (as compared with more rigorous, test-based measures of teacher effectiveness like value-added and student growth percentiles).

Although NYC does not appear to have been as proactive about creating dialogue with stakeholders as Denver and New Haven, the district did spend a couple of years piloting certain aspects of their teacher evaluation system before the first full year of implementation. Nonetheless, the local teacher’s union was not entirely supportive of the proposed evaluation plan. In fact, because the district could not come to an agreement with the local teacher’s union, NYC had to return more than half of a $46 million Race to the Top grant the district was awarded by the federal government in 2010 (McCann, 2012; Zubrzycki, 2012).[16] In general, policy-makers in New York state and NYC have faced considerable pushback from teacher unions throughout the teacher evaluation design and implementation process. As a result of these ongoing negotiations, policy makers had to make several compromises and ultimately implemented a teacher evaluation system that seems to privilege flexibility over rigor. While the state established the broad legal parameters for teacher evaluation, districts were given a great deal of discretion to design the specific features of the teacher evaluation system. Additionally, NYC allows school-based committees to determine the measures of student learning (MOSL) that will account for half of teachers’ overall evaluation. This localized decision-making process was likely intended to allow schools to tailor teacher evaluation for their context, but the end result is little standardization and a lack of transparency across locations.

In contrast with the district-level discretion afforded to Denver, New Haven and NYC by their respective state governing bodies, NC, NM, and TN were more actively involved in the design of teacher evaluation. That is, rather than creating a broad legal structure within which districts would design their own teacher evaluation systems, these three states designed complete teacher evaluation system allowing districts some discretion on select features of the evaluation system. For instance, districts in NM can choose one of two observation plans (i.e., one formal observation versus three formal observation), but the state determined which components would be used to evaluate teachers and how these components would be combined to generate final ratings. Similarly, TN allows districts to include student perception surveys as part of teacher evaluation, but districts must submit a request to the state and the state defines how student

[16] To be clear, NYC was not the only district that had to return part of a Race to the Top grant; Milwaukee and Chicago were also unable to secure the support needed from the local teacher union. In response to these unforeseen challenges, the U.S. Department of Education revised the Race to the Top application process for future cohorts, “requiring applicants to produce evidence of teacher and principal collaboration from the start,” rather than allowing a “yearlong period in which districts could obtain stakeholder buy-in” (Zubrzycki, 2012).

surveys will be used to generate final ratings.[17] Additionally, both TN and NC allow local districts to submit alternative evaluation plans but requires that these alternative evaluation plans align with state mandates. Ultimately, few districts appear to have taken advantage of this flexibility. Indeed, more than 80% of districts in TN adopted the state-designed teacher evaluation system (Ehlert, Pepper, Parsons, Burns, & Springer, 2013, p. 13).[18]

DCPS is a rather unique context in which to consider design and implementation. Most districts are bound solely by state law. DCPS does have oversight by a state agency (Office of State Superintendent of Education) but also is subject to federal congressional oversight unlike any other state or district. Furthermore, at the time DCPS was designing its teacher evaluation system, the district was not contractually obligated to negotiate with the teacher union. As a result, DCPS was able to design and implement teacher evaluation with very few legal or contractual limitations. This may help explain the fairly ambitious nature of teacher evaluation in DCPS. For instance, DCPS initially allowed value-added to account for 50 percent of a teacher's final rating.[19] As we discuss elsewhere in this report, the district also required five observations of all teachers; three by a school-based leader and two by an expert, outside evaluator. Perhaps most controversially, the district created a highly consequential system: high-performing teachers are eligible for some of the largest financial incentives available in the nation, and low-performing teachers are dismissed.

Despite DCPS's unique context, a communality with TN emerged in our analysis: although neither school system appears to have been proactive about engaging with stakeholders prior to implementation,[20] both DCPS and TN have continued to fine-tune teacher evaluation in response to feedback and analysis. On the other hand, NM seems to have done little to engage stakeholders either before or after implementation.[21] Indeed, the state continues to confront legal challenges brought by the NM teacher union.

In the end, there is no right way to go about designing and implementing a teacher evaluation system. That being said, the manner in which teacher evaluation systems evolve, from initial design to on-the-ground implementation, has important implications for the integrity of the system. Indeed, the variation observed across the seven school systems in our analysis speaks to the value of having a nuanced understanding of the key stakeholders involved in the process, as well as the potential implications of involving stakeholders at critical points along the way.

Implementation: Rigor of standards-based observations. Standards-based observations (SBOs) are a predominant feature in virtually all teacher evaluation systems (Steinberg & Donaldson, 2016) but the rigor with which SBOs are implemented varies greatly across school systems. At a minimum, rigorous implementation of SBOs requires careful rubric construction,

[17] In AY1314, 19 districts were approved to use student perception surveys (SPS) as 5 percent of teacher evaluation. The state established the weight assigned to SPS and determined that SPS would reduce the weight assigned to standards-based observations (SBOs).

[18] We were unable to obtain a comparable statistic for NC.

[19] The weight assigned to value-added was later reduced to 35 percent.

[20] We did not find any resources documenting TN's efforts to increase buy-in prior to implementation. In one review of DCPS' implementation process, the author notes, "The IMPACT team held 50 to 60 focus groups across the district by job category to gather input from teachers and other school-based staff to inform the design of IMPACT" (Curtis, 2011, p. 12).

[21] Some news reports reference the New Mexico Teacher Leader Network, a group of 15 educators who recommended revisions to the evaluation system (Burgess, 2017a; Hutchinson, 2017), but there is little additional documentation of state efforts to engage with stakeholders.

thorough observer training, protocols to ensure observer reliability, and thoughtful planning regarding the number and type (i.e. formal versus informal) of observations that can truly support instructional growth for teachers at different points in their career trajectory. School systems must also decide who will conduct observations, what arrangements must be made to enable observations (e.g., ensuring school administrators have enough time to conduct the required observations), and how and when teachers will receive feedback from each observation.

However, several of the systems we examined had limited time prior to implementation to address these issues in a comprehensive manner. (See Table 7 for a summary of the rigor with which systems implemented SBOs.) Specifically, DCPS, NYC, NC, NM, and TN seem to have had about a year to design and implement their teacher evaluation system (including rubric development and observer training, among many other logistical considerations). Perhaps owing to this somewhat compressed timeline, observers only received a few days of training that focused almost exclusive on obtaining certification. Although observers were all required to demonstrate proficient use of the observation rubric, training does not appear to have been sufficiently comprehensive. Indeed, one report from TN notes that, in retrospect, the scope of observer training may have been too narrow; in addition to providing reliable observation scores, observers must learn how to provide meaningful feedback and must also be prepared to help explain how the different components of the evaluation system come together (Reform Support Network, 2012).

In addition to initial training and certification, school districts must develop protocols to ensure rater reliability. While all seven systems we examined require annual observer certification, none appear to engage raters in regularly-schedule calibration and reliability exercises. For instance, Denver only mentions that raters will have at least once chance each year to norm as a school leadership team through Instructional Leadership Team (ILT) calibration sessions offered by the district (Denver Public Schools, 2016, p. 4). A report describing DCPS' implementation process notes that "part of each of the monthly, full day, Principals Academy sessions focused on [the observation rubric]," (Curtis, 2011, p. 10). Similarly, NYC requires that all evaluators receive support from Teacher Development and Evaluation Coaches (TDECs). However, none of these systems appear to require raters to demonstrate accurate scoring on a regular basis. As a result, observations may lack the desired consistency over time and across teachers to accurate identify teachers' strengths and weaknesses.

While there seems to be a shortage of formal calibration/reliability training across all seven systems we examined, TN uses teacher evaluation data to provided targeted evaluator coaching. Specifically, the state examines the relationship between teacher value-added scores and teacher observation scores.[22] Evaluation coaches were asked to support schools with a high percentage of teachers who had more than a 2-point difference between their value-added score and overall observation score.[23] As a result of this process, seven coaches supported 58 schools

[22] An additional report prepared by Tennessee's Department of Education identifies raters as "non-discriminating" if the rater scores more than 90 percent of the observation indicators within adjacent levels on the district's 5-point rubric (i.e., 90 percent of indicators were scored 4 and 5). Although the report recommends evaluator coaching as a means of improving the evaluation process, it is unclear whether the state uses this metric to provide targeted evaluator coaching. (Pratt, 2014)

[23] "For example, if a large number of teachers in a school had an individual value-added score of 1 and an observation average of 4 or higher, that school would have been eligible for state support." (Tennessee Department of Education, 2015, p. 27)

in 33 districts during the 2013-14 school year (Tennessee Department of Education, 2015, p. 27). Although TN does not appear to require raters to demonstrate accurate scoring on a regular basis, this data-driven approach to targeted evaluator coaching stands in contrast with processes observed in the other school systems we examined (e.g., the non-discretionary use of TDECs in NYC). In other words, while all seven systems we examined could do more to strengthen rater calibration/reliability practices, TN's targeted evaluator coaching model seems like an effective strategy to leverage evaluation data and limited resources.

New Haven is another example of a system that has strategically allocated limited resources to strengthen SBOs, as evidenced by the district's targeted use of external observers. Rather than using external observers in all schools, New Haven requires external observers to validate observations for teachers at the tails of the performance distribution. That is, external observers must confirm the scores of high-performing teachers who may be eligible for leadership opportunities as well as the scores of low-performing teachers who may be subject to dismissal. In this manner, the district guards against unobservable factors that might lead principals to assign certain teachers inaccurate observation scores (e.g., we can imagine that principals might assign higher scores to teachers with whom they have developed positive professional relationships).

In contrast with New Haven's targeted use of external observers, DCPS used to require external observers to observe all teachers twice annually. However, DCPS eliminated the external observer role at the end of the 2015-16 school year, likely due in large part to the cost of employing a cadre of external observers. The elimination of the external observer role in DCPS is unfortunate, as this was one feature of the system that probably boosted the rigor of SBOs. As noted in a MET Project publication, the use of external observers can help guard against in-school bias (Cantrell & Kane, 2013). The lead authors of the MET Project publication go on to suggest that schools might use external observers for a sample of teachers, much more in keeping with the practice employed in New Haven rather than DCPS. Ultimately, across the seven systems we examined, only New Haven appears to be employing external observers in a manner than supports rigorous SBOs.

We close this section by noting the number of observations teachers typically receive in each of the seven systems we examined. With the exception of Denver, the districts we examined required novice and low-performing teachers to receive more observations than high-performing teachers. In this manner, most systems are providing more frequent feedback to teachers who need more support. Nonetheless, considering that classroom observations are a critical avenue through which teacher evaluation is theorized to improve teacher instructional practice, the number of formal observations required in each district is somewhat low. For instance, Denver requires one only formal observation and one informal observation. DCPS and NM each require a maximum of three formal observations. NYC's most intensive evaluation plan requires only one formal observation and three informal observations; New Haven's most intensive evaluation plan requires two formal observations and three informal observations; NC's most intensive evaluation plan requires three formal observations and one peer observation; and TN's most intensive evaluation plan requires six observations each year. Thus, with the exception of TN, low-performing and novice teachers receive 1-3 formal observations each year. Although there is little guidance regarding the ideal number of observations, school systems must consider whether they are providing teachers with enough feedback to support meaningful improvement in instructional practices.

Implementation: Role of coaching and professional development. As noted in the Theory of Change (Figure 1), one avenue through which a teacher evaluation system is theorized to improve student outcomes is by driving improvements in teacher practice. Indeed, all of the systems we examined identified teacher growth as one of the overarching goals of their evaluation program. To achieve this goal, teacher evaluation could be used to provide targeted professional development (PD), thereby amplifying any improvements that may stem from observation alone. For instance, school systems could use teacher evaluation data to identify and prioritize the needs of low-performing teachers.[24] While this practice seems to hold great promise as a means of driving improvements in teacher effectiveness, of the seven systems we examined only NC and New Haven have clearly articulated and comprehensive PD plans that dovetail with their teacher evaluation plans. (See Table 8 for a summary of the use of professional development.)

There are three dominant features of New Haven's PD plan. First, all teachers establish annual goals for professional growth and, working with an administrator, develop a personalized PD plan to realize those goal. Second, teachers may be placed on a "Structured Support Plan" at any point during the school year to address a specific area of need. Third, low-performing teachers are flagged for more comprehensive support in the form of a "Developmental Plan" or an "Intensive Improvement Plan." This PD approach uses information gathered through teacher evaluation to provide targeted and personalized support to teachers at different points in their career trajectories.

Similar to New Haven, NC requires all teachers to create a growth plan at the end of each school year, but the system is tiered according to teacher performance. Specifically, high-performing teachers create Individual Growth Plans and implement these plans with little oversight. On the other hand, low-performing teachers must create a Monitored Growth Plan or a Directed Growth Plan. As these names suggest, Directed Growth Plans are required for the lowest-performing teachers and are more prescriptive than Monitored Growth Plans. In this manner, NC utilizes its teacher evaluation system to provide more focused supports for low-performing teachers.

Although NYC, TN, and NM also use their teacher evaluation to tailor PD, their approach is less comprehensive than NC's and New Haven's. Specifically, NYC and NM require professional growth plans for low-performing teachers, but it is not clear how these two school systems support teachers at other points in the effectiveness distribution. In collaboration with a team of researchers, TN conducted a randomized pilot study of a school-based peer-collaboration model. For the study, 7 elementary schools and 7 middle schools in one mid-sized school district were randomly assigned to treatment and control conditions. Using indicator-level classroom observation data from 2012-13, low-performing teachers in treatment schools were partnered with high-performing teachers at the same school and partners were encouraged to work together throughout the 2013-14 school year. Initial results from the pilot study suggest that student achievement increased as a result of peer-collaboration, as did teacher performance (Papay,

[24] This is not to say that low-performing teachers should be the sole recipients of professional development. Indeed, high-performing teachers may very well be proactive in seeking out opportunities to continue honing their craft. One possibility would be to allow high-performing teachers more flexibility in choosing professional development activities that would enrich their instructional practices. On the other hand, low-performing teachers might receive intensive coaching and/or select from professional development opportunities designed to target their greatest areas of need, as determined using teacher evaluation data.

Taylor, Tyler, & Laski, 2016).[25] Given these initial findings, TN intends to scale-up the program over the coming years. While the results from the pilot study are certainly promising, how well the program performs once implemented at scale remains to be seen.

In the absence of targeted professional development, there was little evidence pointing to a systematic and intensive approach to instructional coaching and/or professional development in the seven school systems we examined. One exception is DCPS, which recently implemented a school-based, content-specific coaching program called LEAP. Through LEAP, all teachers are placed in a small, content-specific group and assigned a content-expert for the year. Throughout the school year, these groups meet on a weekly basis to review student data, lesson plan, discuss teaching strategies, and engage in other activities intended to support ambitious instruction. Teachers also receive non-evaluative feedback from their assigned content-expert on a weekly or bi-weekly basis. Although these is little evidence linking professional development to improvements in student outcomes (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), LEAP aligns with what little is known regarding best practices for professional development (Garet, Porter, Desimone, Birman, & Yoon, 2001; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Specifically, LEAP provides "intensive, sustained, job-embedded PD focused on the content of the subject that teachers teach [which] is more likely to improve teacher knowledge, classroom instruction, and student achievement" (Wayne et al., 2008, p. 470).

Effects of Evaluation in Seven Teacher Evaluation Systems

Despite their significant efforts to improve teacher evaluation, few of the seven systems we examined appear to meaningfully differentiate teacher effectiveness. As shown in Figure 7, in all systems except New Haven, at least 75 percent of teachers were rated proficient or higher. Moreover, although New Haven appears to identify more teachers as low-performing, we were unable to find data more recent than 2012-13; a more recent distribution of teacher ratings from New Haven could very well look more like the distributions from the other systems we examined. While there is no ideal distribution of teacher effectiveness, it makes good sense that the outcomes from rigorous teacher evaluation system might align more closely with student outcomes. However, only 30 percent of 4th graders in DC and 26 percent of 4th graders in NYC scored at or above proficient on the 2015 National Assessment of Education Progress (NAEP) math exam.[26] These results contrast quite starkly with teacher evaluation outcomes that rank 80 percent of DCPS teachers and 88 percent of NYC teachers at or above proficient (see Figure 7).

The following merits repetition: while one should not expect to see a one-to-one correlation between teacher evaluation ratings and student performance ratings,[27] school systems invested in implementing rigorous teacher evaluation systems might take substantial misalignment between these outcomes as an indication that further refinements are necessary. Indeed, in 2012-13 DCPS introduced a fifth rating category, "Developing," that was created by assigning the lower half of the range of scores previously labeled "Effective" as" Developing,"

[25] See Appendix A for effect sizes.

[26] https://www.nationsreportcard.gov/profiles/districtprofile?chort=1&sub=RED&sj=XQ&sfj=NL&st=MN&year=2015R3

[27] Indeed, as we note in our theory of change, there are many factors that shape student outcomes. Furthermore, teacher evaluation is only one of several policies that can be used to improve teacher effectiveness and, ultimately, student outcomes.

thereby nearly halving the number of teachers rated Effective. In DCPS, Developing teachers must improve within two years or they will be dismissed. This decision was made because policymakers believed the previous Effective category encompassed too wide a range of teacher performance and because it was at odds with student performance.

Our search for high-quality empirical literature focusing on teacher evaluation in the seven systems featured in our analysis left us relatively empty-handed. The few studies we have uncovered to date (summarized below) focus on specific features of teacher evaluation (i.e., high-stakes consequences) or on the effects of broader reforms (e.g., the effect of Denver's financial rewards, which only overlap slightly with teacher evaluation). To an extent, the paucity of rigorous analyses of teacher evaluation systems reflects the complex design and implementation processes that precede these systems. For instance, because teacher evaluation has often been implemented at scale (i.e., throughout an entire school system), there are few opportunities to examine specific features of the teacher evaluation system and it can also be quite challenging to identify credible comparison groups for quasi-experimental studies. To the extent that it is feasible, school systems that have an opportunity to implement a novel teacher evaluation system would do well to use randomized, controlled trials (RCTs) to pilot specific features of the evaluation system. This would allow the school system to refine each aspect of the system before running a final RCT that implements the evaluation system as a whole.

In the absence of RCTs, researchers have employed descriptive and quasi-experimental methods to evaluate teacher evaluation and compensation reforms. Studies focusing on Denver's teacher compensation system, ProComp, find that financial incentives are associated with improved teacher retention (Fulbeck, 2014) and student outcomes (Goldhaber & Walch, 2012). Two studies also found that teachers who participated in ProComp were higher-performing than non-participating teachers (Goldhaber & Walch, 2012; Wiley, Spindler, & Subert, 2010). The difference between participating and non-participating teachers may have arisen through two pathways that are not mutually exclusive: one possibility is that ProComp attracted higher-performing teachers to the district; another possibility is that the district improved its hiring and/or selection process.

Despite these positive effects, it is important to note that ProComp is a system that awards fourteen different financial incentives; only two of these are related to teacher evaluation outcomes. It is therefore difficult to parse the role of teacher evaluation in driving the observed changes. Furthermore, the few findings related to teacher evaluation in Denver are not particularly encouraging. Specifically, Goldhaber and Walch (2012) found that the financial incentives related to teacher evaluation were only weakly associated with math teacher effectiveness (as measured by the authors' own value-added calculation). As a result, Denver's performance-based financial incentives are unlikely to drive improvements in teacher practice, because they may not be well-targeted. On the other hand, the authors also note that student achievement improved among non-participating teachers, which suggests that reforms other than ProComp may have contributed to improvements in student academic outcomes. As the authors conclude, "This finding is important in that it provides evidence that it is not only the compensation aspect of the program that is driving the general increase in student achievement, which is not terribly surprising given the array of system changes implemented to support ProComp. For instance, […] there were significant changes to teacher evaluations and performance feedback systems put in place under ProComp that affected all teachers."

(Goldhaber & Walch, 2012, p. 1077) In brief, there is little evidence to connect Denver's teacher evaluation system with improved teacher or student outcomes.

DCPS is another system that has been evaluated using quasi-experimental methods. Dee and Wyckoff (2015) use a regression discontinuity design (RDD) to examine differences in outcomes for teachers who are near consequential score thresholds. Although the RDD has a strong causal warrant, this type of study can only draw inferences about the contrast in incentives/sanctions at specific points in the teacher effectiveness distribution. That is, the paper is not an assessment of DCPS' teacher evaluation system writ large. Nonetheless, the authors found that, among teacher near consequential score thresholds, teacher performance improved. Moreover, value-added improved among teachers near the dismissal threshold, which suggests that student achievement may also have improved for this subset of teachers. Finally, the authors also note that low-performing teachers near the dismissal threshold were more likely to leave the district, which could lead to positive changes in the teacher workforce (conditional on finding more effective teachers to replace exiting teachers). To this end, a subsequent study examined the relationship between teacher turnover and student achievement in DCPS. Using quasi-experimental methods, the authors found DCPS was able to replace low-performing teachers with teachers who improved student math and reading achievement by about a half a year of learning (Adnot, Dee, Katz, & Wyckoff, 2017). This study is not an assessment of DCPS' teacher evaluation system, as there may have been other reforms (e.g., changes in teacher hiring and selection practices) that affected the observed outcomes.

It is worth returning to the issue of potential unintended consequences is assessing the effects of teacher evaluation in DCPS. DCPS has recently come under scrutiny for inappropriately graduating students who had not met graduation requirements in an effort to improve graduation rates, a widely-cited measure of educational success, and one that can play a small role in DCPS principal evaluations. School leaders were also caught manipulating—or pressuring their teachers to manipulate—student attendance and course credit data to meet school-level performance targets (Balingit & Tran, 2018; McGee, 2018; Brown, Strauss, & Stein, 2018). These allegations, while notable and troubling, are not directly salient for IMPACT. Graduation rates, attendance rates, and credit accumulation are not a component of teachers' IMPACT scores. Instead, IMPACT heavily weights classroom observations intended to induce teachers to improve diverse pedagogical skills and behaviors.

In theory, the emphasis on TLF could encourage manipulation by principals who want to support teachers' ratings. However, the presence of additional TLF ratings by external evaluators and the corresponding system of principal accountability suggest that such manipulation is unlikely.[28] Allegations of cheating on the high-stakes student achievement test in DCPS received extensive coverage in the press prior to 2012-13. There are several reasons we believe these allegations are not empirically relevant for the results presented here, not least is that test cheating is relevant for fewer than 20 percent of the teachers and fewer than a dozen classrooms were alleged to have cheated. Dee and Wyckoff (2015) test for the effects of test manipulation and find no evidence that such manipulation influenced outcomes.

[28] The variability in principals' TLF ratings is also inconsistent with widespread manipulation. Dee and Wyckoff (2015) also find that IMPACT incentives generated similar increases in the TLF ratings by principals and external evaluators.

The last bit of evidence available to date relates to a pilot study conducted in conjunction with TN's teacher evaluation system. In 2012-13, TN offered high-performing teachers in low-performing schools a $5,000 retention bonus that was conditional upon returning to a low-performing school for the following (2013-14) school year. Using a fuzzy RDD, the authors found that while the retention bonus did not affect the retention of high-performing teachers on average, high-performing teachers of tested-subjects were approximately 20 percent more likely to continue teaching in a low-performing school when compared with tested-subject teachers just below the retention bonus eligibility threshold (Springer, Swain, & Rodriguez, 2016). As with the other studies we have discussed, this paper is not an assessment of TN's teacher evaluation system, but rather an examination of a retention bonus. Moreover, despite the encouraging results, TN does not appear to have maintained or expanded the retention bonus program after the pilot study.

Conclusion

The evidence compiled in this review points to substantial revision in the design of teacher evaluation in the United States. In some ways, school systems appear to have made a good faith effort to design teacher evaluation systems that align with emerging best practices. For instance, nearly all teacher evaluation systems incorporate multiple measures of teacher effectiveness. However, other features of teacher evaluation systems do not seem to adhere with empirically-driven recommendations. For example, most teacher evaluation systems assign a preponderance of weight to classroom observations. As noted in reports from the MET study, placing a majority of weight on any single measure of teacher effectiveness is likely to shape teachers' behavioral response to the evaluation system.

There is a growing sense that teacher evaluation systems have not realized their potential. That impression is largely based on circumstantial evidence. For example, many teacher evaluation systems continue to rate nearly all teachers as effective or better. Thus, even though nearly all of these systems have at least four ratings categories, only rarely are more than five percent of the teachers rated as Unsatisfactory or Needs Improvement (names typically associated with the bottom two categories in a four category system).

This may not be problematic if these ratings accurately reflect teachers' effectiveness. Available evidence suggests this is not the case. For example, a survey of teacher classroom evaluators (e.g., principals or assistant principals) during 2014-15 in one school district finds they give about 6 percent of their teachers summative ratings of Unsatisfactory or Needs Improvement when they actually believe that about 19 percent of teachers deserve such ratings. Evaluators may "over rate" teachers for a variety of reasons including the administrative time costs associated with documenting relatively poor performance or concern over imposing consequences on teachers identified as poorly performing (Kraft & Gilmour, 2017). Regardless, the promise of teacher evaluation as a mechanism to improve teacher quality has been compromised by the unintended responses of actors to the systems.

One of the purported strengths of the revised evaluation systems was the inclusion of student achievement as a key component for teachers in tested grades and subjects. However, as implemented, more than half of all of systems that include student growth as an evaluation component allow these teachers to receive a rating of Effective or better despite having a less than effective score from a student growth measure (Walsh et al., 2017).

While potentially troubling, neither the paucity of poorly rated teachers nor the ability to work around a measure of student achievement growth means that revised teacher evaluation systems aren't improving the quality of teaching and student achievement. While few systems rate many teachers as Unsatisfactory or Needs Improvement, there is much greater dispersion between the top two categories of a four category system. In many states, this distinction may have few explicit consequences; it is nonetheless a distinction that may matter to teachers and thus motivate them to improve.

The U.S. system of teacher compensation which rewards experience and formal education is at best only loosely aligned with teacher productivity. As a result, there have been recurring attempts to explore the potential for various merit pay or pay-for-performance designs as a mechanism to improve student achievement.[29] However, rigorous evaluation of these efforts has been limited until recently. Within the last five years several studies with differently structured incentives, applied in different contexts have generally found null effects (Fryer, 2013; Glazerman & Seifullah, 2012; Springer et al., 2012). It is tempting to conclude from these studies that incentives are ineffective. We believe this would be premature. Most of these studies were one-time, relatively short-lived experiments, which explore the hypothesis that teachers have the resources and understand how to improve but are poorly motivated to do so. As Dee and Wyckoff (2015) conjecture, this hypothesis is overly simplistic and fails to account for the complexity of teaching. Well-designed evaluation systems incorporate individual diagnostics of performance with clear feedback on development.

Direct evidence on the effects of teacher evaluation is limited and rarely provides insights on the mechanisms or components of teacher evaluation. Taylor and Tyler (2012) employ the phase-in of teacher evaluation on Cincinnati to show that otherwise similar teachers who experienced a rigorous evaluation improved student achievement by 10 percent of a standard deviation compared to teachers who had not be evaluated even though the evaluation carried no meaningful consequences. Similarly, Steinberg and Sartain (2015) employ a phased rollout (that was effectively random) of teacher evaluation in Chicago to find students of teachers who experienced evaluation improved by 0.10 of a standard deviation more than students of teachers who had not been evaluated. These effects were concentrated among the highest performing students and effects for math were not statistically significant.

As we describe more fully elsewhere, IMPACT, the District of Columbia Public Schools teacher evaluation system was designed and implemented consistent with the best available evidence for teacher evaluation. Several steps were taken to attend to the issues of validity, reliability and transparency. Assessing the effects of IMPACT, Dee and Wyckoff (2015) find that that the dismissal threats associated with poor performance as measured by IMPACT increased the voluntary attrition of low-performing teachers by 11 percentage points (i.e., more than 50 percent) and improved the performance of teachers who remained by 0.27 of a teacher-level standard deviation. They also find that financial incentives available to high-performing teachers if they perform well again improved the performance of these teachers (effect size = 0.24).

As the prevalence of teacher evaluation has increased, so too has rigorous research regarding the design and effects of teacher evaluation. While there are still many unanswered

[29] For a good summary of this literature see Springer (2009).

questions, best practices are emerging to provide guidance as school systems continue to refine their teacher evaluation program.

Teacher Evaluation Promising Practices

There is little rigorous empirical analysis of teacher evaluation systems and certainly none to our knowledge that allows any systematic comparative analysis of the optimal structure of evaluation systems. We summarize our discussion by proposing promising practices that synthesize various evidence with from the experience of districts and states.

Multiple measures. Effective teaching is a multidimensional process. As a result, no single component, e.g., value-added, is likely to effectively capture each of these components. Moreover, any single measure is subject to measurement error; combining measures increases the reliability of the system.

Value-added. There are many reasons to include value-added as a meaningful component of a teacher evaluation system. It is typically one of the few measures linked to a measure of student outcomes that is consistent across schools. As has been discussed, it is the best predictor of a teacher's future effectiveness. In addition, even though value-added is typically available for only about 20 to 25 percent of teachers, we believe that including value-added provides a benchmark against which other measures can be compared. For example, for value-added teachers, if their VAM score is inconsistent with their SBO score, it raises questions about the validity of the SBOs that may have important implications for non-value-added teachers.

Rigorous standards-based observations. SBOs can provide teachers with insightful and actionable information if they are implemented in ways that are valid and reliable, but doing so is likely expensive. For example, in the first seven years of its system, the District of Columbia Public Schools evaluated teachers five times a year, three times by a school administrator and twice by a master educator. Each of these observers were normed on the observation rubric to insure they reliably rated teachers. This intensive approach minimizes errors in assessing effectiveness and provides teachers with frequent feedback on their performance. This approach requires careful development of the measures and the rubrics and intensive training of the observers, all of which requires time and financial investments.

Linking outcomes to consequential action. Evaluating teachers without connecting those evaluations to some consequential outcome misses important opportunities for teacher development. These actions need not be high-stakes. For example, using the evaluation outcomes to link teachers to targeted professional development is more likely to yield improvement than more generic professional development. This targeting may reflect both substance and intensity depending on the evaluation outcomes.

Stay the course. For many teachers and administrators, teacher evaluation systems are time consuming and stressful. Implementing these systems well requires trust between teachers and administrators. In some schools, this process may proceed quickly and smoothly, in others the process may take longer. Most of these systems develop over time as administrators and teachers develop an approach that works well for them. It is also the case that these systems will inevitably be adjusted over time to enhance their effectiveness in specific contexts. For all these reasons, policymakers are advised to persist in the face of initial resistance but should be prepared to make adjustments that increase the acceptance of the system without compromising its effectiveness in improving teacher quality.

Dogged implementation. Our observations of these seven systems suggest that the effectiveness of the system is often determined by the attention devoted to implementation. Effective communication of the systems goals and details, design of materials, training of participants, creation of aligned feedback and PD, and appropriate recognition for exemplary performance can all contribute to teachers and administrators treating evaluation as an opportunity to improve teacher quality and student outcomes rather than compliance with another regulation.

Questions

Finally, we close with a few questions, evidence for which is crucial in better designing teacher evaluation systems:

- What is the effect of various design elements on teacher performance and student outcomes?
    - What is the best set of component measures—validity, reliability, face validity, transparency, etc.
    - What role do stakes play and are the effects of stakes symmetric at the low and high end of performance?
    - Is feedback without coaching, which is expensive, sufficient to realize most of the gains?
- How do these systems mature? Can the focus of the system shift more toward development after an initial focus on the composition of teachers?
- How do the different components correlate with different outcome measures, e.g., are classroom observations more highly correlated to student socio-emotional outcomes?
- How does evaluation change the nature of teachers attracted and voluntarily retained?
- What is the policy tolerance for different types of errors in the classification of teachers? Are policymakers equally tolerant of dismissing a teacher labeled as ineffective who is not as they are of keeping an ineffective teacher who is labeled effective?

As our summary of the redesign of teacher evaluation in the U.S. has made clear, teacher evaluation has a variety of design elements which afford states and school districts a many choices. Our questions reflect, to some extent, this complex landscape. Much more will be known about the optimal design of teacher evaluation systems over the next few years.

Figures

Figure 1

Teacher Evaluation Theory of Change

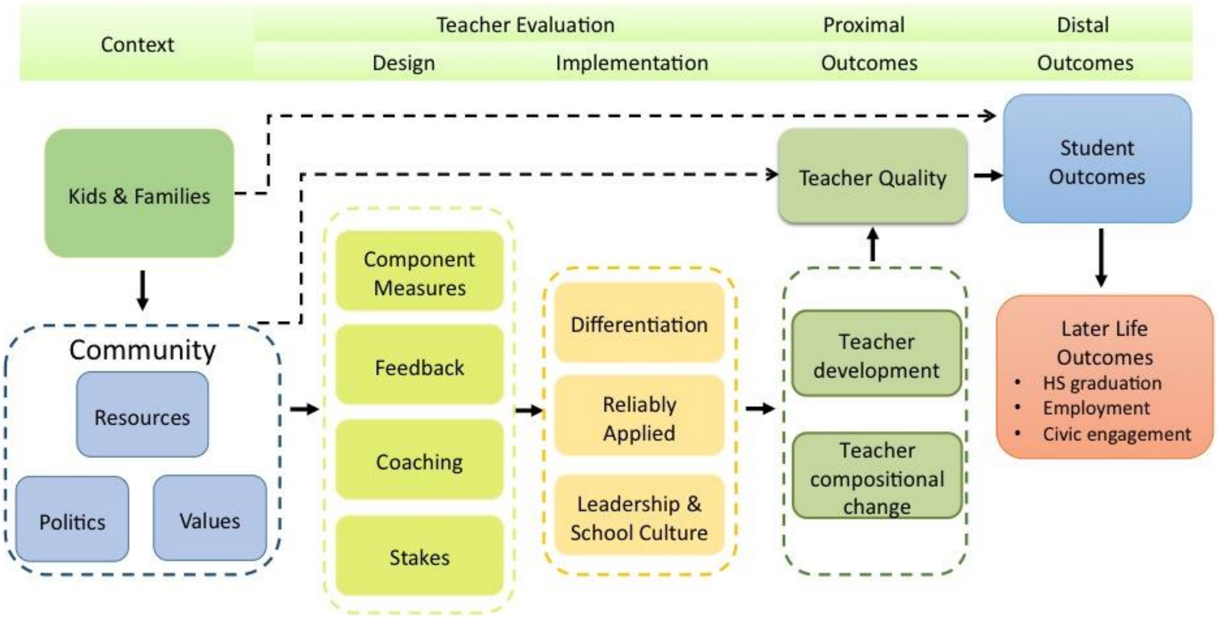
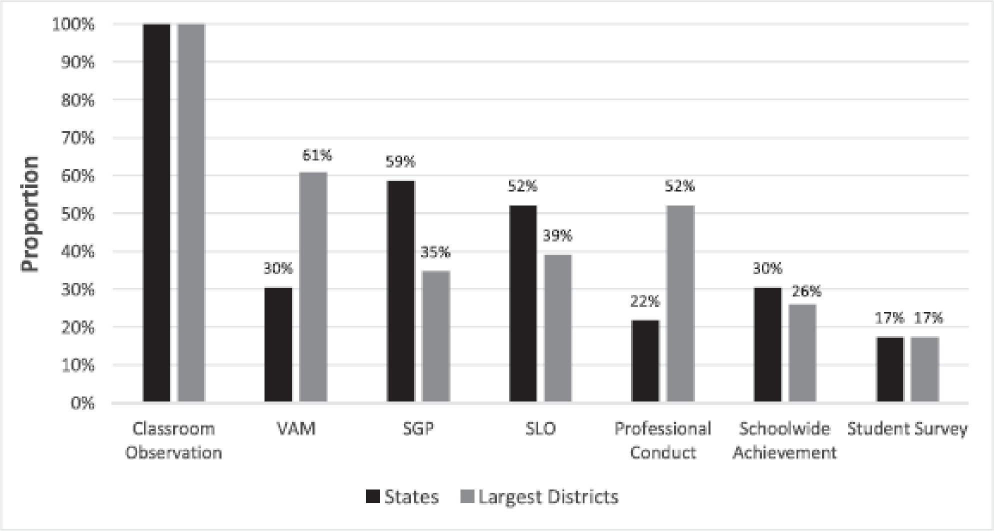


Figure 2

Frequency of Components of Teacher Evaluation Systems

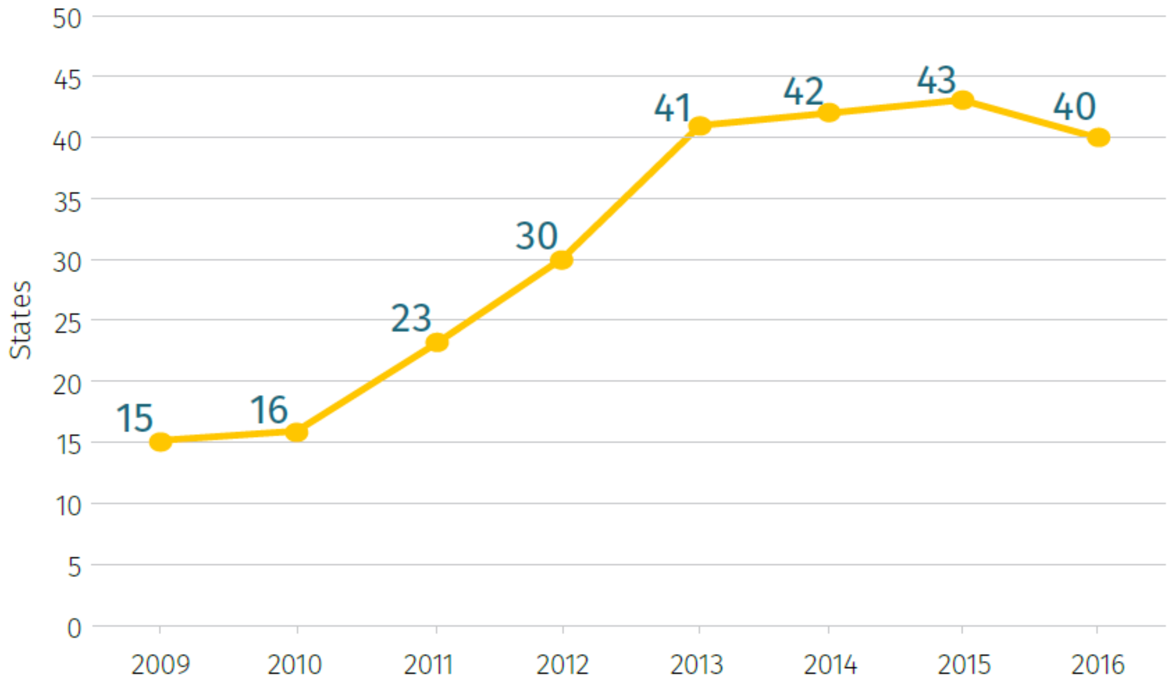


Notes: For states, the proportion using an evaluation component is out of 46; for the largest districts, the proportion using an evaluation component is out of 23 (including DC). See tables 1 and 2 for more detail on the weights associated with each measure.

Source: Steinberg & Donaldson, 2016, Figure 3

Figure 3

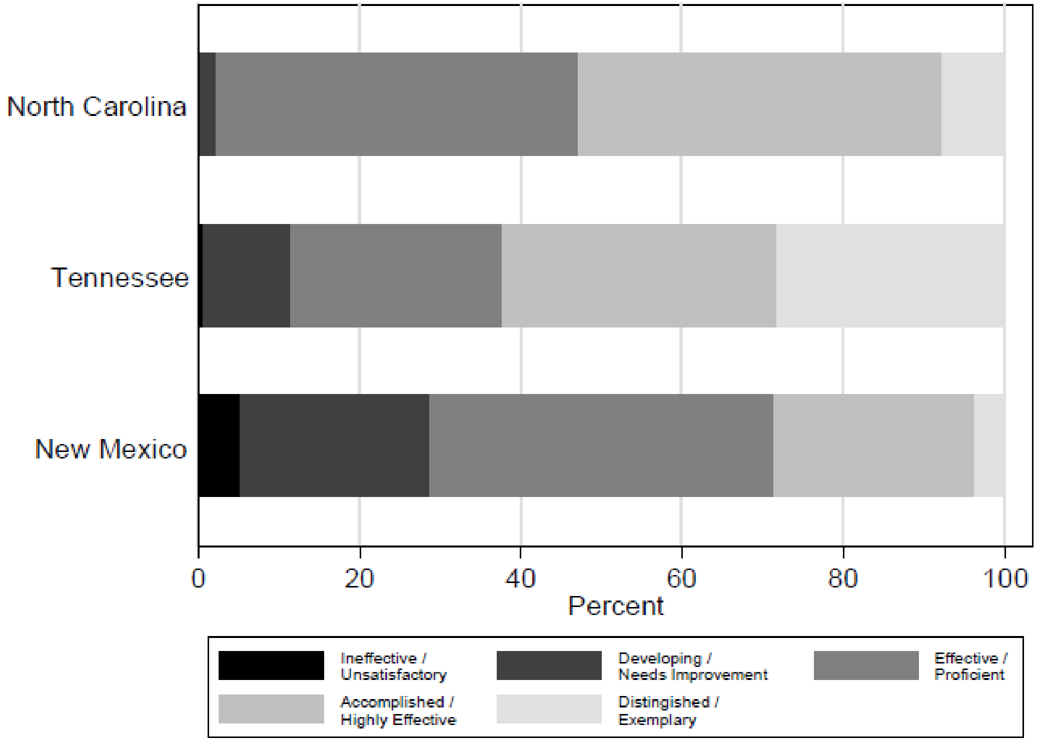
States Requiring Evidence of Student Learning in Teacher Evaluation



Source: Walsh et al., 2017, Figure 1

Figure 4

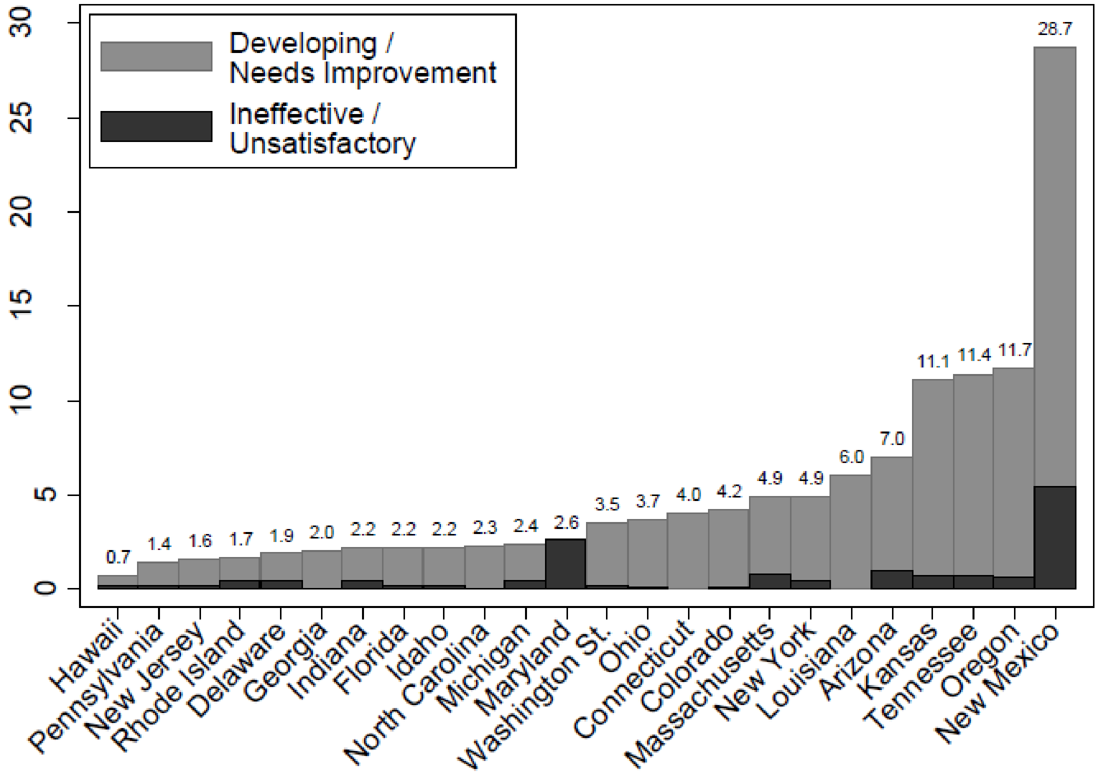
Distribution of Teacher Evaluation Ratings Across States with Five Rating Categories



Source: Kraft & Gilmour, 2017, Figure 3, Panel B

Figure 5

Percentage of Teachers Rated Below Proficient Across 24 State Evaluation Systems



Source: Kraft & Gilmour, 2017, Figure 1

Figure 6

Summary of Component Weights, Seven Teacher Evaluation Systems

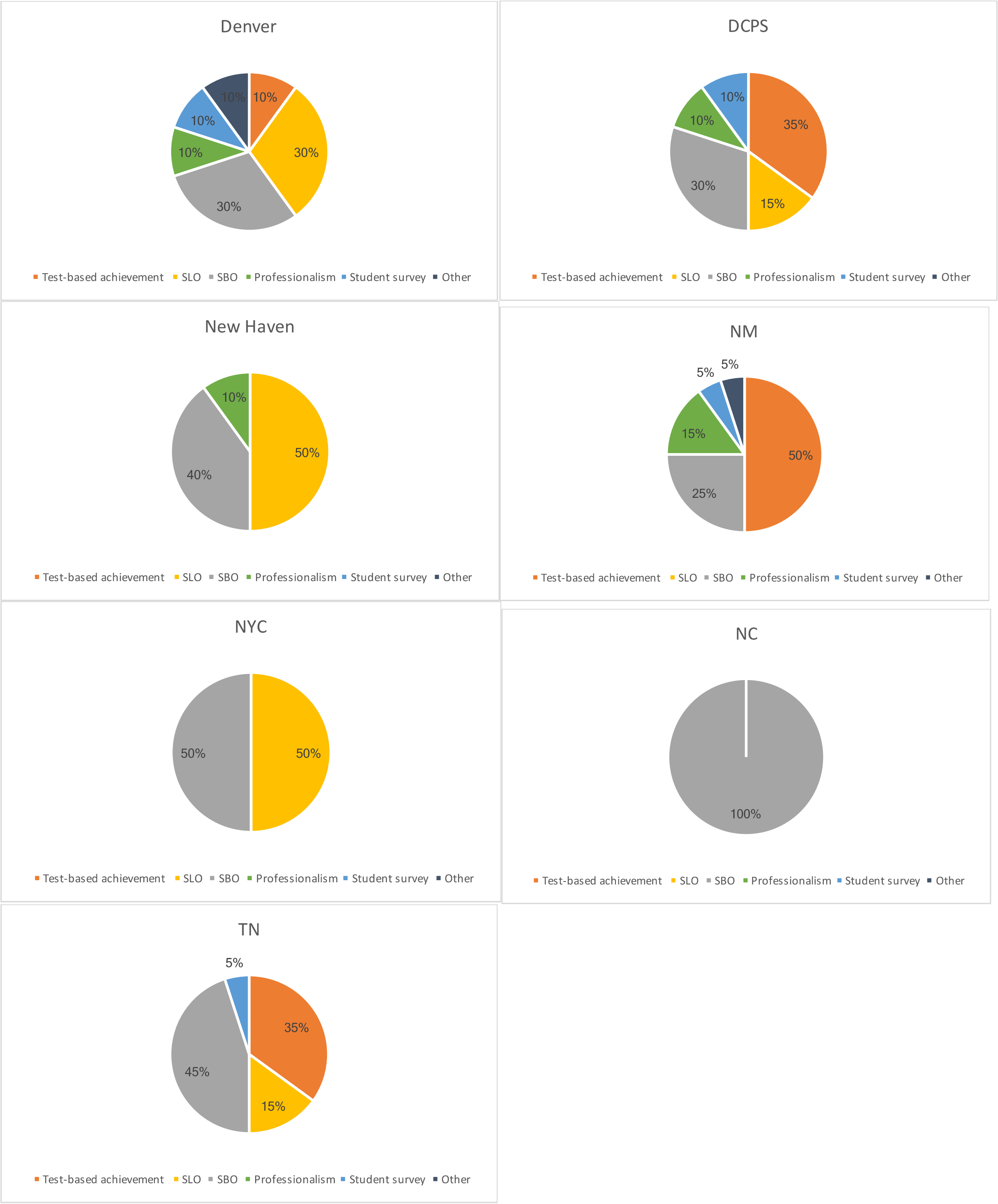
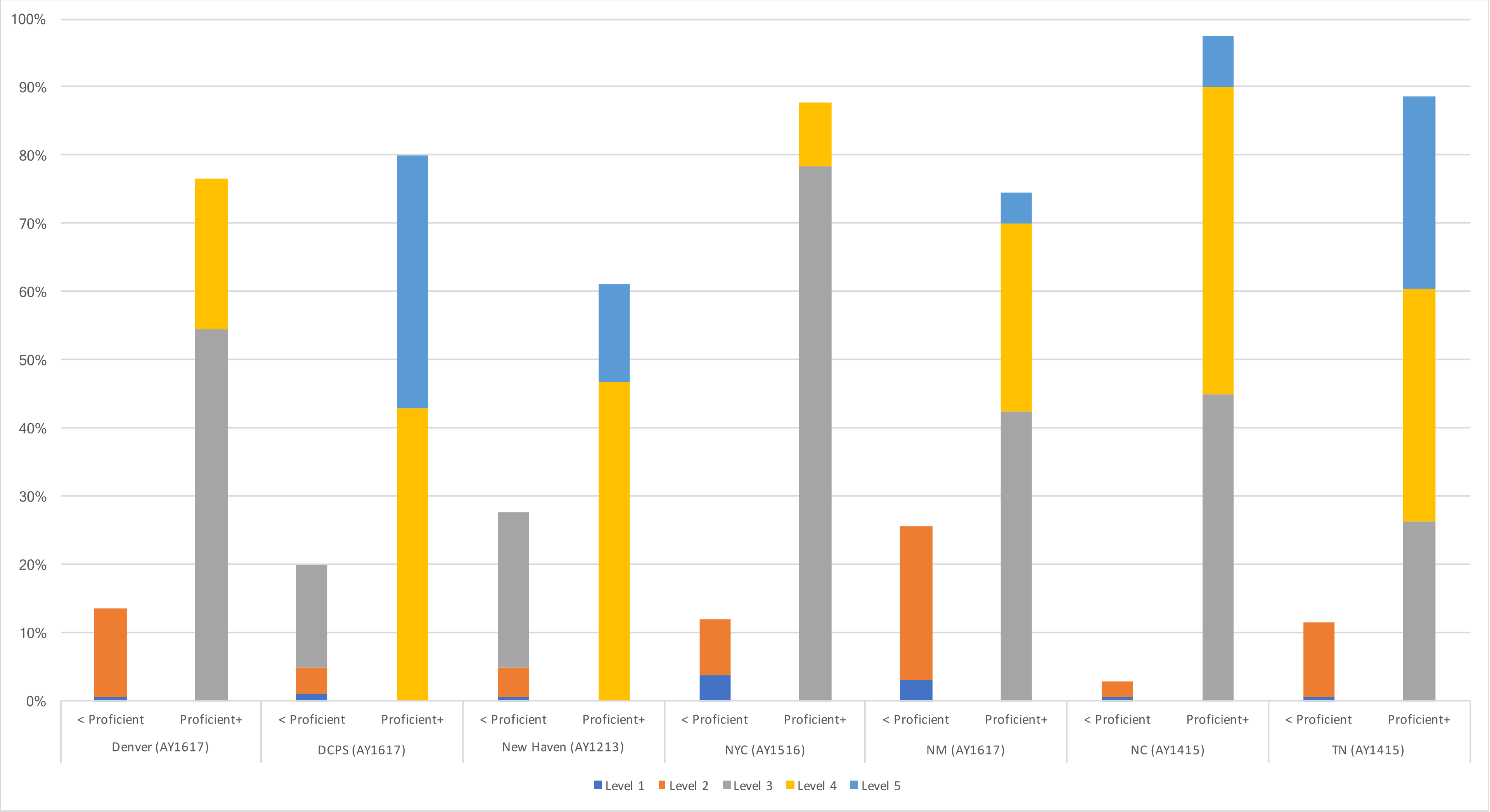


Figure 7

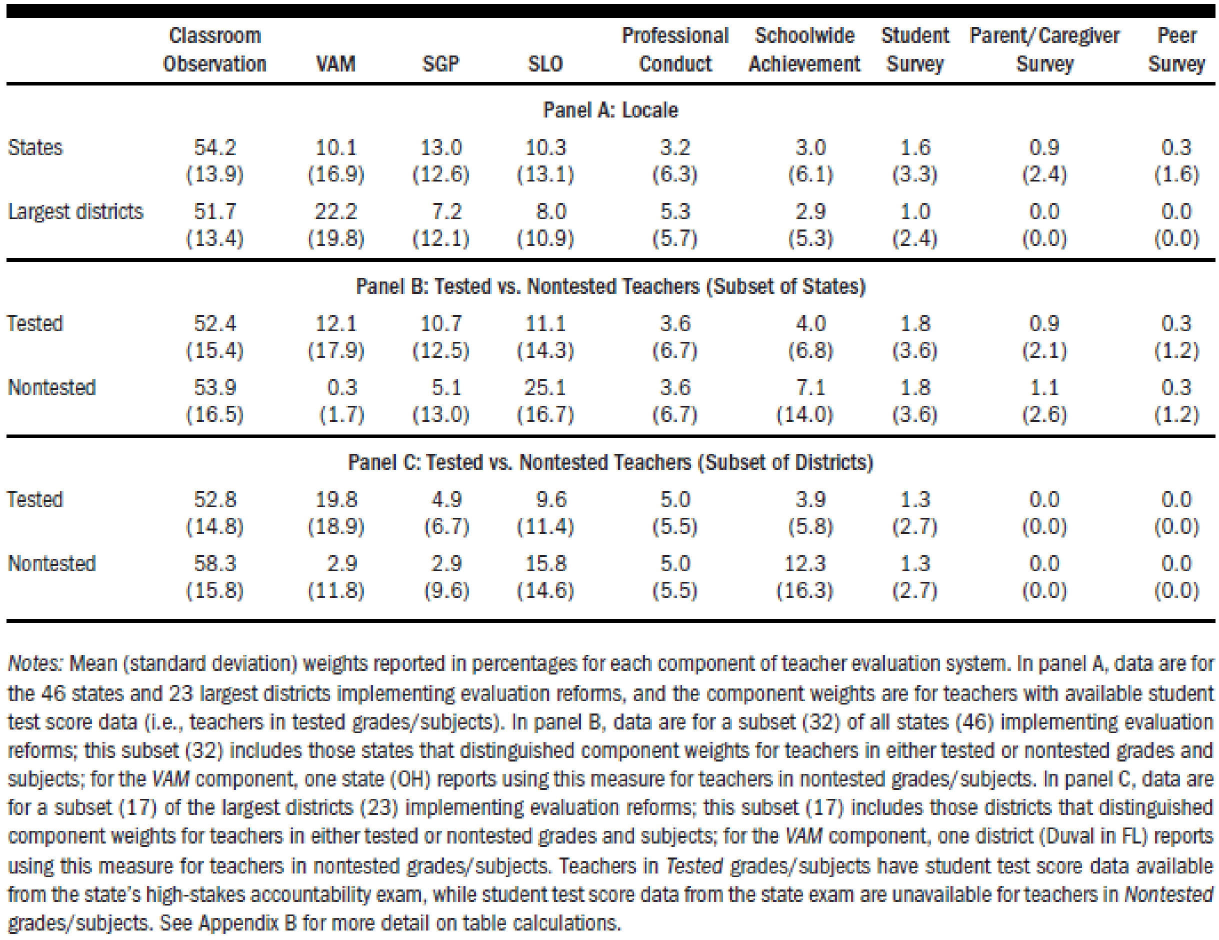
Distribution of Final Ratings, Seven Teacher Evaluation Systems



Tables

Table 1

System-Level Component Weights of New Teacher Evaluation Systems



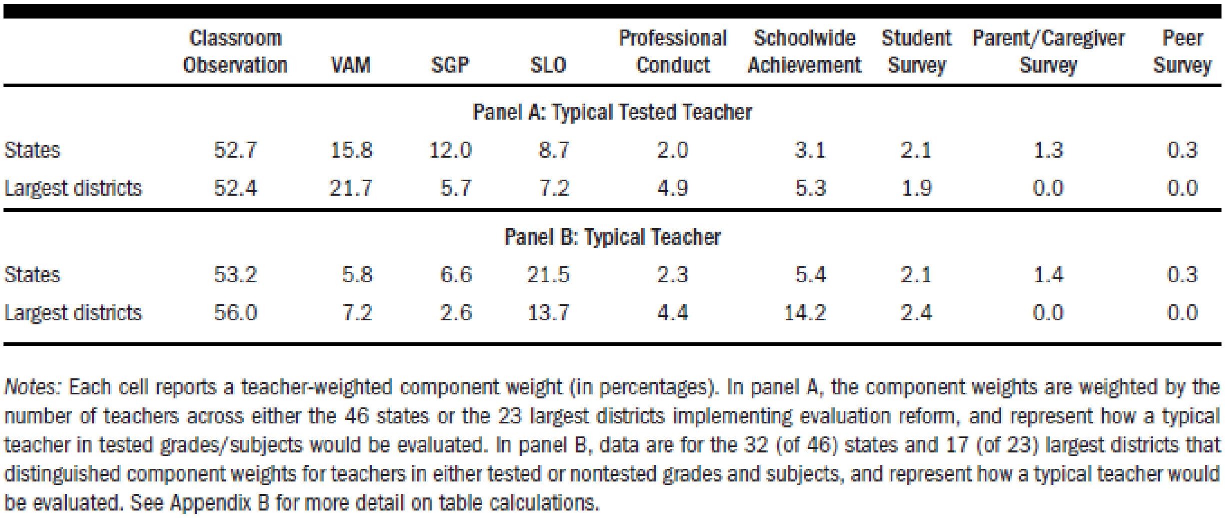
	Classroom Observation	VAM	SGP	SLO	Professional Conduct	Schoolwide Achievement	Student Survey	Parent/Caregiver Survey	Peer Survey
					Panel A: Locale				
States	54.2 (13.9)	10.1 (16.9)	13.0 (12.6)	10.3 (13.1)	3.2 (6.3)	3.0 (6.1)	1.6 (3.3)	0.9 (2.4)	0.3 (1.6)
Largest districts	51.7 (13.4)	22.2 (19.8)	7.2 (12.1)	8.0 (10.9)	5.3 (5.7)	2.9 (5.3)	1.0 (2.4)	0.0 (0.0)	0.0 (0.0)
					Panel B: Tested vs. Nontested Teachers (Subset of States)				
Tested	52.4 (15.4)	12.1 (17.9)	10.7 (12.5)	11.1 (14.3)	3.6 (6.7)	4.0 (6.8)	1.8 (3.6)	0.9 (2.1)	0.3 (1.2)
Nontested	53.9 (16.5)	0.3 (1.7)	5.1 (13.0)	25.1 (16.7)	3.6 (6.7)	7.1 (14.0)	1.8 (3.6)	1.1 (2.6)	0.3 (1.2)
					Panel C: Tested vs. Nontested Teachers (Subset of Districts)				
Tested	52.8 (14.8)	19.8 (18.9)	4.9 (6.7)	9.6 (11.4)	5.0 (5.5)	3.9 (5.8)	1.3 (2.7)	0.0 (0.0)	0.0 (0.0)
Nontested	58.3 (15.8)	2.9 (11.8)	2.9 (9.6)	15.8 (14.6)	5.0 (5.5)	12.3 (16.3)	1.3 (2.7)	0.0 (0.0)	0.0 (0.0)

Notes: Mean (standard deviation) weights reported in percentages for each component of teacher evaluation system. In panel A, data are for the 46 states and 23 largest districts implementing evaluation reforms, and the component weights are for teachers with available student test score data (i.e., teachers in tested grades/subjects). In panel B, data are for a subset (32) of all states (46) implementing evaluation reforms; this subset (32) includes those states that distinguished component weights for teachers in either tested or nontested grades and subjects; for the *VAM* component, one state (OH) reports using this measure for teachers in nontested grades/subjects. In panel C, data are for a subset (17) of the largest districts (23) implementing evaluation reforms; this subset (17) includes those districts that distinguished component weights for teachers in either tested or nontested grades and subjects; for the *VAM* component, one district (Duval in FL) reports using this measure for teachers in nontested grades/subjects. Teachers in *Tested* grades/subjects have student test score data available from the state's high-stakes accountability exam, while student test score data from the state exam are unavailable for teachers in *Nontested* grades/subjects. See Appendix B for more detail on table calculations.

Source: Steinberg & Donaldson, 2016, Table 1

Table 2

Teacher-Level Component Weights of New Teacher Evaluation Systems



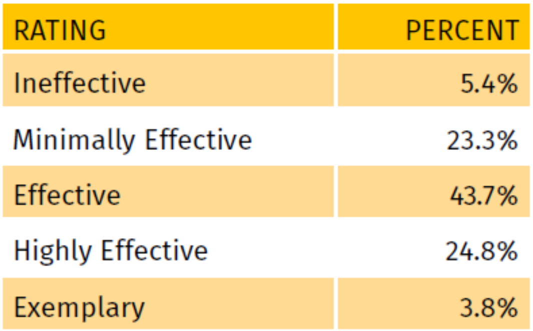
	Classroom Observation	VAM	SGP	SLO	Professional Conduct	Schoolwide Achievement	Student Survey	Parent/Caregiver Survey	Peer Survey
				Panel A: Typical Tested Teacher					
States	52.7	15.8	12.0	8.7	2.0	3.1	2.1	1.3	0.3
Largest districts	52.4	21.7	5.7	7.2	4.9	5.3	1.9	0.0	0.0
				Panel B: Typical Teacher					
States	53.2	5.8	6.6	21.5	2.3	5.4	2.1	1.4	0.3
Largest districts	56.0	7.2	2.6	13.7	4.4	14.2	2.4	0.0	0.0

Notes: Each cell reports a teacher-weighted component weight (in percentages). In panel A, the component weights are weighted by the number of teachers across either the 46 states or the 23 largest districts implementing evaluation reform, and represent how a typical teacher in tested grades/subjects would be evaluated. In panel B, data are for the 32 (of 46) states and 17 (of 23) largest districts that distinguished component weights for teachers in either tested or nontested grades and subjects, and represent how a typical teacher would be evaluated. See Appendix B for more detail on table calculations.

Source: Steinberg & Donaldson, 2016, Table 2

Table 3

New Mexico's 2015-2016 Teacher Evaluation Ratings

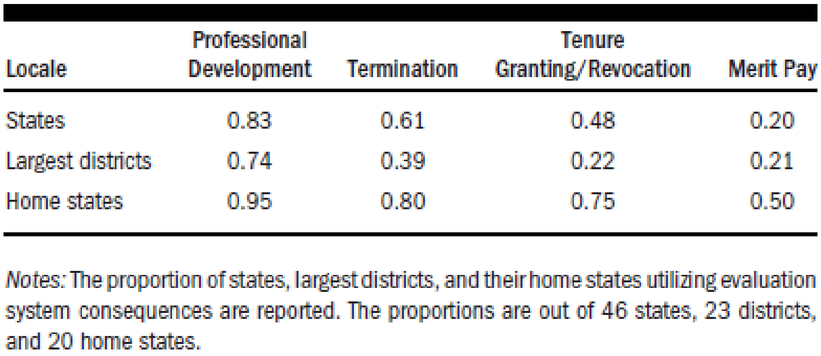


RATING	PERCENT
Ineffective	5.4%
Minimally Effective	23.3%
Effective	43.7%
Highly Effective	24.8%
Exemplary	3.8%

Source: Walsh et al., 2017, Figure 3

Table 4

Teacher Evaluation System Consequences



Locale	Professional Development	Termination	Tenure Granting/Revocation	Merit Pay
States	0.83	0.61	0.48	0.20
Largest districts	0.74	0.39	0.22	0.21
Home states	0.95	0.80	0.75	0.50

Notes: The proportion of states, largest districts, and their home states utilizing evaluation system consequences are reported. The proportions are out of 46 states, 23 districts, and 20 home states.

Source: Steinberg & Donaldson, 2016, Table 5

Table 5

Use of Sanctions and Rewards, Seven Teacher Evaluation Systems

		Denver	DCPS	New Haven	NM	NYC	NC	TN
Sanctions	Performance improvement plan				✔			
	Salary freeze	✔	✔	✔				✔
	Change in probationary status	✔						
	Dismissal		✔	✔			✔	
Rewards	Career advancement	✔	✔	✔		✔	✔	
	Performance-based salary increase	✔	✔					
	Performance-based bonus		✔		✔			

Table 6

Summary of Implementation Process, Seven Teacher Evaluation Systems

	Denver	DCPS	New Haven	NM	NYC	NC	TN
Efforts to increase buy-in prior to implementation	High	Medium	High	Low	Medium	Medium	Low
Responsiveness after initial implementation	Low	High	Medium	Low	Low	Low	High
District autonomy	Medium	High	Medium	Low	Medium	Low	Low

Table 7
Rigor of SBO Implementation, Seven Teacher Evaluation Systems

	Denver	DCPS	New Haven	NM	NYC	NC	TN
Use of external observers	None	Discontinued	Targeted	Optional	None	None	None
Annual observer certification	✔	✔	✔	✔	✔	✔	✔
Ongoing efforts to improve observer reliability	Once per year	No evidence	No evidence	No evidence	Untargeted observer coaching	No evidence	Targeted observer coaching
Maximum number of formal observations	1	3	2	3	1	3	6

Table 8
Use of Professional Development, Seven Teacher Evaluation Systems

	Denver	DCPS	New Haven	NM	NYC	NC	TN
Teacher evaluation informs PD	No evidence	No evidence	Yes, for all teachers	Yes, for low-performing teachers	Yes, for low-performing teachers	Yes, for all teachers	Yes; limited pilot study
Comprehensive PD plan	No evidence	Yes	Yes	No evidence	No evidence	Yes	No evidence

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. https://doi.org/10.1086/508733

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76. https://doi.org/10.3102/0162373716663646

Allen, J. P., Pianta, R., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science*, *333*(6045), 1034–1037. https://doi.org/10.1126/science.1207998

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do First Impressions Matter? Predicting Early Career Teacher Effectiveness. *AERA Open*, *1*(4), 2332858415607834. https://doi.org/10.1177/2332858415607834

Bailey, M. (2011, September 13). New Eval System Pushes Out 34 Teachers. *New Haven Independent*. Retrieved from http://www.newhavenindependent.org/index.php/archives/entry/new_eval_system_pushes_34_teachers_out/

Bailey, M. (2012, October 10). In 2nd Year, Evals Ease Out 28 Teachers. *New Haven Independent*. Retrieved from http://www.newhavenindependent.org/index.php/archives/entry/t-vals_year_two/

Bailey, M. (2014, February 26). Evals Push Out 20 More Teachers. *New Haven Independent*. Retrieved from http://www.newhavenindependent.org/index.php/archives/entry/new_haven_teacher_evals_3rd_year/

Balingit, M. & Tran, A. B. (2018, January 6). Before a graduation scandal made headlines, teachers at D.C.'s Ballou High raised an alarm. *Washington Post*. Retrieved from: https://www.washingtonpost.com/local/education/before-a-graduation-scandal-made-headlines-teachers-at-dcs-ballou-high-raised-an-alarm/2018/01/06/ad49f198-df6a-11e7-89e8-edec16379010_story.html?utm_term=.7a95adfa3e20

Brown, E., Strauss, V., & Stein, P. (2018, March 10). It was hailed as the national model for school reform. Then the scandals hit. *The Washington Post*. Retrieved from: https://www.washingtonpost.com/local/education/dc-school-scandals-tell-me-that-its-not-great-and-that-youre-dealing-with-it/2018/03/10/b73d9cf0-1d9e-11e8-b2d9-08e748f892c0_story.html?utm_term=.52f11dc57cbf

Burgess, K. (2017a, April 2). PED alters teacher evaluation system after pushback. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/981045/ped-alters-teacher-evaluation-system-after-pushback.html

Burgess, K. (2017b, July 23). Teachers union lawsuit against PED delayed. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/1037067/teachers-union-lawsuit-against-ped-delayed.html

Cantrell, S., & Kane, T. J. (2013). *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study*. Retrieved from http://k12education.gatesfoundation.org/resource/ensuring-fair-and-reliable-measures-of-effective-teaching-culminating-findings-from-the-met-projects-three-year-study/

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Curtis, R. (2011). *District of Columbia Public Schools: Defining Instructional Expectations and Aligning Accountability and Support*. Washington, D.C.: The Aspen Institute Education & Society Program. Retrieved from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=1509&download

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105–115. https://doi.org/10.1037/h0030644

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. *Review of Educational Research*, *71*(1), 1–27. https://doi.org/10.3102/00346543071001001

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

Denver Public Schools. (2016, August). LEAP Basic Fairness Guide 2016-17. Retrieved from http://thecommons.dpsk12.org/cms/lib/CO01900837/Centricity/Domain/103/2016-17%20LEAP%20Basic%20Fairness%20Guide%20-%2009_07_16.pdf

Desimone, L. M., & Garet, M. S. (2015). Best Practices in Teachers' Professional Development in the United States. *Psychology, Society, & Education*, *7*(3), 252–263.

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (Second Edition, pp. 174–193). New York, NY: Routledge.

Dynarski, M. (2016, December 8). Teacher observations have been a waste of time and money. Retrieved October 18, 2017, from https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/

Ehlert, M. W., Pepper, M. J., Parsons, E. S., Burns, S. F., & Springer, M. G. (2013). *Educator Evaluation in Tennessee: Initial Findings from the 2013 First to the Top Survey*. Nashville, TN: Tennessee Consortium on Research, Evaluation & Development.

Fryer, R. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373–427.

Fulbeck, E. S. (2014). Teacher Mobility and Financial Incentives: A Descriptive Analysis of Denver's ProComp. *Educational Evaluation and Policy Analysis*, *36*(1), 67–82. https://doi.org/10.3102/0162373713503185

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What Makes Professional Development Effective? Results From a National Sample of Teachers. *American Educational Research Journal*, *38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Glazerman, S., & Seifullah, A. (2012). *An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report* (Mathematica Reference No. 06736-520). Mathematica Policy Research, Inc. Retrieved from http://eric.ed.gov/?id=ED530098

Goldhaber, D. (2015). Teachers Clearly Matter, But Finding Effective Teacher Policies Has Proven Challenging. In *Handbook of Research in Education Finance and Policy* (2nd ed., pp. 157–173). New York, NY: Routledge.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review*, *31*(6), 1067–1083. https://doi.org/10.1016/j.econedurev.2012.06.007

Goldin, C. D., & Katz, L. F. (2009). *The Race between Education and Technology*. Harvard University Press.

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015). A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance. *Statistics and Public Policy, 2*(1), 1–11. https://doi.org/10.1080/2330443X.2015.1034820

Harris, D. N. (2011). *Value-Added Measures in Education*. Cambridge, MA: Harvard University Press.

Hutchinson, J. (2017, August 20). Teacher evaluations can be a tool for success. *Las Cruces Sun-News*. Retrieved from http://www.lcsun-news.com/story/opinion/columnists/2017/08/20/teacher-evaluations-can-tool-success/583680001/

Jerald, C. J. (2013). *Beyond Buy-In: Partnering with Practitioners to Build a Professional Growth and Accountability System for Denver's Educator*. Washington, D.C.: The Aspen Institute Education & Society Program. Retrieved from http://careers.dpsk12.org/wp-content/uploads/2016/10/Stakeholder-Engagement-white-paper.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project*. Bill & Melinda Gates Foundation. Retrieved from https://eric.ed.gov/?id=ED540959

Klafehn, A. (2015). *An Analysis of Teacher Evaluation Data and Teacher Characteristics* (Legislative Brief). Nashville, TN: Tennessee Comptroller of the Treasury: Offices of Research and Education Accountability.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). *The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence* Review of Educational Research 88(4) 547-588.

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.

Lazear, E. P. (1995). *Personnel Economics*. Cambridge, MA: MIT Press.

McCann, C. (2012, September 12). Districts Return Teacher Incentive Fund Dollars, Unable to Reach Agreement with Unions. *Ed Money Watch*.

McGee, K. (2018, January 25). Most DCPS teachers feel pressure to pass students, teacher union survey says. *WAMU*. Retrieved from: https://wamu.org/story/18/01/25/dcps-teachers-feel-pressure-pass-students-teacher-union-survey-says/

New Mexico Public Education Department. (2017). *NMTEACH: New Mexico Educator Effective System Technical Guide Business Rules and Calculation 2016-2017 School Year Teacher Summative Report*. Santa Fe, NM: New Mexico Public Education Department. Retrieved from http://ped.state.nm.us/ped/NMTeachDocs/2017/NMTEACH%20Technical%20Guide%202016-2017.pdf

Papay, J., Taylor, E. S., Tyler, J., & Laski, M. (2016). *Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data* (Working Paper No. 21986). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w21986.pdf

Peterson, K. D., Wahlquist, C., Bone, K., Thompson, J., & Chatterton, K. (2001). Using More Data Sources To Evaluate Teachers. *Educational Leadership*, *58*(5), 40–43.

Pratt, T. (2014). *Making Every Observation Meaningful: Addressing Lack of Variation in Teacher Evaluation Ratings*. Nashville, TN: Tennessee Department of Education: Office of Research and Policy. Retrieved from http://team-tn.org/wp-content/uploads/2013/08/rpt_non-differentiating_observers1.pdf

Public Schools of North Carolina. (2013, May). North Carolina Professional Teaching Standards. Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/standards/prof-teach-standards.pdf

Reform Support Network. (2012). *Transition to TEAM: First-Year Reflections on Implementing a New Teacher Evaluation System in Tennessee*. Retrieved from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/tle-sap-tn.pdf

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

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, *94*(2), 247–252.

SAS EVAAS. (2015). *Technical Documentation for 2015 TVAAS Analyses*. Cary, NC: SAS Institute Inc.

SAS EVAAS. (2017). *Technical Documentation of 2017 TVAAS Analyses*. Cary, NC: SAS Institute Inc. Retrieved from https://tvaas.sas.com/support/TVAAS-TechnicalDocumentation-2017.pdf

Slotnick, W. J., & Smith, M. D. (2004). *Catalyst for Change: Pay for Performance in Denver Final Report*. Boston: Community Training and Assistance Center. Retrieved from http://www.ctacusa.com/wp-content/uploads/2013/11/CatalystForChange.pdf

Springer, M. G. (Ed.). (2009). *Performance Incentives*. Washington, D.C.: Brookings Institution Press.

Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., … Stecher, B. M. (2012). *Final Report: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective Teacher Retention Bonuses: Evidence From Tennessee. *Educational Evaluation and Policy Analysis*, *38*(2), 199–221. https://doi.org/10.3102/0162373715609687

Stecher, B., Holtzman, D., Garet, M., Hamilton, L., Engberg, J., Steiner, E., Robyn, A, Matthew D., Gutierrez, I., Peet, E., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunger, G., and Chambers, J. (2018) Improving teaching effectiveness final report: The intensive partnerships for effective teaching through 2015-16. Policy Report. Santa Monica, CA: RAND Corporation.

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, *11*(3), 340–359. https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Kraft, M. A. (2017). *The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems* (Working Paper).

Steinberg, M. P., & Sartain, L. (2015). Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago’s Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Tagami, T. (2017) “Educators lose appeal in Atlanta test-cheating case” Atlanta Journal-Constitution recovered from https://www.ajc.com/news/state--regional-education/educators-lose-appeal-atlanta-test-cheating-case/slNEtdrkh4Mk4Gw8BzKx1J/

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

Tennessee Department of Education. (2015). *Teacher and Administrator Evaluation in Tennessee: A Report on Year 3 Implementation*. Nashville, TN: Tennessee Department of Education. Retrieved from http://team-tn.org/wp-content/uploads/2013/08/rpt_teacher_evaluation_year_31.pdf

Tennessee Department of Education. (2016). *Teacher and Administrator Evaluation in Tennessee: A Report on Year 4 Implementation* (Policy Brief). Nashville, TN: Tennessee Department of Education. Retrieved from http://team-tn.org/wp-content/uploads/2013/08/TEAM-Year-4-Report1.pdf

The New Haven Model for Teacher Evaluations. (2010, May 2). *The New York Times*. Retrieved from https://www.nytimes.com/2010/05/03/opinion/03mon3.html

Walsh, K., Joseph, N., Ross, E., & Lubell, S. (2017). *Running in Place: How New Teacher Evaluations Fail to Live Up to Promises*. Washington, D.C.: National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsStage.do?fn=Final_Evaluation_Paper

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting With Teacher Professional Development: Motives and Methods. *Educational Researcher*, *37*(8), 469–479. https://doi.org/10.3102/0013189X08327154

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. Second Edition*. New York. Retrieved from http://eric.ed.gov/?id=ED515656

Wiley, E. W., Spindler, E. R., & Subert, A. N. (2010). *Denver ProComp: An Outcomes Evaluation of Denver's Alternative Teacher Compensation System*. Boulder, CO: University of Colorado at Boulder, School of Education. Retrieved from http://valueadded.cmswiki.wikispaces.net/file/view/Denver-ProCompOutcomesEvaluationApril2010final.pdf/215405160/Denver-ProCompOutcomesEvaluationApril2010final.pdf

Wyckoff, J., & Katz, V. (2017). Policies to Improve Teacher Quality. In *International Handbook of Teacher Quality and Policy* (pp. 97–114). New York, NY: Routledge.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033* (Issues & Answers Report, REL 2007 No. 033). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://eric.ed.gov/?id=ED498548

Zubrzycki, J. (2012, August 12). Big-City Districts Bail on Teacher-Incentive Grants. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2012/08/22/01tif_ep.h32.html