THE NEW EDUCATIONAL ACCOUNTABILITY: UNDERSTANDING THE LANDSCAPE OF TEACHER EVALUATION IN THE POST-NCLB ERA

Matthew P. Steinberg
(corresponding author)
Graduate School of Education
University of Pennsylvania
Philadelphia, PA 19104
steima@gse.upenn.edu

Morgaen L. Donaldson
Neag School of Education
University of Connecticut
Storrs, CT 06269-3093
Morgaen.donaldson@uconn.edu

Abstract

In the past five years, teacher evaluation has become a preferred policy lever at the federal, state, and local levels. Revisions to teacher evaluation systems have made teachers individually accountable for student achievement to a greater extent than ever before. We describe and analyze the components, processes, and consequences embedded in new teacher evaluation policies in all fifty states, the twenty-five largest school districts, and Washington, DC. We contextualize these policies by basing our analysis in prior research on teacher evaluation, and examining key comparisons between state and district policies, including their treatment of teachers in tested and untested subjects with career and beginning teachers. We find notable differences in how states and the largest districts have structured evaluation policies for all teachers and, in particular, for early career teachers compared with their more veteran counterparts, and for teachers in nontested grades and subjects compared with those in tested grades and subjects.

doi:10.1162/EDFP_a_00186

INTRODUCTION

In the last five years, teacher evaluation has become the primary approach for holding individual teachers accountable for their students' performance. A growing research base has demonstrated teachers' impact on student achievement (Goldhaber 2002; Rockoff 2004; Rivkin, Hanushek, and Kain 2005) and revealed substantial within-school heterogeneity in teacher effectiveness (Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007). Despite this, traditional teacher evaluation systems have done a poor job of differentiating teacher effectiveness (Weisberg et al. 2009). New models seek to rectify this problem, using student performance measures and standards-based observations to produce more rigorous assessments of teachers' practice, and hold teachers accountable for their students' learning. We present early evidence on the design of new evaluation systems in all fifty states, the twenty-five largest school districts, and Washington, DC. We describe how the components, process, and consequences of evaluation vary. Our findings reveal important differences in how evaluation is conducted across settings and for teachers with varied career status and assignment.

BACKGROUND ON TEACHER EVALUATION

Most new teacher evaluation systems incorporate measures of student achievement and observations of classroom instruction to assess teacher performance (NCTQ 2013; Hallgren, James-Burdumy, and Perez-Johnson 2014). The espoused goal of these new evaluation systems is to more closely tie the work of teachers to improvements in student learning (Darling-Hammond, Wise, and Pease 1983; Murphy, Hallinger, and Heck 2013). There are two approaches to satisfying the system's fundamental goal of improvement in student outcomes: (1) developing teachers' skills to improve student performance, and (2) evaluating teacher effectiveness for accountability purposes related to tenure, rewards, and dismissal.

COMPONENT MEASURES OF TEACHER PERFORMANCE

In the past, there was little indication that the measures used for teacher evaluation consistently supported instructional improvement or personnel decisions. Teacher evaluation has historically relied on administrators' observations of teachers' instruction using instruments that were not grounded in theory or research (Porter, Youngs, and Odden 2001) and student achievement played a limited role in teachers' evaluations (Peterson 2004).

Newly implemented teacher evaluation systems differ from their predecessors in two important ways. First, today's systems include standards-based observation protocols. Such protocols—most notably, Charlotte Danielson's Framework for Teaching (1996) and the National Board of Professional Teaching Standards—were developed and adopted by some states and districts in the 1990s (Milanowski 2004). Since 2009, these models have become much more widespread (Loup et al. 1996; Brandt et al. 2007). Observations of teachers' practices are included in evaluation systems based on the logic that evaluation should provide teachers with information to improve their instructional practice. The goal of these classroom observation measures is to more directly link teacher practice to instructional standards. Doing so may then increase student achievement (see, e.g., Danielson and McGreal 2000; Pianta and Hamre 2005).

Second, student outcomes based on state standardized tests and other achievement metrics are now included in many teacher evaluation systems (NCTQ 2013; Hallgren, James-Burdumy, and Perez-Johnson 2014). The inclusion of such measures in teachers' evaluations—through value-added measures (VAMs) which estimate a teacher's contribution to student achievement growth by controlling for prior student achievement and other student covariates (or student growth percentiles [SGPs], which compare a student's achievement growth to his/her peers with similar pre-test scores)—is based on the logic that teachers should be assessed on the extent to which their actions produce student learning.

Beginning in 2009, Race to the Top required that states (and later districts) weigh student outcomes heavily in teacher evaluation. In response, many of these sites now devote a much larger share of teachers' evaluation scores to student outcomes than they did a decade ago (NCTQ 2013). Some states have chosen to use VAMs in their systems. The benefits of VAMs have been discussed at length (see, e.g., Glazerman et al. 2010). Chetty, Friedman, and Rockoff (2014a, b) find that VAMs predict students' long-term outcomes, including college attendance and adult earnings. The drawbacks of such measures, including their instability and limited usefulness for new teachers, have also been documented (see, e.g., Baker et al. 2010; Papay 2011). A substantial number of states use teacher-developed measures of student performance known as student learning objectives (SLOs) (Lacireno-Paquet, Morgan, and Mello 2014). SLOs are subject- and grade-specific student learning goals based on state or national standards or teacher- or district-established goals. A student's progress toward meeting SLOs may be evaluated through a variety of measures, including externally developed or district-developed tests or other school-based tests and classroom assessments (Lacireno-Paquet, Morgan, and Mello 2014). Policy makers have argued that such measures increase teacher buy-in and have greater potential to motivate teachers than do VAMs (Locke and Latham 2002).

Single measures of instructional quality differ in their ability to identify effective teachers (see, e.g., Garrett and Steinberg 2015; McCaffrey et al. 2009; Goldhaber and Hansen 2010; Papay 2011). New evidence from the Measures of Effective Teaching study, however, suggests composite measures of teacher performance offer promise. When used in combination, classroom observation scores, student surveys of teacher practice, and value-added scores based on student achievement data (with VAM scores receiving the greatest weight in the composite measure) have been shown to identify effective teachers (Kane et al. 2013).

THE PROCESS OF EVALUATION

Under past evaluation systems, tenured teachers were rarely observed, and thus evaluation did not generally provide them with feedback that might improve their practice (Peterson 2004; Weisberg et al. 2009). Moreover, the infrequency of evaluations, especially for tenured teachers, interfered with their accountability function. New evaluation systems include more frequent observation and observers are required to provide structured feedback based on observation rubrics (Hallgren, James-Burdumy, and Perez-Johnson 2014). These features may better support the use of evaluation not just for instructional improvement but also for personnel decisions. More frequent observation of teacher practice using observation rubrics can provide more information

about a teacher's instructional performance than ever before. Although some research suggests that principals' managerial activities have a stronger relationship with school performance than do their instructional leadership tasks (Grissom and Loeb 2011), this information can help school leaders target professional development, improve school performance (Steinberg and Sartain 2015), and inform retention decisions (Sartain and Steinberg 2016).

THE CONSEQUENCES OF EVALUATION

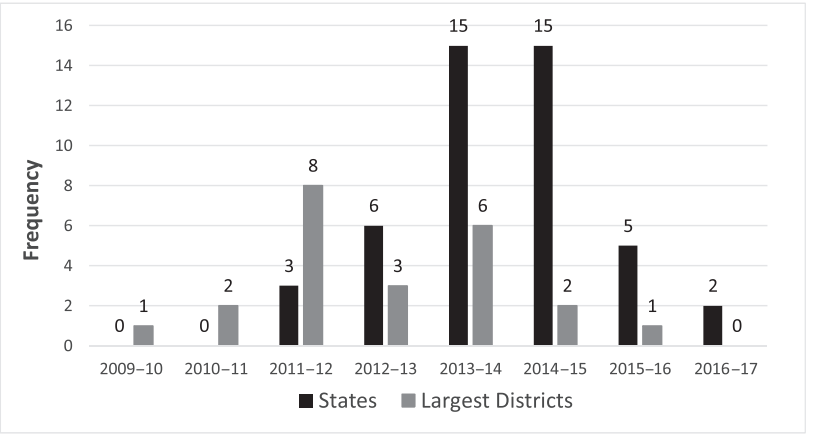
In the past, there was little evidence that evaluation produced tangible consequences for teachers or schools (Peterson 2004). Most district and school leaders did not link teacher evaluation and professional development, thus limiting instructional improvement (Stronge and Tucker 2003), and evaluation identified few teachers for termination or nonrenewal (Bridges 1992; Jacob 2011; Weisberg et al. 2009). Evidence on the relationship between pay for performance and student achievement is mixed (see, e.g., Figlio and Kenny 2007; Fryer 2011; Goldhaber and Walch 2012; Imberman and Lovenheim 2014; Springer et al. 2010). Policy makers have attempted to couple evaluation ratings with professional development, termination, or merit pay (Center on Great Teachers and Leaders, 2014; Hallgren, James-Burdumy, and Perez-Johnson 2014). We know little about the extent to which states have attached these consequences to evaluation reform, however.

Observation systems grounded in standards-based evidence of instruction and supported by frequent observations can improve teacher practice and student performance (Steinberg and Sartain 2015; Taylor and Tyler 2012). Early evidence from settings incorporating multiple measures of teacher performance indicates that such systems can lead to improvements in teacher performance while attaching real consequences to teachers' ratings (Dee and Wyckoff 2013), although there is little systematic evidence regarding the content, process, and consequences of new evaluation systems.

DOCUMENTING TEACHER EVALUATION REFORMS

We analyze teacher evaluation policies from the fifty states, largest twenty-five districts, and Washington, DC. We chose to analyze policies from all states that have recently implemented evaluation reform. Policy debates on teacher evaluation have also focused on systems developed in large cities, such as Washington, DC, New York, and Chicago (see, e.g., Sawchuk 2013). Thus, we chose to sample the twenty-five largest districts (based on student enrollment) and Washington, DC, an early, prominent adopter of evaluation reform (see tables A.1 and A.2).

Data collection proceeded in an iterative fashion (Creswell 2013). Beginning in December 2013, we reviewed policy scans, major reviews of the research on teacher evaluation in U.S. schools, and recent research on this topic to develop a matrix that incorporated salient aspects of the components, process, and consequences of teacher evaluation likely to be present in state and district policies (Darling-Hammond, Wise, and Pease 1983; Peterson 2004; Donaldson 2009; Weisberg et al. 2009; NCTQ 2013; Donaldson and Papay 2015; Steinberg and Sartain 2015). We then piloted the matrix by gathering key data from teacher evaluation documents pertaining to Connecticut's state teacher evaluation policy (see www.connecticutseed.org) and documents describing the



Notes: The year reflects the first year of scheduled statewide or district-wide implementation of teacher evaluation reform. There are 46 states and 23 districts included. AL, CA, IA, and TX are excluded from the state time trend; Cypress-Fairbanks (TX), Montgomery County (MD), and San Diego Unified (CA) are excluded from the largest districts trend.

Figure 1. Implementation Timing of Teacher Evaluation Reforms.

teacher evaluation policy of New Haven, Connecticut (see www.nhps.net/node/1082). Through this pilot process, we added categories and refined our matrix for broader use. We then used the matrix to gather and analyze data from publicly available documents pertaining to teacher evaluation, including state statutes, teacher evaluation policy Web sites, and policy handbooks, which often detailed the new policy for educators. At the district level, we also examined school board policies. This process occurred between December 2013 and May 2014, and involved reading and re-reading all data sources, verifying interpretations of data with multiple members of the research team and, where excerpts were vague, confirming interpretations with state or district policy makers.

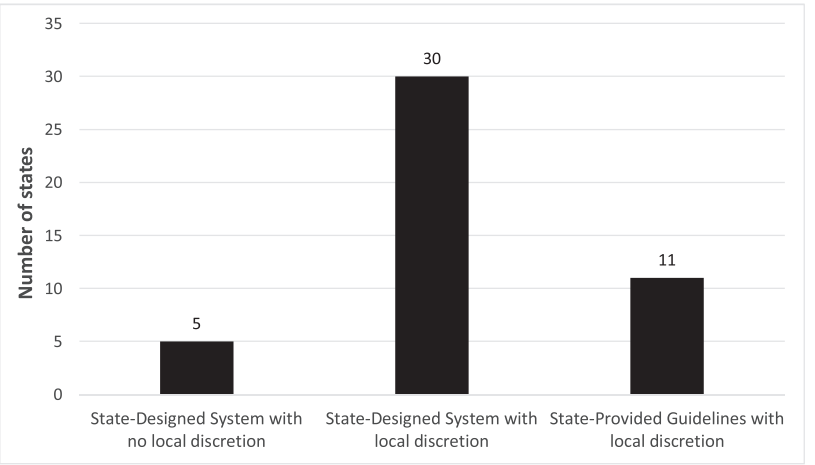
FINDINGS

Overall, we find that classroom observation remains the dominant method of evaluation in U.S. schools. We also find key differences in the content, processes, and consequences of teacher evaluation between states and the largest districts, between tenured and nontenured teachers, and between teachers of tested and nontested grades and subjects.

IMPLEMENTATION OF TEACHER EVALUATION SYSTEMS

The timing of state and district-level implementation of new teacher evaluation systems varies. The largest school districts are implementing teacher evaluation reforms earlier than state-level implementation requirements. By the end of the 2013–14 school year, 77 percent (20 of 26) of the largest school districts and Washington, DC implemented teacher evaluation reforms, whereas only 48 percent (24 of 50) of states had done so (see figure 1).

Figure 2 summarizes the distribution of states across three categories capturing whether the state designed a formal evaluation model and the extent of local flexibility for implementation. Of the 46 newly implemented state evaluation systems, most



Notes: There are 46 states included in the figure. The following states have designed new teacher evaluation systems that local school districts are mandated to implement without local discretion: DE, GA, HI, MS, and WV. The following states have designed teacher evaluation systems but districts have some local discretion in the choice of evaluation components, processes, and/or consequences: AZ, AR, CO, IN, KS, KY, LA, MD, MA, MI, MN, MO, MT, NV, NH, NJ, NM, NY, NC, OH, OK, PA, RI, SC, SD, TN, UT, WA, WI, and WY. The following states have provided guidelines for evaluating teachers but districts have discretion in determining the design of teacher evaluation reforms: AK, CT, FL, ID, IL, ME, NE, ND, OR, VT, and VA.

Figure 2. State Policy Design and Local District Discretion.

states (35) designed a formal model evaluation system. Of these 35 states, only 5 states mandated that districts implement the state-designed system without local discretion.[1] A majority of states (30) that designed model systems offered some degree of district discretion.[2] Fewer states (11) provided districts with a set of guidelines within which districts could design their systems.[3]

COMPONENTS OF TEACHERS' SUMMATIVE EVALUATION RATINGS

We construct two approaches to examine the weights that states and districts assign to the component measures of teachers' summative evaluation ratings. The first is a system-level approach that averages the component weights across states and across the largest districts (see table 1). The second is a teacher-level approach whereby the component weights are weighted by the number of teachers at the state and at the

1. For example, beginning in the 2012–13 school year, all school districts and charter schools in Delaware were mandated to implement the Delaware Performance Appraisal System, and 2013–14 marked the first statewide implementation of Hawaii's Education Effectiveness System.
2. For example, Act 82 of the Pennsylvania School Code mandated that 20 percent of a teacher's summative evaluation be based on elective data (such as nationally recognized standardized tests, student portfolios, or student projects) that are locally developed and selected by the school district from a list approved by the Pennsylvania Department of Education.
3. For example, the Code of Virginia, which requires that school boards' procedures for evaluating teachers incorporate student academic progress, allows local school boards to determine how this requirement is satisfied. Moreover, Virginia provides (via its *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers*), a list of recommended—though, importantly, not mandated—components of a teacher's summative evaluation rating.

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

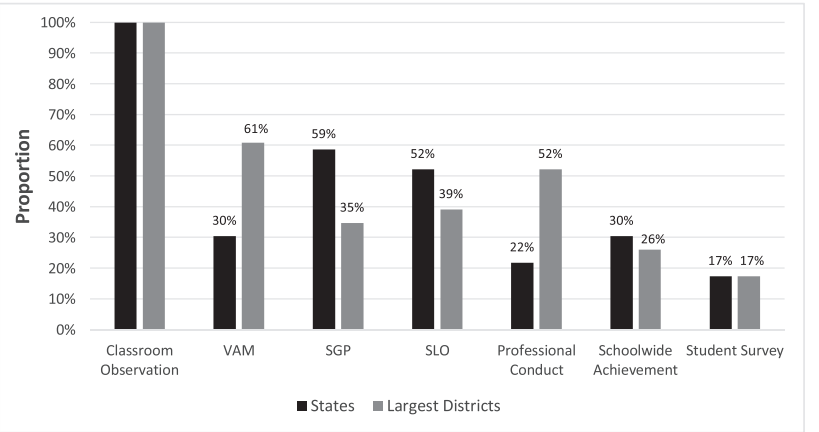
Table 2. Teacher-Level Component Weights of New Teacher Evaluation Systems

	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.

district levels (see table 2). This teacher-level approach reflects how the typical teacher across states and across the largest districts is evaluated (please see Appendix B for details on the system- and teacher-level calculations).

We find that the classroom observation score is the most frequently used measure of teacher performance (see figure 3). Of the 46 states and 23 districts implementing new teacher evaluation systems, all incorporate classroom observation as a component of a teacher's summative evaluation rating. Classroom observation scores also represent the largest share of a teacher's summative rating. Across state and district settings,



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.

Figure 3. Frequency of Components of Teacher Evaluation Systems.

on average, 54 percent and 52 percent, respectively, of a teacher's rating is based on observation scores (see table 1, panel A). For the typical teacher in a tested grade/subject nationwide, observation scores represent 53 percent of their final rating; among the largest school districts, observation scores account for 52 percent of a typical tested teacher's summative rating (see table 2, panel A).

Classroom observation rubrics such as the Danielson Framework for Teaching[4] include a professional conduct domain as part of a teacher's classroom observation score. Fifty percent of the largest districts and approximately 20 percent of states incorporate a separate measure of a teacher's professional conduct into the summative evaluation rating. Across state and district settings, on average, 3 percent and 5 percent, respectively, of a teacher's rating is based on a separate measure of professional conduct. Nationwide, the typical teacher in a tested grade/subject will have 2 percent of her summative rating based on professional conduct; for the typical tested teacher teaching in one of the largest school districts, professional conduct represents 5 percent of her summative evaluation rating.

Consistent with the nationwide emphasis on incorporating student test score data into a teacher's summative evaluation score, 80 percent of both states and largest districts use one or more measures of teacher performance based on student test score data.[5] Notably, the largest districts appear to prefer VAMs of teacher performance whereas states are more likely to use SGPs. Approximately 60 percent (14) of the largest districts and only 30 percent of states use VAMs, whereas 35 percent (8) of the largest districts and 60 percent (27) of states use SGPs. States and the largest districts also vary in the weights they assign to each measure. Across states, 10 percent

4. The Danielson Framework for Teaching reserves Domain 4 for "Professional Responsibilities," of which one component (4f) evaluates teachers based on "Showing Professionalism."
5. Of the 46 states with newly implemented evaluation systems, 9 use VAM only, 22 use SGP only, and 5 use both VAM and SGP to estimate student achievement growth in teachers' evaluations. Of the 23 districts, 12 use VAM only, 6 use SGP only, and 2 use both VAM and SGP.

of a teacher's summative evaluation rating, on average, is based on VAMs, whereas VAMs account for 22 percent, on average, of a teacher's summative evaluation rating across the largest school districts. For the typical tested teacher nationwide, 16 percent of their summative rating will depend on VAMs, compared with 22 percent for the typical tested teacher in the nation's largest school districts. These patterns are reversed for SGP, which contributes more to the summative evaluation ratings of tested teachers in states (13 percent on average) than the largest districts (7 percent on average).

Examining SLOs, we find that 52 percent (24) of states and 39 percent (9) of the largest districts use these measures, of which 9 states (and 1 district) use SLOs as their only method of estimating student growth. Across states, 10 percent of a teacher's summative evaluation rating, on average, is based on SLOs, and SLOs account for 8 percent, on average, of a teacher's summative evaluation rating across the largest school districts.

The component weights described here pertain to teachers who teach in tested grades and subjects (for example, math and reading in grades 3–8) where students take a state-mandated, end-of-year exam. Upward of 70 percent of teachers nationwide teach in nontested grades or subjects (Watson, Kraemer, and Thorn 2009), however, and, in principle, will not have student test score data from which to generate a VAM or SGP.[6] We find similar patterns in how states and districts assign component weights for teachers of tested and nontested grades and subjects.[7] In both cases—state and district comparisons of teachers in tested and nontested grades and subjects—a larger share of a nontested teacher's summative evaluation rating depends on SLOs and a schoolwide achievement measure[8] than for teachers in tested grades/subjects. At the state level, whereas 11 percent and 4 percent, on average, of tested teachers' summative evaluation ratings depend on SLOs and schoolwide achievement, respectively, these weights increase to 25 percent and 7 percent for nontested teachers (see table 1, panel B). For the largest districts, 10 percent and 4 percent, on average, of tested teachers' summative evaluation ratings depend on SLOs and schoolwide achievement, respectively, and these weights increase to 16 percent and 12 percent for nontested teachers (see table 1, panel C). And, although states weigh the classroom observation score similarly for tested (52 percent) and nontested teachers (54 percent), the largest districts give

6. We do find that some states and largest districts require VAM or SGP for teachers in nontested grades/subjects, and use either locally developed or nationally normed exams of student achievement to generate VAM or SGP. See table 1 for more detail on which states and districts utilize these measures for nontested teachers.
7. First, we looked at the component weights for teachers in tested and nontested grades/subjects across the 46 states. We note here that there are 14 states that do not specify whether (and how) the component weights differ for teachers in tested versus nontested grades/subjects. In particular, these states either do not provide any detail on how the component weighting should change, leave the decision up to the local district, or have not yet made a final determination regarding the state's approach to measuring performance for teachers in nontested grades/subjects. We then compared tested and nontested teachers across 17 (of 23) districts that distinguished the component measures of a teacher's summative evaluation rating for these two groups of teachers.
8. In Pennsylvania, for example, the schoolwide achievement measure includes factors such as indicators of academic achievement (such as the percent proficient or advanced on state exams), indicators of closing the achievement gap (for all students and by subgroups, such as race and socioeconomic status), academic growth based on the state's VAM (Pennsylvania Value Added Assessment System), other academic indicators such as graduation rates, promotion rates, attendance, Advanced Placement or International Baccalaureate, PSAT participation, as well as credit for advanced achievement (on state exams).

Table 3. Frequency of Classroom Observations, by Locale and Career Status

	Career Teachers		Beginning Teachers	
Locale	Formal Observations	Informal Observations	Formal Observations	Informal Observations
States	1.7 (0.72)	2.1 (1.31)	2.1 (0.78)	2.2 (1.33)
Largest districts	1.8 (1.08)	2.7 (2.19)	2.5 (1.19)	3.0 (2.28)
Home states	1.6 (0.49)	2.3 (1.25)	2.3 (0.64)	2.5 (1.5)

Notes: Mean (standard deviation) number of observations reported. *Beginning* teachers are early career teachers in their first three years of teaching and *Career* teachers are in at least their fourth year of teaching. We include data, where available, from all 46 states, all 23 of the largest districts, and the 10 home states (FL, GA, IL, MD, NV, NY, NC, PA, TN, and VA) that have implemented new teacher evaluation systems. We report the weighted means for the largest districts' home states (the weighted means account for the fact that some districts reside in the same state; e.g., if one state houses five districts, the state is counted five times when calculating the mean component weightings). For states and districts reporting a range for a given observation type, we used the midpoint of the range to calculate the mean occurrences, by locale and career status.

greater weight (on average, 58 percent) to the classroom observation score for nontested teachers compared with their tested counterparts (with, on average, 53 percent of their summative rating based on classroom observation scores).

The inclusion of student test score data in teachers' summative evaluation ratings has received significant attention from both scholars and the public. Our findings suggest that test-score–based measures of teacher performance play a limited role in the evaluation of a typical teacher nationwide and in the largest school districts, with no more than 13 percent of the evaluation rating of a typical teacher nationwide depending on VAMs and/or SGPs. For the typical teacher teaching in one of the nation's largest school districts, at most 10 percent of their evaluation rating will reflect student test score performance (see table 2, panel B).

OBSERVATION AND EVALUATION PROCESS

As discussed earlier, classroom observations of a teacher's instructional practice are a component of the summative evaluation rating in all sites (see figure 3). States with new evaluation systems require, on average, two formal observations for both career and beginning teachers; states also require, on average, two informal observations for both career and beginning teachers (see table 3).[9] The largest districts require about the same number of formal observations (two) for career teachers as do the states, but approximately 0.5 more formal observations for beginning teachers than the states require. Moreover, the largest districts require nearly one more informal observation for both career and beginning teachers than the states require. Therefore, on average, the largest districts require a total of 4.5 observations for career teachers, and 5.5 observations for beginning teachers, and states require, on average, approximately 4 total observations for both career and beginning teachers.

9. In most cases, a formal observation lasts at least thirty minutes, and observers gather and record evidence of a teacher's instructional practice guided by an observation rubric. Informal observations can range from fifteen-minute drop-ins to short walkthroughs, and data collection is often not required.

Table 4. Teacher-Observer Conferences, by Locale and Career Status

	Career Teachers				Beginning Teachers			
Locale	Mid-Year Conference	Summative Conference	Pre-Observation Conference	Post-Observation Conference	Mid-Year Conference	Summative Conference	Pre-Observation Conference	Post-Observation Conference
States	0.26	0.54	0.50	0.54	0.28	0.54	0.50	0.54
Largest districts	0.22	0.57	0.52	0.70	0.22	0.57	0.57	0.70
Home states	0.45	0.70	0.35	0.30	0.50	0.70	0.35	0.30

Notes: The proportion of states, largest districts, and their home states utilizing teacher-observer conferences are reported. The state proportions are out of 46, the district proportions are out of 23, and the home states proportions are out of 20. The home states calculations only apply to the 10 home states with newly implemented evaluation policies, as follows: FL has 7 of the largest districts; MD has 3; GA and NC each have 2; and IL, NV, NY, PA, TN, and VA each have one of the largest districts. We exclude TX and CA, which have 2 and 1, respectively, of the largest districts with new teacher evaluation policies.

Table 5. 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.

Emerging evidence suggests that ongoing conferences between observers and teachers are critical components of the evaluation process, providing opportunities for the type of formative feedback that can improve instruction and student achievement (Taylor and Tyler 2012; Steinberg and Sartain 2015). Within states and the largest school districts, we find that there is a consistent distribution of conferences, by type (mid-year, summative, pre-observation, post-observation), across beginning and career teachers (see table 4). Whereas states and the largest districts require summative conferences at very similar rates (54 percent of states and 57 percent of the largest districts), however, a larger share of districts (70 percent) requires teacher-observer post-observation conferences than do states (54 percent).

EVALUATION SYSTEM CONSEQUENCES

A major criticism of traditional evaluation systems is that few formal consequences resulted from teachers' summative evaluation ratings. Our findings suggest that policy makers take a developmental stance toward evaluation. Most state (83 percent) and district (74 percent) policies link professional development to a teacher's summative evaluation rating (see table 5). This applies to all teachers, but the requirements for underperforming teachers are particularly explicit. These professional development opportunities are often structured using a professional development plan, crafted by the teacher and their observer (usually the school's principal) that delineates targeted goals for the teacher. Professional development is often a consequence of underperformance and low summative ratings, but some states require professional development for

teachers who underperform on individual components of the evaluation process.[10] A much smaller share of states (20 percent) and districts (21 percent) provide merit-based rewards for a teacher's summative evaluation rating.

DISCUSSION

Our findings reveal how states and districts have emphasized particular features of teacher evaluation reform efforts. First, like others (Whitehurst, Chingos, and Lindquist 2014), we find that classroom observation remains a major component in district and state evaluation systems but constitutes a smaller share of teachers' overall ratings than in the past (Peterson 2004). Second, consistent with recent policy mandates, we find that almost all states and districts that revised their evaluation system include some measure of student performance in teachers' evaluations.

Moreover, we find there are important differences between the evaluation systems of the largest districts in the United States and those in place in smaller districts. First, large districts tended to implement evaluation reform earlier—in some cases considerably so—than states. Indeed, a determined and resourceful superintendent may be able to enact district-level reforms more easily than a state commissioner of education who works with a broader, more diverse, and more diffuse constituency. Second, the process of evaluation differs across sites. Districts require more informal and formal observations than do states. They also emphasize post-observation conferences more than states. These demands on administrators may be particularly challenging in urban districts, which may have fewer resources than their nonurban counterparts.

We find striking differences in how teachers of tested and nontested grades/subjects are evaluated. Specifically, school average performance and SLOs constitute a greater proportion of nontested teachers' ratings than they do for tested teachers' ratings. The difference in weights assigned to SLOs for tested versus nontested teachers is greatest in the state sample, and the difference in weights assigned to schoolwide achievement for tested versus nontested teachers is greatest in the largest districts.

A schoolwide measure of teacher effectiveness may have little (if anything) to do with nontested teachers' own individual performance. Though the weight assigned to schoolwide achievement for nontested teachers in the largest school districts (12.3 percent) appears modest in magnitude, it represents the third most important component of their summative evaluation ratings. Additionally, depending on how schoolwide measures are calculated, the performance ratings for highly effective nontested teachers in large districts may reflect much poorer quality instruction than their own. Further, more-disadvantaged, lower-performing schools may find it more difficult to retain higher-performing teachers, given the potential adverse effect on the ratings of high-performing teachers. Finally, because there is no within-school variation on a schoolwide achievement measure, it provides no additional information about teacher performance. Therefore, other components, such as the observation score, will

10. In Delaware, an early RTT grantee, an Improvement Plan is required for teachers who receive one of two (of four) lowest summative evaluation ratings (e.g., "Ineffective" or "Needs Improvement") and who underperform on one of the components (e.g., classroom observation, VAM) of their summative evaluation, irrespective of their overall summative evaluation rating. In addition, Delaware's plan encourages evaluators to develop an improvement plan for teachers whose performance is unsatisfactory during an individual classroom observation.

represent a larger effective share of nontested teachers' summative evaluation ratings than for teachers in tested grades/subjects.

These findings raise a number of important questions about the design of new evaluation systems. First, policy makers might give additional consideration to the relationship between measurement and motivation (Firestone 2014). For example, what are the implications of increasing the weight given to schoolwide measures of student performance for nontested teachers? A schoolwide measure of student performance may be relatively simple to generate, but may have little influence (or even a negative impact) on teachers' motivation. Policy makers might also give additional thought to the practical implications of these new systems. Time is one of the most consistently cited obstacles to high quality evaluation (Krajewski 1978; Tucker 1997; Peterson 2000; Kimball 2002; Donaldson et al. 2014). New systems, particularly those in large districts, place considerable demands on evaluators to carry out the tasks involved with evaluation. Do districts have the human capital to implement the systems as intended, let alone to leverage these systems to drive instructional improvement?

Our findings also suggest areas for further research. First, our finding that the components of teacher evaluation and their weights vary suggests that research should investigate the implications of these decisions. How do measures of teacher performance based on student achievement, such as VAM, SGP, and SLO, differentially influence a teacher's summative evaluation rating? How do their psychometric properties compare? How do they perform as sources of formative instructional feedback? How do school leaders use them in their work to support and motivate teachers to improve? How do teachers of tested and nontested subjects view the fairness of the ratings formula? Second, our analysis underscores key differences in the evaluation process specified by newly implemented evaluation systems. Little research investigates the efficacy of these processes. Is it more efficacious to prioritize ongoing observation debriefs (such as post-observation teacher-observer conferences), as the large districts do, or the summative conference, as the states emphasize? What number of observations is most effective? Should the number differ for tenured and nontenured teachers? There has been little research on these questions. Third, studies on the consequences of new teacher evaluation systems are warranted. How often do these new systems lead to a teacher's dismissal or revocation of tenure? How is professional development best structured to meet individual teachers' needs while supporting schoolwide and districtwide improvement? Lastly, our study reveals key differences between the policies of the largest districts and the states. Research on new teacher evaluation systems has tended to focus on a small number of large districts (i.e., Taylor and Tyler 2012; Dee and Wyckoff 2013; Steinberg and Sartain 2015); our research suggests that studies in more typical districts are important.

CONCLUSION

In this policy brief, we present one of the first systematic policy analyses of newly developed and implemented teacher evaluation systems. We examine the components, processes, and consequences of teacher evaluation enshrined in the policies of all states, the largest twenty-five school districts, and Washington, DC. In so doing, we highlight how states and the largest districts in the country have built different assumptions

into their evaluation systems, emphasizing different features of an evaluation system aimed at generating the greatest returns in terms of teacher performance and student achievement. This work provides a foundation for policy makers and practitioners to assess various systems, and for researchers to develop important lines of inquiry to inform the nascent and ongoing implementation of teacher evaluation reform.

ACKNOWLEDGMENTS

The authors thank John Papay and William Firestone for valuable feedback on earlier versions of this paper, and Filippo Bulgarelli for excellent research assistance. We gratefully acknowledge funding from University of Pennsylvania's Undergraduate Urban Research Colloquium (UURC). Authors contributed equally to this article.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25(1):95–135. doi:10.1086/508733

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A Shepard. 2010. Problems with the use of student test scores to evaluate teachers. Washington, DC: Economic Policy Institute Briefing Paper 278.

Brandt, Chris, Carrie Mathers, Michelle Oliva, Melissa Brown-Sims, and Jean Hess. 2007. *Examining district guidance to schools on teacher evaluation policies in the Midwest Region*. Available http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_2007030.pdf. Accessed 25 February 2008.

Bridges, Edwin M. 1992. *The incompetent teacher: Managerial responses*, 2nd ed. Philadelphia, PA: Falmer.

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

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

Center on Great Teachers and Leaders. 2014. *National picture: A different view*. Available www.gtlcenter.org/sites/default/files/42states.pdf. Accessed 1 December 2015.

Creswell, John W. 2013. *Qualitative inquiry and research design*. Los Angeles, CA: Sage.

Danielson, Charlotte. 1996. *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Danielson, Charlotte, and Thomas L. McGreal. 2000. *Teacher evaluation to enhance professional practice*. Alexandria, VA: ASCD.

Darling-Hammond, Linda, Arthur E. Wise, and Sara R. Pease. 1983. Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research* 53(3):285–328. doi:10.3102/00346543053003285

Dee, Thomas, and James Wyckoff. 2013. Incentives, selection, and teacher performance: Evidence from IMPACT. NBER Working Paper No. 19529.

Donaldson, Morgaen L. 2009. *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Washington, DC: Center for American Progress.

Donaldson, Morgaen L., and John Papay. 2015. Teacher evaluation for accountability and development. In *Handbook of research in education finance and policy*, edited by Helen F. Ladd and Margaret E. Goertz, pp. 174–193. New York: Routledge.

Donaldson, Morgaen L., Casey D. Cobb, Kimberly LeChasseur, Rachael Gabriel, Richard Gonzales, Sarah Woulfin, and Aliza Makuch. 2014. *An evaluation of the pilot implementation of Connecticut's system for educator evaluation and development*. Storrs, CT: Center for Education Policy Analysis.

Figlio, David N., and Lawrence W. Kenny. 2007. Individual teacher incentives and student performance. *Journal of Public Economics* 91(5–6):901–914. doi:10.1016/j.jpubeco.2006.10.001

Firestone, William A. 2014. Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher* 43(2):100–107. doi:10.3102/0013189X14521864

Fryer, Roland G. 2011. Teacher incentives and student achievement: Evidence from New York City public schools. NBER Working Paper No. 16850.

Garrett, Rachel, and Matthew P. Steinberg. 2015. Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis* 37(2):224–242. doi:10.3102/0162373714537551

Glazerman, Steven, Dan Goldhaber, Susanna Loeb, Stephen Raudenbush, Douglas Staiger, and Grover J. "Russ" Whitehurst. 2010. *Evaluating teachers: The important role of value-added*. Available www.brookings.edu/research/reports/2010/11/17-evaluating-teachers. Accessed 31 December 2015.

Goldhaber, Dan. 2002. The mystery of good teaching. *Education Next* 2(1):50–55.

Goldhaber, Dan, and Michael Hansen. 2010. Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. CALDER Working Paper No. 31, Urban Institute.

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

Grissom, Jason A., and Susanna Loeb. 2011. Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal* 48(5):1091–1123. doi:10.3102/0002831211402663

Hallgren, Kristen, Susanne James-Burdumy, and Irma Perez-Johnson. 2014. *State requirements for teacher evaluation policies promoted by Race to the Top*. Available www.mathematica-mpr.com/~/media/publications/PDFs/education/rtt_ies_brief.pdf. Accessed 1 December 2015.

Imberman, Scott A., and Michael F. Lovenheim. 2014. Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics* 29(2):364–386. doi:10.1162/REST_a_00486.

Jacob, Brian A. 2011. Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis* 33(4):403–434. doi:10.3102/0162373711414704

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kimball, Steven M. 2002. Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education* 16(4):241–268. doi:10.1023/A:1021787806189

Krajewski, Robert J. 1978. Secondary principals want to be instructional leaders. *Phi Delta Kappan* 60(1):65.

Lacireno-Paquet, Natalie, Claire Morgan, and Daniel Mello. 2014. *How states use student learning objectives in teacher evaluation systems: A review of state websites* (REL 2014–013). Available http://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=380. Accessed 1 December 2015.

Locke, Edwin A., and Gary P. Latham. 2002. Building a practically useful theory of goal setting and work motivation: A 35-year odyssey. *American Psychologist* 57(9):705–717. doi:10.1037/0003-066X.57.9.705

Loup, Karen S., Joanne S. Garland, Chad D. Ellett, and John K. Rugutt. 1996. Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education* 10(3):203–226. doi:10.1007/BF00124986

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. The inter-temporal variability of teacher effect estimates. Nashville, TN: National Center for Performance Incentives Working Paper Series No. 2009–03.

Milanowski, Anthony. 2004. The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education* 79(4):33–53. doi:10.1207/s15327930pje7904_3

Murphy, Joseph, Philip Hallinger, and Ronald H. Heck. 2013. Leading via teacher evaluation: A case of the missing clothes? *Educational Researcher* 42(6):349–354. doi:10.3102/0013189X13499625

National Council on Teacher Quality (NCTQ). 2013. *State of the states 2013 connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: NCTQ.

Papay, John P. 2011. Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal* 48(1):163–193. doi:10.3102/0002831210362589

Peterson, Kenneth D. 2000. *Teacher evaluation: A comprehensive guide to new directions and practices*, 2nd ed. Thousand Oaks, CA: Corwin.

Peterson, Kenneth. 2004. Research on school teacher evaluation. *NASSP Bulletin* 88(639):60–79. doi:10.1177/0192636504088639O6

Pianta, Robert C., and Bridget K. Hamre. 2005. *Classroom assessment scoring system, secondary manual*. Charlottesville, VA: Teachstone Training.

Porter, Andrew C., Peter Youngs, and Allan Odden. 2001. Advances in teacher assessment and their uses. In *Handbook of research on teaching*, edited by Virginia Richardson, pp. 259–297. New York: Macmillan.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2):417–458. doi:10.1111/j.1468-0262.2005.00584.x

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement. *American Economic Review* 94(2):247–252. doi:10.1257/0002828041302244

Sartain, Lauren, and Matthew P. Steinberg. 2016. Forthcoming. Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources*. doi:10.3368/jhr.51.3.0514-6390R1.

Sawchuk, Stephen. 2013. *Chicago teachers see value in new evaluations, but eschew test scores*. Available http://blogs.edweek.org/edweek/teacherbeat/2013/09/chicago_teachers_see_value_in.html. Accessed 1 December 2015.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Santa Monica, CA: RAND Corporation.

Steinberg, Matthew, and Lauren Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy* 10(4):535–572. doi:10.1162/EDFP_a_00173

Stronge, James H., and Pamela D. Tucker. 2003. *Teacher evaluation: Assessing and improving performance*. Larchmont, NY: Eye on Education.

Taylor, Eric S., and John H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7):3628–3651. doi:10.1257/aer.102.7.3628

Tucker, Pamela D. 1997. Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education* 11(2):103–126. doi:10.1023/A:1007962302463

Watson, J. G., S. B. Kraemer, and C. A. Thorn. 2009. *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Whitehurst, Grover J. (Russ), Matthew M. Chingos, and Katherine Lindquist. 2014. Evaluating teachers with classroom observations: Lessons learned in four districts. Washington, DC: Brown Center on Education Policy, Brookings Institution.

APPENDIX A: STATE AND DISTRICT SUMMARY

Table A.1. State Summary

State	Race to the Top (RTT) Status	Teachers	Student Enrollment	Year Teacher Evaluation Reform Implemented
Alabama (AL)	Applicant (Phase 1 & 2)	47,723	744,621	2009+
Alaska (AK)	Non-applicant	8,088	131,167	2015
Arizona (AZ)	RTT ($25,080,554)	50,800	1,080,319	2013
Arkansas (AR)	Applicant (Phase 1)	33,983	483,114	2014
California (CA)	Finalist (Phase 2)	268,689	6,287,834	1971+
Colorado (CO)	RTT ($17,946,236)	48,078	854,265	2013
Connecticut (CT)	Applicant (Phase 1 & 2)	43,805	554,437	2013
Delaware (DE)	RTT ($119,122,128)	8,587	128,946	2012
Florida (FL)	RTT ($700,000,000)	175,006	2,668,156	2011

Table A.1. Continued.

State	Race to the Top (RTT) Status	Teachers	Student Enrollment	Year Teacher Evaluation Reform Implemented
Georgia (GA)	RTT ($399,952,650)	111,133	1,685,016	2014
Hawaii (HI)	RTT ($74,934,761)	11,458	182,706	2013
Idaho (ID)	Applicant (Phase 1)	15,990	279,873	2013
Illinois (IL)	RTT ($42,818,707)	131,777	2,083,097	2016
Indiana (IN)	Applicant (Phase 1)	62,339	1,040,765	2012
Iowa (IA)	Applicant (Phase 1 & 2)	34,658	495,870	2001+
Kansas (KS)	Applicant (Phase 1)	37,407	486,108	2014
Kentucky (KY)	RTT ($17,037,544)	41,860	681,987	2014
Louisiana (LA)	RTT ($17,442,972)	48,657	703,390	2012
Maine (ME)	Applicant (Phase 2)	14,888	188,969	2015
Maryland (MD)	RTT ($249,999,182)	57,589	854,086	2013
Massachusetts (MA)	RTT ($250,000,000)	69,342	953,369	2013
Michigan (MI)	Applicant (Phase 1 & 2)	86,997	1,573,537	2015
Minnesota (MN)	Applicant (Phase 1)	52,832	839,738	2014
Mississippi (MS)	Applicant (Phase 2)	32,007	490,619	2014
Missouri (MO)	Applicant (Phase 1 & 2)	66,252	916,584	2013
Montana (MT)	Applicant (Phase 2)	10,153	142,349	2014
Nebraska (NE)	Applicant (Phase 1 & 2)	22,182	301,296	2014
Nevada (NV)	Applicant (Phase 2)	21,132	439,634	2014
New Hampshire (NH)	Applicant (Phase 1)	15,049	191,900	2014
New Jersey (NJ)	RTT ($37,847,648)	109,719	1,356,431	2013
New Mexico (NM)	Finalist (Phase 1 & 2)	21,957	337,225	2013
New York (NY)	RTT ($696,646,000)	209,527	2,704,718	2012
North Carolina (NC)	RTT ($399,465,769)	97,308	1,507,864	2011
North Dakota (ND)	Non-applicant	8,525	97,646	2015
Ohio (OH)	RTT ($400,000,000)	107,972	1,740,030	2013
Oklahoma (OK)	Applicant (Phase 1 & 2)	41,349	666,120	2013
Oregon (OR)	Non-applicant	26,791	568,208	2014
Pennsylvania (PA)	RTT ($41,326,299)	124,646	1,771,395	2013
Rhode Island (RI)	RTT ($75,000,000)	11,414	142,854	2012
South Carolina (SC)	Finalist (Phase 2)	46,782	727,186	2014
South Dakota (SD)	Applicant (Phase 1)	9,247	128,016	2015
Tennessee (TN)	RTT ($500,741,220)	66,382	999,693	2011
Texas (TX)	Non-applicant	324,282	5,000,470	2004+
Utah (UT)	Applicant (Phase 1 & 2)	25,970	598,832	2014
Vermont (VT)	Non-applicant	8,364	89,908	2014
Virginia (VA)	Applicant (Phase 1)	90,832	1,257,883	2012
Washington (WA)	Applicant (Phase 2)	53,119	1,045,453	2013
West Virginia (WV)	Applicant (Phase 1)	20,247	282,870	2013
Wisconsin (WI)	Applicant (Phase 1 & 2)	56,245	871,105	2014
Wyoming (WY)	Applicant (Phase 1)	7,847	90,099	2016

Notes: Data on a state's RTT application and award status retrieved from U.S. Department of Education (www2.ed.gov/programs/racetothetop/index.html). The total award for RTT grantees listed in parentheses. The number of teachers represents the number of full-time equivalent (FTE) teachers for the 2011–12 school year and was retrieved from the National Center for Education Statistics (http://nces.ed.gov/ccd/elsi/). Student enrollment is for the 2011–12 school year, includes public school students in grades preK–12, and was also retrieved from NCES. The *Year Teacher Evaluation Reform Implemented* refers to the first year of scheduled statewide rollout of the new teacher evaluation system or state-mandated requirements that all districts implement a teacher evaluation system for all teachers (2010 refers to the 2010–11 school year). There are four states (+) that have not implemented or mandated district-implementation of new teacher evaluation systems: Alabama, California, Iowa, and Texas. For Alabama, although documentation indicates a new teacher evaluation policy was implemented in 2009–10, the state provides no description of the actual policy, and so we have chosen to exclude it from all calculations related to teacher evaluation system design, process, and consequences. Note that some states incorporate additional evaluation components (such as student test scores) and system consequences in years after the initial, statewide rollout.

Table A.2. Largest District Summary

District	State	Race to the Top District (RTT-D) Status	Teachers	Student Enrollment	Year Teacher Evaluation Reform Reform Implemented
New York City	New York	Applicant	62,368	968,143	2013
Los Angeles Unified	California	Applicant	28,769	659,639	2013
Chicago	Illinois	Finalist	22,460	403,004	2012
Dade	Florida	Non-applicant	21,117	350,239	2011
Clark County	Nevada	Non-applicant	14,822	313,398	2015
Broward	Florida	Applicant	14,533	258,478	2011
Houston	Texas	RTT-D ($29,999,782)	10,920	203,066	2011
Hillsborough	Florida	Applicant	13,862	197,041	2010
Orange	Florida	Applicant	11,308	180,000	2011
Fairfax County	Virginia	Non-applicant	13,878	177,606	2012
Palm Beach	Florida	Non-applicant	11,682	176,901	2011
Gwinnett County	Georgia	Non-applicant	10,324	162,370	2013
Dallas	Texas	Applicant	10,277	157,575	2014
Shelby County	Tennessee	Non-applicant	10,064	157,375	2011
Philadelphia	Pennsylvania	Applicant	9,299	154,262	2013
Wake County	North Carolina	Applicant	9,440	148,154	2010
Montgomery County	Maryland	Non-applicant	9,622	146,459	2001+
Charlotte-Mecklenburg	North Carolina	Applicant	8,791	141,728	2012
San Diego Unified	California	Non-applicant	6,706	131,044	Pre-1999+
Prince George's County	Maryland	Non-applicant	7,796	123,833	2013
Duval	Florida	Applicant	7,589	125,429	2011
Cypress-Fairbanks	Texas	Non-applicant	6,243	107,960	1999+
Cobb County	Georgia	Non-applicant	7,342	107,291	2014
Baltimore County	Maryland	Finalist	7,219	105,153	2013
Pinellas	Florida	Applicant	7,289	103,776	2011
District of Columbia (DC)	-	RTT ($74,998,962)	3,472	44,618	2009

Notes: Data on a district's RTT application and award status retrieved from U.S. Department of Education (www2.ed.gov/programs/racetothetop-district/index.html). The District of Columbia (DC) was awarded a RTT grant under the state (and not district) competition. The total award for RTT grantees listed in parentheses. The number of teachers represents the number of full-time equivalent teachers for the 2011–12 school year and was retrieved from the National Center for Education Statistics (http://nces.ed.gov/ccd/elsi/). Student enrollment is for the 2011–12 school year, includes public school students in grades preK–12, and was also retrieved from NCES. For Shelby County (TN), the number of teachers and student enrollment is inclusive of the former Shelby County Schools and Memphis City Schools districts, which were consolidated in 2013. The *Year Teacher Evaluation Reform Implemented* refers to the first year of scheduled districtwide rollout of the new teacher evaluation system for all teachers (2010 refers to the 2010–11 school year). For Duval (FL), we used system data from Florida's state-provided guidelines for evaluating teachers, which mandated that all districts implement new evaluation systems across the state for the first time in the 2011–12 school year. There are three districts (+) that have not implemented new teacher evaluation systems: San Diego Unified (CA), Montgomery County (MD), and Cypress-Fairbanks (TX), so we exclude these districts from all calculations related to teacher evaluation system design, process, and consequences. For San Diego, any formal updates to the district's evaluation system occurred prior to the 1999–2000 school year. For Montgomery County, the existing evaluation system, Peer Assistance and Review, has been in place since 2001.

APPENDIX B: CALCULATING COMPONENT WEIGHTS

Of the 46 states and 23 districts implementing evaluation reform, state (district) i reported a weight for component j in one of three ways: (a) $w_i^j = x_o$, where x_o is a specific value; (b) $x_o^{lower} < w_i^j < x_o^{upper}$, where x_o^{lower} is the value of the lower bound and x_o^{upper} is the value of the upper bound of state (district) i's reported range of weights for component j; or (c) $w_i^j > 0$. For states (districts) reporting a range of values (as in b), we calculate w_i^j in the following way: $w_i^j = \frac{x_o^{lower} + x_o^{upper}}{2}$. For states (districts) reporting

$w_i^j > 0$, we assign the following value: $w_i^j = \bar{w}^j = \sum_{i=1}^{N} \frac{w_i^j}{N_{i \in (a,b)}}$. For states (districts) that did not report a weight for component j, we assign a value of zero.

SYSTEM-LEVEL COMPONENT WEIGHTS

In table 1, panel A, we calculate system-level component weights in the following way: (i) $\bar{W}^j_{States} = \frac{\sum_{i=1}^{46} w_i^j}{46}$; (ii) $\bar{W}^j_{Largest\ Districts} = \frac{\sum_{i=1}^{23} w_i^j}{23}$. For panel B, we recalculate $\bar{W}^j_{States}$ by tested and nontested teacher status in the following way: (i) $\bar{W}^j_{States,\,Tested} = \frac{\sum_{i=1}^{32} w^j_{i,\,Tested}}{32}$; (ii) $\bar{W}^j_{States,\,NonTested} = \frac{\sum_{i=1}^{32} w^j_{i,\,Nontested}}{32}$. For panel C, we recalculate $\bar{W}^j_{Largest\ Districts}$ by tested and nontested teacher status in the following way: (i) $\bar{W}^j_{Largest\ Districts,\,Tested} = \frac{\sum_{i=1}^{17} w^j_{i,\,Tested}}{17}$; (ii) $\bar{W}^j_{Largest\ Districts,\,Nontested} = \frac{\sum_{i=1}^{17} w^j_{i,\,Nontested}}{17}$.

TEACHER-LEVEL COMPONENT WEIGHTS

In table 2, panel A, we calculate teacher-level component weights for the "typical tested teacher" across states and the largest districts implementing evaluation reform in the following way: (i) $\bar{\hat{W}}^j_{States} = \sum_{i=1}^{46} \hat{w}_i^j$, where $\hat{w}_i^j = \left(\frac{T_i}{\sum_{i=1}^{46} T_i}\right) * w_i^j$; (ii) $\bar{\hat{W}}^j_{Largest\ Districts} = \sum_{i=1}^{23} \hat{w}_i^j$, where $\hat{w}_i^j = \left(\frac{T_i}{\sum_{i=1}^{23} T_i}\right) * w_i^j$. For both (i) and (ii) in panel A, T_i is the total number of full-time equivalent (FTE) teachers in state (district) i during the 2011–12 school year (source: NCES; http://nces.ed.gov/ccd/elsi/). In table 2, panel B, we calculate teacher-level component weights for the "typical teacher" across states and the largest districts, accounting for the different component weights assigned to teachers in both tested and nontested grades/subjects, in the following way: (i) $\tilde{W}^j_{States} = \sum_{i=1}^{32} \ddot{w}_i^j$, where $\ddot{w}_i^j = \left(\frac{T_i}{\sum_{i=1}^{32} T_i}\right) * (0.3 w^j_{i,\,Tested} + 0.7 w^j_{i,\,Nontested})$; (ii) $\tilde{W}^j_{Largest\ Districts} = \sum_{i=1}^{17} \ddot{w}_i^j$, where $\ddot{w}_i^j = \left(\frac{T_i}{\sum_{i=1}^{17} T_i}\right) * (0.3 w^j_{i,\,Tested} + 0.7 w^j_{i,\,Nontested})$. For (i) and (ii) in panel B, we assume that 70 percent of FTE teach in nontested grades/subjects and 30 percent of FTE teach in tested grades/subjects (Watson, Kraemer, and Thorn 2009). In both (i) and (ii), T_i is again the total number of FTE teachers in state (district) i during the 2011–12 school year (source: NCES; http://nces.ed.gov/ccd/elsi/).