

The New Educational Accountability:

Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era

Matthew P. Steinberg (A)
Assistant Professor
Graduate School of Education
University of Pennsylvania
3700 Walnut Street, #408
Philadelphia, PA 19104
Email: steima@gse.upenn.edu
Phone: (215) 898-9336

Morgaen L. Donaldson (B)
Associate Professor
Neag School of Education
University of Connecticut
249 Glenbrook Road, Unit 3093
Storrs, CT 06269-3093
Email: morgaen.donaldson@uconn.edu
Phone: (860) 486-4438

August 19, 2014

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.

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. In this policy brief, we describe and analyze the components, processes, and consequences embedded in new teacher evaluation policies in all 50 states, the 25 largest school districts, and Washington DC. We contextualize these policies by grounding our analysis in prior research on teacher evaluation and examining key comparisons between state and district policies and between their treatment of teachers of tested and untested subjects and 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 to their more veteran counterparts and for teachers in non-tested grades and subjects compared to those in tested grades/subjects. We conclude by discussing the implications of our findings for policy, research, and practice.

Introduction

In the last 5 years, teacher evaluation has become the primary approach for holding individual teachers accountable for their students' performance. In addition to demonstrating teachers' impact on student achievement (Goldhaber 2002; Rockoff 2004; Rivkin, Hanushek, and Kain 2005), a growing research base has revealed substantial within-school heterogeneity in teacher effectiveness (Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007). Despite this, traditional teacher evaluation systems relying solely on principal observations of teacher practice did a poor job differentiating teacher effectiveness, with the vast majority of teachers rated "satisfactory" or above (Weisberg et al. 2009). However, new evaluation systems 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.

Evidence on the design of newly developed teacher evaluation systems, however, remains limited. In this policy brief, we present some of the first evidence on newly developed and implemented evaluation systems, addressing the following questions: (1) What are the components of newly developed evaluation systems? Do the components (and their associated weights) vary by locale and for teachers in tested and non-tested grades and subjects? (2) What is the frequency of observations and evaluation conferences? Does the process vary by locale and teacher tenure (e.g. beginning versus career teachers)? (3) What are the consequences of evaluation? Do the consequences vary by locale? Our findings reveal important differences in how evaluation is conducted across states and the largest school districts and for teachers by career status and for those in tested rather than non-tested grades and subjects.

Background on Teacher Evaluation

Recent policy scans suggest that most new teacher evaluation systems incorporate measures of student achievement and observations of classroom instruction to assess teacher performance (National Council on Teacher Quality 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 et al. 1983; Murphy, Hallinger and Heck 2013). There are two potentially competing approaches to satisfying the system's fundamental goal of improvement in student outcomes: developing teachers' instructional skills to improve student performance and evaluating teacher effectiveness for accountability purposes related to tenure, rewards, and dismissal. What is the evidence on whether the components, process and consequences of teacher evaluation have produced these outcomes?

Component Measures of Teacher Performance

Historically, there is little indication that the component measures used for teacher evaluation have consistently supported instructional improvement or personnel decisions. Generally, teacher evaluation has relied on administrators' observations of teachers' instruction using observational instruments that were not grounded in theory or research (Porter, Youngs, and Odden 2001). Moreover, evaluation instruments were not linked to student learning or achievement (Peterson 2004). Under traditional evaluation systems, student achievement played a limited role in teachers' evaluations, with little evidence that teacher evaluation was used to hold teachers accountable for producing student learning. Although there have been forays into assessing teachers based on their students' performance, such experiments have remained the exception rather than the norm (Peterson 2004; Stronge and Tucker 2003).

Newly implemented teacher evaluation systems differ from their predecessors in two important ways. First, today's systems include standards-based observation protocols. In contrast to past systems, current evaluation models are more likely to be based on instructional standards and employ observation rubrics, such as the widely used Framework for Teaching (FFT), developed by Charlotte Danielson (Donaldson and Papay forthcoming). Evaluation systems include observations of teachers' practice based on the logic that evaluation should provide teachers with information to improve their instructional practice. The goal of these classroom observation measures in new evaluation systems is to more directly link teacher practice to instructional standards. Doing so may then increase student achievement (see, e.g., Danielson 2000; Pianta and Hamre 2005). Second, to tighten the link between evaluation and accountability, student outcomes based on state standardized tests are now included in many teacher evaluation systems (NCTQ 2013; Hallgren et al. 2014). The inclusion of student performance measures in teachers' evaluations – through value-added measures (VAM) of teacher performance or other test-based measures such as student growth percentiles (SGP) – 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 (RTTT) required that states (and later districts) weigh student outcomes heavily in teacher evaluation. In response, many states and districts 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 value-added measures in teacher evaluation systems. The benefits and drawbacks of value-added measures have elsewhere been discussed at length (see, e.g. Baker et al. 2010; Donaldson and Papay forthcoming; Glazerman et al. 2010). Many states use other measures of student performance for these teachers and some

states have chosen not to include value-added measures in their evaluation systems at all, opting instead for teacher-developed measures of student performance known as student learning objectives (SLO). Policymakers have argued that such measures increase teacher buy-in and may be more motivating for teachers than value-added measures (Locke and Latham 1990).

While single measures of instructional quality may be limited in their ability to identify effective teachers (see, e.g. Garrett and Steinberg forthcoming; McCaffrey, Sass, Lockwood, and Mihaly 2009; Goldhaber and Hansen 2010; Papay 2011), new evidence from the Measures of Effective Teaching (MET) study suggests that composite measures of teacher performance offer promise. Indeed, when used in combination, multiple measures of teacher performance – classroom observation scores, student surveys of teacher practice, and value-added scores based on student achievement data – have been shown to identify effective teachers (Kane, McCaffrey, Miller, and Staiger 2013).

The Process of Evaluation

Under past evaluation systems, the evaluation process was limited in both its influence on teacher practice as well as its utility for personnel decisions. 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.

In newly developed evaluation systems, the frequency of observation has increased relative to prior systems and observers are required to provide structured feedback based on observation rubrics (Hallgren, et al. 2014). The frequency of observation and evaluation may better support the use of evaluation not just for instructional improvement but also for personnel decisions. More frequent observation of teacher practice structured by standards-based

observation rubrics is likely to provide more information about a teacher's instructional performance than ever before. Evidence suggests that this additional information can help school leaders target professional development, generate improvements in school performance (Steinberg and Sartain forthcoming) and inform retention decisions (Sartain and Steinberg 2014).

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 the extent to which teachers' instruction improved through this mechanism (Stronge and Tucker 2003). Similarly, research suggests that evaluation identified few teachers for termination or non-renewal (Bridges 1992; Jacob 2011; Weisberg et al. 2009). The challenges of using merit pay to compensate teachers are well documented (Johnson 1986; Murnane and Cohen 1986). 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; Springer et al. 2010). Policymakers have attempted to couple new evaluation systems and the ratings they produce to professional development, termination, or merit pay (CGTL 2014; Hallgren et al. 2014). However, we know little about the extent to which states have attached these consequences to evaluation reform.

Recent research suggests that observation systems grounded in standards based evidence of instructional practice and supported by frequent observations can improve teacher practice and student performance (Steinberg and Sartain forthcoming; Taylor and Tyler 2012), while early evidence from settings incorporating multiple measures of teacher performance can lead to improvements in teacher performance while attaching real consequences to teachers' ratings (Dee and Wyckoff 2013). However, despite the scope and scale of teacher evaluation reform in

the wake of RTTT, there is little systematic evidence regarding the content, process, and consequences of new evaluation systems. To address this gap, we next present evidence from our examination of state and local teacher evaluation reforms.

Documenting Teacher Evaluation Reforms

We analyze teacher evaluation policies from the 50 states, largest 25 districts and Washington DC. We chose to analyze new 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 City, and Chicago (see, e.g., Sawchuk 2013). For this reason, we chose to sample the 25 largest districts (based on student enrollment) and Washington DC, an early and prominent adopter of evaluation reform (see Tables A1 and A2).

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 which 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; Donaldson 2009; Donaldson and Papay forthcoming; NCTQ 2013; Peterson 2004; Weisberg et al. 2009; Steinberg and Sartain forthcoming). We then piloted the matrix by gathering key data from teacher evaluation documents pertaining to Connecticut's state teacher evaluation policy (<http://www.connecticutseed.org/>) and documents describing New Haven, Connecticut's teacher evaluation policy (<http://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 as source material,

including state statutes, teacher evaluation policy websites, and policy handbooks, which often detailed the new policy for teachers and administrators; 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 policymakers.

Findings

We find key differences in the content, processes, and consequences of teacher evaluation between states and the largest districts, between tenured and non-tenured teachers, and between teachers of tested and non-tested subjects and grades. We document these differences below.

Implementation of Teacher Evaluation Systems

There is variation in the timing of state and district-level implementation of new teacher evaluation systems. The largest school districts are implementing teacher evaluation reforms earlier than state-level implementation requirements for all districts. Indeed, by the end of the 2013-14 school year, 77% (20 of 26) of the largest school districts and D.C. will have implemented teacher evaluation reforms, while 48% (24 of 50) of states will have revised their teacher evaluation systems (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 district flexibility for implementation. Of the 46 newly implemented state evaluation systems, most states (35) designed a formal model evaluation system. Of these 35 states, only 5 states have mandated that

districts implement the state-designed system without local discretion.¹ However, a majority of states (30) that designed model evaluation systems offered some degree of district discretion.² Fewer states (11) provided districts with a set of guidelines within which districts could locally design their systems.³

Components of Teachers' Summative Evaluation Ratings

Observation scores of a teacher's classroom practice are the most frequently used measures of teacher performance currently being incorporated into a teacher's summative evaluation rating (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 final evaluation rating. In addition, classroom observation scores represent the largest share of a teacher's summative rating; across states and the largest districts, approximately 50 percent of a teacher's rating is based on observation scores (see Table 1, Panel A). While observation rubrics, such as the Danielson Framework for Teaching (FFT) instrument, incorporate a professional conduct domain into a teacher's classroom observation score,⁴ 50% of the largest districts (with new evaluation policies) and approximately 20% of states identify a teacher's professional conduct as a separate measure contributing to a teacher's summative evaluation rating. For districts using this separate measure of professional conduct, on average, 10% of a teacher's

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, while 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% of a teacher's summative evaluation be based on elective data (such as nationally recognized standardized tests, student portfolios or projects) that are locally developed and selected by the school district from a list approved by 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.

⁴ The Danielson FFT reserves Domain 4 for “Professional Responsibilities,” of which one component (4f) evaluates teachers based on “Showing Professionalism.”

summative rating is based on professional conduct; for states, the average weight given to this measure is 15%.

Consistent with the nationwide emphasis on incorporating student test score data into a teacher's summative evaluation score, 80% 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 value-added measures (VAM) of teacher performance while states are more likely to use student growth percentiles (SGP). Indeed, for those states and districts having implemented evaluation reforms, approximately 60% (14) of the largest districts and only 30% of states use VAM, while 35% (8) of the largest districts and 60% (27) of states employ SGP. States and the largest districts vary not just in their preferences for VAM or SGP but also in the weights they assign to each measure. For the states and districts utilizing VAM, approximately one-third of a teacher's summative rating, on average, is based on VAM. For states and districts using SGP, approximately 20% of a teacher's summative rating depends on this measure.

The use of student learning objectives (SLOs) reflects, in many states, the provision of local discretion in determining the component measures of a teacher's summative evaluation score. For the 52% (24) of states and 39% (9) of largest districts using SLOs, 20% of a teacher's summative rating, on average, depends on these locally developed measures. Nine states (and one district) use SLOs as their only method of estimating student growth.

The component weightings just described pertain to teachers who teach in tested grades and subjects (for example, math and reading teachers in grades 3-8 and 11) where students take a state-mandated, end-of-year exam. However, upwards of 70 percent of teachers nationwide teach in non-tested grades or subjects (Watson et al. 2009), and, in principle, will not have student test

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.

score data in which to generate a VAM or SGP.⁶ Does the composition of a teacher's summative evaluation score differ for those teaching in tested versus non-tested grades/subjects?

We find similar patterns in how states and districts assign component weights for teachers of tested and non-tested grades and subjects.⁷ In both cases – state and district comparisons of teachers in tested and non-tested grades and subjects – a larger share of a non-tested teacher's summative evaluation rating depends on SLOs and a schoolwide achievement measure⁸ than for teachers in tested grades/subjects. At the state level, while 20% and 10%, on average, of tested teachers' summative evaluation ratings depend on SLOs and schoolwide achievement, respectively, these weights increase to 32% and 17% for non-tested teachers (see Table 1, Panel B). For the largest districts, 20% and 11%, on average, of tested teachers' summative evaluation ratings depend on SLOs and schoolwide achievement, respectively, and these weights increase to 25% and 26% for non-tested teachers (see Table 1, Panel C). And, while states weigh the classroom observation score similarly for tested and non-tested teachers (approximately 50%), the largest districts give greater weight (on average, 58%) to the classroom observation score for non-tested teachers compared to their tested counterparts (with, on average, 53% of their summative rating based on classroom observation scores).

⁶ We do find that some states and largest districts require VAM or SGP for teachers in non-tested grades/subjects, and employ either locally-developed or nationally normed exams of student achievement to generate VAM or SGP. Please see Table 1 for more detail on which states and districts employ these measures for non-tested teachers.

⁷ First, we looked at the component weights for teachers in tested and non-tested 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 non-tested 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 non-tested grades/subjects. We then compared tested and non-tested 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 value-added measure (PVAAS, or Pennsylvania Value Added Assessment System), other academic indicators such as graduation rates, promotion rates, attendance, Advanced Placement (AP) or International Baccalaureate (IB), PSAT participation, as well as credit for advanced achievement (on state exams).

Observation and Evaluation Process

All states and districts with newly implemented teacher evaluation systems employ classroom observations of a teacher's instructional practice as a component of the summative evaluation rating (see Figure 3). To what extent do these observations vary by type (e.g. formal versus informal) and by a teacher's career status (e.g. beginning versus career teachers⁹). We summarize these results in Table 2.

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.¹⁰ Comparing the largest districts to the national landscape of evaluation reform, we find that the largest districts require about the same number of formal observations (2) for career teachers, but approximately 25% more formal observations for beginning teachers. Moreover, the largest districts require nearly 1 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 (informal and formal) for career teachers, and 5.5 observations for beginning teachers, while states require, on average, approximately 4 total observations for both career and beginning teachers.

Emerging evidence suggests that ongoing, pre- and post-observation conferences between observers and teachers are critical components of the evaluation process, providing opportunities for the type of formative and ongoing feedback that is necessary for improving instruction and student achievement (Taylor and Tyler 2012; Steinberg and Sartain forthcoming). We next examined the extent to which states and the largest districts require teacher-observer conferences

⁹ Beginning teachers are early career teachers in their first three years of teaching and Career teachers are those in at least their fourth year of teaching.

¹⁰ In most cases, a formal observation lasts at least 30 minutes, and observers gather and record evidence of a teacher's instructional practice guided by an observation rubric. Informal observations can range from 15 minute drop-ins to short walkthroughs, and data collection is often not required.

to better understand the evaluation process embedded in new evaluation systems. Table 3 summarizes these results.

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. Indeed, the evaluation and observation process within locale (e.g. nationally across states or across large urban districts) and across teachers looks almost identical. Across locales, however, there are some notable differences. Although states and the largest districts required summative conferences at very similar rates (54% of states and 57% of the largest districts), a larger share of districts (70%) requires teacher-observer post-observation conferences than do states (54%).

Evaluation System Consequences

A major criticism of traditional evaluation systems is that few formal consequences resulted from teachers' summative evaluation ratings. To better understand the extent to which states and districts are embedding consequences into their new evaluation systems and linking them to a teacher's summative evaluation score, we consider the following: (a) professional development opportunities; (b) merit based pay awards; (c) employment termination; and (d) revocation of tenure. Table 4 summarizes these findings.

Our findings suggest that policymakers take a developmental stance towards evaluation. Most state (83%) and district (74%) policies link professional development to a teacher's summative evaluation rating. This applies to all teachers, but the requirements for under-performing teachers are particularly explicit. These professional development opportunities are often structured using a professional development plan, crafted by the teacher and his/her 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, while some states require professional development for teachers who underperform on individual components of the evaluation process.¹¹ A much smaller share of states (20%) and districts (21%) provide merit based rewards for a teacher's summative evaluation rating. The fact that states and districts are emphasizing professional development over merit-based pay rewards is particularly notable in light of the evidence on formative assessment and the benefit of targeted, professional development opportunities compared to the effect of merit-based pay on student achievement.

Discussion

Our findings reveal the ways in which states and districts have emphasized particular features of teacher evaluation. First, we find that classroom observation is a major component in district and state evaluation systems, but constitutes a smaller share of teachers' overall ratings than in past evaluation systems (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' summative evaluations.

Moreover, our findings suggest that there are important differences between the evaluation systems of the largest districts in the United States and those in place in smaller districts throughout the country. First, large districts tended to implement evaluation reform earlier—in some cases considerably so—than states. Second, the process of evaluation in large districts differs substantially from that of states. Districts require more informal and formal observations than do states. They also emphasize post-observation conferences more than states

¹¹ In Delaware, an early RTTT 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.

do. These demands on administrators may be particularly challenging in urban districts, which may have fewer resources than their counterparts in non-urban settings.

We find striking differences in how teachers of tested and non-tested grades/subjects are evaluated. Specifically, school average performance and student learning objectives constitute a greater proportion of non-tested teachers' ratings than they do for tested teachers' ratings. In light of the fact that schoolwide achievement measures constitute a much larger proportion of non-tested teachers' ratings than those of tested teachers, non-tested teachers are being assessed on a measure of effectiveness that has little (if anything) to do with their own individual performance. Depending on how schoolwide measures are calculated, this may be particularly problematic for highly effective non-tested teachers, whose performance ratings 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 in these settings, given the potential adverse effect on the ratings of high-performing teachers. Finally, since there is no within-school variation on a schoolwide achievement measure, it provides no additional information about teacher performance for teachers without student test score data. Therefore, other components – such as the observation score – will represent a larger effective share of non-tested 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, policymakers 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 non-tested 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.

Policymakers 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 (Donaldson et al. 2014; Kimball 2002; Krajewski 1978; Peterson 2000; Tucker 1997). 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 untested subjects view the fairness of the ratings formula, particularly given the disproportionate emphasis on schoolwide achievement for teachers in non-tested grades/subjects? Second, our research 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 non-tenured 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 often do these systems institute

structures that support struggling teachers and thus avoid a dismissal or tenure revocation? 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 in the U.S. and the states. Research on new teacher evaluation systems has tended to focus on a small number of large districts (i.e. Dee and Wyckoff 2013; Steinberg and Sartain forthcoming; Taylor and Tyler 2012); 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 25 school districts, and Washington D.C. 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 policymakers 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.

References

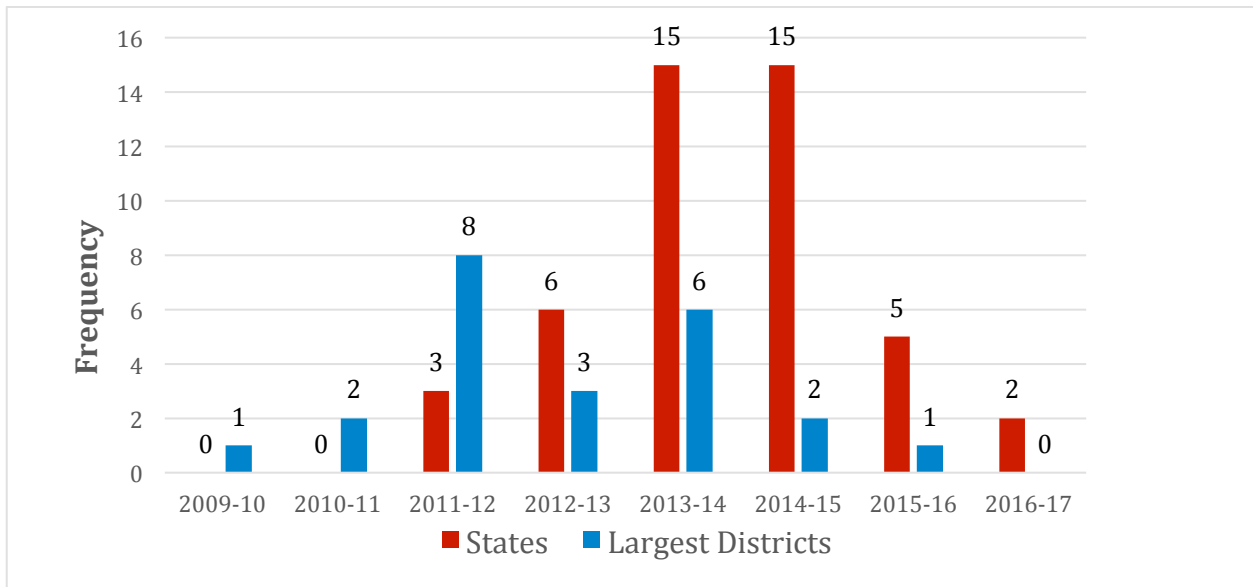
- Aaronson, D., Barrow, L., & Sander, W. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1): 95-135.
- Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., & Shepard, L.A. 2010. Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper 278. *Economic Policy Institute*.
- Bridges, E.M. 1992. *The Incompetent Teacher*. (2nd ed.). Philadelphia: Falmer.
- Creswell, J. 2013. *Qualitative Inquiry and Research Design*. Los Angeles: Sage.
- Danielson, C. 2007. *Enhancing Professional Practice: A Framework for Teaching*. ASCD.
- Darling-Hammond, L., Wise, A.E., & Pease, S.R. 1983. Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, 53(3): 285-328.
- Dee, T., & Wyckoff, J. 2013. Incentives, selection, and teacher performance: Evidence from IMPACT. National Bureau of Economic Research.
- Donaldson, M.L. 2009. *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Center for American Progress.
- Donaldson, M.L. & Papay, J. forthcoming. Teacher evaluation for accountability and development. In (Eds.) H.F. Ladd and M.Goertz, *Handbook of Research in Education Finance and Policy*.
- Donaldson, M.L., Cobb, C., LeChasseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. 2014. *An Evaluation of the Pilot Implementation of Connecticut's System for Educator Evaluation and Development*. Storrs, CT: Center for Education Policy Analysis.
- Figlio, D. N., & Kenny, L. W. 2007. Individual teacher incentives and student performance. *Journal of Public Economics*, 91: 901–914.
- Firestone, W. 2014. Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*, 43(2): 100-107.
- Fryer, R. G. 2011. Teacher incentives and student achievement: Evidence from New York City public schools. NBER working paper 16850.
- Garrett, R., & Steinberg, M. forthcoming. Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. Forthcoming in *Educational Evaluation and Policy Analysis*.
- Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. 2010. *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy, Brookings Institution.
- Goldhaber, D. & Hansen, M. 2008. Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. National Center for Analysis of Longitudinal Data in Education Research.

- 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: 1067-1083.
- Goldhaber, Dan. 2002. The mystery of good teaching. *Education Next* 2(1): 50-55.
- Hallgren, K., James-Burdumy, S. & Perez-Johnson, I. 2014. *State Requirements For Teacher Evaluation Policies Promoted By Race To The Top*. NCEE Evaluation Brief. Washington DC: Institute For Education Sciences. Retrieved June 13, 2014 At http://www.mathematica-mpr.com/publications/pdfs/education/rtt_ies_brief.pdf?spsmailingid=8478820&spuserid=mtu3odmznji4oas2&spjobid=292664207&spreportid=mjkynjy0mja3s0.
- Jacob, B.A. 2011. Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis*, 33(4): 403-434.
- Johnson, S. M. 1986. Incentives for teachers: What motivates, what matters. *Educational Administration Quarterly*, 22(3): 54-79.
- 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*.
- Kimball, S. 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.
- Krajewski, R. 1978. Secondary principals want to be instructional leaders. *Phi Delta Kappan*, 60: 65.
- Locke, E.A., & Latham, G.P. 2002. Building a practically useful theory of goal setting and work motivation: A 35 year odyssey. *American Psychologist*, 57: 705-717.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. 2003. *Evaluating value-added models for teacher accountability*. RAND Corporation.
- Murnane, R.J., & Cohen, D.K. 1986. Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, 56(1): 1-17.
- Murphy, J., Hallinger P. & Heck, R. 2013. Leading via teacher evaluation: A Case of the missing clothes? *Educational Researcher*, 43(6): 349-354.
- National Council on Teacher Quality 2013. *State of the States 2013 Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.
- Papay, J.P. 2011. Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates across Outcome Measures. *American Educational Research Journal*, 48(1): 163-193.
- Peterson, K. 2000. *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.) Thousand Oaks: Corwin.
- Peterson, K. 2004. Research on School Teacher Evaluation. *NASSP Bulletin* 88(639), 60-79.

- Pianta, Robert C., and Bridget K. Hamre. 2005. *Classroom Assessment Scoring System, Secondary Manual*. Teachstone Training: Charlottesville, VA.
- Porter, A.C., Youngs, P., & Odden, A. 2001. Advances in teacher assessment and their uses. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th edition) (pp.259-297). New York: Macmillan.
- Rivkin, S.G., Hanushek, E.A., Kain, J.F. 2005. Teachers, schools, and academic achievement. *Econometrica*, 73(2): 417-458
- Rockoff, J.E. 2004. The impact of individual teachers on student achievement. *American Economic Review, Papers & Proceedings*, 94(2): 247-252.
- Sartain, L. & Steinberg, M. 2014. Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. Working paper.
- Sawchuk, S. 2013. Chicago Teachers See Value in New Evaluations, But Eschew Test Scores. Retrieved from:
http://blogs.edweek.org/edweek/teacherbeat/2013/09/chicago_teachers_see_value_in.html
- Springer, M. G., Ballou, D., Hamilton, L. S., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., et al. 2010. Teacher pay for performance: Experimental evidence from the project on incentives in teaching. Nashville, TN: National Center on Performance Incentives.
- Steinberg, M. & Sartain, L. forthcoming. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. Forthcoming in *Education Finance and Policy*.
- Stronge, J.H., & Tucker, P D. 2003. Teacher evaluation. Assessing and improving performance. Larchmont, NY: Eye on Education.
- Taylor, E.S. & Tyler, J.H. 2012. The effect of evaluation on teacher performance. *American Economic Review*, 102(7): 3628-3651.
- Tucker, P.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.
- Watson, J.G., Kraemer, S.B. & Thorn, C.A. 2009. *The Other 69 Percent*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education.
- Weisberg, D., Sexton, S., Keeling, D. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.

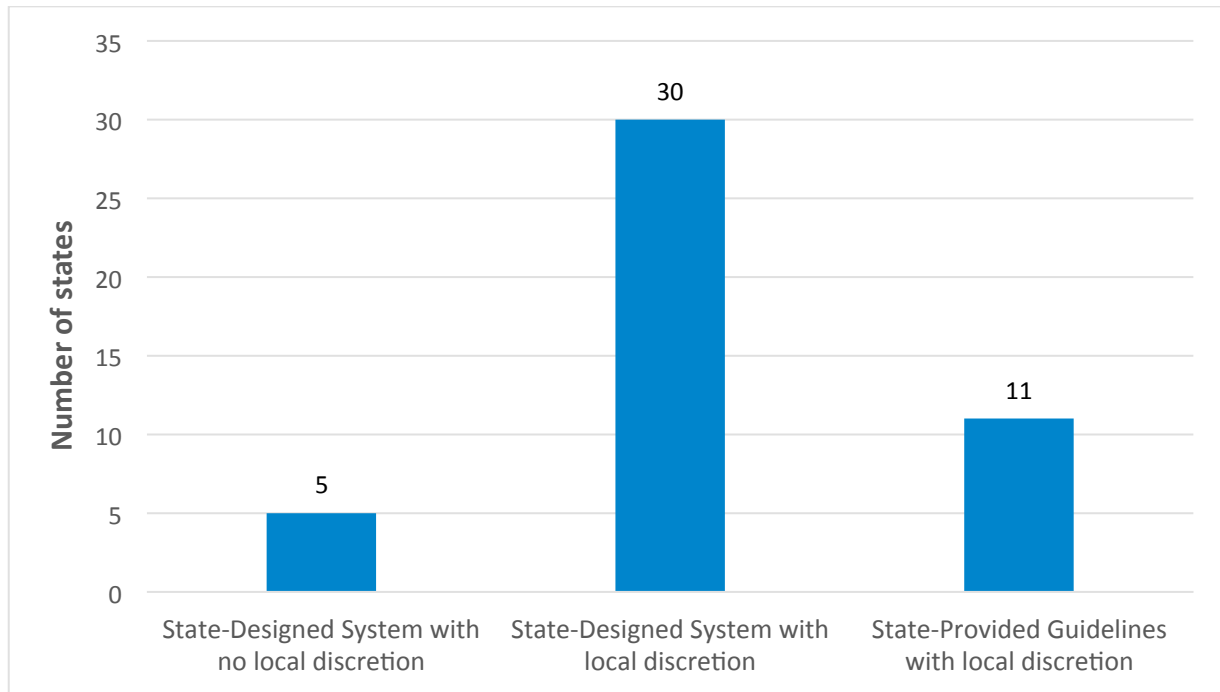
Figures & Tables

Figure 1. Implementation Timing of Teacher Evaluation Reforms



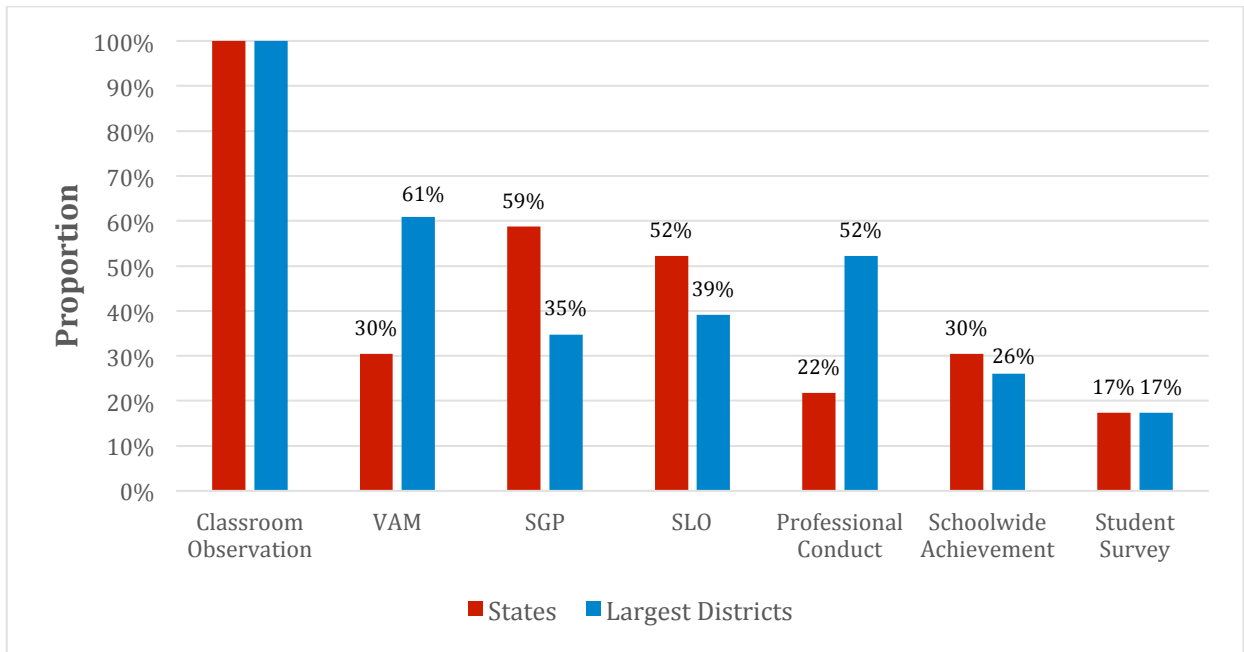
Notes. The year reflects the first year of scheduled statewide or districtwide implementation of teacher evaluation reform. There are 46 states and 23 districts included. AL, CA, IA and TX are excluded from the state time trend; Cypress-Fairbanks (TX), Montgomery County (MD) and San Diego Unified (CA)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

Figure 2. State Policy Design and Local District Discretion



Notes. There are 46 states included in the figure. The following states have designed new teacher evaluation systems which 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 3. 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). Please see Table 1 for more detail on the weights associated with each measure.