Revisiting the Widget Effect:
Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness

Matthew A. Kraft
Brown University

Allison F. Gilmour
Vanderbilt University

February 2016

Abstract

In 2009, TNTP's The Widget Effect documented the failure of U.S. public education to recognize and act on differences in teacher effectiveness. We revisit these findings by compiling teacher performance ratings across 19 states that have adopted major reforms to their teacher evaluation systems. In a majority of these states, less than 3% of teachers are rated below Proficient. We also find substantial differences in the percentage of teachers rated below Proficient and Highly Effective across states. We present original survey data from an urban district illustrating that evaluators perceive more than three times as many teachers in their schools as below Proficient than they actually rate as such. Interviews with principals reveal several potential explanations for these patterns.

Suggested Citation:
Kraft, M.A. & Gilmour, A.F. (2016). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. Brown University Working Paper.

We are grateful to Tim Drake, Heather Hill, Susan Moore Johnson, and Jal Mehta for their feedback on earlier drafts. We thank Mariela Mannion and Melissa Lovitz for their excellent research assistance. All errors and omissions are our own.

The failure of evaluation systems to provide accurate and credible information about individual teachers' instructional performance sustains and reinforces a phenomenon that we have come to call the Widget Effect. The Widget Effect describes the tendency of school districts to assume classroom effectiveness is the same from teacher to teacher. This decades-old fallacy fosters an environment in which teachers cease to be understood as individual professionals, but rather as interchangeable parts.

- *The New Teacher Project, 2009*

In 2009, The New Teacher Project (TNTP) characterized the failure of U.S. public education to recognize and respond to differences in teacher effectiveness as "the Widget Effect" (Weisberg et al., 2009). The study highlighted the discrepancy between formal teacher evaluation ratings and perceptions about the actual distribution of teacher effectiveness. The authors found that, in most districts, less than 1% of teachers were rated as unsatisfactory, but 81% of administrators and 57% of teachers could identify a teacher in their school who was ineffective. Researchers at TNTP were not alone in their view that evaluation systems failed to differentiate among teachers. Over a decade earlier, Tucker (1997) labeled the U.S. education system's failure to recognize "incompetent" teaching as the "Lake Wobegon Effect" – referring to Garrison Keillor's fictitious town where "all the children are above average." Several other studies characterized teacher evaluation in the first decade of the 21st century as a superficial exercise that failed to assess instructional quality or to inform teacher professional development and personnel decisions (Donaldson, 2009; Toch & Rothman, 2008).

New research documenting the central importance of teacher effectiveness (e.g. Rockoff, 2005; Rivkin, Hanushek, & Kain, 2005) combined with growing recognition of the broken teacher evaluation system helped to generate momentum for evaluation reforms (Donaldson & Papay, 2015). The U.S. Department of Education's Race to the Top (RTTT) competition and state waivers for regulations in the No Child Left Behind Act compelled states

to make sweeping changes to these systems. Applicants were required to replace binary checklists with systems that included multiple rating categories and differentiated teachers by performance (U.S. DOE, 2009; 2012). The combination of these federal policy initiatives along with local reform efforts have led to substantial changes in teacher evaluation.

Today, almost every state has designed and adopted new teacher evaluation systems (see Steinberg & Donaldson [in press] for a survey of reform efforts and Donaldson & Papay [2015] for a summary of new evaluation systems features). Some scholars view the recent focus on high-stakes evaluation systems as misguided (Fullen, 2011; Hallinger, Heck, & Murphy, 2014; Metha & Fine, 2015). Even those who see evaluation reforms as promising do not agree on *how* these systems should be used to improve the teacher workforce. Some scholars (Hanushek, 2009) and journalists (Thomas, Wingert, Conant, & Register, 2010) emphasize the importance of identifying low performing teachers in order to increase their effort on the job or to dismiss them. Others see evaluation as central to supporting teachers' professional growth by providing teachers with individualized feedback and identifying areas for targeted professional support (Almy, 2011; Curtis & Weiner, 2012; Papay, 2012). Critically, both approaches require an evaluation system that differentiates among teachers and accurately reflects the quality of their instruction.

In this paper, we revisit the Widget Effect by examining the degree to which new teacher evaluation systems differentiate among teachers. Research on evaluation reforms has primarily focused on the properties of performance measures (e.g. Grossman, Loeb, Cohen, & Wyckoff, 2013; Kane, McCaffrey, Miller, & Staiger, 2013 and the March 2015 special issue of Educational Researcher), the effect evaluation systems have had on student achievement (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2013), and principals' use of value-

added measures (Goldring et al., 2015; Rockoff, Staiger, Kane, & Taylor, 2012). Research suggests that principals are capable of distinguishing between low and high performing teachers (Harris & Sass, 2014; Jacob & Lefgren, 2008), but that they do not always do so on high-stakes evaluation ratings (Grissom & Loeb, forthcoming). We have little evidence about the degree to which these reforms have fundamentally changed the distribution of teacher performance ratings. Such evidence is essential given increased state autonomy to design evaluation systems under the Every Student Succeeds Act. Thus, we ask: Have teacher evaluation reforms resulted in meaningful variation in teacher performance ratings? Does the distribution of teacher performance ratings better reflect perceptions about the true distribution of teacher effectiveness? And, if not, what explains why evaluation reforms have not resulted in greater differentiation and more accurate ratings?

We examine these questions by drawing on quantitative and qualitative data collected over the course of three years. We begin by presenting data on the distribution of teacher evaluation ratings across states. An exhaustive search produced data for 19 states that have implemented teacher evaluation reforms with multiple performance categories. We complement these state-level data with a case study of the distribution of teacher evaluation ratings in one large urban school district. Specifically, we leverage original survey data linked to evaluation records to compare evaluators' perceptions of the true distribution of teacher effectiveness in their schools with both their predictions of what the ratings distribution will be, as well as the actual end-of-year ratings. We conduct this comparison in both the first year of the new evaluation system as well as the third to examine how perceptions changed as evaluators gained more experience with the new system. We then discuss findings from in-depth interviews with a random sample of principals in the district that help to explain why differences existed between

evaluators' perceptions, predictions and the actual distribution of teacher effectiveness ratings. Together, these data provide new insights about the potential and pitfalls of improving the quality of the teacher workforce through teacher evaluation reforms.

Data and Methods

State Teacher Evaluation Ratings

We compiled data on state distributions of teacher evaluation ratings following a systematic search process. Our target sample included 38 states that had either piloted or fully implemented a new teacher evaluation system by the 2014/15 school year. We began by reviewing RTTT annual performance reports. We then searched for reports published on state education agency and teacher evaluation system websites. Finally, we used the advanced search features in Google and Google Scholar to search for studies, reports, and news articles containing information on teacher evaluation ratings.[1] Our search produced data on the distribution of teacher effectiveness for 19 states including 14 RTTT winners. Among these states, 13 rated teachers across four performance categories, five use five categories and one used three categories. The primary sources for these data include 14 state department of education reports, three research papers, three news articles, and one Institute for Education Sciences report. We provide detailed information about rating systems and source data for each state in Appendix A.

District Case-Study of Teacher Evaluation

Our case-study analyses focused on teacher evaluation ratings in a large urban district in the northeast. Hispanic and African American students make up approximately 75% of the district student body, while the remaining 25% of students are predominantly Caucasian and Asian American. Over 70% of students in the district are eligible for free or reduced price lunch and nearly half speak a language other than English as their first language.

[1] When multiple years of data were available we used the most recent year.

For many years in the district, evaluation consisted of administrators checking off whether or not teachers met the standard on a list of classroom practices and professional standards. Evaluations were infrequent and many teachers went unevaluated. In 2012/13, the district implemented a new evaluation system that was adapted from the state's new framework in partnership with the local teacher's union. This new evaluation process is centered on a cycle of assessment using a rubric that captures observable standards related to teaching effectiveness.

Principals and select members of their administrative teams (e.g. Assistant Principals, Directors of Instruction) are responsible for providing teachers with a mid-year formative assessment and an end-of-year summative assessment. Evaluators primarily use evidence from classroom observations as well as artifacts and progress towards teacher-defined Student Learning Goals to inform their overall ratings. Teachers rated as Proficient or Exemplary proceed on a cycle of self-directed growth while those who are rated as Needs Improvement or Unsatisfactory are placed on more structured evaluation plans, which, after repeated low evaluations, can result in dismissal (See Kraft & Gilmour [2015] for a detailed description of the evaluation system).

Evaluator surveys. We worked with district officials to administer a brief survey to evaluators in the summer/early fall of 2012. Two questions on the survey are central to this study. These questions asked evaluators (1) to rate the percentage of teachers in their school that *in their judgement* were in each of the four performance categories and (2) to predict the percentage of teachers in their school they thought *will actually be rated* at each of these levels (see Appendix B for survey items). District officials administered paper copies of the survey at district-wide meetings and followed up with an email link for completing the survey on-line. We collected survey responses from a total of 161 of the 340 evaluators in 2012/13. We re-

administered these same two questions to evaluators participating in a training program during the fall/winter of 2014/15. As part of a larger ongoing study, evaluators were randomly assigned to attend required training sessions in either 2013/14 or 2014/15. Of the 177 evaluators who attended the training in 2014/15, 172 completed the survey among this randomly selected subset of district evaluators.

We then linked evaluators' survey responses with the actual performance ratings in their schools. We calculated the distribution of performance ratings for classroom teachers at each school by collapsing individual evaluation records to the school-level in each year. We focused our analysis on teachers' overall summative evaluation ratings. We restricted our final analytic dataset to those evaluators whose survey responses totaled to 100% and were successfully linked to schools with valid evaluation data.[2] This resulted in an analytic sample of 107 evaluators across 58 schools in 2012/13 and 157 evaluators across 66 schools in 2014/15. Although we cannot rule out the possibility of differential selection into the survey sample across years, in supplemental analyses we find that the patterns we report below remain the same when we restrict our data to include only schools for which we have survey responses in both years.

Principal interviews. In the summer of 2013, we conducted interviews with a stratified random sample of principals in the district to understand their experiences implementing the new teacher evaluation system. Twenty-four out of the 46 principals we contacted agreed to be interviewed. These principals worked at a range of small and large elementary, middle, and high schools, and were diverse in both demographic characteristics and administrative experience. We find no statistically significant differences in the demographic and school characteristics for

[2] For teachers whose responses total to within plus or minus 1 percentage point of 100 we round up their estimates in the top ratings category to reduce data loss due to minor computational error. Evaluation data is not available for several schools in the district that are not required to use the district designed evaluation system.

those principals in the district we interviewed and those we did not (for full details see Kraft & Gilmour, 2015).

We interviewed each principal for 45-60 minutes using a semi-structured interview protocol (See Appendix C for protocol). We audio-recorded and transcribed each interview and then drafted thematic summaries to identify potential codes (Strauss & Corbin, 1998). We developed and refined our codes using an iterative process that built on both the scholarly literature and themes that emerged from our data (Miles & Huberman, 1994). Each author coded two manuscripts, reviewed the other author's codes, and discussed discrepancies (See Appendix D for codes). After reaching coding agreement and developing the final codebook, we coded each transcribed interview and then analyzed these data by organizing codes around broad themes. In this paper, we focus our discussion on themes related to principals' experiences and perspectives on assigning teachers a below Proficient performance rating.

Findings

Distribution of Teacher Evaluation Ratings

In Figure 1, we present the percentage of teachers in the ratings categories that fall below Proficient/Exemplary ("Proficient" going forward) among the 19 states in our analytic sample. We find that the median percentage of teachers rated below Proficient is 2.7%. This represents a meaningful increase compared to the "fraction of a percentage" of teachers that were rated below Proficient five years earlier, but not a landmark change in ratings. Figure 1 also illustrates how the percentage of teachers rated as below Proficient varies substantially across states. Five states identified approximately 5% of teachers or more as below Proficient, ten states rated between 2% and 4% of teachers as below Proficient, and in four states less than 2% of teachers received below Proficient ratings. The majority of these teachers fall in the Developing/Needs

Improvement category; just three states (Louisiana, Maryland, and New Mexico) rated more than 1% of teachers as Ineffective/Unsatisfactory.

We present the corresponding percentage of teachers rated in the performance category (or categories) above Proficient in Figure 2. The median percentage of teachers rated above Proficient is 39% but varies considerably from 3% in Georgia to 73% in Tennessee. In fact, a majority of teachers are rated above Proficient in four states, while less than 20% of teachers are rated above Proficient in four other states.

In Figures 3A and 3B, we present the full distributions of teacher evaluation ratings for states with four and five performance categories, respectively. For states with four rating categories, the primary differentiation among teachers is between the two highest performance categories (i.e. Effective vs. Highly Effective). Teacher evaluation ratings in states with five rating categories appear to differentiate slightly more by distributing teachers across the three top rating categories. The exception to this generalization is Florida, which we omit from Figure 3B because it classifies teachers into three categories below Proficient and only one above. In Florida, 97.6% of teachers are rated as either Effective or Highly Effective.

Overall, these data show that some new teacher evaluation systems do differentiate among teachers, but most only do so at the top end of the ratings spectrum. These findings suggest that exchanging binary rating systems for multiple rating categories is not a guarantee of a more differentiated ratings distribution. Although states with five performance categories tend to rate more teachers as top performers, we do not observe any clear relationship between the number of rating categories and the percentage of teachers rated below Proficient. More rating categories does not appear to translate into greater differentiation at the lower end of the rating scale.

Evaluators' Perceptions of the Distribution of Teacher Quality

We next present data from our district case-study on the degree to which evaluators' perceptions of the effectiveness of teachers in their schools aligned with the actual performance ratings they assigned. On average, the evaluators who participated in our survey in 2012/13 estimated that 27.8 percent of all teachers in their schools' were performing at a level below Proficient. As shown in Figure 4A, this estimate is more than four times the percentage of teachers who were actually rated below Proficient. Figure 4A also demonstrates that evaluators anticipated that fewer teachers would be rated below Proficient than they thought were performing at these levels (27.8% perceived vs. 24.3% predicted below Proficient). However, these same evaluators substantially underestimated the degree to which their actual ratings would be inflated upwards (6.5% actual below Proficient).[3]

Evaluators may not have fully understood or anticipated the challenges associated with rating teachers below Proficient in 2012/13, the first year of district-wide implementation of a new teacher evaluation system. We examined this question by re-administering our survey to evaluators in 2014/15, the third year of the new evaluation system. Again, we found similar patterns as shown in Figure 4B where they perceived over three times as many teachers as below Proficient than they rated as such (19.1% perceived vs. 6.3% actual below Performance). Evaluators also continued to overestimate the proportion of teachers they would eventually rate in one of the two lowest performance categories (13% predicted), but less so than in 2012/13.

These findings are telling in several ways. We see that in both years, evaluators who were responsible for assigning final summative ratings in their schools predicted that they would assign fewer teachers below Proficient ratings than they perceived were warranted. These

[3] Our estimate of the actual percentage of teachers rated below Proficient is weighted by the number of evaluators who participated in the survey from each school to make it directly comparable to our survey results. The unweighted exact statistic is 6.7% in 2012/13 and 5.7% in 2014/15.

differences illustrate how evaluation ratings reflect more than just teacher performance; they are a product of a complex evaluation process with multiple purposes. Further, evaluators appeared to become more aware that the performance ratings they would eventually assign would not accurately reflect their perceptions of teachers' performance. This suggests that persistent implementation challenges and misaligned incentives are more likely to explain these patterns than short-term difficulties associated with adopting a new evaluation system.

Why Few Teachers Receive Below Proficient Ratings

In-depth interviews with principals provide several explanations for why so few teachers receive below Proficient ratings across states as well as why the ratings evaluators assigned teachers did not reflect their perceptions of teachers' actual performance in the district we studied.

Time constraints. Fourteen principals told us that a lack of time was the most frequent reason for not giving a teacher a low rating. Rating a teacher as below Proficient required intensive amounts of time to document their performance and to provide support for their professional growth. Several principals even questioned whether they could collect sufficient evidence in a few observations to justify a rating below Proficient. As a middle school principal with nine years of experience put it, "I just feel like sometimes you have to have a lot of detail before you can give somebody a Needs Improvement." A high school principal explained that both observations and support were major constraints, "When you have an unsatisfactory teacher, it takes a lot of time to observe that teacher, to give true honest-to-goodness feedback."

Several principals felt as if it was unfair to rate teachers as below Proficient if they did not have the capacity to provide these teachers with support. A middle school principal described this tension as follows:

> It's not possible for an administrator to carry through on ten unsatisfactories simultaneously. I mean once somebody is identified as unsatisfactory, the amount of work, the amount of observation, the amount of time and attention that it requires to support them can become overwhelming. There is a threshold… otherwise I'm not providing that person with the quality coaching and feedback that they need to improve.

The new evaluation system required evaluators to conduct up to four unannounced formal observations and write improvement plans for teachers whom they rated as unsatisfactory. This led some principals to use low ratings selectively. An elementary school principal explained:

> There were some areas that they could have been needs improvement. Because I was focusing on two or three other teachers who really needed needs improvement. I gave them Proficient in those areas. I did it because I couldn't tackle that many teachers at the same time as far as writing prescriptions and then following through on the work that I would need to do.

This principal took a triage approach to evaluating and supporting teachers. He reserved Needs Improvement ratings for those teachers that needed the most help because of the increased workloads these ratings would trigger.

Teachers' potential and motivation. Principals reported that they sometimes factored in teachers' potential when assigning an evaluation rating. For example, one principal spoke about giving new teachers more leeway:

> A first year teacher, I tend to give a little more the benefit of doubt. Like, give you a little time, the opportunity to improve, here are some suggestions…Sometimes someone's who fairly new teaching in the building, they are more apt to accept that feedback.

Principals felt that new teachers were still learning and that it was unfair to rate new teachers as below Proficient if they were working to improve their practice. A principal from a large high school said he wanted "to give people opportunities, give people chances." Other principals used this approach more broadly for teachers they viewed as just below Proficient. "They're not bad teachers. They need a little more time to develop and become better," explained a high school principal. They were "good enough." Assigning a Proficient rating was seen as a way to recognize teachers' efforts to improve.

Many of these principals also felt that giving a low rating to a potentially good teacher could be counterproductive to a teacher's development. For example, one middle school principal said he "will give [teachers] a Proficient rating to keep them on board and to keep them moving in a direction," rather than risk losing a potentially good teacher. An elementary school principal with 15 years of experience described how low ratings could cause teachers to become less receptive to feedback:

> There's one teacher who I probably should have given an overall 'does not meets' or whatever it is now in the new one. Instead, I gave her a subcategory…. I think she's somebody that I could support into being a stronger teacher. I don't think I can do that as well if I give an overall 'unsatisfactory,' get the union involved, and get the teacher taking my feedback in a very different way.

It was easier to address and remediate poor performance outside of the evaluation system. Principals, in some cases, shied away from using the lowest ratings for summative evaluations because it caused teachers to shift their focus from what they could do to improve to the consequences of the rating itself.

Personal discomfort. Six principals touched on how difficult it was to have emotional conversations with teachers whom they rated as below Proficient. These principals suggested that this might cause some evaluators to be reluctant to assign ratings that were below Proficient. One experienced principal nearing retirement articulated this view clearly:

> The most difficult part of the job is probably to deliver those difficult messages, and not everyone is capable of that. That's where administrators actually fall down is when they're unable to deliver those type of messages.

Principals spoke about how there was "definitely emotion" involved in assigning below Proficient ratings. A middle school principal told us, "I was pretty communicative and still people would be crying, or, 'I can't believe you think that.'" In his experience, some teachers reacted poorly to their low ratings despite his efforts to be transparent throughout the evaluation process.

Principals were also keenly aware that an unsatisfactory rating could eventually lead to teachers losing their jobs. Many principals saw this as an unfortunate but important responsibility, while others were less comfortable with initiating the dismissal process. A first year high school principal said:

> The last thing I think I wanna do as a human being is to watch another human being walk out with their head down; dejected, because they just lost their job because they couldn't do it. This is something that they wanted to do. That's a little bit harsh, you know?

This new principal did not want to expose teachers to the consequences that low ratings carried. Not surprisingly, neither this principal nor any other said they had personally chosen to rate a teacher as Proficient in order to avoid a challenging conversation or to shield a teacher from the

threat of dismissal. But on more than one occasion principals, such as an experience middle school principal, stated bluntly that "People shy away from difficult conversations." This suggests that some principals may have been willing to give slightly higher ratings to those teachers on the margin to avoid the discomfort of discussing a low rating and its potential consequences if teachers' did not improve.

Additional reasons. Several additional factors that could explain the discrepancy between perceived and actual teacher evaluation ratings emerged from our interviews including racial tensions, concerns about the quality of replacements, the burdensome dismissal process, and exchanging higher ratings for voluntary departures. Three principals mentioned concerns that a disproportionate number of non-White teachers would receive low ratings. An experienced elementary school principal told us that evaluation "became a racial issue, and it was huge." Two principals expressed their hesitancy to initiate dismissals with low ratings for fear of having human resources assign them a teacher from the excess pool. An experienced high school principal described how she chose to rehire a teacher:

> He's a problem, but he's my problem, and he's one that I can really work with. Relative to the problems that were ringing my doorbell, I thought, "I haven't begun to see how low it can go."

In her own words, "The one you know is better than the one you don't." A second principal's initial experience with dismissing teachers led her to be wary of assigning low ratings:

> If there's someone who's bad, you can evaluate them out, but you risk getting someone who's worse. When I first started, that happened to me twice with the same position. I had a math teacher who was terrible, I evaluated her out, I got one actually worse.

A few principals mentioned that they also sought to avoid the "long, laborious, legal, draining process" of evaluating out a teacher. Similarly, two principals found it easier to remove teachers outside of the evaluation process. As one principal stated frankly:

> I didn't give her a negative evaluation in certain terms of then having to evaluate her out. That would've meant that she would have to stay in my school for another year and I had to go through the whole long process thing. She was clearly not going to work out anyway and she was going to leave. She agreed to leave."

Here, it was more expedient for the principal to trade a Proficient evaluation for a teacher's voluntary departure.

Policy Implications

Six years after the Widget Effect's "call to action," teacher evaluation reforms have fundamentally changed the ways in which teachers are evaluated in U.S. public schools. In many states, observations are more frequent and focused on instruction, student achievement results are considered, and teachers are rated on scales with multiple performance categories. These reforms, however, have not guaranteed that new evaluation systems necessarily reflect meaningful differences in teacher effectiveness. The authors of the Widget Effect argued that "school districts must begin to distinguish great from good, good from fair, and fair from poor." Our results show that many states now do distinguish great teaching from good, but that they rarely make distinctions between good, fair, and poor instruction.

Although we cannot know the true distribution of teacher effectiveness in each state, the wide variability in teacher ratings across states suggests that ratings reflect more than just true differences in teacher performance. Regional labor markets and the quality of teacher pre-service and in-service programs certainly play a role, but it seems unlikely these or other outside factors

could fully explain why 1% or fewer teachers are below Proficient in Hawaii and Delaware but 26.2% are below Proficient in New Mexico. Differences in performance standards must at least partially explain why only 3% of teachers in Georgia and 8% of teachers in Massachusetts are above Proficient but 73% meet this higher standard in Tennessee. Furthermore, we show that in one district that has adopted a new teacher evaluation system, the ratings evaluators' assigned to teachers differed substantially from what they perceived to be the true distribution of teacher effectiveness at their schools. Only New Mexico's evaluation system has resulted in ratings that come close to the perceived distribution of performance reported by the evaluators we surveyed, and this system is currently facing a series of legal challenges (Brown, 2015).

Our exploratory case-study reveals how the failure to differentiate between teachers is a product of conscious choices by evaluators as they navigate implementation challenges, competing interests, unintended consequences, and perverse incentives. These findings exemplify Michael Lipsky's (1980) seminal observation that policies are ultimately made by the "street-level bureaucrats" who implement them rather than the policymakers who design them. Increased follow-up requirements for below Proficient ratings may distort evaluators' rating decisions. Holding evaluators to a higher standard of evidence and follow-up support for teachers rated as below Proficient makes sense from a measurement and professional development perspective, but it creates strong incentives for evaluators to rate no more than a few teachers as below Proficient given the many other demands on their time. Telling someone they are not Proficient at their job can be a difficult and unpleasant thing to do. This may cause some evaluators to judge teachers more favorably than they would otherwise.

Several of the principals we spoke with argued that rating struggling teachers as low performers in a high-stakes evaluation system could be counterproductive to improving the

teacher workforce. For some teachers, a low rating may motivate them to invest in their own professional growth or pressure them to work harder. For others, it may raise their defenses, causing them to be less receptive to feedback on how to improve. Assigning low ratings can undercut relational trust that is essential for mobilizing collective effort (Bryk & Schneider, 2002) and sometimes led to racial tensions. Some principals remained doubtful that the time and effort spent navigating the dismissal process would ultimately result in removing a teacher and finding a better replacement.

The design of teacher evaluation systems has changed substantially over the last five years, but it appears that evaluation norms and practices are proving much more difficult to change. Going forward, it will be critical for future research to examine the degree to which the challenges evaluators faced and the choices they made are common across districts, as well as how to address these challenges productively. Important questions also remain about how the design features of evaluation systems such as the performance measures, choice of evaluators, weights, and cut points affect the distribution of evaluation ratings. Realizing the full potential of teacher evaluation systems as tools for improving the teacher workforce requires that we recognize teachers as professionals with individual strengths and weakness, not interchangeable parts.

References

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Washington, DC: Education Trust.

Kraft, M.A. & Gilmour, A. (2015) Can principals promote teacher development as evaluators? A case study of principals' views and experiences. Brown University Working Paper. http://scholar.harvard.edu/files/mkraft/files/principals_as_evalutors_rr_final_-_unblinded.docx.pdf?m=1447816036

Brown, E. (2015, February 28). Contentious teacher-related policies moving from legislatures to the courts. *The Washington Post*.

Bryk, A., & Schneider, B. (2002). *Trust in Schools: A Core Resource for Improvement: A Core Resource for Improvement*. Russell Sage Foundation.

Curtis, R., & Wiener, R. (2012). *Means to an end: A guide to developing teacher evaluation systems that support growth and development.* Washington, DC: Aspen Institute.

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

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

Donaldson, M.L. & Papay, J.P. (2015). Teacher evaluation for accountability and development. In H.F. Ladd & M.E. Goertz, eds. Handbook of Research in Education Finance and Policy. New York: Routledge.

Fullan, M. (2011). *Choosing the wrong drivers for whole system reform*. East Melbourne: Centre for Strategic Education.

Goldring, E., Grissom, J. A., Ruben, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*, 96-104.

Grissom, J. A., & Loeb, S. (forthcoming). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy*.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The Relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*, *119*, 445–470.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement. An analysis of the evidence. *Educational Assessment, Evaluation, and Accountability, 26*, 5-28.

Hanushek, E. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (165-180). Washington, DC: Urban Institute Press.

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, 40, 183-204.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*, 101–136.

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.* MET Project. Seattle: Bill & Melinda Gates Foundation.

Kraft, M.A. & Gilmour, A. (2015) Can principals promote teacher development as evaluators? A

case study of principals' views and experiences. Brown University Working Paper.

Lipsky, M. (1980). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service: Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

Mehta, J., & Fine, S. (2015). Bringing values back in: How purposes shape practices in coherent school designs. *Journal of Educational Change*, *16*(4), 483-510.

Miles, M. & Huberman, M. (1994). *Qualitative data analysis: A expanded sourcebook* (2nd ed.). Thousand Oaks: Sage Publications.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

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: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*, 3184-3213.

Steinberg, M. P., & Donaldson, M. L. (in press) The new educational accountability: Understanding the landscape of teacher evaluation in the post NCLB era. *Education Finance and Policy*.

Steinberg, M. P., & Sartain, L. (in press). Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy*.

Strauss, J. & Corbin, A. (1998). *Basics of qualitative research: Grounded theory procedures and*

techniques. (2nd Ed.). Thousand Oaks, CA: SAGE Publications.

Taylor, E. S., & Tyler, J. H. (2013). The effect of evaluation on teacher performance. *American Economic Review, 102*, 3628-3651.

Thomas, E., Wingert, P., Conant, E., & Register, S. (2010). Why we can't get rid of failing teachers. *Newsweek, 155*(11), 24-27.

Toch, T., & Rothman, R. (2008). *Rush to judgement: Teacher evaluatino in public education*. Washington, D. C.: Education Sector.

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.

U.S. Department of Education (2009). Race to The Top program executive summary. Washington, D.C.: U.S. Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education (2012). ESEA flexibility. Washington, D.C.: U.S. Department of Education. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html.

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

Figures

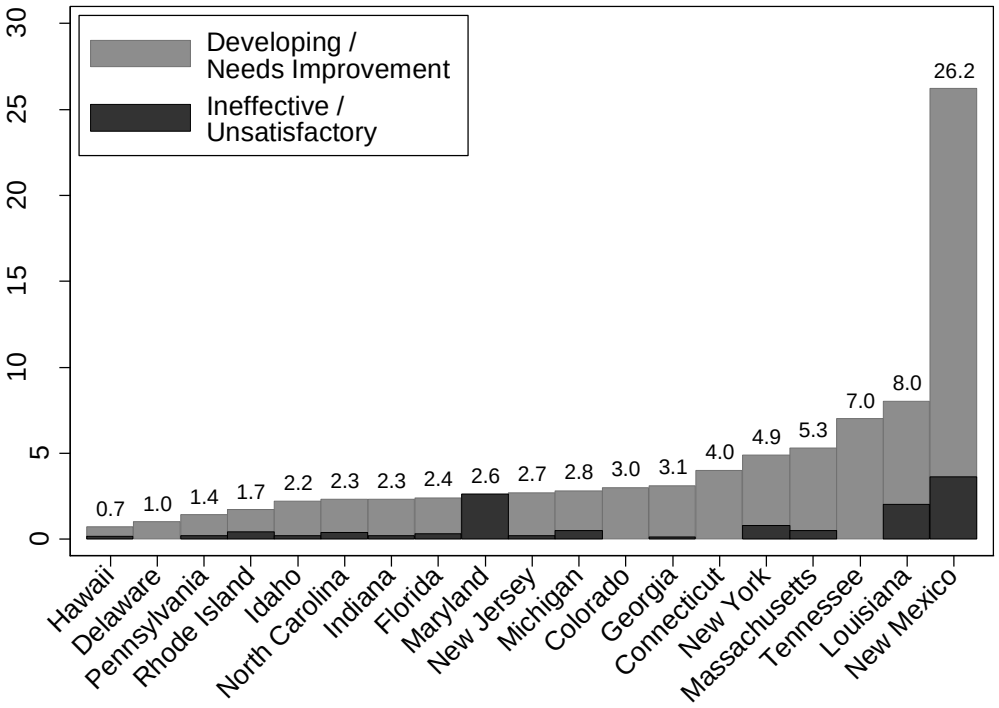


Figure 1: The percentage of teachers rated below Proficient across 19 state evaluation systems.

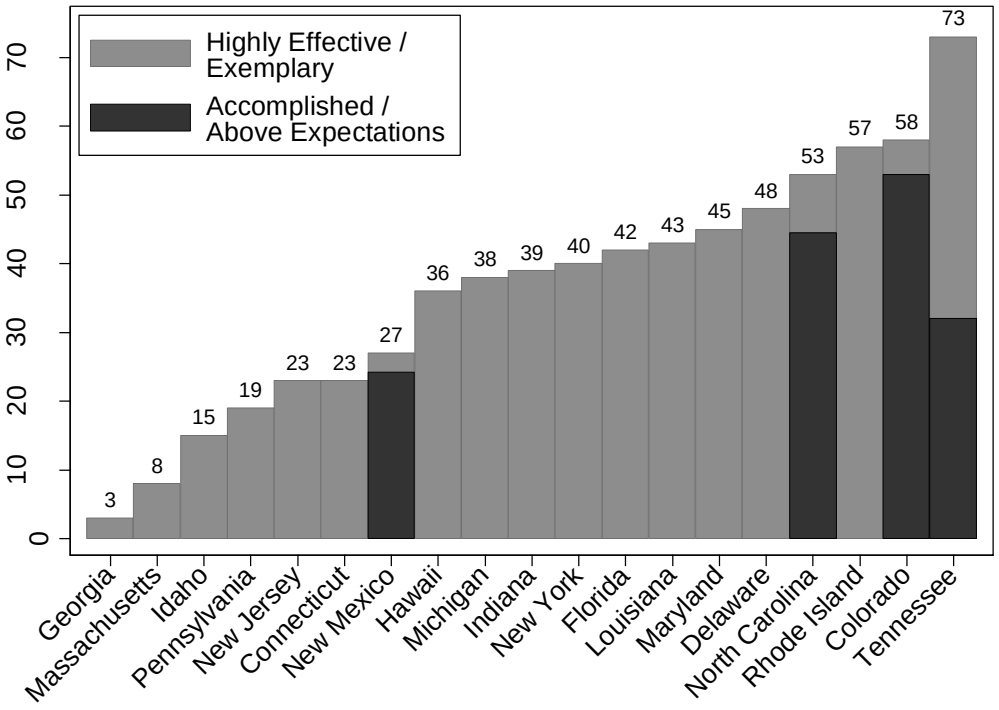
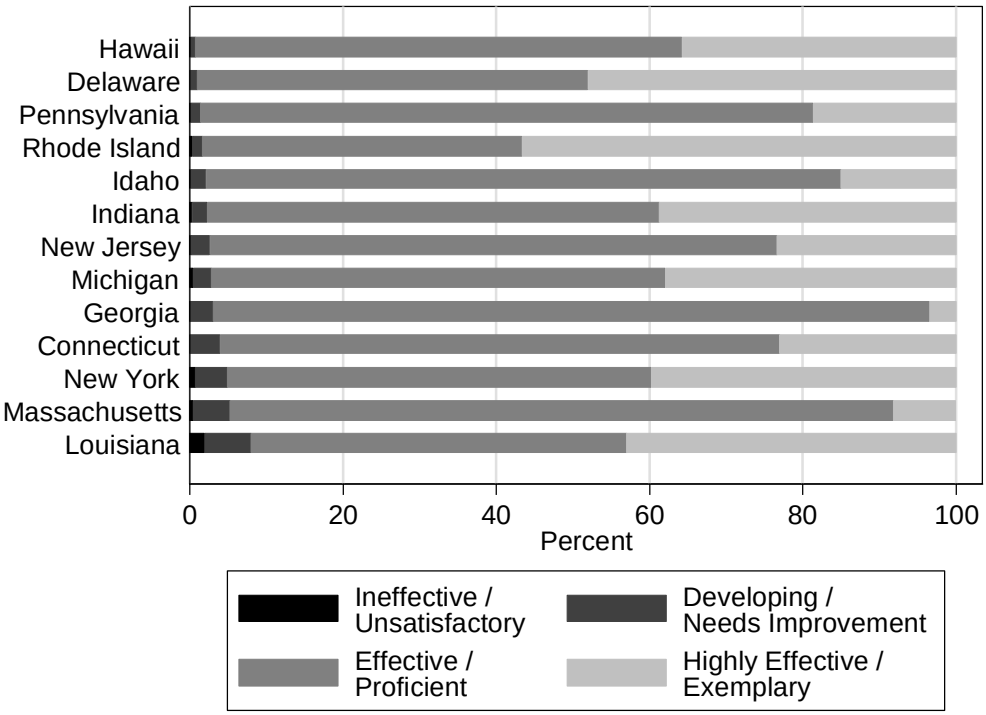


Figure 2: The percentage of teachers rated above Proficient across 19 state evaluation systems.

Panel A: States with four performance categories



Panel B: States with five performance categories

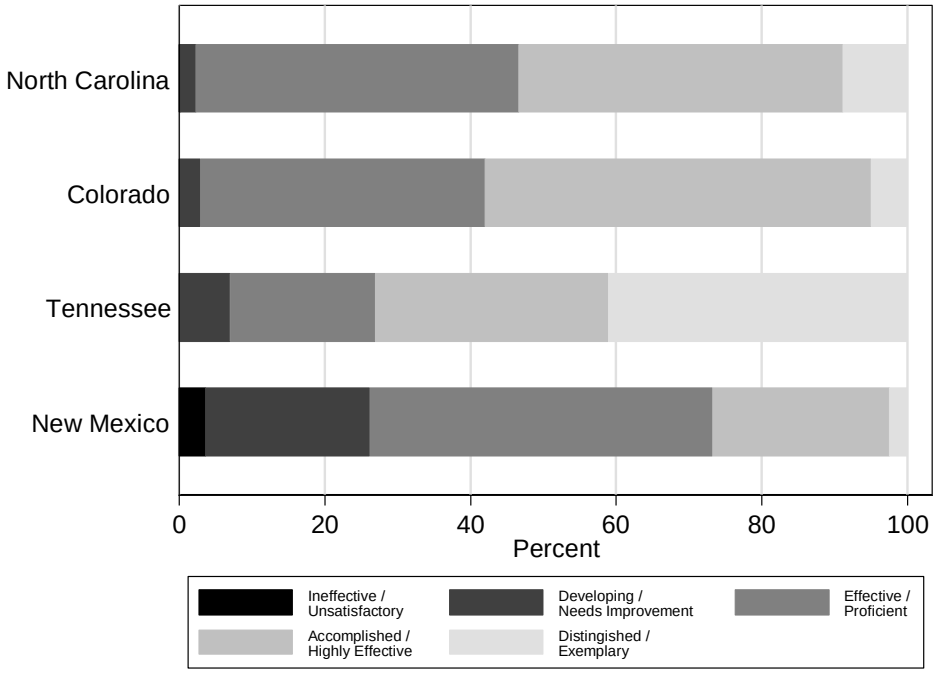
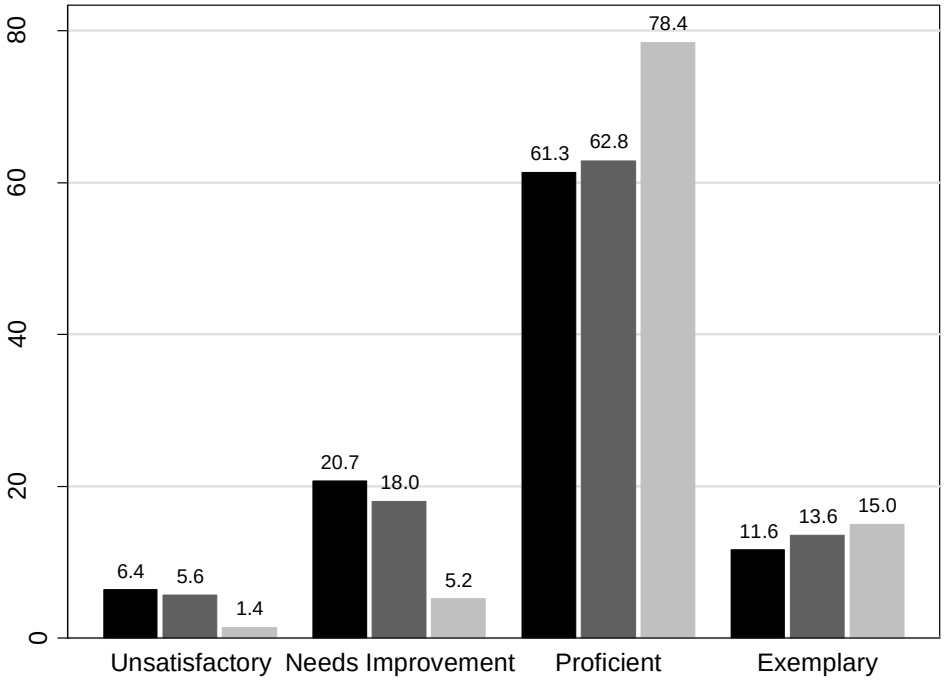


Figure 3: The distribution of teacher evaluation ratings across states with four (Panel A) and five (Panel B) rating categories.
Note: We exclude Florida from Panel B because its performance categories do not align with other states.

Panel A: 2012/13



Panel B: 2014/15

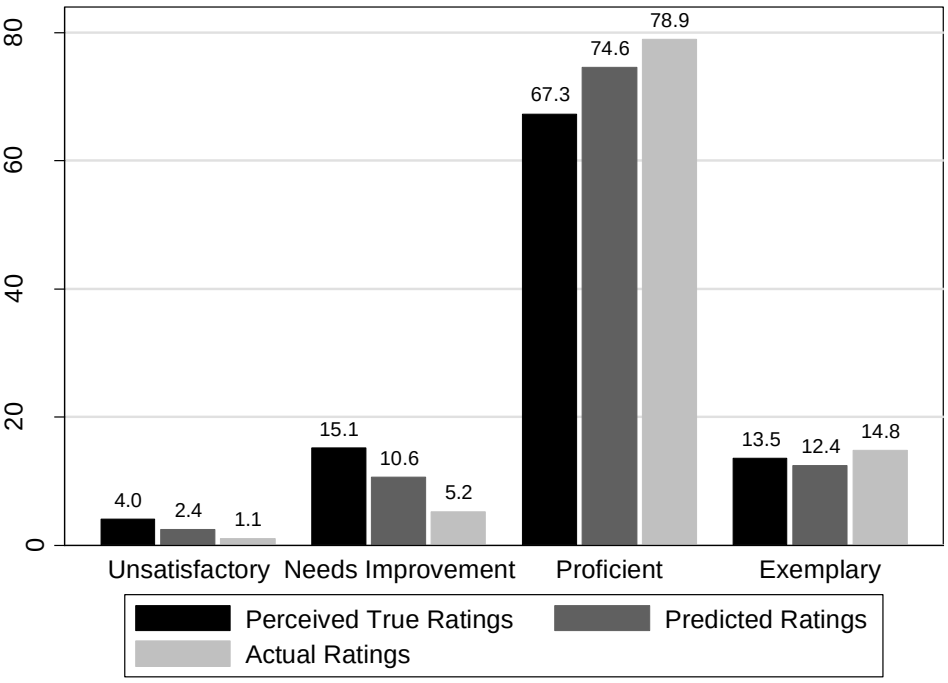


Figure 4: The perceived, predicted and actual distribution of teacher evaluation ratings in evaluators' schools in the first (Panel A) and third (Panel B) year of a new teacher evaluation system.

Note: Perceived true ratings are evaluators' assessments of the actual effectiveness of all classroom teachers in their school. Predicted ratings are evaluators' estimates of the summative evaluation ratings teachers in their school will receive at the end of the school year. Actual ratings are the summative evaluation ratings assigned to teachers in their school at the end of the school year. Bars for perceived and predicted ratings represent averages across all evaluators who had complete survey data and could be linked to school evaluation data. Bars for actual evaluation ratings represent a weighted average of the percentage of teacher to receive a given performance evaluation rating across the schools represented in our evaluator sample. Weights are derived based on the number of evaluators per school that completed the survey. This approach allows for a direct comparison between evaluators' average perceptions and predictions to the actual performance ratings. The samples consisted of 107 evaluators in 2012/13 and 157 evaluators in 2014/15.

Appendix A

Appendix Table A1: Background Information on State Evaluation Systems

	Colorado	Connecticut	Delaware	Florida	Georgia	Hawaii	Idaho	Indiana	Louisiana	Maryland
Year of Full Implementation	2015-2016	2016-2017	2012-2013	2011-2012	2014-2015	2013-2014	2013-2014	2012-2013	2012-2013	2013-2014
School Year	2013-2014*	2012-2013*	2013-2014	2013-2014	2012-2013	2014-2015	2013-2014	2013-2014	2013-2014	2014-2015
Evaluation System Structure	statewide /approved alternative	statewide /approved alternative	statewide	district designed w/ state criteria	statewide	statewide	district designed w/ state criteria	district designed w/ state criteria	statewide /approved alternative	statewide /approved alternative
Race to the Top Winner	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Number of Performance Categories	5	4	4	5	4	4	4	4	4	3
% Below Proficient	3.0	5.0	1.0	2.4	3.1	0.7	2.2	2.3	8.0	2.6
% Above Proficient	58.0	23.0	48.0	41.9	3.4	35.7	15.0	97.7	43.0	44.6
% in Category 1 (bottom)	0.0	<1	0.0	0.3	0.1	0.2	0.2	0.4	2.0	2.6
% in Category 2	3.0	4.0	1.0	0.7	3.0	0.6	2.0	2.0	6.0	52.8
% in Category 3	39.0	73.0	51.0	1.4	93.5	63.5	82.8	59.0	49.0	44.6
% in Category 4	53.0	23.0	48.0	55.7	3.4	35.7	15.0	38.7	43.0	
% in Category 5 (top)	5.0			41.9						
Name of Category 1 (bottom)	Basic	Below Standard	Ineffective	Unsatisfactory	Ineffective	Uneffective	Unsatisfactory	Ineffective	Ineffective	Ineffective
Name of Category 2	Partially Proficient	Developing	Needs Improvement	3 yrs- Developing	Needs Development	Marginal	Basic	Improvement Necessary	Effective: Emerging	Effective
Name of Category 3	Proficient	Proficient	Effective	Needs Improvement	Proficient	Effective	Proficient	Effective	Effective: Proficient	Highly Effective
Name of Category 4	Accomplished	Exemplary	Highly Effective	Effective	Exemplary	Highly Effective	Distinguished	Highly Effective	Highly Effective	
Name of Category 5 (top)	Exemplary			Highly Effective						

Notes: * Represents states for which only pilot data are available. Data on evaluation system structures are from state Department of Education reports the National Council on Teacher Quality 2015 State Teacher Policy Yearbook. See state performance evaluation data sources below for specific sources.

Appendix Table A1 Continued: Background Information on State Evaluation Systems

	Massachusetts	Michigan	New Jersey	New Mexico	New York	North Carolina	Pennsylvania	Rhode Island	Tennessee
Year of Full Implementation	2013-2014	2018-2019	2013-2014	2013-2014	2012-2013	2011-2012	2013-2014	2012-2013	2011-2012
School Year	2013-2014	2013-2014*	2013-2014	2014-2015	2013-2014	2013-2014	2013-2014˜	2013-2014	2013-2014
Evaluation System Structure	statewide /approved alternative	district designed w/ state criteria	district designed w/ state criteria	district designed w/ state criteria	district designed w/ state criteria	statewide /approved alternative	statewide /approved alternative	statewide /approved alternative	statewide /approved alternative
Race to the Top Winner	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Number of Performance Categories	4	4	4	5	4	5	4	4	5
% Below Proficient	5.3	2.8	2.7	26.2	5.0	2.3	1.4	1.7	7.0
% Above Proficient	8.1	37.9	23.4	26.7	39.8	53.3	18.6	56.6	73.0
% in Category 1 (bottom)	0.5	0.5	0.2	3.6	0.8	0.2	0.2	0.4	0.0
% in Category 2	4.8	2.3	2.5	22.6	4.2	2.1	1.2	1.3	7.0
% in Category 3	86.5	59.3	73.9	47.1	55.3	44.4	80.0	41.7	20.0
% in Category 4	8.1	37.9	23.4	24.2	39.8	44.5	18.6	56.6	32.0
% in Category 5 (top)				2.5		8.8			41.0
Name of Category 1 (bottom)	Unsatisfactory	Ineffective	Ineffective	Ineffective	Ineffective	Not Demonstrated	Failing	Ineffective	Sig. Below Expectations
Name of Category 2	Needs Improvement	Minimally Effective	Partially Effective	Minimally Effective	Developing	Developing	Needs Improvement	Developing	Below Expectations
Name of Category 3	Proficient	Effective	Effective	Effective	Effective	Proficient	Proficient	Effective	At expectations
Name of Category 4	Exemplary	Highly Effective	Highly Effective	Highly Effective	Highly Effective	Accomplished	Distinguished	Highly Effective	Above Expectations
Name of Category 5 (top)				Exemplary		Distinguished			Sig. Above Expectations

Notes: * Represents states for which only pilot data are available. ˜ indicates data were calculated with incomplete information. Data on evaluation system structures are from state Department of Education reports the National Council on Teacher Quality 2015 State Teacher Policy Yearbook. See state performance evaluation data sources below for specific sources.

State Performance Evaluation Data Sources

Colorado

Colorado DOE. (2015, January). Starting the Journey: Progress Report on Colorado's Educator Evaluation and Support System. 2. Retrieved from https://www.cde.state.co.us/educatoreffectiveness/reportcoevaluationsystem

Connecticut

Connecticut DOE. (2015). 2015 SEED Handbook: Connecticut's System for Educator Evaluation and Development. Retrieved from http://www.connecticutseed.org/wp-content/uploads/2015/11/2015_SEED_Handbook_11_24_15.pdf

Donaldson, M., Cobb, C., LeChausseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. (2014, January 1). An Evaluation of the Pilot Implementation of Connecticut's System of Educator Evaluation and Development. *University Of Connecticut Center for Education Policy Analysis, Neag School of Education*, 63. Retrieved from https://s3.amazonaws.com/s3.documentcloud.org/documents/1010013/neag-teacher-evaluation-report-january-2014.pdf

Delaware

Delaware DOE. (2014, October). Performance Matters: A Report on "Year Two" of the Revised DPAS - II Educator Evaluation System. 29. Retrieved from http://www.doe.k12.de.us/cms/lib09/DE01922744/Centricity/domain/271/present%20and%20reports/DPAS-II_Year_2_Report_2014.pdf.

Florida

Florida DOE. (n.d.). Annual Legislative Report on Teacher Evaluation. Retrieved from http://www.fldoe.org/core/fileparse.php/7503/urlt/1314AnnualLegisReportTeacherEval.pdf.

Florida DOE. (n.d.). Personnel Evaluation. 2. Retrieved from http://www.fldoe.org/core/fileparse.php/7503/urlt/1314EduEvalRatings.pdf

Georgia

Barge, J. (2014, February 21). Overview/Executive Summary of the 2012-2013 TKES and LKES Evaluation Report. *Georgia Department of Education,* 29. Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/FINAL%20Year%203%20Report%20_2-21-2014_FORMATTED%202-23-2014.pdf.

Hawaii

Hawaii DOE. (2015). Educator Effectiveness System. Retrieved from http://www.hawaiipublicschools.org/TeachingAndLearning/EducatorEffectiveness/EducatorEffectivenessSystem/Pages/home.aspx.

Kalani, N. (2015, October 26). Teachers' Evaluation Grade. *Star Advertiser*. Retrieved from http://www.pressreader.com/usa/honolulu-star-advertiser/20151026/282621736573997/TextView.

US DOE. (2015, April). Race to the Top: Hawaii Report Year 4: 2013-2014. Retrieved from http://www2.ed.gov/programs/racetothetop/phase1-report/hirttyrrpt42015.pdf.

- Note: Percentages are calculated based on the total amount of teachers rated as reported by a news article.

Idaho

Corbin, C. (2015, June 12). Teachers Got Identical Evaluations Across 32 Idaho Districts, Complicating Career Ladder Pay Plan. *Idaho Statesman.* Retrieved from http://www.idahostatesman.com/2015/06/12/3848480/teachers-got-identical-evaluations.html.

Indiana

Ritz, G. (n.d.). Staff Performance Evaluation Results. *Indiana Department of Education*, 5. Retrieved from http://www.in.gov/sboe/files/ER_Data_Presentation_to_SBOE_-_v._12.30.14.pdf.

- Note: the original data presented in the DOE report included distribution percentages that included all teachers, not just ones who were evaluated, and excluded other administrators such as principals. 8.64% of the teachers were not evaluated. Data in our study was adjusted to only include evaluated teachers by dividing the percentage value of teachers in each rating category by the percentage value of teachers evaluated, 91.36.

Louisiana

Louisiana DOE. (n.d.). Improving Teaching & Leadership. 3. Retrieved from https://www.louisianabelieves.com/docs/default-source/teaching/2013-2014-compass-annual-report.pdf?sfvrsn=2.

- Note: The DOE report presents the distribution of evaluation scores for both teachers and leaders. We only used data for teachers.

Maryland

Smith, J.R. (2015, October 27). Teacher and Principal Evaluation Data: Effectiveness Rating from SY 2014-2015. 6. Retrieved from http://www.marylandpublicschools.org/stateboard/boardagenda/10272015/Tabs-F1-F2-TeacherPrincipalEvaluationReportUpdate.pdf.

Massachusetts

Massachusetts Department of Elementary of Secondary Education. (2014). *2013-14 Educator Evaluation Performance Statewide Report* (Data File). Retrieved from http://profiles.doe.mass.edu/state_report/educatorevaluationperformance.aspx.

- Note: We took the evaluation data for the group type "teachers."

Michigan

Michigan DOE. (n.d.). Educator Evaluations & Effectiveness in Michigan. 18. Retrieved from http://www.michigan.gov/documents/mde/2013-14_Educator_Evaluations_and_Effectiveness_485909_7.pdf?20160115090609.

New Jersey

New Jersey DOE. (n.d.). 2013 - 2014 Final Educator Evaluation Implementation Report. 13. Retrieved from http://www.nj.gov/education/AchieveNJ/resources/201314AchieveNJImplementationReport.pdf.

New Mexico

New Mexico Public Education Department. (2015, May 4). 2015 Teacher Evaluation Results: Statewide Overview. 4. Retrieved from http://ped.state.nm.us/ped/NMTeachDocs/Toolbox/2015%20NMTEACH%20Results_Overview%2005%2004%2015_Final.pdf.

New York

New York State Education Department. (2015). *2013-2014 Teacher Evaluation Database* (Data File). Retrieved from http://data.nysed.gov/downloads.php.

- Note: Data accessed is titled "2013-2014 Teacher Evaluation Database" located under the "Downloads" tab in the NYSED data website, http://data.nysed.gov. The dataset is a spreadsheet of every New York teacher evaluated and the rating category they received, from ineffective to effective. The percentage distribution was calculated using the number of teachers in each rating category divided by the total number of teachers in the downloaded dataset. Evaluated principals were removed from the dataset and not included in the calculations.

North Carolina

North Carolina Department of Public Instruction. (2014). *North Carolina Educator Effectiveness Data* (Data File). Retrieved from http://apps.schools.nc.gov/pls/apex/f?p=155:5:0::NO

- Note**:** North Carolina data is presented as the number and percent distribution of the five rating scores for each evaluation standard. There are five evaluation standards in total. Our data is the average of the percent values of each five standards in a given category. For example, the percentage of teachers classified as "Not Demonstrated" is taken as an average of the percentage of teachers who scored "Not Demonstrated" in the five standards.

Pennsylvania

Chute, E. (2015, June 15). How Qualified are Pennsylvania's Teachers? The Numbers Say Extremely. *Pittsburgh Post-Gazette*. Retrieved from http://www.post-gazette.com/news/education/2015/06/15/New-rating-system-finds-nearly-all-Pennsylvania-teachers-are-qualified/stories/201506080003.

- Note: Data is for the 2013-2014 SY and may not be exact due to inconsistent reporting. The data in our spreadsheet is an estimation based on the information reported in this article. The article reports the number of teachers evaluated as distinguished, needs improvement, and failing and reports the estimated percentage of teachers evaluated as proficient ("over 80%"). It does not, however, report the total number of teachers

evaluated. Our data is calculated by estimating the number of teachers rated as proficient by calculating what number of proficient teachers would garner the 80% figure presented in the article. Once we calculated the estimated number of proficient teachers, we could estimate the total number of teachers and calculate the percentage distribution of rating scores.

Rhode Island

Rhode Island DOE. (2014, October). RI Educator Evaluation Systems: Improving Teaching and Learning. 4. Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Education-Eval-Main-Page/FER_Year2_Report_Final.pdf.

Tennessee

Koedel, C., Li, J., & Springer, M.G. (2014, October). "The Impact of Performance Ratings on Job Satisfaction for Public School Teachers." *Tennessee Consortium on Research, Evaluation and Development*, 31. Retrieved from http://www.tnconsortium.org/data/files/gallery/ContentGallery/The_Impact_of_Performance_Ratings_on_Job_Satisfaction_for_Public_School_Teachers_.pdf.

US DOE. (2015, April). Race to the Top: Hawaii Report Year 4: 2013-2014. Retrieved from http://www2.ed.gov/programs/racetothetop/phase1-report/tnrttyrrpt42015.pdf.

Supplemental Source

National Council on Teacher Quality. (2015). State Teacher Policy Yearbook. Retrieved from http://www.nctq.org/statePolicy/2014/policyIssueFindings.do?policyIssueId=6&masterGoalId=11&stateId=23.

Appendix B

Evaluator Survey Items

1) In your opinion, what percent of teachers ***at your school*** perform at an Unsatisfactory, Needs Improvement, Proficient, or Exemplary level as defined by the XXXX rubric? Your total must add up to 100%.

Unsatisfactory	____%
Needs Improvement	____%
Proficient	____ %
Exemplary	____%
Total	100%

Note: The name of the evaluation system is redacted to maintain the confidentiality of the district.

2) Based on your best guess, what percent of teachers ***at your school*** will receive summative performance ratings of Unsatisfactory, Needs Improvement, Proficient, or Exemplary at the end of the academic year? Your total must add up to 100%.

Unsatisfactory	____%
Needs Improvement	____%
Proficient	____ %
Exemplary	____%
Total	100%

Appendix C:

Interview Protocol

Narrative about Research Project and Framework of Interview (Read to Interviewee):

Hi my name is XXXXXX and I'm a member of a research team from Brown and Vanderbilt studying the experiences of principals in implementing new evaluation systems. We are interested in your opinions about, and experiences with, the teacher evaluation system in XXXX. I'll ask you a series of questions meant to give you the opportunity to share your thoughts about the transition from the old to the new teacher evaluation system. We particularly hope to learn about whether this change has made a difference in your work. We are also interested in how you decide which ratings to give to teachers under this new system and whether/how the new system supports professional growth and development among teachers. The interview should last approximately 50 minutes.

The information you share is completely confidential. No individuals or schools will be identified in any written reports or presentations. This information will be the basis of a scholarly article and a set of recommendations we provide to XXXXXXX on how to improve the Educator Evaluation System.

I would like to record the conversation so I can focus on what we discuss rather than taking detailed notes, is that ok with you?

Personal & School Background:

Step 1: Briefly review the information on the demographic questionnaire to be sure it is correct.

1. What makes your school unique compared to other schools in XXX?
2. What is the biggest challenge you face as a principal at your school?

Evaluation Background:

1. Were you responsible for evaluating teachers under the old teacher evaluation system? If yes . . .
2. What do you think was the primary purpose of teacher evaluation under the old system?
3. What did you view as the strengths and weakness of this old system?

Current Evaluation System:

1. Do you think the primary purpose of teacher evaluation has changed under the new teacher evaluation system? If so, how and why?
2. What are the strengths and weaknesses of the new evaluation system?
3. What are the opportunities and challenges associated with being both a supervisor and instructional leader/coach as part of the new teacher evaluation system?
4. In your experience, does your relationship with a teacher affect how you deliver feedback and what feedback you provide? If so, can you please provide an example?

5. Are there certain grades or subjects in which you feel more comfortable evaluating teachers? If so, why . . .
6. Research studies suggest that teachers receive positive evaluations even when their performance is unsatisfactory or in need of improvement.
    a. Did this happen in XXXXX under the old evaluation system? If so, why . . ?
    b. Does this happen in XXXXX under the new evaluation system? If so, why . . ?
7. In your experience, does the new evaluation system makes it easier or more difficult to rate a teacher as unsatisfactory (or needs improvement)? Please explain . .
8. Are there ever situations when you were unable or unwilling to give a low rating? Can you give an example?
9. Were any teachers at your school rated as unsatisfactory? If so, why?
10. Were any teachers at your school rated as needs improvement? How do you use this rating?
11. Did rating a teacher as unsatisfactory or needs improvement affect your relationship with other teachers in the school? If so , how?
12. What proportion of your time completing evaluations do you spend observing & collecting data vs. writing & entering in feedback, vs. meeting with teachers to discuss feedback?
13. Do teachers rated as proficient or exemplary receive the same amount and type of feedback (written vs. in person) as those rated as unsatisfactory or needs improvement?
14. In your experience, does the feedback teachers receive via the evaluation process help teachers improve their practice? How? please provide a specific example.
15. Research suggests that teachers can be reluctant to act on the feedback they receive, what strategies do you use to communicate feedback effectively and build teacher's buy-in?

<u>Evaluation & Improvement</u>:

1. What do you think the primary purpose of teacher evaluation systems should be?
2. What training and support would be most useful to you to help you improve your ability to provide feedback to teachers about how to improve their practice?
3. If you could change anything about the Educator Evaluation System, how would you change it?
4. What do you think is the best way to improve instruction at your school?

<u>Closing Question</u>:

1. Are there any other issues or points you would like to raise before we conclude the interview?

Appendix D:

Original and Final Codes

Original Codes	Final Codes
Category: Evaluation systems	**Category: Evaluation**
Modified old evaluation system	Old
Old system- not enough feedback	New
Old system- easy to complete	General
Old system- not everyone evaluated	
Old system- easy to use, predictable	**Category: Pro**
Old system- flexible	Online
Old-system- teachers did not look at evaluation	Efficient/flexible
Found old system useful	Multiple rating categories
Compliance	Rubric/evidence
Evaluation for dismissal	Teacher involvement
Evaluation to improve struggling teachers	Other
Evaluation to improve teacher practice	Online
More time on low rated teachers	Efficient/flexible
Weakness of binary system	Multiple rating categories
Binary system hard to rate teachers accurately	Rubric/evidence
Average teachers with areas to improve	Teacher involvement
Collaborative	Other
Four categories	
Likes online system	**Category: Challenges**
Dislikes online system	Binary
Time consuming	Focus on compliance
Gave low rating	Time consuming/number of people to evaluate
Did not improve with help	Rubric
Needs improvement to help mediocre teacher	Proficient teachers who want exemplary
Developing skills	Distinguishing categories
Negative reaction to NI rating	Other
NI in area but not overall	
More time with low rater teachers	**Category: Time allocation**
Teacher leadership	Distribution of time across ratings
District assistance with evaluation	
Punitive	**Category: Experience as instructional coach**
Rubric is helpful	Harder to coach outside of expertise
Artifacts/evidence is helpful	Tailoring feedback
Artifacts/evidence is not helpful	No time for in person feedback
Identifies mediocre teachers	Professional practice goals
No time for conferences	Supervision vs. instructional leader
Writing is time consuming	Focus on pedagogy
Counseled out	Focus on a few dimensions only
Provided help but did not improve	Other
Evaluation too serious for use as PD	
Isolating experience of being evaluator	
More comfortable evaluating teachers in familiar subjects	

Evaluated pedagogy even if not familiar with subject
Evaluator accountability
Group goals
Equal feedback not dependent on rating
Does not like deadlines
Self-evaluation/assessment
Setting goals
Thoughtful practice
Frustration with lack of flexibility
Bureaucratic
Clearer standards for non classroom teachers
Teacher involvement
Proficient teachers who were upset not rated exemplary
More attention to those rated low
Leverage certain standards
Feedback harder when not in experience area
Good pressure

Category: Why not rated unsatisfactory	**Category: Why not rated unsatisfactory**
Time constraints	Time consuming & barriers to removal
Challenge of dismissing teacher	Arbitration
New system, more accurate ratings	Hard conversations
Easier to counsel out	Rate on potential
Hard to give low rating to someone previously rated satisfactory	Not enough data
Rated based on potential	Receive worse teacher
Time required with low rating	Binary
Avoiding arbitration	Experience of not rating unsat/NI when they should
No problem giving a low rating	Uncomfortable making that assessment
Easier to give low rating with new system	Other
Challenge of delivering negative feedback	
Duty to give low rating	**Category: Experience giving a low rating**
Get a worst teacher as a replacement	Teacher improved
Hard to be the cause of someone losing job	Teacher did not improve
Contractual obligations that make it hard to dismiss	Teacher focused on rating
Rates differently based on relationship with teacher	Teacher did not return to school
No low ratings because of autonomy to hire	Other
No improvement, gets an unsat rating	
Giving low score in sub-category	
Race	
Not enough data	
Bully teachers	

Category: Supports needed	**Category: Supports needed**
How to manage time	Operational
How to use goals	Calibration
Eliminate the managerial parts of job/operational support	How to provide feedback
Calibrate ratings	Get feedback
Observing other principals	Content area coaches

Coaching other admin on the system Better technology Best practices for giving feedback How to remove teacher Better definition/modeling of exemplary	Other **Category: Strategies for improving instruction** Peer observation Teacher collaboration/teams Student data Coaching/feedback PD

Category: Purpose of teacher evaluation	**Category: Purpose of teacher evaluation**
Dismissal Instructional coaching Teacher collaboration Peer evaluation End classroom isolation (egg-crate mentality) Data for improvement Wants tool for dismissal New system- faster to dismiss Need separate system for evaluation and dismissal Better preservice preparation School environment Time Selection (reward and punishment) Distinguishing between bad and those who can grow	Removal Improvement Both