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

Matthew A. Kraft
Brown University

Allison F. Gilmour
Vanderbilt University

July 2016

Abstract

In 2009, TNTP's The Widget Effect documented the failure to recognize and act on differences in teacher effectiveness. We revisit these findings by compiling teacher performance ratings across 24 states that adopted major reforms to their teacher evaluation systems. In the vast majority of these states, the percentage of teachers rated Unsatisfactory remains less than 1%. However, the full distributions of ratings vary widely across states with 0.7% to 26% rated below Proficient and 3% to 62% rated above Proficient. 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 rate as such. Interviews with principals reveal several potential explanations for these patterns.

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

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. The Widget Effect was not the first or only study to draw attention to the failure 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 studies also characterized teacher evaluation as a superficial exercise that failed to assess instructional quality or to inform teacher professional development and personnel decisions (Donaldson, 2009; Toch & Rothman, 2008).

Growing recognition of the broken teacher evaluation system amplified by new research documenting the importance of teacher effectiveness (e.g. Rockoff, 2005; Rivkin, Hanushek, & Kain, 2005) 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 created strong incentives for 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 initiatives along with local reform efforts 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 this focus on high-stakes evaluation systems as misplaced (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. One camp of scholars (Hanushek, 2009) and journalists (Thomas, Wingert, Conant, & Register, 2010) emphasize the importance of differentiating among teachers in order to motivate them through performance incentives and to dismiss those judged to be low performing. 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). Both of these theories of action require an evaluation system that differentiates among teachers and accurately assesses 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 on teacher satisfaction (Koedel, Li,

& Springer, 2015) and 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. Policymakers have assumed that the sweeping changes to evaluation system features would result in greater differentiation, ignoring Lipsky's (1980) seminal observation that policies are ultimately made by the "street-level bureaucrats" who implement them. History shows that the success of policy initiatives depends on the will and capacity of local actors to implement reforms (Honig, 2006). This is particularly true in the decentralized U.S. education system where local practice is often decoupled from central policy (Spillane & Kenney, 2012). Guided by this lens, we ask: What is the distribution of teacher performance ratings in states that have adopted reforms to their teacher evaluation systems? Does the distribution of teacher performance ratings reflect principals' perceptions about the distribution of teacher effectiveness? And, if not, what are principals' explanations for why teacher evaluation reforms have not resulted in greater differentiation in performance ratings?

We examine these questions with quantitative and qualitative data collected over the course of three years. We begin by presenting data on the distribution of teacher evaluation ratings across 24 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 distribution of teacher effectiveness with both their predictions and actual ratings. 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 actual performance ratings. Throughout the paper we focus much of our analyses and discussion on the percentage of performance ratings below and above Proficient given the high-stakes incentives and consequences attached to these ratings in many districts (e.g. Dee & Wyckoff, 2015). 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 and outreach 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 studies, reports, and news articles containing information on teacher evaluation ratings using Google's advanced search features as well as academic databases such as ERIC and Academic Search Premier. Finally, we reviewed information on state education agency websites and directly contacted agency staff to request data. Our search produced data on the distribution of teacher effectiveness for 24 states including 14 RTTT winners. Among these states, 17 rated teachers across four performance categories, six use five categories and one used three categories. The primary sources for these data include 16 state department of education reports, four data requests, three

news articles, and one research paper. 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 focus 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.

For many years in the district, evaluation consisted of administrators completing binary checklists. Evaluations were infrequent and many teachers went unevaluated. For example, 83% of non-tenured teachers and 77% of tenured teachers were unevaluated in 2008/09. In 2012/13, the district implemented a new evaluation system that was adapted from the state's new framework for evaluation. Under the new system, principals and select members of their administrative teams (e.g. Assistant Principals, Directors of Instruction) are responsible for conducting formative and summative evaluations of teachers. Evaluators consider evidence from classroom observation ratings on the district rubric as well as artifacts and progress towards teacher-defined Student Learning Goals. They then assign teachers an overall performance rating based on their holistic assessment of the evidence rather than a weighted sum of multiple measures. Performance measures based on standardized tests such as value-added scores or student growth percentiles were not calculated or incorporated into the evaluation system at the time of this study.

Throughout the paper we focus on the overall summative (and formative) ratings evaluators' assigned to teachers rather than any specific rating component on the district rubric.

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 structured evaluation plans with frequent observations, which, after repeated low evaluations, can result in dismissal. There are no formal incentives in place for receiving an exemplary rating (See Authors [2016] for a more detailed description of the evaluation system). The distribution of performance ratings in the district is broadly similar to the state's distribution but slightly skewed upward with several percentage points fewer teachers rated as proficient and more rated as Exemplary.

Evaluator surveys. We worked with district officials to administer a 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 judgment* were in each of the four performance categories and (2) to predict the percentage of teachers in their school that *will receive* overall summative evaluation ratings 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 linked evaluators' survey responses with the actual summative 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

weighted all calculations based on the actual distribution of performance ratings by the number of evaluators who completed our survey in each school to make them directly comparable to our survey results. 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.[1] 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 available upon request 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. We created six strata based on school size and level. 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 those principals in the district we interviewed and those we did not (for full details see Authors, 2016).

We interviewed each principal for 45-60 minutes using a semi-structured interview 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. 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/Effective among the 24 states in our analytic sample. The median percentage of teachers rated below Proficient is 3.05% while the weighted average across these states is 4.60% (5.38% unweighted) where weights are based on the number of public school teachers in each state in 2013/14 (Glander, 2015). Figure 1 illustrates how the percentage of teachers rated as below Proficient varies substantially across states. Eight states identified 5% of teachers or more as below Proficient, eleven states rated between 2% and 5% of teachers as below Proficient, and in five states less than 2% of teachers received below Proficient ratings. The majority of these teachers fall in the Developing/Needs Improvement category. Across all states, the weighted average of teachers rated Unsatisfactory/Ineffective is 0.49% (0.56% unweighted); only two states, Maryland and New Mexico, rated more than 1% of teachers in the lowest category.

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 38.5% (with a weighted and unweighted average of 34.47% and 35.02%), but varies considerably from 3% in Georgia to 62% 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 five 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. Proficient vs. Exemplary). 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, 98.4% 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 of the ratings spectrum. These findings suggest that exchanging binary rating systems for multiple rating categories does not guarantee more differentiated ratings. 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).[2]

Evaluators may not have fully 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 examine this possibility with survey data from 2014/15, the third year of the new evaluation system. Again, we find similar patterns as shown in Figure 4B where evaluators 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 again overestimated the proportion of teachers they would rate in one of the two lowest performance categories (13% predicted), but less so than in 2012/13.

We extend these analyses by comparing the distributions of formative and summative performance ratings across the schools included in our survey sample. As shown in Figure 5, evaluators appear more likely to assign lower formative ratings. Twice as many teachers received formative Needs Improvement ratings compared to summative ratings while only half as many teachers received formative Exemplary ratings compared to summative ratings. Improvement in teacher practice over the course of the year is unlikely to fully account for these changes given that this pattern is consistent across both years. Instead, teachers' formative ratings in 2014/15 shift back to levels similar to the 2012/13 formative ratings rather than more closely resembling the 2012/13 summative ratings distribution.

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

Further, comparing survey results across both years suggests evaluators became 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 competing tradeoffs 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 most 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 who's 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 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 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 experienced elementary school principal 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' ... 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.

Principals sometimes 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 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 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 of low ratings. 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." Relatedly, 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." 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 or addressing the underlying causes of inequitable performance ratings along racial lines.

The challenges of removing and replacing teachers. Several principals mentioned that they also sought to avoid the "long, laborious, legal, draining process" of evaluating out a teacher. Although the evaluation reforms implemented by the district aimed to streamline the dismissal process, it is unclear the degree to which these principals' perceptions were accurate or a justification for not utilizing the new process. 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.

Two principals expressed their hesitancy to initiate dismissals due to fear of having to hire an even lower-quality replacement. A secondary school 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.

An experienced high school principal described how she chose to rehire a low-performing 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."

This principal explained that she wanted to avoid the possibility that human resources would assign her a teacher from the excess pool at all costs. In her words, "The one you know is better than the one you don't."

Policy Implications

Spurred in part by The Widget Effect's "call to action," teacher evaluation reforms have changed the ways in which teachers are evaluated in U.S. public schools fundamentally. 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. Importantly, these changes have increasingly focused educators' attention on classroom instruction. However, we find that new evaluation systems have not consistently resulted in greater differentiation among teacher performance ratings. 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 most states now distinguish great teaching from good, but that they make distinctions between good, fair, and poor instruction must less consistently. Just as TNTP found, only a "fraction of a percentage" of teachers are rated Unsatisfactory in the vast majority of new systems. At the same time, in the majority of states we studied 3% or more of teachers were rated in categories below Proficient. In our view this represents a meaningful departure from ratings under old evaluation systems, but not a landmark change in the distribution.

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 play a role, but it seems unlikely these 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 evaluation metrics and performance standards must at least partially explain why only 3% of teachers in Georgia and 9% of teachers in Massachusetts are above Proficient but 62% meet this higher standard in Tennessee (Authors, 2016). Furthermore, we show that in one district, the ratings evaluators assigned to teachers differed substantially from what they perceived to be the true distribution of teacher effectiveness at their schools.

The variation in performance ratings across states is also likely a product of the differing sets of rewards and sanctions attached to these ratings. For example, pay-for performance programs with fixed budgets necessarily constrain the number of teachers it is feasible to rate as top performers. Given the wide range of incentive structures across districts in the same state, it is difficult to formally assess the role that high-stakes consequences play in shaping evaluation ratings at the macro level. Our exploratory case-study helps to illustrate how incentives shape evaluators' decisions at the micro level as they navigate implementation challenges, competing interests, unintended consequences, and high-stakes decisions. 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 extra

work this creates and the many other demands on their time. Telling someone they are not Proficient at their job can be a difficult thing to do. 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. These findings exemplify Lipsky's (1980) theory of street-level bureaucracy where policies are ultimately written by the people who implement them rather than the policymakers who design them.

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 cause 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. These reasons likely explain why more than twice as many teachers in the district received formative ratings below Proficient than did on their summative ratings. If formative evaluations serve to more accurately identify and address teachers' weaknesses then differentiation in the summative ratings may be less important. Moving towards evaluation systems that ask, "How is a teacher effective?" rather than "How effective is a teacher?" could serve to more accurately engage with the full range of teachers' strengths and weaknesses rather than a narrow focus on differentiating among them using a single rating scale.[3]

The limitations of the present study point to several areas for future research. Our focus is on describing ratings distributions under new teacher evaluation systems. These data provide a snapshot in time rather than a longitudinal trend or a causal framework for analyzing how evaluation reforms have affected the distribution of performance ratings. Important questions

also remain about how the design features of evaluation systems such as the performance measures, choice of evaluators, weights and cut points, as well as high-stakes decisions affect the distribution of evaluation ratings. Our survey and qualitative data are from a single district. There is reason to believe that evaluators' perceptions of the true distribution of teacher performance in this district are not necessarily reflective of state-wide distributions. Large urban districts often draw from different labor markets and serve different student populations than other types of districts in the same state. Future surveys attempting to capture perceptions about the true distribution of teacher effectiveness would benefit from examining perspectives across multiple contexts and, when possible, asking educators to rate individual teachers rather than estimating the full ratings distribution.

Teacher evaluation policies have undergone major reforms in recent years, but professional norms and practices around evaluation are proving much more difficult to change. The decisions and actions of administrators and educators themselves will ultimately determine whether teachers are treated as professional or interchangeable parts.

Endnotes

1. For principals 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.

2. The unweighted exact statistic for the average percent of teachers rated below proficient in these schools is 6.7% in 2012/13 and 5.7% in 2014/15.

3. We thank an anonymous reviewer for the suggesting the language used in this sentence.

References

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

Authors (2015).

Authors (2016).

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*, 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.

Glander, M. (2015). *Selected statistics from the public elementary and secondary education universe: School year 2013-14. First look. NCES 2015-151*. Washington, DC: National Center for Education Statistics.

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.

Koedel, C., Li, J., & Springer, M. G. (2015). *The impact of performance ratings on job satisfaction for public school teachers*. Retrieved from https://my.vanderbilt.edu/matthewspringer/files/2013/02/Koedel-et-al-Job-Satisfaction-january1.pdf

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

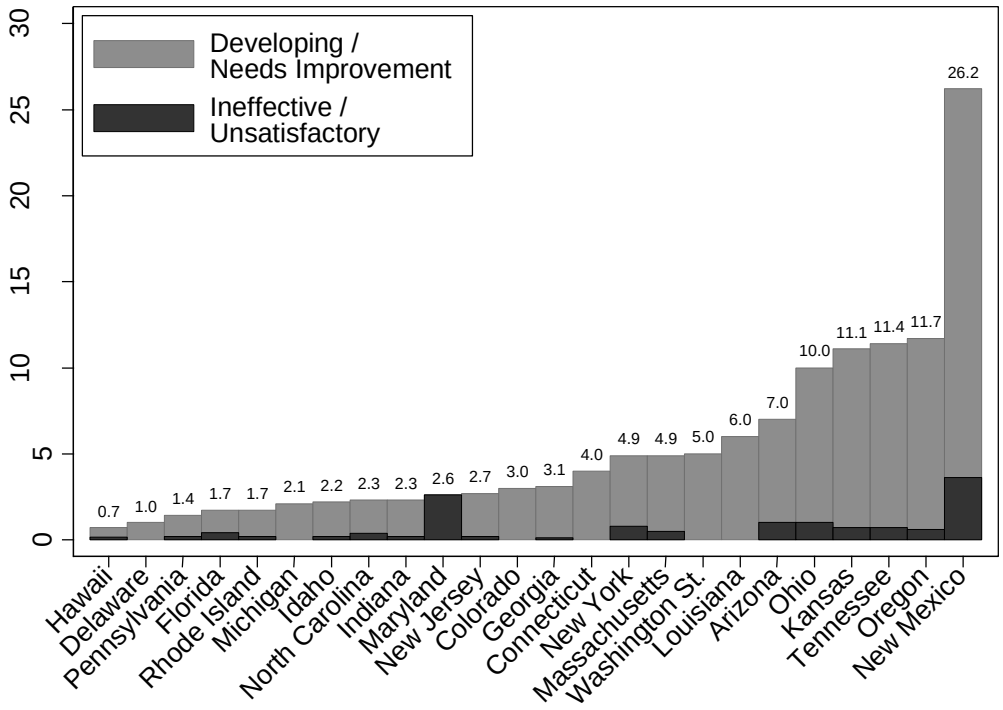

Figure 1: The percentage of teachers rated below Proficient across 19 state evaluation systems.

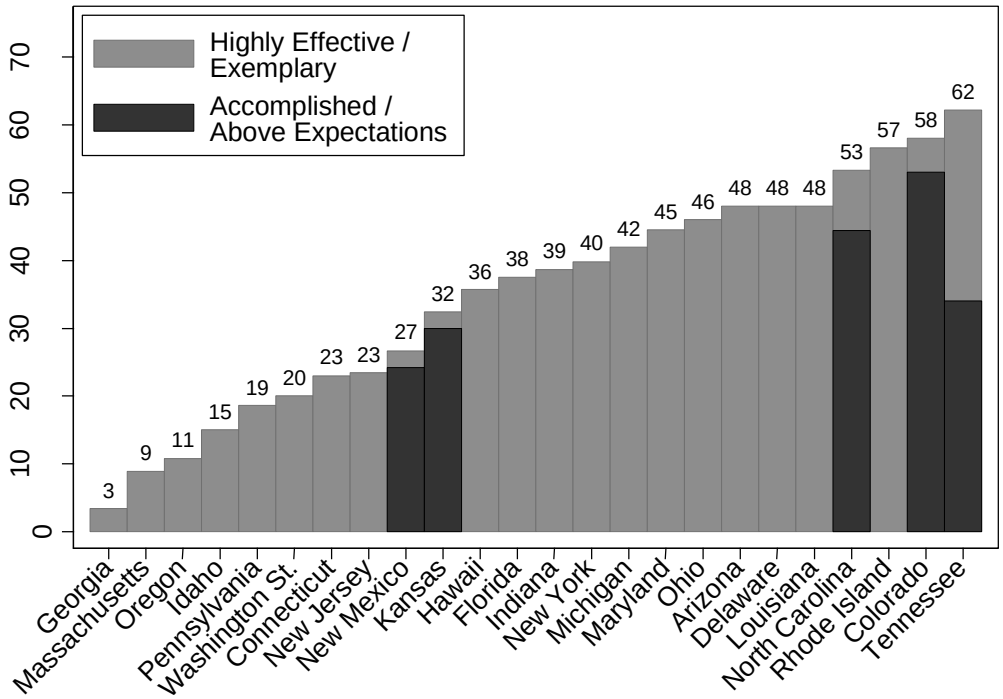

Figure 2: The percentage of teachers rated above Proficient across 19 state evaluation systems.

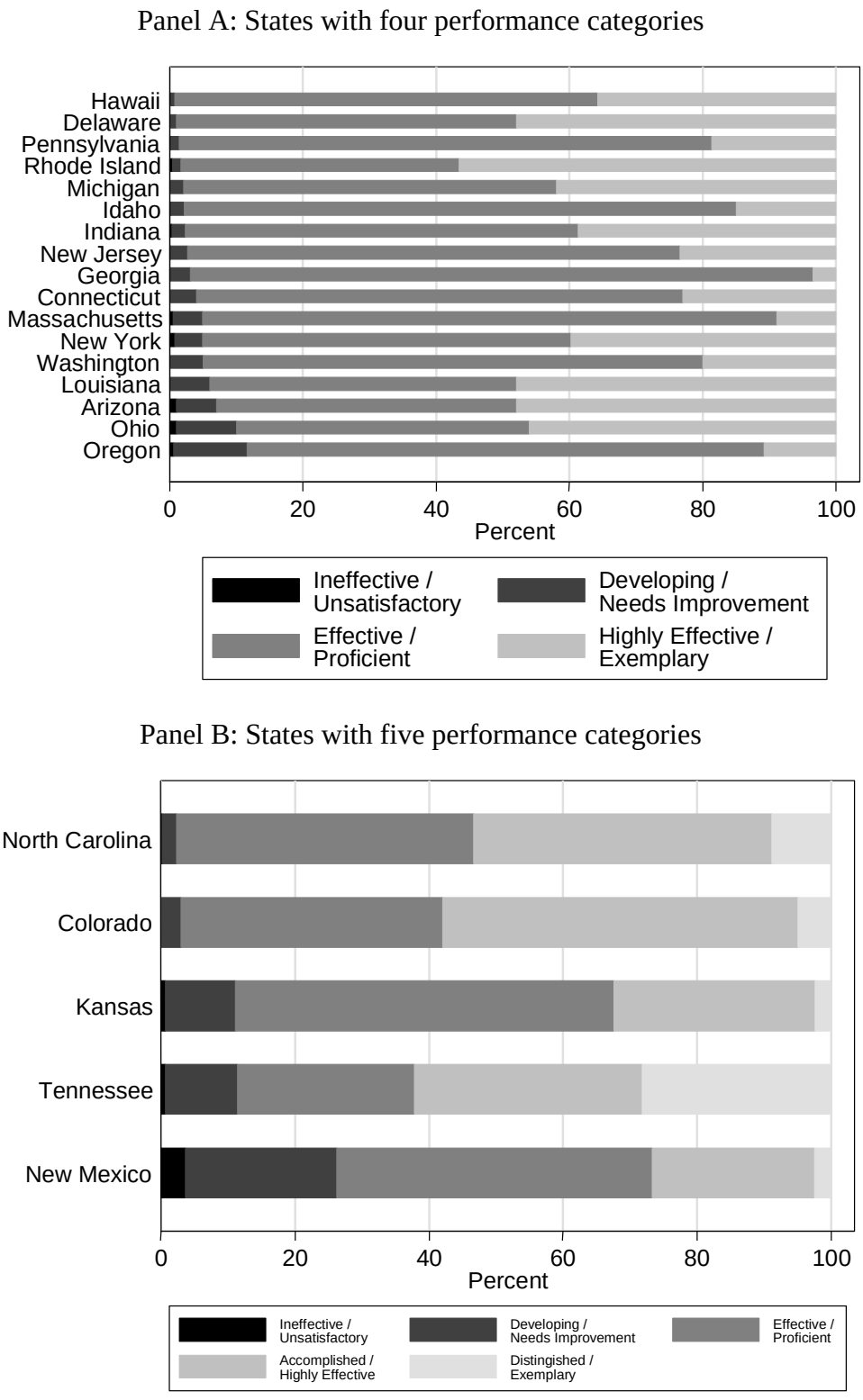

Figure 3: The distribution of teacher evaluation ratings across states with four (Panel A) and five (Panel B) rating categories.

Note: We exclude Florida from Panel B because its performance categories do not align with other states.

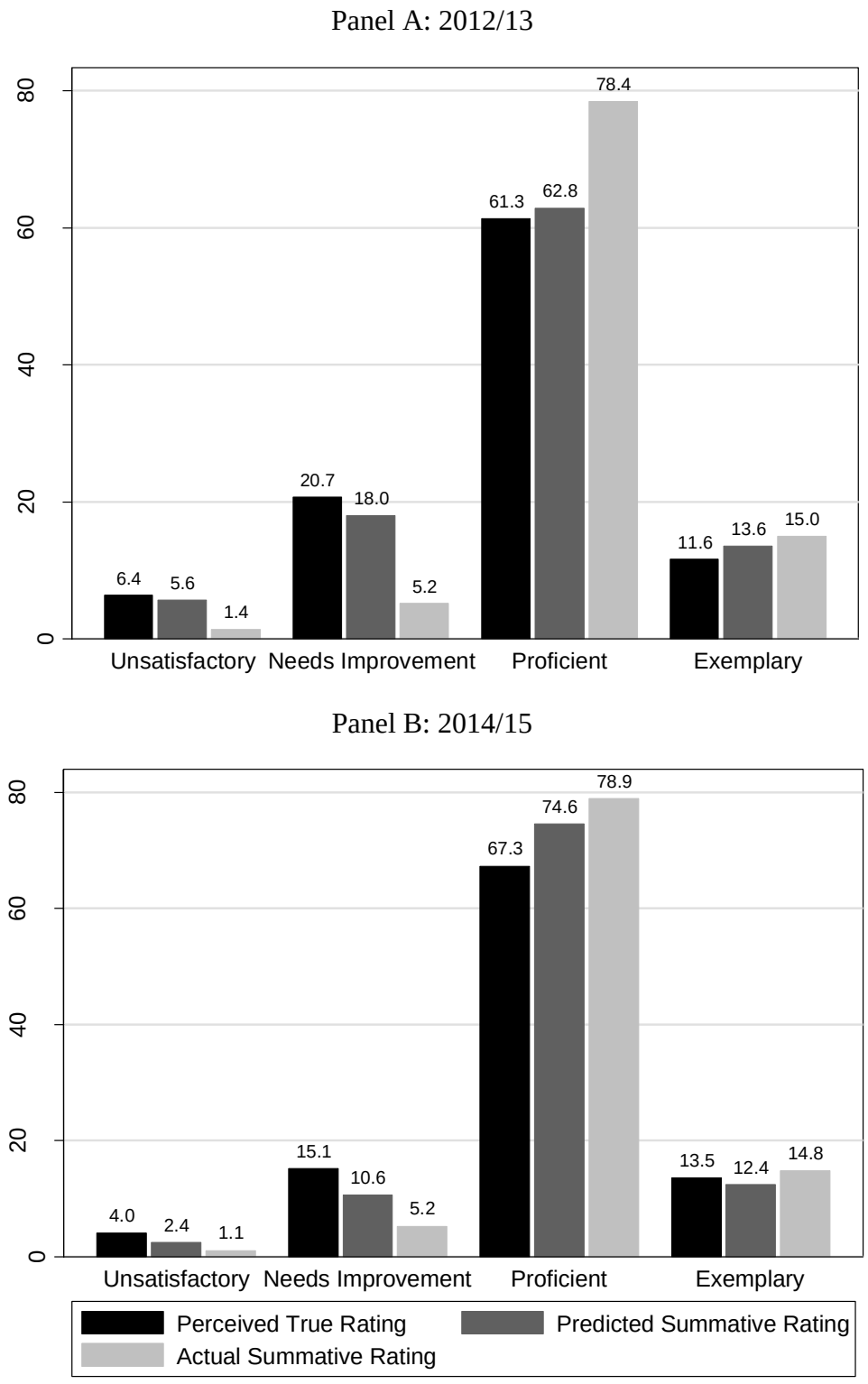

Figure 4: The perceived, predicted and actual distribution of teacher evaluation ratings in evaluators' schools in the first (Panel A) and third (Panel B) year of a new teacher evaluation system.

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.

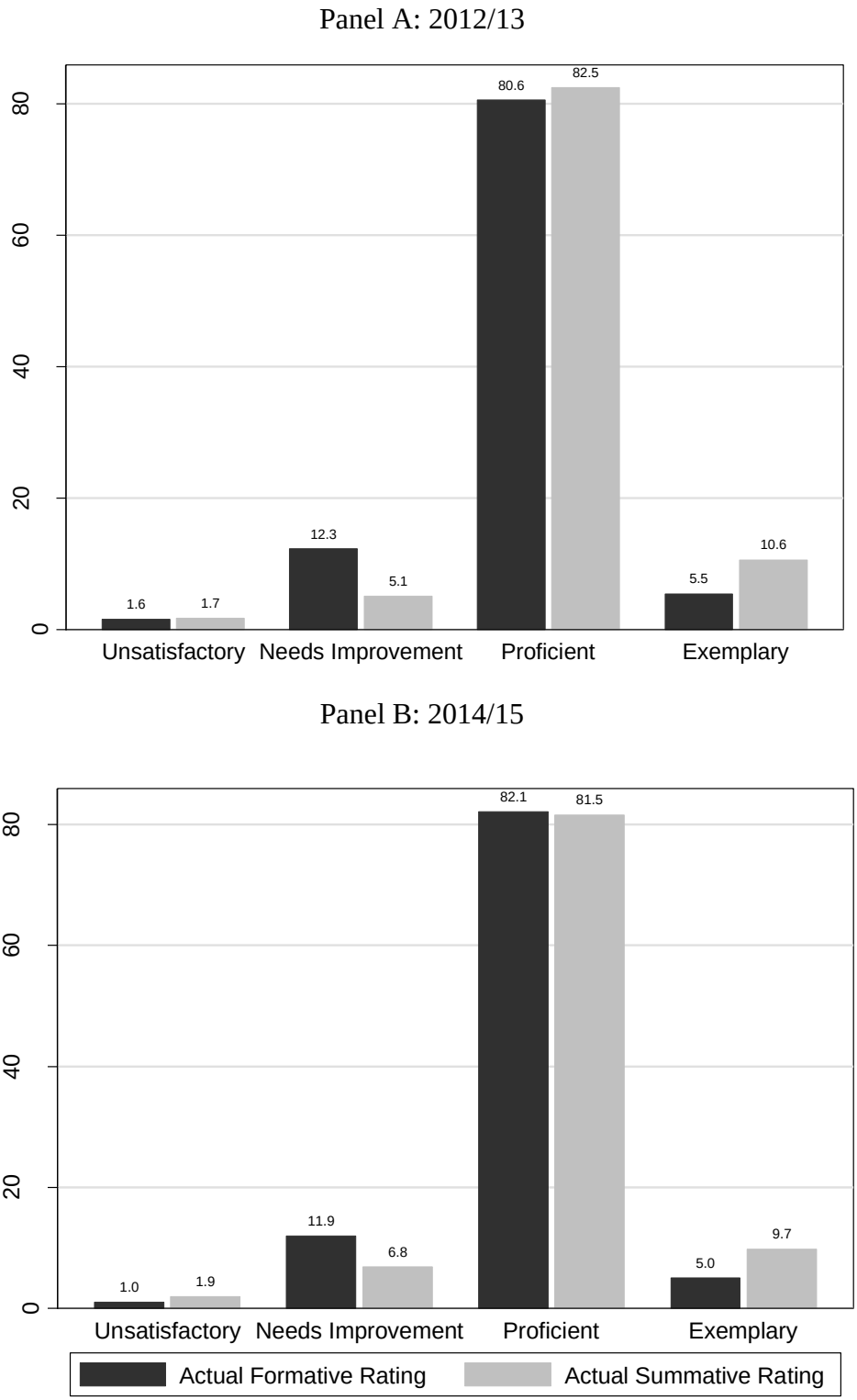

Figure 5: The actual distribution of formative and summative teacher evaluation ratings in evaluators' schools in the first (Panel A) and third (Panel B) year of a new teacher evaluation system among all teachers who received both rating.

Note: The distribution of summative ratings does not match in Figure 4 and Figure 5 because Figure 5 uses a restricted sample of teachers who have both formative and summative ratings. In 2012/13 79% of teachers received both formative and summative ratings. In 2014/15 only 58% of teachers received both ratings. See notes for Figure 4 for further details.