NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

CALDER

AIR

Performance Evaluations as a Measure of Teacher Effectiveness when Standards Differ: Accounting for Variation across Classrooms, Schools, and Districts

James Cowan

Dan Goldhaber

Roddy Theobald

WORKING PAPER No. 197-0618-2 • August 2020

Performance Evaluations as a Measure of Teacher Effectiveness when Standards Differ: Accounting for Variation across Classrooms, Schools, and Districts

James Cowan
American Institutes for Research

Dan Goldhaber
American Institutes for Research
University of Washington

Roddy Theobald
American Institutes for Research

Contents

Acknowledgments

This research was supported by a grant from the Massachusetts Department of Elementary and Secondary Education (ESE) and by the National Center for Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. We wish to thank ESE for providing the data we utilize, Bingjie Chen for excellent research assistance, and Claire Abbott, Matthew Steinberg, and attendees of the 2017 APPAM Fall Conference for comments that improved the manuscript. All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of ESE, the study's sponsors or the institutions to which the author(s) are affiliated. Any errors are attributable to the authors.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

Performance Evaluations as a Measure of Teacher Effectiveness when Standards Differ: Accounting for Variation across Classrooms, Schools, and Districts

James Cowan, Dan Goldhaber, Roddy Theobald

CALDER Working Paper No. 197-0618-2
August 2020

Abstract

We use statewide data from Massachusetts to investigate teacher performance evaluations as a measure of teaching effectiveness. Consistent with prior research, we find that assignment to lower achieving classrooms reduces teachers' performance ratings. But after adjusting for these and other observable differences between classroom assignments, we show that regression-adjusted performance measures can reliably predict future evaluation ratings as teachers move across grades and subjects within the same school. However, we also document substantial unexplained variation in ratings across schools and districts in the state. In particular, districts vary substantially both in the extent to which they differentiate between teachers and in the sensitivity of performance ratings to differences in teacher effectiveness as measured by value added. As a result, even after regression adjustment, teacher evaluation ratings generally provide unreliable predictions of future teacher evaluations after teachers switch schools. These findings suggest that policymakers and researchers should use caution in using performance evaluation ratings to make comparisons between teachers in different contexts.

1. Introduction

The passage of the Every Student Succeeds Act (ESSA) in 2015 represents a scaling back of federal involvement in teacher evaluations, particularly as the inclusion of student growth measures in the Obama Administration's waiver policies under No Child Left Behind (NCLB) essentially made their use a requirement for states. Since ESSA's adoption, at least 10 state legislatures have considered or implemented laws reducing the role of standardized achievement tests in teacher evaluations (Education Commission of the States, 2018). Consequently, observational and other qualitative measures of teacher performance may become relatively more important components of evaluation systems. Although this in part represents a return to policy before the advent of widespread standardized testing, the role of teacher evaluation in determining compensation, promotion, and tenure has changed significantly in the interim (Aldeman, 2017). Yet there is only a nascent literature about the properties, sensitivity, and validity of observational teacher evaluations in public schools.

The central difference between qualitative measures of teacher effectiveness and those derived from student outcomes is their reliance on human judgment. School administrators have substantial information about the proficiency of their teachers and they are likely to provide more reliable assessments than measures based solely on test scores (Ho & Kane, 2013). Principals also assess a wider range of teaching skills than those measured by standardized tests alone (Harris & Sass, 2014). On the other hand, subjective evaluations may be susceptible to various biases. Some studies of commonly used classroom observation tools suggest that teachers earn higher ratings when working in classrooms with higher-achieving students (Campbell & Ronfeldt, 2018; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). In addition, some analyses of hiring decisions or qualitative evaluations in other fields suggest that they may be sensitive to stereotypes based on race or gender (Bertrand &

Mullainathan, 2004; Grissom & Bartannen, 2020; Neumark, Blank, & Van Nort, 1996; Ouazad, 2018). These kinds of subjective biases could systematically affect certain teachers, although a recent random assignment experiment suggests these shortcomings may be overcome by adjusting for observable student characteristics (Bacher-Hicks, Chin, Kane, & Staiger, 2017).

Qualitative rating systems further differ from quantitative measures in the role they reserve for local leaders in their design and implementation. Unlike value-added measures, which apply a single statistical algorithm to standardized, statewide data, qualitative evaluation systems often rely on inputs that are developed or interpreted at the local level. This is partly by design, as it allows districts flexibility to adjust evaluation systems to local needs (McGuinn, 2012). However, many implementation choices might affect reliability or sensitivity to differences in teacher quality. For instance, there is considerable variation across districts in the number of observations conducted, the intensity of rater training, and the types of evidence collected (Chambers, Reyes, & O'Neil, 2013; U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2016). Districts may also have different standards for awarding performance ratings or weigh different teaching skills more heavily in their evaluations.

In addition to their use in evaluation systems, researchers and policymakers are increasingly using teacher performance evaluations as descriptive measures of teaching effectiveness. Examples include studies of the effects of in-service or pre-service training on teacher outcomes (Chen et al., 2019; Ronfeldt et al., 2018), the effectiveness of teachers with different credentials (Bastian, 2019; Cowan et al., 2017), and systems for ranking preparation programs (Bastian et al., 2018; Tennessee State Board of Education, 2019). Their usefulness for these purposes clearly depends on the extent to which they reflect true differences in teacher effectiveness rather than differing evaluation standards or classroom context.

In this study, we use statewide data from Massachusetts to investigate teacher performance evaluations as a measure of teaching effectiveness. We first conduct a descriptive analysis of variation in performance ratings across classrooms, schools, and districts. Consistent with prior research on observational ratings, we find that assignment to lower achieving classrooms reduces teachers' performance ratings. However, we also document substantial unexplained variation in ratings across schools and districts. Districts vary both in the extent to which they differentiate among teachers and in the sensitivity of evaluation ratings to differences in teacher effectiveness as measured by value added.

We then consider the implications of these patterns for using performance evaluations to describe teacher effectiveness. Using a variation of the teacher switching design proposed by Chetty et al. (2014), we find that performance measures derived from simple regression adjustment methods can reliably predict evaluations as teachers move across grades and subjects within the same school. However, even after regression adjustment, teacher evaluation ratings generally provide biased predictions of future teacher evaluations after teachers switch schools and districts. Our findings do not speak to *why* ratings standards differ substantially across schools and districts for similarly effective teachers; schools and districts having different conceptions of effective teaching or different preferences for identifying exceptional or struggling teachers are both consistent with the patterns we document. Nonetheless, put together, these findings suggest that policymakers and researchers should use caution in using performance ratings to make comparisons between teachers in different schools and districts, particularly if these comparisons are attached to high-stakes accountability or compensation decisions.

2. Background

2.1 The Massachusetts Educator Evaluation Framework

The teacher performance ratings we study in this paper are a central part of the teacher evaluation, feedback, and professional development processes in Massachusetts. The evaluations are aligned to the state's Standards for Effective Teaching (SET), which describe the expectations for effective teaching in the state. The four standards are: curriculum, planning, and assessment (*Standard 1*); teaching all students (*Standard 2*); family and community engagement (*Standard 3*); and professional culture (*Standard 4*). Together, the standards identify 33 specific elements of teaching practice (Massachusetts Department of Elementary and Secondary Education, 2015). Evaluation under the SET follows a five-step cycle with a timeline that depends on a teacher's career stage and prior evaluation results. The cycle begins with a self-assessment by the teacher and the development of a professional growth plan. During the implementation of the growth plan, teachers receive periodic feedback through a formative assessment process. Finally, the cycle concludes with a summative evaluation of teaching practice. Teachers receive an evaluation for each of the four standards and an overall summative performance rating. The summative evaluation occurs at least annually for beginning and low-performing teachers and at least biennially for teachers previously earning one of the top two ratings.[1]

[1] Teachers on a biennial review cycle receive a formative assessment in the alternating year. We include these formative ratings in the analyses in this paper, although the results are not sensitive to using summative performance ratings only.

Teacher performance on each of the standards is rated on a four-point rating scale: unsatisfactory, needs improvement, proficient, or exemplary. Administrators then use the collected evidence to assign a final rating. Massachusetts does not mandate specific procedures or formulas for aggregating the individual standard scores into a final summative rating. Instead, local evaluators award a final performance rating by reviewing the information (e.g., observational ratings, student surveys, and professional development activities) collected during the evaluation cycle and making a subjective determination about how to weight different components that feed into a teacher's summative evaluation. The state requires only that teachers earning a proficient rating must receive at least a rating of proficient on both Standards 1 and 2.[2]

2.2 Conceptual Model

Prior research has found that several forms of qualitative assessment – including principal evaluations (Harris & Sass, 2014; Jacob & Lefgren, 2008), classroom observations (Araujo et al., 2016; Blazar, 2015; Garrett & Steinberg, 2015; Gill et al., 2016; Grossman et al., 2013; Kane et al., 2013; Kane & Staiger, 2012), and student surveys (Kane & Staiger, 2011) – predict student test score gains. However, because qualitative evaluations rely on human judgment, they may be susceptible to different sources of error than value-added methods, which rely on standardized achievement measures and a consistent application of a statistical algorithm.

We sketch a simple model of teacher performance evaluations to illustrate potential sources of measurement error. Each year, raters evaluate some dimension of unmeasured teacher quality (θ_i).[3] Teachers are observed by observer j in a particular context with features (classroom

[2] Third year teachers (or teachers new to a district for three years) must be rated proficient on all four standards to receive tenure.

[3] This simplified conceptual model assumes a single dimension of teacher quality, but emerging evidence suggests that that teacher effects on non-cognitive outcomes are not highly correlated with their effects on student test scores (Gershenson, 2016; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2019), providing evidence of multiple dimensions of teacher quality.

composition, grade, subject) defined by x_{it}. We assume that observers differ both in the average ratings they provide (i.e., some are more lenient on all teachers) and in the extent to which their ratings differentiate candidates by their underlying effectiveness. We can write the observed evaluation as

$$E_{ijt} = \beta_j + \alpha_j\theta_i + x_{it}\gamma + \epsilon_{ijt}. \tag{1}$$

The $x_{it}\gamma$ describe the effects of assignment characteristics on the rater's perception of the teacher's effectiveness. Prior research has documented several factors that might influence performance ratings. For instance, several studies have found that assignment to lower achieving classrooms reduces observational evaluation scores (Campbell & Ronfeldt, 2018; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014). And Harris et al. (2014) find some evidence that principals award higher ratings to consider elementary school teachers than to teachers of older students. The β_j describe the leniency of raters. Ho and Kane (2013) find that principals rate their own teachers systematically higher than administrators from other schools. Principal evaluators may therefore have higher β_j in Eq. (1). Finally, α_j describes the sensitivity of the evaluators' ratings to differences in teacher quality. Larger α_j might indicate that the evaluator more accurately assesses teaching performance. It might also indicate a willingness to provide especially high or low ratings.

Other school- or district-level practices may affect the strength of the relationship between evaluation ratings and teaching effectiveness. Protocols for classroom observations—which are usually an important component of evaluation systems—are typically governed by provisions in districts' collective bargaining agreements with the teacher's union (Strunk et al., 2018). These provisions often specify the number of required observations, the instruments used to assess teacher quality, and the extent of evaluator training, all of which can significantly affect the reliability of observational ratings and increase the likelihood of misclassifying teachers (Ho & Kane, 2013; Kane & Staiger, 2012). Schools may also have unofficial policies on assigning poor ratings to teachers to avoid mandated sanctions or remediation for low-performing teachers (Kraft & Gilmour, 2017) or mitigate teacher concerns about punitive consequences of the evaluation process (Kraft et al., 2020).

The contributions of assignment characteristics and rater error to observed evaluations can limit their usefulness as measures of teacher effectiveness. Suppose a teacher works in two different environments in consecutive years with different raters. Then

$$E_{i,1,1} - E_{i,0,0} = [(\beta_1 - \beta_0) + (x_{i1} - x_{i0})\gamma] + (\alpha_1 - \alpha_0)\theta_i$$

The first term on the right-hand side combines the difference in rater leniency and assignment difficulty between teaching sites. This term affects all teachers equally. The methods proposed by Bacher-Hicks et al. (2014) may adjust for these influences so that comparisons of the average effectiveness of different groups of teachers will yield the correct sign. The second term, however, depends on the relative discrimination of the two raters and the teacher's underlying effectiveness. Moving to a site with less differentiation between high and low performing teachers (i.e., $\alpha_1 - \alpha_0 < 0$) would be expected to reduce performance ratings for an effective teacher and improve performance ratings for a less effective teacher. As a consequence, the magnitude of differences in performance evaluations between different groups of teachers may depend on the observer. This pattern may be especially important for studies that compare multiple groups of teachers who work in very different environments. For instance, ranking teacher preparation programs by performance ratings may be sensitive to the geographic sorting of candidates to different school systems (Bastian et al., 2018).

3. Data

We construct two samples for the analyses in this paper. Because the analytical methods used in this study rely on comparing multiple measures of teacher performance, we initially limit the sample to grades, subjects, and years in which we observe teacher evaluation scores and can also estimate teacher value-added. In addition, to simplify the analysis of the influence of classroom assignments on teacher evaluations, we restrict the sample to teachers working in self-contained classrooms in grades four and five. To ensure that we identify classrooms that correspond to actual courses, we limit the sample to students with a single teacher (or students who are assigned to co-taught courses) with at least 10 students and exclude English as a Second Language classrooms and supplemental and developmental classes. The sample with valid classroom matches and complete data includes 65% of fourth grade students and 42% of fifth grade students enrolled in public schools between 2014 and 2018.[4]

After identifying valid classrooms, we match teachers to the student data using common course codes. Using the linked student and teacher data from the 2014 to 2018 school years, we estimate teacher value-added on state assessments.[5] The student achievement data come from the standardized Massachusetts Comprehensive Assessment System (MCAS) and Partnership for the Assessment of Readiness for College and Careers (PARCC) end-of-grade tests. As our main test-based measure of teacher effectiveness, we use a composite value-added measure that combines math and ELA tests. Using the testing data stacked over subjects, we estimate value-added models that control for a cubic polynomial in prior achievement, student demographic and

[4] The lower rate in fifth grade is due to the greater prevalence of departmentalized instruction (i.e,. in which students receive math and ELA instruction from different teachers) in this grade.

[5] We estimate value-added models that control for cubic polynomials of lagged math and ELA achievement, student gender, race, subsidized lunch status, learning disability status, participation in English language learner programs, and the means of each of these variables at the school and classroom level.

program participation information, test type and test mode, and the school and classroom means of these variables. Each of these variables is interacted with the test subject. Because we use the teacher value-added data in regressions models to adjust ratings for differences in teacher effectiveness, we follow Chetty et al. (2014) and estimate jack-knife value-added measures that exclude data from a teacher's current students and shrinks estimates from other years according to their predictive power for the year in question (Stepner, 2013).[6]

In the second part of the paper, we consider common methods for using performance evaluations to measure teacher effectiveness. For this analysis, we use a broader set of teachers in core academic subjects (ELA, math, science, and social studies). We restrict the sample to teachers working in teaching assignments in a single school, who are matched to a classroom with between 10 and 75 students, and who receive a teaching evaluation during the school year.

We combine these data with records on teacher performance ratings between 2014 and 2018. The data include both formative and summative assessments on each of the four standards and an overall rating. We code the overall performance rating on a four-point scale. In some analyses, we aggregate ratings on each of the four standards. We also encode these ratings on a four-point scale and then follow Kraft et al. (2019) and aggregate the four standards into a single measure using a graded response model. The graded response model specifies the likelihood that teacher j with unobserved effectiveness θ_j, $\theta_j \sim N(0,1)$, receives a rating of at least k $(k = 1, 2, 3, 4)$ on standard i as

$$\Pr\left(E_{ij} \geq k \mid \theta_j\right) = \frac{\exp\{\alpha_i\left(\theta_j - b_{ik}\right)\}}{1 + \exp\{\alpha_i\left(\theta_j - b_{ik}\right)\}}$$

[6] To the extent that shocks to teacher value added and teacher performance ratings are correlated, controlling for contemporaneous value-added measures may absorb part of the effect of classroom assignments. We therefore rely on teacher value-added data from other years.

The model permits the standards to vary in their difficulty (b_{ik}) and in their relationship with true unobserved teacher quality (α_i). We use the estimated θ_j as an aggregated measure of teacher quality. In Appendix Table B.1, we estimate correlations between the performance ratings and value-added in the self-contained classroom sample. We find that the aggregated ratings have a slightly higher correlation with teacher value-added than the overall ratings.

We present summary statistics for the self-contained classroom sample in Table 1. The sample includes 6,471 teachers and 17,195 classrooms. The mean rating for the full sample is 3.1 on a 4 point scale (3 corresponds to proficient). The sample sizes for columns 2-4 demonstrate that 84.9% of the ratings are at the proficient level, 3.5% are below proficient (unsatisfactory or needs improvement), and 11.7% are exemplary. Formative evaluations, which are not consequential, account for 36.1% of the sample. The descriptive statistics do indicate that teachers with lower performance ratings have lower-achieving and less-advantaged students, although this may result from the assignment of less effective teachers to these classrooms. On average, teachers earning ratings below proficient were assigned to classes with predicted average achievement 0.25 standard deviations below the mean; teachers earning exemplary ratings had students expected to score 0.06 standard deviations above the mean.

The full core subject sample includes 57,038 unique teachers working in 756,974 classrooms. We present summary statistics for this sample in Table 2. The general patterns follow the sample of elementary teachers, with low income students, English language learners, and students with disabilities overrepresented among teachers receiving lower ratings. Also, consistent with Harris et al. (2014), we find that elementary teachers are overrepresented among those earning exemplary ratings.

In Figure 1, we show the distribution of performance ratings for the core subject sample across all districts performing at least 100 evaluations between 2014 and 2018. Each vertical stripe depicts the distribution across the four performance categories in a single district. Three trends are apparent from this figure. First, consistent with prior evidence of teacher evaluations (Kraft & Gilmour, 2017; Weisberg et al., 2009), most teachers in the state (85%) receive a proficient rating and very few (<1%) ever receive the lowest possible evaluation (unsatisfactory). Second, districts vary substantially in the extent to which they differentiate between teachers in their performance ratings. Districts on the left side of Figure 1 give practically every teacher a proficient rating, while districts on the right side give this rating to only about half of their teachers. About 10% of the teachers in this sample are in districts rating at least 95% of their teachers as proficient. Another 10% of teachers are in districts assigning this rating to fewer than 75% of their teachers. Finally, even among districts that substantially differentiate between their teachers, districts differ in whether they use the exemplary or needs improvement rating to distinguish teachers from the proficiency category.

4. Teaching Assignments and Performance Ratings

Researchers and policymakers have long understood the possibility that classroom assignments may affect observational or value-added measures of teacher effectiveness. Classroom observations and other subjective evaluations often include student work or assess classroom environment and other features of classrooms that may be jointly influenced by students and teachers (Campbell & Ronfeldt, 2018; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014). Disentangling the contributions of classroom characteristics and teacher quality is challenging. There is substantial evidence of positive matching between students and teachers, as students with higher achievement appear to be systematically assigned to more effective teachers (Goldhaber et al., 2018; Mansfield, 2015). Simple regressions of evaluations on student characteristics, which conflate both the patterns of teacher assignments and the effects of classroom characteristics on evaluations, are therefore unlikely to provide unbiased estimates of the causal effects of interest.

We estimate models that regress teacher performance ratings on other measures of teacher quality or teacher fixed effects. Our most basic approach relies on proxies for teacher quality to control for the non-random assignment of more effective teachers to high achieving classrooms. Specifically, we estimate regressions of ratings on classroom characteristics C_{jst}, teacher quality measures T_{jt}, and school fixed effects α_s:

$$E_{\mathrm{jst}} = C_{\mathrm{jst}}\delta + T_{\mathrm{jt}}\beta + \alpha_s + \epsilon_{\mathrm{jst}}. \tag{2}$$

We include teacher value added and experience in T_{jt}. The main limitation of these models is their use of a small set of characteristics of teachers to control for non-random assignment of teachers to classrooms. In particular, observable characteristics and teacher value added appear to have limited explanatory power for some of the teaching skills a performance evaluation system might consider (Gershenson, 2016; Harris & Sass, 2014; Jackson, 2018). If these unobserved teaching skills are also positively correlated with classroom characteristics, then estimates using proxies for teacher quality would overstate the effects of classroom assignments.

Our primary empirical strategy therefore replaces proxies for teacher quality with teacher fixed effects. We consider how individual teachers' evaluation results change when they teach in different types of classrooms. Following the approach in Whitehurst et al. (2014) and Steinberg & Garrett (2016), we estimate variants of the following teacher fixed effects model:

$$E_{\mathrm{jst}} = C_{\mathrm{jt}}\delta + \mathrm{Exp}_{\mathrm{jt}}\beta + \ \alpha_{js} + \lambda_t + \epsilon_{\mathrm{jst}}. \tag{3}$$

In Eq. (3), we replace the teacher quality proxies (except the experience indicators) with teacher-by-school fixed effects. Assuming that teacher skill does not vary systematically across classrooms, any variation in performance evaluations for a given teacher is likely a reflection of the classroom characteristics rather than of the teacher herself. In particular, this research design assumes that principals do not reward teachers who have had especially good years with better teaching assignments. There have been limited tests of this assumption in empirical investigations of teacher evaluation measures. However, evidence from other sources indicates that classroom assignments may be responsive to changes in teacher quality in ways that could bias our results (Kalogrides et al., 2013; Player, 2010). Our final strategy therefore relies on idiosyncratic variation in student characteristics across cohorts of students. We instrument classroom average prior achievement with the average prior achievement for each cohort (school-grade-year cell) and estimate Eq. (3) by 2SLS. We include teacher-by-school fixed effects, so that our only source of identifying variation is changes in characteristics across cohorts of students attending the same school. Similar research designs have been used to study the effects of class size (Hoxby, 2000) and assignment to high value-added teachers (Chetty et al., 2014).

The regression results, in Table 4, suggest that classroom average prior achievement is associated with performance ratings. In column 1, we estimate regressions of ratings on achievement levels without any controls for teacher quality. These models adjust only for grade, school year, and whether the evaluation is a formative assessment. The point estimate suggests that increasing average predicted student achievement by one standard deviation improves ratings by 0.09 points or about 23% of a standard deviation. However, this estimate conflates classroom composition effects with the assignment of more effective teachers to higher achieving classrooms. In the next three columns, we add teacher effectiveness measures and teacher fixed effects, respectively. Controlling for teacher value added and experience or teacher fixed effects reduces the point estimate on classroom prior achievement to about 0.06 after adjusting for teacher quality proxies and to 0.04 when we include teacher fixed effects. Notably, the IV estimates in column 4 are nearly identical to the teacher fixed estimates in column 3. All results are statistically significant, with the exception of the IV results, which are significant at the 10% level.

Finally, and in preparation for our exploration of variation in performance ratings across schools and districts discussed later, we account for variation in performance ratings across different schools and districts by including nested random effects at the district, school, and teacher level:

$$E_{\mathrm{jsdt}} = C_{\mathrm{jst}}\delta + T_{\mathrm{jt}}\beta_1 + \bar{T}_{\mathrm{st}}\beta_2 + \alpha_j + \alpha_s + \alpha_d + \epsilon_{\mathrm{jst}}.$$

We show results from this specification in column 5 of Table 4. Notably, the effects of class achievement are nearly identical to the teacher FE and IV specifications. Taken together, the results in Table 4 suggest that assignment to lower achieving classrooms reduces teachers' evaluation ratings. The difference between the 10th percentile classroom and 90th percentile classroom is about one standard deviations in predicted achievement, suggesting an increase in average evaluations of about 0.04 points.

To put these estimates into greater context, and to test for differences in rating standards across schools and districts, we estimate binary models where the dependent variables are earning a sub-proficient (unsatisfactory or needs improvement) or exemplary rating using similar specifications as in Table 4. We present estimates of the marginal effects of key variables from these analyses in Table 5. We consistently estimate small and statistically insignificant effects of predicted classroom achievement on the probability of an exceptional rating (Panel A). In linear probability models with school-by-teacher fixed effects (column 4), the marginal effect of a one standard deviation increase in predicted classroom achievement is 0.004 and statistically insignificant. Estimates using random effects logit models yield larger point estimates, but none is statistically significant.

We do find that assignment to lower achieving classrooms increases the likelihood of needs improvement or unsatisfactory ratings (Panel B). Using the random effects models, we estimate that a one standard deviation increase in predicted achievement reduces the likelihood of a low performance rating by about 1.5 percentage points; the estimate from the linear probability model with teacher fixed effects is about 3 percentage points. Thus, moving a teacher from near the bottom of the predicted achievement distribution to the top changes the likelihood of a needs improvement or unsatisfactory rating by about 1.5–3 percentage points, which is also approximately the proportion of teachers receiving one of these ratings.

In column 3, we estimate models with nested district, school, and teacher random effects, the teacher quality controls included in Table 4, and random coefficients on the teacher value-added measures at the district level. We find qualitatively similar effects of classroom achievement (positive but insignificant for the exemplary rating and negative and significant for unsatisfactory/needs improvement) as in the other regression models. Conditional on the teacher quality proxies, the school and district random effects are all statistically significant, suggesting variation in rating standards across sites. Because the variance components are on the logit scale, which does not have a natural interpretation, we focus on the relative magnitudes of the estimates. Two patterns are notable. First, the variance across schools is about evenly split between variation in the average ratings across school districts and variation in the average ratings across schools within a district. Second, the variation in performance across teachers in the same school is about three to four times greater than the variance in ratings across either schools in the same district or across school districts.

It is important to note that, unlike the previous analysis, estimation of the variance components in Eq. (4) is descriptive and does not necessarily provide estimates of the causal effects of schools on teacher performance ratings. In particular, we must assume that unobserved teacher quality (conditional on value-added and experience) is uncorrelated with school and district effects on performance ratings. Although there is far more variation in teacher quality within schools than across schools (Mansfield, 2015; Rivkin et al., 2005), this assumption is unlikely to hold strictly. Nonetheless, the estimated variance components provide some evidence about the potential importance of school or district factors on teacher evaluations.

Variance in the slope coefficient across school districts indicates that some districts differ in how much underlying performance matters for assigning final ratings. Those districts with coefficients on teacher quality closer to zero will tend to have more compressed ratings (i.e., more proficient ratings and fewer exemplary or needs improvement/unsatisfactory), while those with larger coefficients will tend to assign more teachers to the high or low ratings categories. Ideally, we would like to assess differences across districts in the sensitivity of their performance ratings to true teacher quality. Because this is not feasible, we instead allow the coefficient on teacher value-added, our main proxy for teacher quality, to differ across school districts.

Because the variance of the district intercepts and slope coefficients on value-added are estimated on the logit scale, they do not have a clear interpretation. We therefore plot the estimated probabilities from models in column 3 by teacher value-added for each of the largest 10 school districts in the analytic sample from Table 5. To ensure comparability across sites, we estimate probabilities at the means of the covariate distribution using only variation in the random portion of the model. Thus, differences in estimated probabilities are identified from variation in the district random effects and random coefficients on teacher value-added.[7]

We plot estimated probabilities by district in Figure 2. For fixed levels of value-added, the proportion of teachers receiving either high or low ratings differs significantly across school districts. The 75th and 90th percentile of the estimated value-added distributions are 0.15 and 0.23, respectively. The probability of a teacher with 75th (90th) percentile value-added earning an exemplary rating varies from 0.1% (0.1%) to 14.7% (20.1%) across the 10 districts in this sample. Similarly, the 10th and 25th percentile value-added estimates are -0.19 and -0.11. The probability of a teacher with these levels of value-added earning a needs improvement or unsatisfactory rating ranges from 0.2% (0.2%) to 7.2% (3.3%) across these districts.

The primary limitation of this analysis is that value-added measures only one aspect of teacher effectiveness and is not strongly correlated with other skills – especially those under the engagement and professional culture standards – included in the state evaluation framework (Gershenson, 2016; Jackson, 2018). If school districts differ in the importance they place on different teaching skills when determining final summative scores, we would expect to see differences in the relationship between value-added measures and performance ratings even for evaluation systems that produce identical ratings distributions. Although this analysis is only suggestive, we provide additional evidence in the next section that variation in the extent to which schools differentiate teachers of varying effectiveness contributes at least partially to these findings.

[7] Among the largest districts, we estimate that marginal effects on teacher value-added range from about 0.15 to 0.20. Similarly, the estimated marginal effects for receiving a low rating range from -0.03 to -0.14.

5. Describing Teacher Effectiveness using Performance Evaluations

We have found that performance evaluations differ across districts, schools, and classroom contexts. But they also contain useful information about teacher effectiveness. We therefore consider how these measures can be used to provide inferences about teacher quality. We follow prior work on observational ratings (Bacher-Hicks et al., 2017; Whitehurst et al., 2014) and construct regression-adjusted performance measures similar to value-added estimates from student achievement data. We then consider how well the adjusted performance measures predict teacher performance out-of-sample using a version of the teacher switching analysis proposed by Chetty et al. (2014).

In this section, we use all teachers in core academic subjects (ELA, math, science, and social studies) who can be linked to student classroom assignments between 2014 and 2018. Because we have found that the performance ratings aggregated across standards are more highly correlated with teacher value-added, we focus on those measures in this section. However, the results are similar using the overall performance ratings instead.

Using the classroom data, we estimate regression models intended to adjust the performance ratings for the teaching context. The regression analyses address three potential influences on teacher performance ratings: observable features of the teacher's classroom environment (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016), differences in performance standards across districts and schools, and differences in performance standards or teaching difficulty across subjects and grade levels (Harris et al., 2014). Our preferred model includes observed classroom covariates, subject-by-grade-level indicators, and school and year fixed effects:

$$E_{cjst} = X_{cjst}\beta + \gamma_t + \theta_s + \epsilon_{jcst}. \tag{6}$$

In Eq. (6), the indices are over classrooms, teachers, schools, and years, respectively. The classroom covariates and subject-by-grade indicators adjust the performance ratings for differences in the difficulty of teaching assignments, while the school effects are intended to capture differences in the standards employed by evaluators suggested by the analyses in Section 4.

We then form leave-out predictions of teacher performance ratings using residuals from Eq. (6) following the approach of Chetty et al. (2014) that accounts for drift in true teacher performance over time. For each year *t*, we average residuals over teaching assignments for each teacher and year. Using the average residualized performance ratings, we construct measures of teacher effectiveness that optimally weight each year of data using the approach described in Chetty et al. (2014).[8] Because we test forecast bias using annual changes in teacher staffing, we omit year *t* and either the prior year or the next year of teacher performance data to construct the estimates.

Once we construct the two-year leave-out estimates of teacher performance, we aggregate the data to the school-subject-grade-year level and regress changes in actual teacher performance ratings on changes in predicted teacher performance:

$$\Delta\alpha_{jst} = \Delta\,\widehat{\alpha}_{jst}\,\gamma + \lambda_t + \eta_{jst}. \tag{7}$$

Because the predictions rely on several years of data, the movement of teachers across assignments and schools explains the majority of the annual variation in predicted teacher performance. The conceptual model in Eq. (1) demonstrates why teacher switches are a useful

[8] We describe the procedure for constructing leave-out predictions of teacher performance in more detail in Appendix A.

test of forecast bias: they generate variation in true teacher quality, but plausibly do not affect either the difficulty of the teaching assignments or the rating standards of school principals.

We present estimates of the forecast bias in Table 6. In column 1, we show observational estimates of the forecast bias using classroom-level data. We regress observed performance ratings on the one-year leave-out predictions and the same covariates included in the performance ratings value-added models. The estimated forecast coefficient is potentially sensitive to the same sorts of biases as the estimates of teacher performance, but also more precisely estimated than those derived from the teacher switching design. The observational estimate of the forecast bias is 3%, which is statistically significantly different than zero. Column 2 shows the baseline switching model using data on teachers with non-missing performance predictions only. We estimate a forecast bias of 1.5%, which is not statistically different from zero. In the next column, we account for schoolwide shifts in rating standards that might be correlated with teacher quality by including a school-by-year fixed effect in the estimation of Eq. (7). The estimated bias is nearly unchanged.

In column 4, we isolate changes in teacher quality arising from teachers switching across schools. Although our adjusted ratings measures control for average differences in performance ratings across schools through the inclusion of school fixed effects, we found that schools and districts differ significantly in the sensitivity of their ratings to differences in teacher quality. We therefore instrument for $\Delta\, \widehat{\alpha}_{jst}$ in Eq. (7) with the average effectiveness of teachers who leave the school at the end of year *t-1*. Consistent with the evidence from the prior section, we find a forecast bias of 13.8% that is statistically different than zero. We explore this issue in more detail below.

Finally, in columns 5 and 6, we account for the fact that many teachers (especially novices) lack leave-out predictions because they do not have performance data outside the two-year window. We impute the mean rating for each of these teachers and re-estimate Eq. (7) using the full sample. The results with imputed teacher quality measures are similar to those using the complete cases sample, with estimates of forecast bias of about 0.1 – 0.4%.

In Table 6, we test alternative methods for adjusting teacher performance ratings for class and school factors. In the first column, we show estimates from the baseline teacher switching model; in the second column, we additionally include school-by-year effects when estimating the effects of changes in predicted teacher performance on observed evaluations. In the first row, we form predictions from the unadjusted teacher performance ratings and estimate a forecast bias of about 18%. Adding controls for student characteristics and assignment subject and grade (row 2) only modestly improves forecast accuracy. In rows 3 and 4, we introduce district and district-by-year fixed effects, similar to the models for observational ratings considered by Bacher-Hicks et al. (2019). These reduce forecast bias relative to simple covariate adjustment methods, but we still find a bias of about 7%, which is statistically significantly different from zero. In the last row, we find that models with school-by-year effects perform comparably to models with school effects.

In columns 3 and 4, we use the overall performance rating, rather than the aggregates across standards, to construct teacher effectiveness data. The forecast bias on the unadjusted ratings (row 1) is substantially higher than for the aggregate measures. For the adjusted measures, the forecast bias on the overall rating is much closer to that for the aggregates. Using the preferred adjustment method, which includes school fixed effects, we estimate a bias of about 7% using the overall rating, which is statistically significantly different than zero.

As we demonstrated in Section 4, districts differ substantially in their use of high and low performance ratings. The teacher switching design, which primarily leverages teachers switching between subjects and grades within the same school, may overstate the degree to which adjusted performance measures are comparable across school systems. Indeed, we find more evidence of forecast bias when we isolate annual variation in teacher quality resulting from teachers exiting a school. In Table 7, we further explore the forecast bias of adjusted summative ratings measures in other schools using a method proposed by Bacher-Hicks et al. (2014).

We construct several measures using different sources of data on teacher performance. We first estimate predicted performance ratings measures using only data from the same school. We then construct a measure of teacher effectiveness in *other* schools by removing the same-school portion of teacher performance ratings from the full-teacher measure:

$$\widehat{\alpha}_{jst,other} = \widehat{\alpha}_{jst,all} - \widehat{\alpha}_{jst,same}$$

In columns 1 and 2, we compare forecast bias using data from performance ratings given in the same school and performance ratings given in other schools. We estimate the coefficients in column 1 from a regression of teacher performance on the leave-out predictions and other covariates; in column 2, we include both the same-school and other-school predictions aggregated to the school-subject-grade-year level in the teacher switching design described in Eq. (7). We estimate a forecast bias of about 3% using same-school performance ratings and 31 – 34% using data from other schools. The point estimate suggests that about one third of the variation in teacher performance ratings is not stable across school environments. This is somewhat larger than the similar estimate for teacher value-added measures of 18.3% provided by Bacher-Hicks et al. (2014).

The lower coefficient on other-school performance ratings in column 1 does not necessarily indicate bias caused by different ratings standards documented in Section 2. Jackson (2013) shows that an individual teacher's true productivity varies from school to school, so we would not necessarily expect a forecast coefficient of 1 even if all schools employed the same standards for evaluating teachers. Principals may also differ in their assessments of individual teachers, and these kinds of rater error would also tend to depress the forecast bias coefficient in column 1.

In the remaining columns, we further explore the consequences of different performance standards across schools. The conceptual model and empirical results in Section 4 suggest that differences in the use of high and low performance ratings may complicate comparisons of teacher effectiveness across schools. The conceptual model in Section 2 suggests that the estimated forecast bias using data from schools with very different performance standards masks two potentially offsetting forms of bias. When teachers switch from schools with schools that tend not to differentiate teachers in their performance evaluation to one that does, we would expect a one unit increase in performance in the old school to translate into a larger than one unit increase in the new school. Similarly, when teachers switch to less discriminating schools, we expect their ratings to converge toward the mean of the ratings distribution.

We test this possibility by estimating the bias of performance ratings from low variance schools for predicting teacher performance in high variance schools (and vice versa). We split the sample at the median within-school standard deviation of performance ratings and estimate predictions of teacher performance using data in the same and similar schools and then construct the estimates

$$\widehat{\alpha}_{jst,other\ similar\ variance} = \widehat{\alpha}_{jst,similar\ variance} - \widehat{\alpha}_{jst,same}$$

$$\widehat{\alpha}_{jst,other\ dissimilar\ variance} = \widehat{\alpha}_{jst,all} - \widehat{\alpha}_{jst,similar\ variance}$$

where $\widehat{\alpha}_{jst,similar\ variance}$ is based on data from all schools in the same variance group. For ease of interpretation, we relabel these variables as $\widehat{\alpha}_{jst,high\ variance}$ (for predictions based on data from other high variance schools) and $\widehat{\alpha}_{jst,low\ variance}$ (for predictions based on data from other low variance schools).[9]

The results in columns 3 through 6 are generally consistent with the conceptual model. Predictions of teacher performance using data from schools with high variability in their ratings are significantly attenuated in low variance schools. We estimate a coefficient of 0.37 using observational data (column 3) and 0.28 in the teacher switching design (column 4); in other words, a one standard deviation difference in performance ratings in high variance schools predicts only a 0.28 – 0.37 standard deviation difference in teacher performance at low variance schools. On the other hand, predicted performance is more accurate when the teacher performance data comes from other low variance schools. We estimate forecast coefficients of about 0.87 using the observational data and 1.02 using the teacher switching design.

Among teachers in high variance schools, the prediction forecasts are similar for data from both high and low variance schools. For data from high variance schools, we estimate coefficients of 0.75 and 0.82, which are somewhat higher than the coefficients using data from all other schools (columns 1 and 2). We estimate a forecast coefficient of 0.72 – 0.73 when ratings data from low variance schools are used to form predictions of teacher performance in high variance schools. These coefficients are still less than 1, which is consistent with the existence of teacher-school or teacher-rater match effects, but significantly larger than the

[9] That is, we relabel the $\widehat{\alpha}_{jst,similar\ variance}$ prediction as $\widehat{\alpha}_{jst,low\ variance}$ for teachers in low variance schools and as $\widehat{\alpha}_{jst,high\ variance}$ for teachers in high variance schools. Because we estimate these regressions separately by variance group, this redefinition has no effect on the estimated coefficients.

corresponding estimates for predicting rating in low variance schools using prior ratings data from high variance schools.[10]

In Panel B, we repeat the analyses above using standardized performance ratings to account for differences in rating patterns across schools. We standardize the teacher estimates from the graded response model prior to estimating regression-adjusted teacher effects. Standardizing the teacher performance ratings reduces some of the disparities in predictive accuracy between high and low variance schools. The forecast coefficients for performance ratings from all other schools is somewhat lower than with the unstandardized data (0.56 – 0.62 compared to 0.66 – 0.69). But the forecast coefficients for predictions from high and low variance schools are more tightly clustered around the overall other-school forecast estimate. For teachers in low variance schools, we estimate forecast coefficients of about 0.6 using prior performance data from high variance schools and 0.72 – 0.88 using prior data from other low variance schools. The forecast coefficients for performance ratings in high variance schools using data from low variance schools (columns 5 and 6) are about 0.5. Although we estimate that about 40-50% of the variation in teacher performance is not stable across schools, the predictive accuracy of the standardized performance measures appears to be more similar across school types.

6. Discussion

We document significant variation in teacher performance ratings across classrooms, schools, and districts in Massachusetts. Ratings are sensitive to the prior performance of a

[10] We report similar analyses using data from other school districts in Appendix Table A.1. Given the relatively smaller number of district-to-district transitions, the estimates are significantly less precisely estimated than those using data from other schools.

teacher's students within a given classroom, and much of the variation across schools and districts remains after controlling for proxies for teacher effectiveness and school and district characteristics. Although we cannot rule out the possibility that these results are driven by the sorting of teachers to schools and districts along unobservable dimensions, it appears that school systems vary meaningfully in how they interpret standards and implement evaluation systems.

These patterns have several important implications, both for the statistical properties of teacher evaluation scores and for education policy. First, we find that simple regression adjustments appear to adequately control for differences in rating standards or assignment difficulty when teachers move between classrooms or subjects within the same schools. Our findings in this regard are similar to other research on classroom observational measures (Bacher-Hicks et al., 2019). However, we find that a significant portion of evaluated teaching effectiveness does not transfer between schools. This suggests caution in the use of teacher performance ratings to make high-stakes comparisons between teachers in different school settings. Part of this forecast bias is driven by the "widget effect" in schools and districts that do not meaningfully differentiate between teachers in their performance ratings (Kraft & Gilmour, 2017; Weisberg et al., 2009). This suggests that extracting more useful information from teacher evaluations may require changing the extent to which these evaluations actually differentiate between teachers.

These patterns also have important implications for the use of performance measures in research on teacher effectiveness. Researchers frequently standardize performance ratings measures, reflecting the fact that their scales are not necessarily meaningful. But our results suggest that schools and districts are not employing a single scale, so that a one standard deviation increase in performance ratings in two sites may not indicate an equivalent increase in the underlying teaching effectiveness. Other implications of the apparent lack of a uniform rating standard across sites likely depends on the application and research design. Because researchers typically standardize these measures, these patterns may not appreciably affect the sign of findings when researchers are comparing two groups of teachers, such as those with or without a master's degree. However, many applications involve rating multiple groups of teachers (e.g., those from different preparation programs or licensure pathways). In these cases, the ordering of individual groups is likely to depend on both the underlying effectiveness of the individual teachers and the extent to which they work in schools or districts that tend to differentiate teachers in their evaluation systems.

We conclude with two important caveats to these findings that suggest directions for future research. First, the descriptive variation in performance ratings across schools and districts is consistent with two potential explanations: schools and district may simply have different conceptions of effective teaching; and/or schools and districts may have different preferences for identifying exceptional or struggling teachers. Follow-up research can illuminate the reasons for these patterns that could inform future revisions to the design of teacher evaluation systems. And secondly, it is *not* necessarily the case that simply increasing differentiation of teacher evaluation ratings would improve the validity of these ratings. That said, if there are districts and states that put into place policies intended to improve the differentiation of performance ratings *under the same evaluation system*, follow-up research could directly test the extent to which this improves the predictive validity of the resulting evaluation ratings.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54–76.

Aldeman, C. (2017). The teacher evaluation revamp, in hindsight. *Education Next*, *2017*(Spring), 61–68.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415–1453.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, *73*, 101919.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in los angeles* (No. 20657). Cambridge, MA: National Bureau of Economic Research.

Bastian, K. C. (2019). A Degree Above? The Value-Added Estimates and Evaluation Ratings of Teachers with a Graduate Degree. *Education Finance and Policy*, *14*(4), 652–678.

Bastian, K. C., Patterson, K. M., & Pan, Y. (2018). Evaluating Teacher Preparation Programs With Teacher Evaluation Ratings: Implications for Program Accountability and Improvement. *Journal of Teacher Education*, *69*(5), 429–447.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, *94*(4), 991–1013.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, *55*(6), 1233–1267.

Chambers, J., Reyes, I. B. de los, & O'Neil, C. (2013). *How much are districts spending to implement teacher evaluation systems?* (No. WR-989-BMGF). Washington, DC: RAND Education; American Institutes for Research.

Chen, B., Cowan, J., Goldhaber, D., & Theobald, R. (2019). From the clinical experience to the classroom: Assessing the predictive validity of the Massachusetts Candidate Assessment of Performance. CALDER Working Paper 221-0819.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review*, *104*(9), 2593–2632.

Education Commission of the States. (2018). *Policy snapshot: Teacher evaluations*. Education Commission of the States.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (No. REL 2017–191). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation; Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the “Rothstein falsification test”: Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*(1), 8–34.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, *36*, 216–228.

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. Public schools. *American Educational Research Journal, 55*(1), 171-201.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver’s ProComp teacher pay initiative. *Economics of Education Review*, *31*(6), 1067–1083.

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*(3), 445–470.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher Value-Added measures. *American Educational Research Journal*, *51*(1), 73–112.

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

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill; Melinda Gates Foundation.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239–1285.

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *The Review of Economics and Statistics*, *95*(4), 1096–1116.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on Non–Test score outcomes. *The Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–135.

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103–123.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* Seattle, WA: Bill; Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2011). *Learning about teaching*.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. Seattle, WA: Bill; Melinda Gates Foundation.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, 54(1), 1-36.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234-249.

Liu, J., & Loeb, S. (2019). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, 1216-8430R3.

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751–788.

McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems*. Center for American Progress.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill; Melinda Gates Foundation.

Neumark, D., Blank, R. J., & Van Nort. (1996). Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics*, *113*(3), 915–941.

Ouazad, A. (2018). *Assessed by a teacher like me: Race, gender, and subjective evaluations*. SSRN.

Player, D. (2010). Nonmonetary compensation in the public school teacher labor market. *Education Finance and Policy*, *5*(1), 82–103.

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does Cooperating Teachers' Instructional Effectiveness Improve Preservice Teachers' Future Performance? *Educational Researcher*, *47*(7), 405–418.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.

Stepner, M. (2013). *Vam.ado [stata program]*. Cambridge, MA.

Tennessee State Board of Education. (2010). *2018 Educator Preparation Report Card*. Nashville, TN: Author.

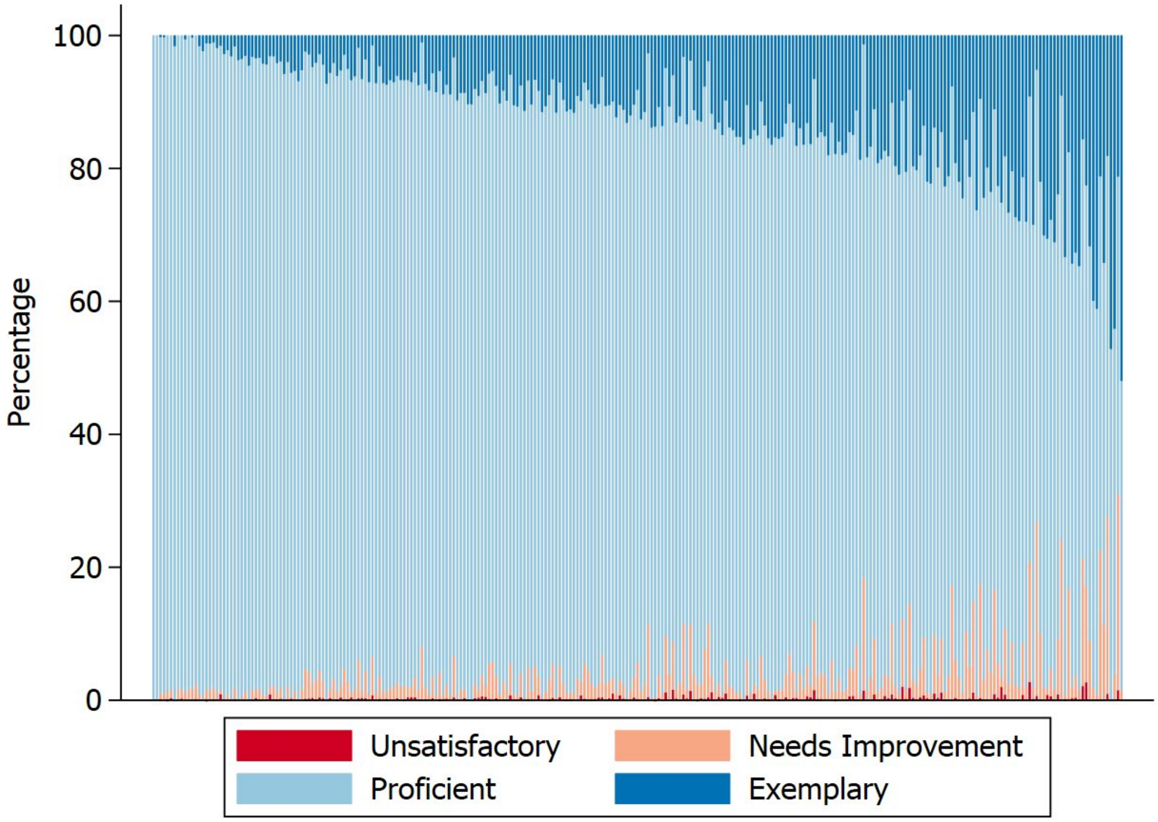
U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. (2016). *Study of emerging teacher evaluation systems*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation; Policy Development, Policy; Program Studies Service.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. TNTP.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Washington, D.C.: Brown Center on Education Policy, Brookings Institution.

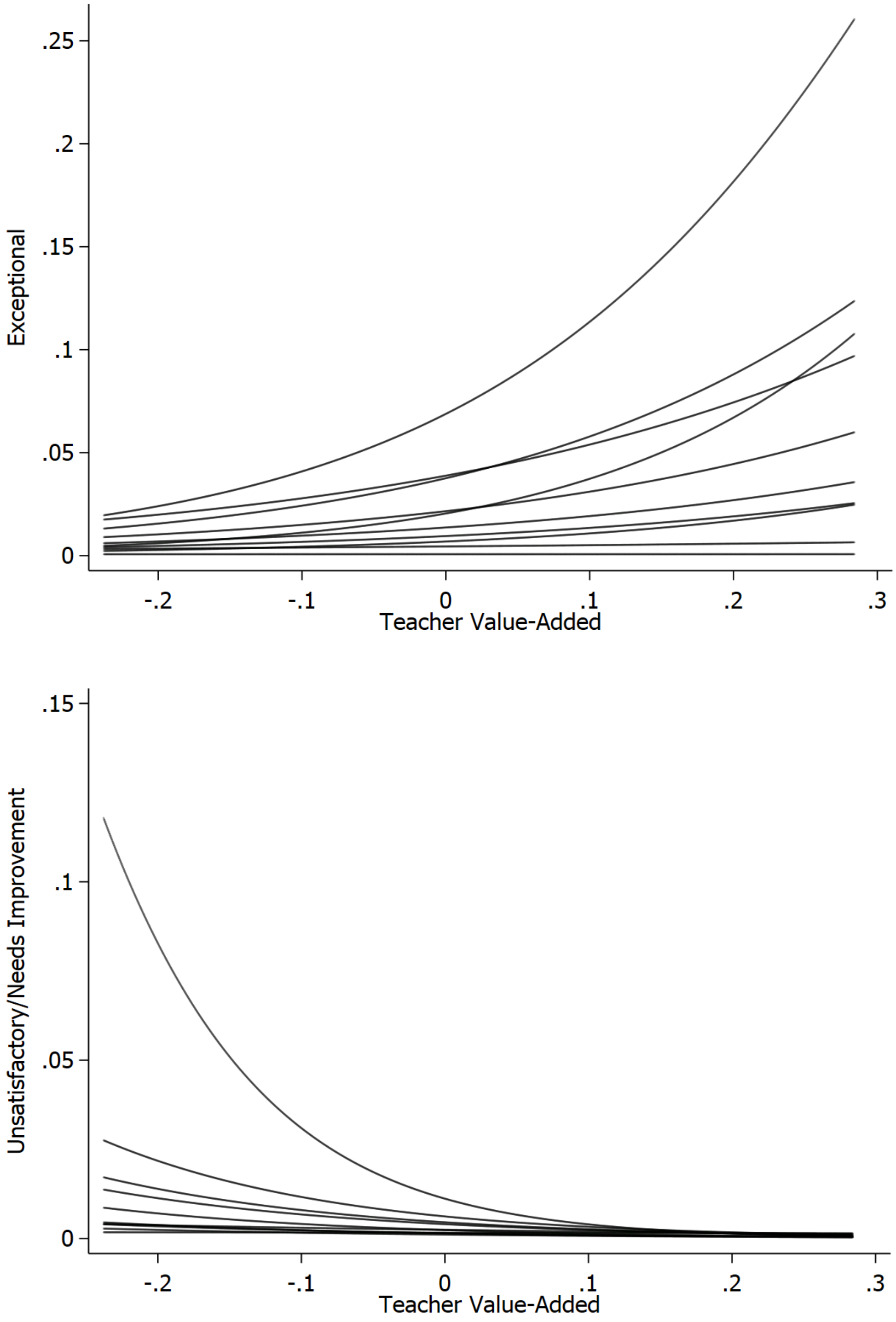
Figures and Tables

Figure 1. Distribution of Teacher Performance Ratings by District



Notes: Distribution of performance ratings (2014-2018) by school district. The sample includes all teachers in core subject classrooms in districts that conduct at least 100 evaluations during the five-year period. Each vertical ribbon represents one district in the state. Figure sorted from left to right by percent of teachers receiving a ranking other than proficient.

Figure 2. Estimated Probability of Ratings by Teacher Value-Added in Large Districts



Notes: Estimated probability of exceptional or unsatisfactory/needs improvement rating by estimated value-added in each of the 10 largest districts in Massachusetts. Estimated probabilities derived from model in column 3, table 5. All covariates, except value-added, are set to sample means; school and teacher random effects are fixed at zero.

Table 1. Summary Statistics (Self-contained Classroom Sample)

	(1) All Teachers	(2) Below Proficient	(3) Proficient	(4) Exemplary
Aggregated ratings	0.027	-2.006	-0.101	1.526
	(0.797)	(0.474)	(0.481)	(0.368)
Formative evaluation	0.361	0.136	0.365	0.394
	(0.480)	(0.343)	(0.481)	(0.489)
Experience	10.813	6.714	10.661	13.076
	(7.730)	(7.719)	(7.666)	(7.547)
Teacher value added	0.009	-0.079	0.005	0.053
	(0.159)	(0.146)	(0.159)	(0.157)
Predicted achievement	0.018	-0.249	0.022	0.063
	(0.401)	(0.443)	(0.398)	(0.376)
Prior math achievement	-0.015	-0.323	-0.011	0.041
	(0.483)	(0.526)	(0.481)	(0.454)
Prior ELA achievement	-0.021	-0.345	-0.016	0.040
	(0.495)	(0.552)	(0.492)	(0.468)
LEP students	0.070	0.130	0.068	0.069
	(0.146)	(0.205)	(0.143)	(0.140)
FRL-eligible students	0.356	0.588	0.350	0.333
	(0.325)	(0.357)	(0.321)	(0.318)
SPED students	0.167	0.158	0.167	0.172
	(0.139)	(0.162)	(0.138)	(0.133)
N	17,195	573	14,604	2,018

Notes: Summary statistics for teachers in the sample of self-contained 4th and 5th grade classrooms between 2014 and 2018. All observations are at the classroom (teacher-year) level.

Table 2. Summary Statistics (Teacher Switching Sample)

	(1) All Teachers	(2) Below Proficient	(3) Proficient	(4) Exemplary
Aggregated ratings	-0.022	-2.079	-0.113	1.506
	(0.826)	(0.508)	(0.488)	(0.367)
Formative evaluation	0.361	0.141	0.370	0.385
	(0.480)	(0.348)	(0.483)	(0.487)
Experience	10.672	6.656	10.607	12.776
	(7.570)	(7.517)	(7.496)	(7.431)
Grade K-5	0.426	0.346	0.425	0.463
	(0.495)	(0.476)	(0.494)	(0.499)
Grade 6-8	0.259	0.303	0.260	0.232
	(0.438)	(0.459)	(0.439)	(0.422)
Grade 9-12	0.315	0.351	0.315	0.304
	(0.465)	(0.477)	(0.464)	(0.460)
ELA	0.294	0.264	0.290	0.334
	(0.456)	(0.441)	(0.454)	(0.472)
Math	0.228	0.268	0.228	0.213
	(0.420)	(0.443)	(0.420)	(0.410)
Science	0.205	0.239	0.206	0.187
	(0.404)	(0.426)	(0.404)	(0.390)
Social Studies	0.198	0.175	0.199	0.198
	(0.398)	(0.380)	(0.399)	(0.399)
All subjects	0.075	0.055	0.077	0.068
	(0.263)	(0.227)	(0.266)	(0.251)
SPED students	0.148	0.161	0.148	0.147
	(0.175)	(0.203)	(0.175)	(0.166)
LEP students	0.096	0.148	0.093	0.095
	(0.199)	(0.246)	(0.196)	(0.199)
FRL-eligible students	0.381	0.573	0.373	0.366
	(0.322)	(0.331)	(0.319)	(0.320)
N	756,974	33,683	639,546	83,745

Notes: Summary statistics for teachers in the sample of core subject classrooms (ELA, math, science, social studies, and self-contained) between 2014 and 2018. All observations are at the classroom level.

Table 3. Classroom Composition Effects

	(1)	(2)	(3)	(4)	(5)
Class Predicted Achievement	0.089***	0.061***	0.042***	0.043*	0.045***
	(0.013)	(0.020)	(0.016)	(0.026)	(0.012)
Teacher Value-Added		0.387***			0.299***
		(0.042)			(0.036)
N	17,195	15,267	14,656	17,195	15,267
Controls		Y	Y	Y	Y
School FE/RE		FE			RE
Teacher FE/RE					RE
Teacher-School FE			Y	Y	
Cohort Achievement Instrument				Y	

Notes: Estimated effects of assignment to self-contained classrooms in grades 4 and 5 with varying predicted classroom achievement on teacher performance ratings. In each regression, the dependent variable is the teacher's overall performance rating (coded as integers 1-4) and the key independent variable is the average predicted achievement in the classroom using student demographic variables and prior achievement. All models include controls for whether the assessment is formative (rather than summative) and grade-by-year indicators. Teacher controls include leave-out value-added estimated from other years (columns 2 and 5) and indicators for teacher experience (columns 2-5). The regression in column 3 includes teacher-by-school fixed effects. The regression in column 4 includes teacher-by-school fixed effects and instruments for classroom predicted achievement with the mean predicted achievement at the school-grade-year level. The first stage coefficient on the school mean achievement variable is 0.95 and the associated t-statistic is 69.55. The regression in column 5 includes teacher random effects nested within school random effects and the set of covariates included in column 2. Standard errors clustered by school in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 4. Effects of Classroom Composition on Likelihood of High and Low Evaluations

	(1)	(2)	(3)	(4)
Panel A. Exceptional Rating				
Class Predicted Achievement	0.017	0.021	0.021	0.004
	(0.013)	(0.015)	(0.015)	(0.016)
Teacher Value-Added	0.150***	0.193***	0.197***	
	(0.022)	(0.028)	(0.028)	
Variance Components (Logit scale):				
Teacher	22.379	8.611	8.478	
School		2.478	2.476	
District		3.277	2.978	
Value-Added (District)			5.938	
N	15,267	15,267	15,267	14,656
Panel B. Unsatisfactory/Needs Improvement Rating				
Class Predicted Achievement	-0.016**	-0.015**	-0.014**	-0.033***
	(0.007)	(0.007)	(0.007)	(0.011)
Teacher Value-Added	-0.097***	-0.092***	-0.086***	
	(0.017)	(0.019)	(0.017)	
Variance Components (Logit scale):				
Teacher	5.710	4.044	3.842	
School		1.286	1.198	
District		0.964	0.785	
Value-Added (District)			11.814	
N	15,267	15,267	15,267	14,656

Notes: Marginal effects of classroom predicted achievement and teacher value-added on the likelihood of receiving high (exemplary) or low (unsatisfactory or needs improvement) ratings. The regression model in column 4 controls for school-by-teacher fixed effects, teacher experience, grade-by-year effects, and whether the evaluation is a formative assessment. The models in columns 2 and 3 control for teacher value-added, teacher experience, grade-by-year effects, whether the evaluation is a formative assessment, and the means of these variables at the school-year level. The models additionally include nested random effects at the teacher, school, and district levels and (in column 3) random coefficients on teacher value-added at the district level. The model in column 4 includes school-by-teacher fixed effects in place of the random effects and coefficients. Standard errors in parentheses (clustered by school in columns 1 and 3). * p < 0.10, ** p < 0.05, *** p < 0.01

Table 5. Estimated Forecast Bias of Performance Rating Predictions

	(1)	(2)	(3)	(4)	(5)	(6)
Adjusted rating	1.030***	0.985***	0.990***	1.138***	0.999***	1.004***
	(0.005)	(0.018)	(0.020)	(0.046)	(0.018)	(0.020)
Quasi-experimental		Y	Y	Y	Y	Y
School-year FE			Y			Y
Turnover instrument				Y		
Imputed ratings					Y	Y
p-value (forecast = 1)	0.000	0.402	0.611	0.003	0.965	0.830
N	719,238	87,800	87,718	87,800	91,516	91,442

Notes: Estimates of forecast bias from the teacher switching design. The dependent variable is the first difference in observed performance ratings estimated using the graded response model at the school-grade-subject level. Adjusted rating denotes the first difference in predicted performance ratings from the regression model in Eq. 6 using data outside each two-year window. The sample in column 1 includes all teachers with non-missing ratings and prediction data. The regression in column 2 adds school-by-year fixed effects. The regression in column 3 instruments for the change in adjusted ratings using the average predicted ratings of teachers departing in the prior year. The samples in columns 4 and 5 impute the sample mean predicted rating for teachers missing prediction data. Standard errors clustered by school in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 6. Estimates of Forecast Bias in Summative Ratings (Alternative Specifications)

	Aggregated Standard Ratings		Overall Rating	
	(1)	(2)	(3)	(4)
Unadjusted	0.821***	0.835***	0.346***	0.412***
	(0.016)	(0.017)	(0.013)	(0.013)
Covariate adjustment	0.877***	0.885***	0.856***	0.859***
	(0.017)	(0.018)	(0.019)	(0.022)
District fixed effects	0.926***	0.932***	0.890***	0.890***
	(0.017)	(0.019)	(0.020)	(0.022)
District-by-year fixed effects	0.929***	0.932***	0.893***	0.894***
	(0.018)	(0.019)	(0.020)	(0.022)
School fixed effects	0.985***	0.990***	0.931***	0.927***
	(0.018)	(0.020)	(0.021)	(0.023)
School-by-year fixed effects	0.989***	0.992***	0.940***	0.941***
	(0.019)	(0.020)	(0.022)	(0.023)

Notes: Estimates of forecast bias from the teacher switching design using alternative methods to adjust the performance ratings for the effects of school and classroom assignments. The unadjusted measure forms leave-out predictions using only the observed teacher evaluations. The covariate adjustment method controls for student gender, race/ethnicity, economic disadvantage, eligibility for subsidized lunches, special education status, learning disability status, limited English proficiency status, class size, and grade-by-subject and year fixed effects. The remaining models additionally add the indicated fixed effects. Models in columns 1 and 2 aggregate teacher performance evaluations across the four standards. Models in columns 3 and 4 use the final performance rating. Standard errors clustered by school in parentheses.
* p < 0.10, ** p < 0.05, *** p < 0.01

Table 7. Teacher Switching across Schools

	All		Low Variance Schools/Districts		High Variance Schools/Districts	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Adjusted Performance Ratings						
Adjusted rating (same school)	1.031***	0.971***	0.917***	0.873***	1.077***	1.014***
	(0.005)	(0.019)	(0.009)	(0.029)	(0.006)	(0.024)
Adjusted rating (other schools)	0.656***	0.690***				
	(0.035)	(0.079)				
Adjusted rating (high variance schools)			0.367***	0.277***	0.752***	0.819***
			(0.054)	(0.107)	(0.054)	(0.127)
Adjusted rating (low variance schools)			0.866***	1.015***	0.734***	0.715***
			(0.095)	(0.167)	(0.093)	(0.225)
N	694,490	84,752	356,354	42,787	338,136	41,965
Panel B. Adjusted Standardized Performance Ratings						
Adjusted rating (same school)	1.030***	1.006***	0.935***	0.917***	1.103***	1.074***
	(0.005)	(0.019)	(0.009)	(0.030)	(0.005)	(0.024)
Adjusted rating (other schools)	0.624***	0.556***				
	(0.035)	(0.101)				
Adjusted rating (high variance schools)			0.568***	0.613***	0.727***	0.418**
			(0.075)	(0.188)	(0.055)	(0.170)
Adjusted rating (low variance schools)			0.722***	0.877***	0.456***	0.468**
			(0.101)	(0.207)	(0.052)	(0.214)
N	680,296	81,622	343,679	40,045	336,617	41,577
Teacher Switching Quasi-Experiment		Y		Y		Y

Notes: Estimates of forecast bias using different sources of prior data on teacher performance. Odd-numbered columns display coefficients from regressions of performance ratings on predictions and classroom controls. Even-numbered columns display coefficients from teacher switching design. The adjusted rating (same school) is a leave-out prediction of teacher performance using other performance ratings given in the same school. The adjusted rating (other schools and high/low variance schools) use data on ratings in other schools as described in the text. The adjusted standardized ratings standardize the performance rating measures by school and year before estimation. Standard errors clustered by school in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix A. Construction of Leave-Out Estimates of Teacher Performance

We follow a two-step approach to constructing leave-out predictions of teacher effectiveness using the performance evaluation data. In the first step, we use classroom-level data to adjust teacher performance ratings for classroom factors:

$$E_{cjst} = X_{cjst}\beta + \gamma_t + \theta_s + \epsilon_{jcst}. \tag{A.1}$$

The control vector X_{cjst} includes an indicator for whether the evaluation is a formative assessment, classroom demographics (gender, race/ethnicity), the percentages of limited English proficient students, economically disadvantaged students, students qualifying for subsidized lunches, full-inclusion special education students, partial inclusion special education students, substantially separately educated special education students, learning disabled students, class size, and grade level-by-subject effects. We additionally include year (γ_t) and school (θ_s) fixed effects. We then form annual performance measures by averaging the residuals of (A.1) over a teacher's classroom assignments each year:

$$\widehat{\alpha}_{jt} = \frac{1}{N_{jt}} \sum\nolimits_c (E_{cjst} - X_{cjst}\,\hat{\beta} - \widehat{\gamma_t} - \widehat{\theta_s})$$

In the next step, we construct predictions that account for drift in teacher performance over time to optimally predict performance in leave-out years. Let $\widehat{\alpha}_{j,-t}$ be the vector of annual estimates of teacher performance. We construct the empirical Bayes predictions using the following weighting vector (Chetty et al., 2014):

$$\omega_{-\mathrm{t}} = \Sigma_{\alpha_{-t}}^{-1} \lambda_{-t}$$

where $\Sigma_{\alpha_{-t}}^{-1}$ is the covariance matrix for the annual performance measures contributing to the leave-out prediction and λ_{-t} is their covariance with performance in year *t.* Note that the covariance matrices (and, hence, the weights) depend on the number of years of available data.

We estimate the variance of the teacher effectiveness estimates as the sample variance of the estimates and the covariance between year *t* and year *t*+*s* as the sample covariance between performance in year *t* and performance in year *t*+*s*. Thus, for two elements of $\widehat{\alpha}_{j,-t}$, $\widehat{\alpha}_{js}$ and $\widehat{\alpha}_{js\prime}$, we estimate their covariance as the covariance between $\widehat{\alpha}_{jt}$ and $\widehat{\alpha}_{j,t+|s-s'|}$. Similarly, the corresponding entries in λ_{-t} are the covariance between $\widehat{\alpha}_{jt}$ and $\widehat{\alpha}_{j,t+|t-s|}$ and $\widehat{\alpha}_{jt}$ and $\widehat{\alpha}_{j,t+|t-s\prime|}$, respectively. This procedure relies on the stationarity assumptions discussed in Chetty et al. (2014).

Appendix B. Additional Results

We use teacher value-added as an additional measure of effectiveness for teachers in this sample. We estimate the main value-added measures used in the text from a regression model that includes prior achievement, student demographics, and classroom and school means of these variables:

$$A_{ijst} = X_{ijst}\beta + \alpha_j + \epsilon_{ijst}.$$

In the main text, we construct leave-out measures following the method of Chetty et al. (2014). To estimate correlations with performance ratings, we construct annual value-added estimates using residuals from this regression:

$$\widehat{\alpha_{jt}} = \frac{1}{N_{jt}} \sum_i (A_{ijst} - X_{ijst}\,\hat{\beta}).$$

Our preferred method for adjusting performance measures for the effects of teaching assignments includes school fixed effects. For comparability, we construct a similar teacher value-added measure that replaces the teacher fixed effect (α_j) with a school fixed effect (θ_s). We then construct annual value-added estimates in a similar fashion:

$$\widehat{\alpha_{jt}^{FE}} = \frac{1}{N_{jt}} \sum_i (A_{ijst} - X_{ijst}\,\widehat{\beta^{FE}} - \widehat{\theta_s}).$$

In Table B.1, we estimate the correlation between evaluation measures and teacher value-added. Because the correlation between two performance measures in a single year may be affected by the difficulty of a classroom teaching assignment, we additionally use data from multiple years to estimate the correlations in the underlying skills (Goldhaber et al., 2013). For two performance measures, E_{jt}^1 and E_{jt}^2, we estimate

$$\rho_{12} = \frac{cov(E_{jt}^1, E_{j,t+1}^2)}{(cov(E_{jt}^1, E_{j,t+1}^1)cov(E_{jt}^2, E_{j,t+1}^2))^{1/2}}.$$

As shown in Table B.1, performance evaluations are correlated with value-added measures. In columns 1 and 2, we show the relationship between the value-added measure (without school fixed effects) and the unadjusted performance ratings. We estimate a correlation between the overall rating and value-added estimates of about 0.10 – 0.14. When we adjust for measurement error in the underlying measures, these correlations increase to about 0.15 – 0.16.

In columns 3 and 4, we estimate correlations using the adjusted performance ratings (Eq. 6 in the text) and the school fixed effects value-added measures. Using the regression adjusted measures increases the correlations between value-added and teacher performance evaluations to 0.17 – 0.18. The relationship between teacher value-added and evaluations differs somewhat across the domains included in the review process, with the curriculum, planning, and assessment and teaching all students more strongly related to teacher contributions to student achievement. Aggregating the ratings on individual standards using the graded response model suggested by Kraft et al. (2019) increases their correlations with teacher value-added slightly.

Table B.1. Correlation between Performance Ratings and Teacher Value-Added

	Unadjusted Ratings		Adjusted Ratings	
	Unadjusted Corr.	Adjusted Corr.	Unadjusted Corr.	Adjusted Corr.
Panel A. Overall rating				
Math	0.133	0.154	0.115	0.165
ELA	0.126	0.156	0.102	0.167
Stacked	0.144	0.164	0.123	0.176
Panel B. Curriculum, planning, and assessment rating				
Math	0.140	0.170	0.121	0.191
ELA	0.147	0.196	0.120	0.224
Stacked	0.160	0.193	0.138	0.220
Panel C. Teaching all students rating				
Math	0.127	0.159	0.107	0.172
ELA	0.132	0.178	0.102	0.180
Stacked	0.144	0.179	0.120	0.189
Panel D. Family and community engagement rating				
Math	0.083	0.116	0.064	0.111
ELA	0.078	0.104	0.057	0.090
Stacked	0.089	0.118	0.069	0.113
Panel E. Professional culture rating				
Math	0.080	0.097	0.063	0.118
ELA	0.088	0.103	0.068	0.088
Stacked	0.092	0.105	0.073	0.111
Panel F. Overall rating (GRM)				
Math	0.148	0.176	0.125	0.194
ELA	0.153	0.191	0.123	0.196
Stacked	0.168	0.194	0.141	0.210

Notes: Estimated correlations between teacher performance ratings and value-added estimates. The unadjusted ratings use reported performance evaluation data. The adjusted ratings use teacher effects estimated from Eq. (6) in the text. The adjusted correlations account for sampling error and year-to-year fluctuations in teaching performance using the method described in the text. Math and ELA value-added are estimated using end-of-grade tests in each subject. Stacked value-added is estimated from a model using both math and ELA test scores and interacting all variables (except teacher effects) with subject.

Table B.2. Teacher Switching across Districts

	All		Low Variance Schools/Districts		High Variance Schools/Districts	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Adjusted Performance Ratings						
Adjusted rating (same district)	1.028***	0.983***	0.979***	0.918***	1.055***	1.018***
	(0.005)	(0.018)	(0.008)	(0.030)	(0.006)	(0.022)
Adjusted rating (other districts)	0.579***	0.820***				
	(0.044)	(0.115)				
Adjusted rating (high variance districts)			0.512***	0.321	0.648***	-0.522
			(0.076)	(0.221)	(0.069)	(0.322)
Adjusted rating (low variance districts)			0.547***	0.874***	0.549***	1.312***
			(0.101)	(0.330)	(0.108)	(0.315)
N	706,783	86,405	364,883	40,853	341,900	45,552
Panel B. Adjusted Standardized Performance Ratings						
Adjusted rating (same district)	1.027***	1.000***	0.977***	0.931***	1.076***	1.045***
	(0.005)	(0.019)	(0.008)	(0.031)	(0.006)	(0.024)
Adjusted rating (other districts)	0.534***	0.548***				
	(0.046)	(0.132)				
Adjusted rating (high variance districts)			0.829***	0.171	0.556***	-0.028
			(0.115)	(0.335)	(0.070)	(0.342)
Adjusted rating (low variance districts)			0.425***	0.703	0.351***	0.565*
			(0.097)	(0.465)	(0.070)	(0.313)
N	692,673	83,248	352,069	38,174	340,604	45,074
Teacher Switching Quasi-Experiment		Y		Y		Y

Notes: Estimates of forecast bias using different sources of prior data on teacher performance. The performance measures estimated by the graded response model have been standardized by school and year. Odd-numbered columns display coefficients from regressions of performance ratings on predictions and classroom controls. Even-numbered columns display coefficients from teacher switching design. The adjusted rating (same school/district) is a leave-out prediction of teacher performance using other performance ratings given in the same school. The adjusted rating (other schools/districts and high/low variance schools) use data on ratings in other schools/districts as described in the text. Standard errors clustered by school in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.