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

AIR

An Exploration of Sources of Variation in Teacher Evaluation Ratings across Classrooms, Schools, and Districts

James Cowan

Dan Goldhaber

Roddy Theobald

WORKING PAPER No. 197-0618-1 • June 2018

An Exploration of Sources of Variation in Teacher Evaluation Ratings 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

An Exploration of Sources of Variation in Teacher Evaluation Ratings across Classrooms, Schools, and Districts

James Cowan, Dan Goldhaber, Roddy Theobald

CALDER Working Paper No. 197-0618-1
June 2018

Abstract

We investigate sources of variation in teacher evaluation ratings across classrooms, schools, and districts. We show that assignment to high achieving classrooms increases teacher evaluation ratings. We also document significant variation in the sensitivity of performance ratings to value-added measures of teacher effectiveness across districts. Consequently, the probability that high or low performing teachers, as measured by value added, receive the highest or lowest evaluation ratings differs considerably across school districts. Our findings suggest that statewide policies that attach high stakes to performance evaluations are likely to have different consequences across schools and districts.

1. Introduction

The passage of 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 Administrations waiver policies under 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 evaluations 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.

An important difference between qualitative measures of teacher effectiveness and those derived from student outcomes is their reliance on human judgment. There are good reasons to believe that school administrators have substantial information about teachers' productivity and they may provide more reliable assessments than measures based solely on test scores (Ho & Kane, 2013). Principals are also likely capable of assessing a wider range of teaching skills than those measured by standardized tests (Harris & Sass, 2014). On the other hand, subjective evaluations may be susceptible to various biases. In fact, some studies of commonly-used classroom observation tools suggest that teachers earn higher ratings when working in classrooms with higher-achieving students (Steinberg & Garrett, 2016; Whitehurst et al., 2014), although a recent random assignment experiment suggests that observational measures that control for observable student characteristics provide unbiased predictions of teachers' observational scores in other classrooms (Bacher-Hicks et al., 2017). 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; Neumark et al., 1996; Ouazad, 2008). These kinds of subjective biases could systematically affect certain teachers.

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 consistent method 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 et al., 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 particular performance ratings or weight different teaching skills more heavily in their evaluations. Such differences in the application of performance standards may lead to large differences in the relative likelihood of receiving an extreme performance rating.

In this study, we investigate the empirical importance of the local implementation of teacher evaluations by schools and districts. We use statewide data from Massachusetts, which, like many states, provides considerable flexibility to districts to tailor the evaluation systems to local policy objectives. We make two main contributions to the literature on qualitative assessments of teacher quality. First, we show that prior findings on classroom composition effects for observational ratings extend to teachers' final summative ratings. We use a novel research design that relies on variation in student attributes across consecutive cohorts rather than solely on changes in teacher assignments to identify the effects of classroom composition. We then show that districts differ significantly in the sensitivity of their ratings to differences in teacher effectiveness as measured by value-added.

Consequently, the likelihood that teachers receive exceptional ratings differs substantially by school and district. The findings suggest that, in high-stakes accountability or compensation, the likelihood of award or sanction may differ meaningfully across districts for similarly effective teachers.

2. Background and Prior Literature

Subjective evaluations are an important component of teacher evaluation systems, yet there is far less evidence on their performance than on quantitative methods such as value-added modeling.[1] 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 & Staiger, 2012; Kane et al., 2011, 2013), 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.

Although disentangling effects of classroom assignments and patterns of teacher assignments is an empirical challenge, several recent studies have used experimental and quasi-experimental research designs and found that observational measures are sensitive to the characteristics of the classrooms to which teachers are assigned. For instance, Whitehurst et al. (2014) found that incoming student achievement was associated with observational ratings, and that this pattern held even in teacher fixed effects models that compare observational ratings for the same teacher in different years. Steinberg and Garrett (2016) analyzed data from the Measures of Effective Teaching (MET) Project, which randomly assigned teachers to classrooms in a number of school districts. They used a combination of random

[1] For instance, see Bacher-Hicks et al. (2014, 2017), Chetty et al. (2014a, 2017), Goldhaber and Chaplin (2015), and Rothstein (2010, 2017).

assignment and teacher fixed effects designs and also found that an assignment to a classroom with lower incoming ELA achievement reduced teachers' observational ratings. In their re-analyses of the MET data, Gill et al. (2016) and Campbell and Ronfeldt (n.d.) additionally found that teachers in classrooms with higher proportions of minority students earned lower observational ratings. Taken together, these studies suggest that evaluation ratings may systematically differ for teachers working in more disadvantaged environments.

Apart from biases arising from classroom composition effects, districts may also differ in their understanding of professional standards or the rigor with which they apply them. For instance, Kraft et al. (2018) show that high-stakes evaluation reforms reduce the supply of new teachers. Under these systems, hard-to-staff school districts may be reluctant to provide low ratings if they affect tenure status or have other high stakes consequences that drive away teachers (Pogodzkinki et al., 2016). Low ratings also typically trigger professional development requirements. If districts differ in their capacity to provide this training, principals in some locations may intentionally avoid providing low ratings (Kraft & Gilmour, 2016).

District practices can also affect the strength of the relationship between teacher evaluation ratings and other measures of teaching effectiveness. For instance, districts often have considerable control over the protocols for classroom observations, which are usually an important component of evaluation systems. These include the number of 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).

Districts' differing conceptualizations of effective teaching may also influence the relationship between evaluations ratings and direct measures of teacher effectiveness. Teachers' value-added to standardized

achievement tests, for instance, tends to be weakly correlated with teacher effects on students' non-test outcomes or with professional contributions (Harris & Sass, 2014; Jackson, 2018). In this study, we use teacher value-added as a quantitative measure of effectiveness and varying relationships between value-added and summative ratings could reflect district choices about which teaching skills to emphasize. That is, we might observe a weaker relationship between quantitative measures of teaching effectiveness and evaluations in some districts because these particular skills are less valued. Nonetheless, reweighting components of evaluation systems to privilege particular skills can significantly impact the reliability of composite scores and their relationship to other measures of teacher quality (Mihaly et al., 2013; Steinberg & Kraft, 2017). Individual teachers may therefore receive significantly different evaluations in different school districts depending on their fit.

Variation in district policy may therefore have important implications even in the absence of significant bias in the instruments. This is because teacher incentive and accountability provisions often target the extremes of the teacher effectiveness distribution. For instance, high-profile teacher compensation reforms in Washington, DC and Denver provide bonuses to teachers who receive high performance evaluations or terminate employment for low-performing teachers (Adnot et al., 2016; Goldhaber & Walch, 2012). District variation in the reliability of evaluation instruments or in the weighting of various teaching skills may have large effects on the classification of individual teachers even if mean observational scores are similar.

3. Data and Setting

3.1 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). The four standards are: curriculum, planning, and assessment; teaching all students; family and community engagement; and professional culture. 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.

Teacher performance on each of the standards is rated on a four-point rating scale: unsatisfactory, needs improvement, proficient, or exemplary. The state requires that teachers earning a proficient rating must receive at least a rating of proficient on both the curriculum, planning, and assessment and teaching all student standards.[2] Beyond this requirement, the evaluation framework preserves an important role for local evaluators in determining how a teachers' performance informs

[2] 3rd year teachers (or tecahers new to a district for three years) must be rated proficient on all four standards to receive tenure.

the final rating. Specifically, the state does not prescribe a method for combining data from the various sources included in the evaluation framework into a single 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 subjective determination about how to weight different components that feed into a teacher's summative evaluation.

3.2 Data

The analytical methods used in this study rely heavily on comparing multiple measures of teacher performance, so we limit the sample to grades, subjects, and years in which we observe teacher evaluation scores and can also estimate teacher value added. In particular, we restrict the sample to teachers working in self-contained classrooms in grades four and five during the 2013-14 through 2015-16 school years. Furthermore, given the interest in the association between classroom characteristics and teacher evaluations, we limit the study to teachers who can be linked to a single classroom. The Massachusetts administrative data have matched students and teachers through common course codes since the 2010-11 school year. We first identify self-contained courses beginning in 2011 using the student datasets. To ensure that we identify classrooms that correspond to actual courses, we limit the sample to students with a single teacher in each subject (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.

After identifying valid classrooms, we match teachers to the student data using the common course codes. Using the linked student and teacher data from the 2011 to 2016 school years, we estimate teacher value added on state assessments.[3] The student achievement data come from the

[3] We estimate value-added models that control for cubic polynomials of lagged math and ELA achievement, student

standardized Massachusetts Comprehensive Assessment System (MCAS) and Partnership for the Assessment of Readiness for College and Careers (PARCC) end-of-grade tests. We use these test scores to estimate value-added measures that are pooled over these years and use empirical Bayes methods to shrink the estimates toward the mean of the teacher effectiveness distribution inversely with their precision, which ameliorates the attenuation bias that results when they are used as a right-hand-side variable.[4] The value added-models control for a cubic polynomial in prior achievement, student demographic and program participation information, and the school and classroom means of these variables. Finally, we aggregate student demographic and baseline achievement data to the teacher-year level, resulting in data on the characteristics of a teachers' classrooms and annual value-added estimates.

We combine these data with annual data on teacher evaluations. We limit the sample to full-time teachers who received a teacher evaluation between 2014 and 2016. Some of our research designs rely on within-teacher variation in classroom assignments and requires controlling for improvements in teacher practice that come with teaching experience. The Massachusetts data do not measure teacher experience directly, so we instead measure experience as the number of years a teacher has held a valid teaching license in Massachusetts. The final dataset includes 5,849 teachers and 11,563 classrooms. Of these teachers, 2,129 (36%) have a single evaluation, 1,726 (30%) have two years of evaluations, and 1,994 (34%) have evaluations in each year.

---

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.

[4] We also estimate jack-knife value added measures using the method of Chetty et al. (2014) (Stepner, 2013). The estimation approach of Chetty et al. (2014) excludes data from a teacher's current students and shrinks estimates from other years according to their predictive power for the year in question. 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. Nonetheless, the results from analyses using these measures are substantively similar to those using the pooled estimates. Results are available from the authors upon request.

We present summary statistics for the sample in Table 1. 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 85.7% of the ratings are at the proficient level, 3.8% are below proficient (unsatisfactory or needs improvement), and 10.5% are exemplary. Formative evaluations, which are not consequential, account for 35.4% of the sample. While this is not shown in Table 1, teachers are more likely to earn below proficient ratings on the summative evaluations: only 14.7% of these ratings are given on formative evaluations. The performance ratings correspond, at a high level, to the value added measures of teacher effectiveness: the correlation between ratings and math and ELA value-added is 0.17 and 0.16, respectively. While the mean math value-added estimate across the entire sample is 0.01, teachers earning ratings below proficient have an average value-added estimate of -0.10; those earning exemplary ratings have an average estimate of 0.07. On average, teachers earning exemplary ratings also have 1.4 more years of experience than those earning below-proficient ratings.

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 prior average achievement 0.23 standard deviations below the mean on the prior year's test; teachers earning exemplary ratings had students who scored 0.11 standard deviations above the mean.

4. Variation in Ratings across Classrooms, Schools, and Districts

4.1 Classroom Characteristics and Subjective Evaluations

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, classroom environment, or other features of classrooms that may be jointly produced by students and teachers. Assessors' ratings of teaching practice may therefore conflate the quality of instruction with the quality of student work and teachers assigned to high-achieving classrooms may benefit from the academic aptitude of their students (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: students with higher achievement appear to be systematically assigned to more effective teachers (Goldhaber et al., 2016; 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.

In this study, we adopt three general approaches for estimating the effects of classroom characteristics on teacher evaluations. The first of these 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 summative ratings on classroom characteristics C_{jt} as well as teacher quality measures T_{jt}:

$$E_{jt} = C_{jt}\delta + T_{jt}\beta + \epsilon_{jt} \quad (1)$$

We include value added and experience in T_{jt}. We additionally include averages of the proxies at the school and district level to adjust for sorting across these dimensions (Altonji & Mansfield, 2014). 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, any observed classroom composition effects could be driven by unobserved differences in teacher quality. 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; Jackson, 2016; Sass et al., 2014). 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 second empirical strategy therefore replaces proxies for teacher quality with teacher fixed effects. We thus 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 and Garrett (2016), we estimate variants of the following teacher fixed effects model:

$$E_{jt} = C_{jt}\delta + Exp_{jt}\beta + \alpha_j + \lambda_t + \epsilon_{jt}. \quad (2)$$

In Eq. (2), we replace the teacher quality proxies (with the exception of the experience indicators) with teacher fixed effects. The strategy is to control for underlying teacher quality through the inclusion of a teacher fixed effect. The key assumption is that teacher practices are unlikely to vary across classrooms so any variation within teacher in performance evaluations is likely a reflection of the classroom characteristics rather than 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 (Kalogrides et al., 2013; Player, 2010). Thus, by relying on all within-teacher variation in student characteristics, teacher fixed effects models may overstate their relationship with summative ratings. This is because the within-teacher variation includes changes in classroom composition resulting from intentional assignments, which may be related to changes in teacher behavior, and natural year-to-year variation in the composition of student cohorts.

Our final strategy therefore relies on idiosyncratic variation in student characteristics across cohorts of students. We instrument average classroom prior achievement with the average prior achievement for each cohort (school-grade-year cell) and estimate Eq. (2) by 2SLS. We estimate versions with teacher fixed effects, which use both cohort-level variation in achievement as well as grade switches, and teacher-by-grade fixed effects, which use only the former source of variation. The identifying assumption in this case is that trends in an individual teacher's performance, as measured by the summative ratings, are uncorrelated with changes in the composition of cohorts across the three years in our sample. 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., 2014a).

The regression results, in Table 2, suggest that classroom average prior achievement is associated with performance ratings. In column 1, we estimate bivariate regressions of ratings on achievement levels. The point estimate suggests that increasing average prior student achievement by one standard deviation improves ratings by 0.07 points or about 18 percent 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 two 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 0.04 to 0.05.

In columns 4 and 5, we instrument classroom prior achievement with average cohort prior achievement as described above. The point estimates are quite similar to the standard teacher fixed effects estimates. Notably, the results are quite similar to those from the regressions that only include the teacher effectiveness proxies in column 2. Overall, the results suggest that a one-unit increase in the classroom achievement measure would raise ratings by about 0.04 points on the four-point rating scale. The difference between the 10th percentile classroom and 90th percentile classroom is about 1.3 standard deviations in lagged achievement, suggesting an increase in average evaluations of about 0.06 points on the four-point scale. Put another way, when we estimate binary outcome models where the dependent variables are earning a needs improvement/unsatisfactory or exemplary rating, the average marginal effects of a one standard deviation increase in average lagged classroom achievement are -0.022 (*se*=0.004) and 0.012 (*se*=0.009), respectively, although only the former is statistically significant.

In Table 3, we present a number of alternative results. First, in Panel A, we estimate models that consider lagged achievement on ELA tests. The results are substantively similar: teachers in higher achieving classrooms earn higher evaluations. There is somewhat more variation in results across specifications. Using proxies for teacher quality, we estimate that an increase in lagged ELA scores by one standard deviation would improve evaluations by about 0.04 points. The models with teacher fixed effects suggest a one standard deviation increase in lagged achievement would increase ratings by about 0.05 to 0.07 points. In Panels B and C, we exclude formative evaluations, which are not as consequential under the Massachusetts educator evaluation framework. The association between classroom achievement and ratings appears stronger on the consequential summative evaluations. In math, we estimate that an increase in lagged achievement by one standard deviation would increase summative ratings by about 0.05 to 0.08 points (compared to about 0.04 points overall). In ELA, we estimate an increase of about 0.05 to 0.10 points.

Taken together, the results suggest that assignment to lower achieving classrooms reduces teachers' evaluation ratings. As in prior studies, however, our research design focuses primarily on within-school variation in classroom characteristics and holds constant many of the implementation factors discussed above. For the remainder of the analysis, we therefore focus on describing differences in ratings that are not explained by the observable characteristics of classrooms.

4.2. Variation in Ratings across Schools and Districts

Schools and districts in Massachusetts vary considerably in the extent to which they award teachers high or low ratings on their evaluations. We plot this variation for the 25 largest school districts in Massachusetts in Figure 1. These districts account for 3,802 teacher-year observations and 2,037 unique teachers (33 percent and 35 percent of our sample, respectively). In Figure 1, the stacked bars show the proportion of teachers in each district earning each rating. The median large district awards the proficient rating to 87.7 percent of teachers, but this number varies across districts: the 10th and 90th percentile districts award 73.9 percent and 96.6 percent of teachers the proficient rating, respectively. Percentage differences in the use of other rating categories are even more substantial given the lower frequency of these ratings. For example, four of the 25 school districts award no ratings below proficient, while six award at least 10 percent of their teachers these ratings. Similarly, at the top end of the distribution, two districts award no exemplary ratings and five districts award exemplary ratings to at least 10 percent of their teachers. Chi-square tests (not shown) demonstrate that the distribution of ratings is statistically significantly different across school districts ($p<0.001$). However, the variation in average ratings is not strongly associated with differences in teacher value added. In Figure 2, we plot the distribution of value added for the same school districts. We group teachers by quantile of estimated value added and assign them to groups of the same size as the performance ratings in Figure

1 (bottom 0.5 percent, 0.5th percentile to 6.1th percentile, etc.). The rank order correlation between mean performance ratings and mean value added is 0.04.

The graphical evidence suggests that districts may apply different standards in their evaluation systems. However, it is also possible that the different ratings are a reflection of true district-level differences in the effectiveness of teachers in different districts. To explore this issue more formally we estimate models that include teacher, school, and/or district random effects, such as

$$E_{jsdt} = \alpha_j + \theta_s + \gamma_d + \epsilon_{jsdt} \quad (3)$$

where α_j is a teacher random effect, θ_s is a school random effect, and γ_d is a district random effect. We specify Eq. (3) as an ordered probit to reflect the ordinal nature of the ratings data. These models estimate the variation in mean differences in rating scores across schools or districts. Such differences in scores could result from differences in unobserved student characteristics across schools that affect subjective evaluations or from more stringent rating standards in some locations. But we would also expect to observe variation in performance ratings across districts if there are true differences in teacher effectiveness and the graphical evidence in Figure 2 suggests that this is likely to be the case. We control for differences in teacher effectiveness by including teacher value added in math and ELA and teacher experience. As before, it is unlikely that these proxies adjust for all differences in teacher quality across districts as a number of important teaching skills are only weakly correlated with teacher value added (Gershenson, 2016; Jackson, 2016; Sass et al., 2014). We therefore control for the means of these variables, and in some cases student characteristics as well, at the school and district level, in order to adjust for sorting of teachers to schools and districts based on the observed effectiveness measures. Although an imperfect method for controlling for sorting of teachers to districts based on unobserved effectiveness, this approach works similarly to the use of classroom mean student

characteristics to provide additional adjustment for unobserved determinants of student achievement gains in teacher value added models (Altonji & Mansfield, 2014).

We show results from these regressions in Table 4. In columns 1 and 4, we estimate models with the random effects from Eq. (3) and no other covariates. Column 1 demonstrates that the variance of the school effects is about 0.70, while Column 4 shows that the variance across schools is about evenly split between variation in the average ratings across school districts (0.35) and variation in the average ratings across schools within a district (0.36). In columns 2 and 5, we add the proxies for teacher effectiveness. Inclusion of the teacher effectiveness proxies reduces our estimates of the variance in average ratings across schools and districts by about 15 percent in each case, with most of the reduction coming from differences in measured teacher effectiveness across schools. Finally, in columns 3 and 6, we add classroom and school characteristics to the model. These characteristics include average baseline achievement as well as the student demographic and program participation information listed in Table 1. The inclusion of these variables reduces the variance of the total across-school effects by a similar proportion as the inclusion of the teacher proxies in column 2. Nonetheless, the variances of the school and district effects remain significant.

The estimates are on the ordered probit scale and do not have a natural interpretation. To provide some context, we estimate the change in the probability of extreme ratings at the mean of the covariate distribution associated with a one standard deviation increase in the school or district effects using the estimates reported in column 6. The average effect of such a movement, which corresponds to the difference between the median school or district and one at the 68th percentile, is about a one percentage point decrease in the probability of receiving a needs improvement or unsatisfactory rating and about a four-percentage point increase in the probability of receiving an exemplary rating. Given that the mean probability associated with these events is 0.04 and 0.10, respectively, schools and district assignments appear to significantly affect the likelihood of extreme ratings.

4.3 The Relationship between Teacher Evaluation Ratings and Value-Added

The school and district random effects models measure mean differences in ratings conditional on other empirical measures of teacher effectiveness, but they may fail to capture important variation in how the evaluation framework is implemented. In particular, decisions about the number of observations to conduct, the intensity of rater training, or quotas for the proportion of teachers in each rating category likely influence the correlation between true teacher effectiveness and the performance ratings. Differences in the strength of the relationship between performance ratings and teacher effectiveness may not manifest in mean differences in ratings across schools. Instead, there may be offsetting biases for high- and low-performing teachers. That is, even if schools do not differ in their mean rankings, the probability of receiving an extreme rating conditional on effectiveness may differ.

To assess how the relationship between teacher effectiveness and performance ratings varies across districts, we estimate random coefficients models that quantify the variation in the strength of the relationship between teacher value added and performance ratings. These models differ from those in the previous section in that they allow the relationship between teacher value added and performance ratings to vary at the district level. In particular, we estimate the following model as an ordered probit:

$$E_{jsdt} = X_{jsdt}\beta + TVA_{jsdt}\gamma_d + \alpha_j + \theta_s + \gamma_d + \epsilon_{jsdt}. \quad (4)$$

We present the variance components for the random coefficient models in Table 5. In columns 1 and 2, we estimate models that omit teacher random effects. We find that the sensitivity of the performance ratings to teacher value added differs substantially across different districts. Specifically, the mean coefficient on value added is 1.45 and the variance of the district random coefficients is 1.10. Hence, there is considerable variability in the sensitivity of ratings to estimated teacher value added. The point

estimates also suggest that districts with higher average ratings tend to discriminate more on the basis of teacher quality, although the estimate of the covariance between the district random effects and random coefficients is not statistically significant.

In columns 3 and 4, we add teacher random effects to the models. These estimates are closest to the variance components results in Table 4. Note that the variance of the random effects and residual terms in the model are not uniquely identified. By convention, the variance of the error term is fixed at 1. Thus, the change in the magnitude of the coefficients reflects the renormalization of the residual error term to exclude variation in ratings across individual teachers. Nonetheless, the relative magnitude of the variance components remains fairly similar. The school and district random effects are similar to those presented in Table 4. In all cases, we observe statistically significant variation in the sensitivity of the performance ratings to estimated teacher value added.

As before, the interpretation of the variance in the random coefficients is not straightforward. We therefore plot predicted probabilities of below-proficient and exemplary ratings by teacher value added for the 25 largest school districts in Figure 3. To ensure that the comparisons are for similar teachers, we fix each of the other covariates, as well as the teacher and school random effects, at the sample mean and use the estimated district random coefficients and intercepts from the estimates shown in column 3 of Table 5. As with the regression results, there is considerable variation in the probability of earning high or low ratings. At the average school in a district, the probability of a teacher at the 10th percentile of the value-added distribution receives a below-proficient rating varies from 0.03% to 10.80%. We observe similar heterogeneity for exemplary ratings. For a teacher at the 90th percentile of the estimated value added distribution, the probability ranges from 0.16% to 24.67%.

5. Conclusion

In this study, we document significant variation in teacher performance ratings across schools and districts in Massachusetts. Much of this variation remains after controlling for proxies for teacher effectiveness and school and district characteristics. Districts differ meaningfully in the likelihood that they award teachers especially high or low ratings, even conditional on measures of teacher effectiveness. 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, and that relatively little of this variation can be explained by the characteristics of a teacher's classroom. Consequently, similarly effective teachers might expect to earn different ratings depending on where they work.

These findings have a few important policy implications. First, differences in the likelihood of high or low evaluations ratings across schools and districts imply that statewide policies that connect stakes to these ratings may be applied differently across different schools and districts. A potential solution to this problem is to standardize evaluation scores within schools or districts for the purposes of high-stakes decision making, but this implicitly assumes that teacher quality is equitably distributed across different schools and district, which prior evidence suggests is not the case in Massachusetts (Cowan et al., 2017). Thus, it may be difficult to design a system (e.g., for determining teacher tenure) that applies performance evaluation ratings consistently across an entire state. That being said, it is not obvious that perfectly consistent application is a desirable objective, especially given evidence that district and school leaders might adapt rating standards to local needs (Kraft & Gilmour, 2016; Pogodzinski et al., 2016).

One important caveat is that our findings do not identify any specific mechanism (besides average classroom achievement) that explains the variation in ratings across schools and districts in the

state. Although factors such as the choice of evaluation metrics or the number of classroom observations may influence the association between quantitative measures of teacher effectiveness and their performance ratings, districts may also make conscious decisions about how to map teacher performance onto the discrete categories provided by a state evaluation framework and some of these decisions may be responsive to the local policy context. For instance, districts may be reluctant to provide low performance ratings if their conceptualization of effective teaching differs meaningfully from the state standard (MacLeod, 2003), or when they face difficulties attracting high-performing teachers.

This latter possibility underscores the difficulty in connecting these results to trends in the broader teacher labor market. The variability we observe in the likelihood of receiving extreme ratings could theoretically contribute to teacher sorting across school districts. In districts with evaluation systems that are less sensitive to differences in teacher effectiveness, effective teachers are less likely to receive top ratings and less effective teachers are less likely to receive low ratings. Thus, these districts may be more attractive to low performing teachers and less attractive to high performing teachers when high stakes are attached to formal evaluations. On the other hand, these differences might merely reflect pre-existing sorting among teachers. That is, preferences for more or less rigorous evaluation systems might differ among school systems. Future research could use variation in the application of performance ratings over time to tease apart these possibilities, and could provide importance evidence about whether the variation we document in this paper has implications for the sorting of teachers across different schools and districts.

References

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

Altonji, J. G., & Mansfield, R. K. (2014). *Group-average observables as controls for sorting on unobservables when estimating group treatment effects: The case of school and neighborhood effects* (National Bureau of Economic Research Working Paper No. 20781). Cambridge, MA: National Bureau of Economic Research.

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. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. 23478). Cambridge, MA: National Bureau for Economic Research.

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.

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. (n.d.). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*. Forthcoming.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review*, *107*(6), 1685–1717.

Cowan, J., Goldhaber, D., & Theobald, R. (2017). Teacher equity gaps in Massachusetts. Retrieved from http://www.doe.mass.edu/research/reports/2017/10teacher-equity.pdf.

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

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 and 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.

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., & 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* (Measures of Effective Teaching Project). Seattle, WA: Bill and 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.

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.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. 14607). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2011). *Learning about Teaching* (Measures of Effective Teaching Project). Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching* (Measures of Effective Teaching Project). Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, *46*(3), 587–613.

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education, 123*, 615–646.

MacLeod, W. B. (2003). Optimal contracting with subjective evaluation. *American Economic Review*, *93*(1), 216–240.

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 and Melinda Gates Foundation.

Milanowski, A. (2017). Lower performance evaluation practice ratings for teachers of disadvantaged students. *AERA Open*. https://doi.org/10.1177/2332858416685550

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. https://doi.org/10.2139/ssrn.1267109

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

Pogodzinski, B., Lenhoff, S., Mayrowetz, D., Superfine, B., & Umpstead, R. (March 2016). The relationship between district stressors and teacher evaluation outcomes. Paper presented at Association of Education Finance and Policy Annual Conference. Washington, DC.

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. *The American Economic Review*, *102*(7), 3184–3213.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, *107*(6), 1656–1684.

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.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*(7), 378-396.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651.

U.S. Department of Education. (2012). *Evaluations of teacher effectiveness: State requirements for classroom observations*. Washington, DC: Reform Support Network, U.S. Department of Education.

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*(5), 1045–1058.

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.

TABLES

Table 1. Summary Statistics

	(1) All Teachers	(2) Below Proficient	(3) Proficient	(4) Exemplary
Rating	3.064	1.930	3.000	4.000
	(0.383)	(0.256)	(0.000)	(0.000)
Formative evaluation	0.354	0.147	0.363	0.359
	(0.478)	(0.355)	(0.481)	(0.480)
Experience	4.392	3.363	4.394	4.751
	(1.177)	(1.695)	(1.166)	(0.735)
Math value added	0.008	-0.095	0.004	0.068
	(0.175)	(0.162)	(0.173)	(0.169)
ELA value added	0.006	-0.083	0.002	0.065
	(0.168)	(0.155)	(0.166)	(0.171)
Avg. math achievement	0.049	-0.231	0.053	0.112
	(0.516)	(0.593)	(0.512)	(0.481)
Avg. ELA achievement	0.050	-0.257	0.057	0.108
	(0.537)	(0.645)	(0.528)	(0.530)
Male students	0.507	0.520	0.507	0.505
	(0.087)	(0.098)	(0.087)	(0.085)
FRL-eligible students	0.378	0.573	0.372	0.357
	(0.324)	(0.350)	(0.321)	(0.316)
LEP students	0.079	0.143	0.076	0.083
	(0.157)	(0.237)	(0.152)	(0.160)
SPED students	0.174	0.166	0.174	0.182
	(0.143)	(0.170)	(0.141)	(0.144)
Asian students	0.065	0.055	0.067	0.053
	(0.103)	(0.088)	(0.105)	(0.088)
African American students	0.075	0.109	0.074	0.069
	(0.133)	(0.155)	(0.132)	(0.124)
Hispanic students	0.173	0.309	0.167	0.180
	(0.235)	(0.294)	(0.228)	(0.252)
N	11,563	441	9,908	1,214

Notes: Summary statistics for teachers in self-contained classrooms in fourth and fifth grades in 2014-2016. Summative rating is the teacher's final rating on a four-point scale. Teacher value added is estimated over the 2011-2016 period. Observations at the teacher-year level.

Table 2. Classroom Achievement and Teacher Evaluations

	(1)	(2)	(3)	(4)	(5)
Class Lagged Achievement	0.0688***	0.0424***	0.0479***	0.0443*	0.0448**
	(0.0102)	(0.0108)	(0.0156)	(0.0241)	(0.0215)
Formative Assessment	0.0368***	-0.0064	-0.0320***	-0.0320***	-0.0352***
	(0.0078)	(0.0077)	(0.0069)	(0.0076)	(0.0067)
N	11,563	11,563	11,563	11,563	11,563
Controls	N	Y	N	N	N
Teacher FE	N	N	Y	Y	N
Teacher-Grade FE	N	N	N	N	Y
Cohort Achievement Instrument	N	N	N	Y	Y

Notes: Regressions in columns (2) – (5) contain controls for teacher experience (indicators for one to four and five or more years of experience). The regression in column (2) additionally includes controls for teacher value added in math and ELA, school average experience and value added, and district average experience and value added. Standard errors clustered by school in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table 3. Other Classroom Achievement Results

	(1)	(2)	(3)	(4)	(5)
Panel A. Lagged ELA Achievement					
Class Lagged Achievement	0.0642***	0.0369***	0.0700***	0.0613**	0.0473*
	(0.0107)	(0.0112)	(0.0173)	(0.0292)	(0.0275)
Formative Assessment	0.0360***	-0.0067	-0.0321***	-0.0321***	-0.0352***
	(0.0078)	(0.0077)	(0.0062)	(0.0066)	(0.0087)
N	11,563	11,563	11,563	11,563	11,563
Panel B. Summative Evaluations (Math)					
Class Lagged Achievement	0.0842***	0.0519***	0.0712***	0.0658**	0.0815**
	(0.0125)	(0.0133)	(0.0214)	(0.0310)	(0.0346)
N	7,469	7,469	7,469	7,469	7,469
Panel C. Summative Evaluations (ELA)					
Class Lagged Achievement	0.0837***	0.0509***	0.0952***	0.0743**	0.0599
	(0.0128)	(0.0135)	(0.0307)	(0.0373)	(0.0465)
N	7,469	7,469	7,469	7,469	7,469
Controls	N	Y	N	N	N
Teacher FE	N	N	Y	Y	N
Teacher-Grade FE	N	N	N	N	Y
Cohort Achievement Instrument	N	N	N	Y	Y

Notes: In Panel A, we include all evaluations for teachers in grades 4 and 5. The independent variable in Panels A and C is lagged classroom ELA score. In Panels B and C, we limit the sample to years in which teachers received summative evaluations only. Regressions in columns (2) – (5) contain controls for teacher experience (indicators for one to four and five or more years of experience). The regression in column (2) additionally includes controls for teacher value added in math and ELA, school average experience and value added, and district average experience and value added. Standard errors clustered by school in parentheses.

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

Table 4. Variance Components of Teacher Ratings

	(1)	(2)	(3)	(4)	(5)	(6)
Variance of Teacher Effects	2.065***	1.755***	1.701***	2.054***	1.741***	1.690***
	(0.186)	(0.161)	(0.158)	(0.185)	(0.160)	(0.158)
Variance of School Effects	0.701***	0.586***	0.507***	0.354***	0.275***	0.256***
	(0.081)	(0.070)	(0.065)	(0.063)	(0.056)	(0.054)
Variance of District Effects				0.357***	0.344***	0.296***
				(0.071)	(0.071)	(0.065)
Teacher Proxies	N	Y	Y	N	Y	Y
Class/School Char.	N	N	Y	N	N	Y
N	11,563	11,563	11,563	11,563	11,563	11,563

Notes: Ordered probit regressions of summative ratings on teacher characteristics with teacher, school, and district random effects. Regressions in columns (2) – (3) and (4) – (5) contain controls for teacher value added and experience for individual teachers as well as school and district averages. Regressions in columns (3) and (6) additionally contain classroom, school, and district average student characteristics. Standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table 5. Variation in Sensitivity to Teacher Effectiveness across Districts

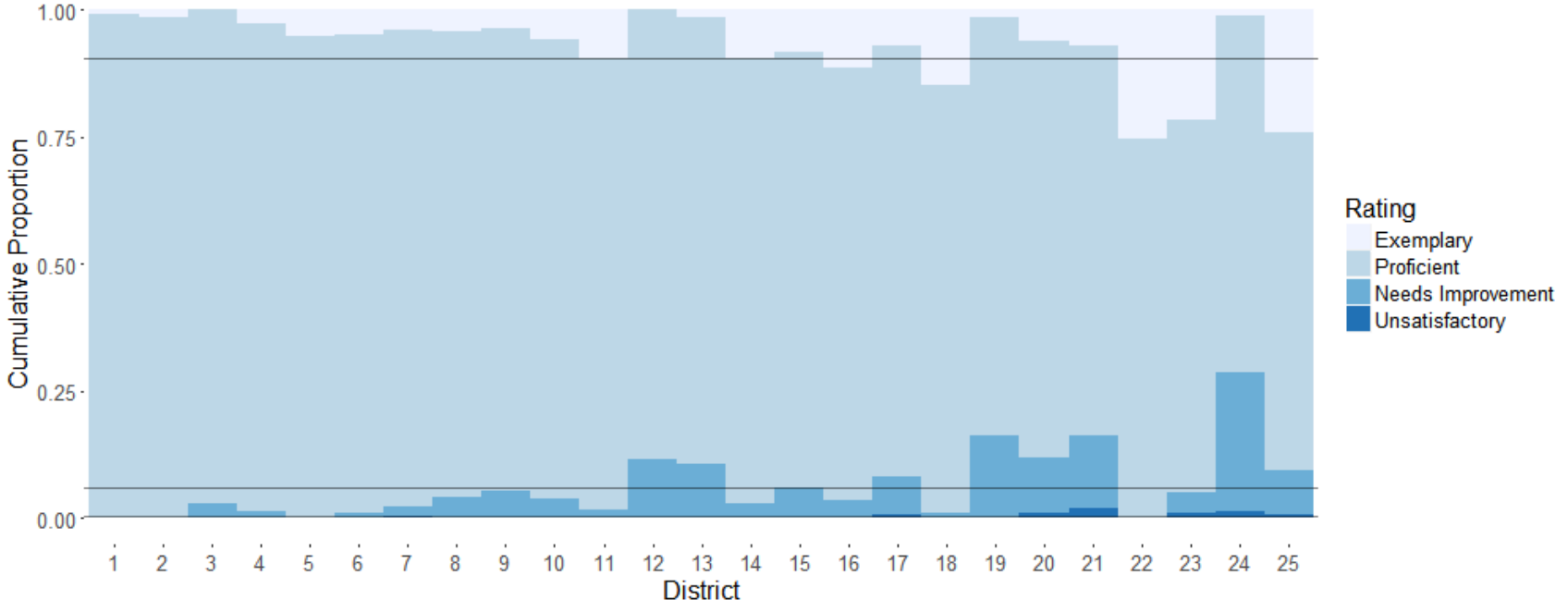
	(1)	(2)	(3)	(4)
Coefficient on Teacher Value Added	1.418***	1.447***	2.313***	2.334***
	(0.133)	(0.133)	(0.224)	(0.222)
Variance of Teacher Effects			1.681***	1.640***
			(0.156)	(0.155)
Variance of School Effects	0.175***	0.168***	0.261***	0.241***
	(0.023)	(0.023)	(0.055)	(0.053)
Variance of District Effects	0.142***	0.129***	0.340***	0.294***
	(0.029)	(0.028)	(0.072)	(0.066)
Variance of District Value Added Coefficients	1.152***	1.096***	1.732***	1.620***
	(0.268)	(0.265)	(0.547)	(0.533)
Covariance between Dist. VA Coefficients and Effects	0.072	0.090	0.116	0.173
	(0.061)	(0.060)	(0.137)	(0.134)
Teacher Proxies	Y	Y	Y	Y
Class/School Char.	N	Y	N	Y
N	11,563	11,563	11,563	11,563

Notes: Ordered probit regressions of summative ratings on teacher characteristics with teacher, school, and district random effects. All regressions contain controls for teacher value added and experience for individual teachers as well as school and district averages. Regressions in columns (2) and (4) additionally contain classroom, school, and district average student characteristics. Standard errors in parentheses.

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

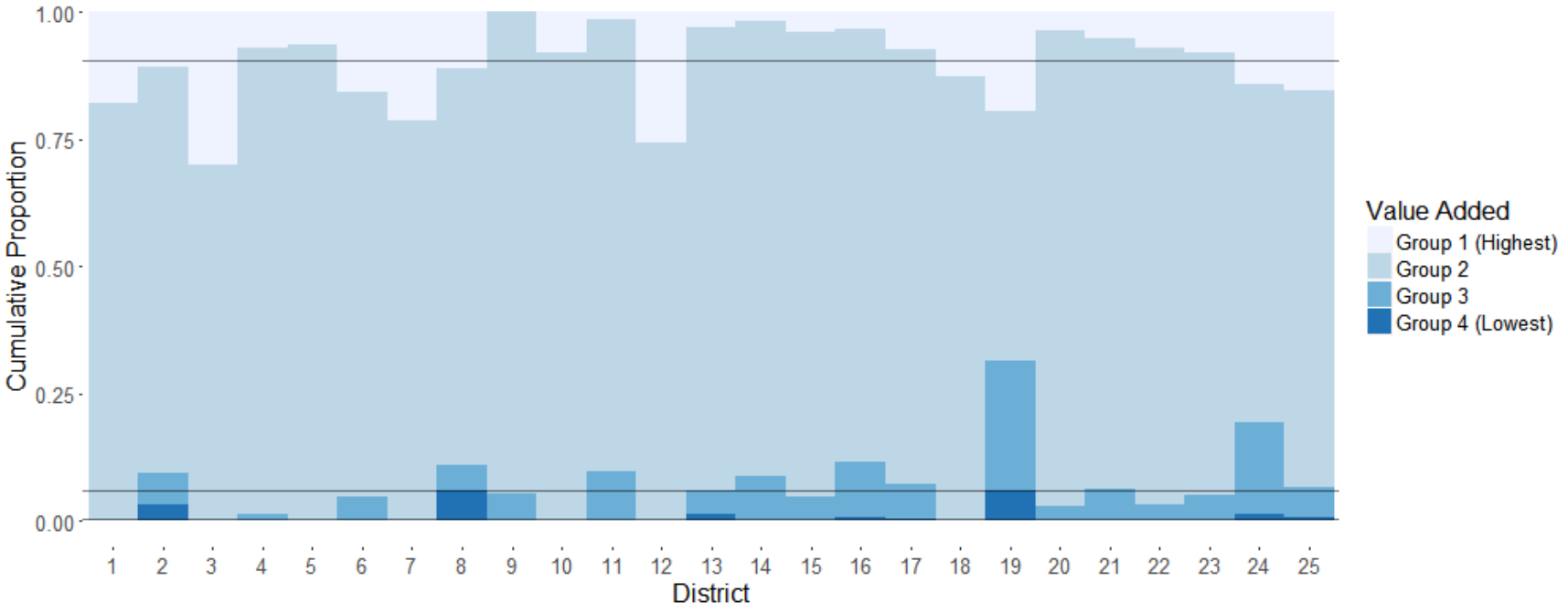
FIGURES

Figure 1. Distribution of Ratings across School Districts



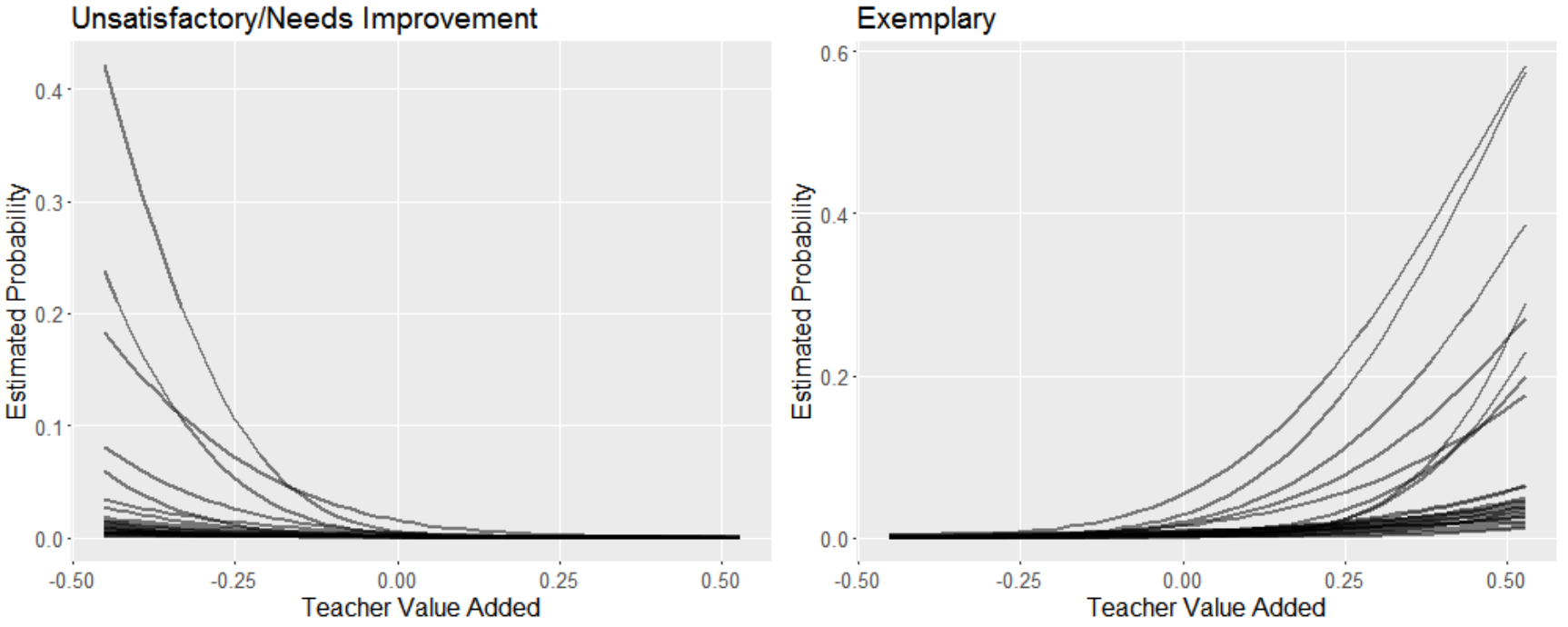
Notes: Proportion of teachers in each of the 25 largest districts receiving the given rating on their performance evaluations.

Figure 2. Distribution of Value Added across School Districts



Notes: Proportion of teachers in the 25 largest school districts receiving value-added ratings in each of the specified groups. We group teachers by quantile of estimated value added and assign them to groups of the same size as the performance ratings in Figure 1 (bottom 0.5 percent, 0.5th percentile to 6.1th percentile, etc.).

Figure 3. Variation in Ratings across Districts



Notes: Estimated probability of extreme ratings by teacher value added for each of the 25 largest districts in the states. The probabilities are estimating using the random coefficient model in column 3 of Table 5 with all teacher, school, and district characteristics fixed at the sample means and varying the teacher value added estimate.

Appendix A. Comparison of Simulated Data from Random Coefficients Model to Actual Data

We use random coefficients models to estimate variation in the sensitivity of ratings to differences in teacher effectiveness across schools. In this appendix, we verify that the regression model provides a reasonable summary of the observed data. To do so, we create 100 simulated datasets by drawing the random effects and then simulating summative ratings using the observed predictors and estimated regression coefficients. Specifically, we first draw teacher, school, and district random effects from the appropriate normal distributions. In the case of the district effects, this involves sampling the random intercepts and coefficients on teacher value added from a bivariate normal distribution. All of the parameters of these distributions are as given in column 3 of Table 5. We then use the observed covariate data and estimated coefficients to draw simulated performance ratings. We average each of the estimates from the simulated data over the 100 iterations.

The simulated ratings data match several features of the actual data (in Table A.1). In the observed data, 85.7% of teachers receive a proficient rating; in the simulated datasets, on average 85.2% of teachers do so. The variation across and within schools is also similar. The standard deviation of ratings within schools is 0.35 in the real data and 0.36 in the simulated data. The models actually slightly understate the variance in ratings across schools; the school-level standard deviation of ratings is 0.16 in the real data and 0.14 in the simulated data. The data also closely matches the year-to-year correlations in individual teacher's performance ratings; this is 0.49 in the actual data and 0.45 in the simulated data. Finally, we calculate correlations in the summative ratings in the year before and after school and district switches in the real and simulated data. The simulated data match the changes in teacher ratings across districts quite well: ratings are correlated at 0.28 across districts in the actual data and 0.30 in the

simulated data. These transitions form the basis for the estimated probabilities in Section 4.3. On the other hand, the simulated data understates the correlation in ratings across schools. We estimate this to be 0.35 in the simulated data, but it is 0.43 in the observed data.

Next, in Figure A.1, we plot the relationship between teacher value added and the likelihood of receiving an exemplary or unsatisfactory/needs improvement rating. We estimate these using local linear regression on the real data and the 100 simulated datasets. The fitted relationships are shown in Figure A.1 for the observed and simulated data.

Table A.1. Comparison of Simulated Data from Random Coefficients Model to Actual Data

	Observed Data	Simulated Data
Percent Unsatisfactory	0.28%	0.31%
Percent Needs Improvement	3.66%	3.47%
Percent Proficient	85.72%	85.20%
Percent Exemplary	10.35%	11.01%
Within-School Standard Deviation of Ratings	0.35	0.36
Across-School Standard Deviation of Ratings	0.16	0.14
Year-to-Year Correlation in Teacher Ratings	0.49	0.45
Correlation in Teacher Ratings Before/After District Move	0.28	0.30
Correlation in Teacher Ratings Before/After School Move	0.43	0.35

Figure A.1. Relationship between Teacher Value Added and Performance Ratings