Teacher Performance Ratings and Professional Improvement Effort

Cory Koedel, Jiaxi Li, Matthew G. Springer, Li Tan

October 2017

Like other public workers, teachers typically receive high and compressed ratings that do little to differentiate them based on performance. Motivated by empirical evidence of substantial variation in effectiveness among teachers, there has been a recent push to develop more informative evaluation systems with greater ratings dispersion. We study one of the first of these new systems, implemented in Tennessee, in order to understand how teachers respond to the provision of new, more-differentiated performance ratings. We focus on whether summative ratings influence teachers' self-directed professional improvement activities as measured by a statewide teacher survey. Using a regression discontinuity design, we find that teachers do not alter their time investments in professional improvement, or adjust their professional improvement activities based on evaluation feedback, in response to their ratings.

Keywords: teacher evaluation; public employee evaluation; compressed employee ratings; rating effects; regression discontinuity

Acknowledgement

Koedel is in the Department of Economics and Truman School of Public Affairs, and Li and Tan are in the Department of Economics, at the University of Missouri. Springer is in the Peabody College of Education and Human Development at Vanderbilt University. This study was supported by the Tennessee Education Research Alliance at Vanderbilt University's Peabody College, a unique research-practice partnership between Vanderbilt University and State of Tennessee. We appreciate helpful comments and suggestions from Dale Ballou, Jason Grissom, Colleen Heflin, Peter Mueser, Michael Podgursky and Nate Schwartz. We would also like to acknowledge the many individuals at the Tennessee Alliance and Tennessee Department of Education for providing data and expert insight to conduct our analysis, in particular, Susan Burns, Erin O'Hara, and Matthew Pepper. The usual disclaimers apply.

1. Introduction

In an influential policy report, Weisberg et al. (2009) document that high and compressed ratings are ubiquitous in teacher evaluation systems across the U.S. Among their findings are that less than 1 percent of teachers receive an unsatisfactory rating, and the evaluations for almost 3 in 4 teachers do not identify any areas for improvement.[1] The situation for the public workforce more broadly is similar. For example, a 2013 report from the U.S. Government Accountability Office (U.S. GAO, 2013) found that 99 percent of permanent federal employees are rated at "fully successful" or above in their annual reviews, and 61 percent are rated as "outstanding" or "exceeds fully successful."[2]

Recent empirical evidence from the rapidly expanding literature on teacher quality shows that there are substantial differences in effectiveness among teachers as measured by their ability to raise student achievement (Koedel, Mihaly and Rockoff, 2015). Chetty, Freidman and Rockoff (2014) further link these differences to longer-term outcomes of interest including college attendance, teenage childbearing, and earnings. The misalignment between what we know to be large and important differences in teacher effectiveness, and the small differences implied by teachers' formal evaluation ratings, has spurred policy efforts to develop more rigorous and informative evaluation systems in the education sector. The federal Race to the Top competition, which was first held in 2010 and incentivizes states to build better systems for teacher evaluation, is a prime example.[3]

Although the implementation of new systems has been mixed (Kraft and Gilmour, 2017), some states and school districts have introduced meaningfully-revised evaluations that differ substantially from historical norms in their depth and rigor. One example is the new statewide system

[1] A follow-up study by Kraft and Gilmour (2017) shows some movement toward more-differentiated teacher evaluations nationally, but generally confirms the basic conditions found by Weisberg et al. (2009). Grissom and Loeb (2017) document that principals are more likely to assign lower ratings to teachers in low-stakes settings than in high-stakes settings.

[2] Rating compression in the public sector is particularly acute, but the phenomenon is not unique to public workers (e.g., see Murphy and Cleveland, 1991; Prendergast, 1999).

[3] For more information, see: https://www2.ed.gov/programs/racetothetop/index.html.

in Tennessee – an initial Race to the Top grant awardee – first implemented during the 2011-12 school year.[4] We use statewide data from Tennessee to assess teachers' responsiveness to more differentiated information about performance. Teachers (and their employers) may respond to differentiated ratings in many ways. We focus specifically on understanding the response of teachers in terms of self-directed professional improvement behaviors. The motivation for this line of inquiry is the concern that the typically high and compressed ratings most teachers receive hide useful information about performance (Almy, 2011; Papay, 2012; Weisberg et al., 2009); if teachers have little in the way of private information about their own effectiveness, those who are not performing well may not know it and will not see the need to improve. We measure teachers' self-directed improvement activities using data from a statewide survey that includes four questions about professional improvement.

Much of the literature to date on the use of more-differentiated teacher performance measures has focused on employer-initiated human resource policies, real or hypothetical, such as selective retention policies (e.g., Hanushek, 2011; Springer, Swain, and Rodriguez, 2016; Staiger and Rockoff, 2010; Winters and Cowen, 2013) or task re-allocation policies (e.g., see Condie, Lefgren and Sims, 2014; Goldhaber, Cowan and Walch, 2013). Other related studies include (a) Papay et al. (2016), who study a peer-to-peer professional development intervention that uses differentiated teacher performance information to pair teachers, and (b) Taylor and Tyler (2012), who study the effect of a deeper and more rigorous evaluation process on teacher effectiveness. Taylor and Tyler (2012) find that the evaluation process they study has large effects on teacher improvement. While their research design is not suited to identify precise mechanisms, they suggest as a possibility that teachers learn new information about their own performance from the evaluation process, and respond by increasing effort and/or developing new skills.

[4] Kraft and Gilmour (2017) show that the new system in Tennessee is among the most aggressive in the country in terms of identifying teachers as below proficient.

We estimate the causal effects of teachers' ratings on their self-directed professional-improvement behaviors using a regression discontinuity (RD) design. Our models compare otherwise similar teachers who differ by the summative signal they receive about performance from the Tennessee system. We find that rating differences alone do not affect teachers' professional improvement behaviors, either in terms of their intensity or responsiveness to evaluative feedback. The rating effects we estimate are statistically insignificant, inconsistent in sign, and precise enough to rule out modest impacts. The null results are robust to a number of measurement and modeling modifications, and indicate that the provision of more-differentiated summative information about performance alone is not sufficient to change professional-improvement behavior.

In a discussion section we contextualize our findings, both with respect to the policy environment in Tennessee and the recent related literature on how ratings affect teacher and student behavior (e.g., Dee and Wyckoff, 2015; Koedel et al., 2017; Papay et al., 2016). We also discuss several limitations of our study. In brief, the most notable are the standard limitations inherent to the regression discontinuity approach and its necessarily local identification, and our focus on just one channel through which the new Tennessee system conveys information about teacher performance – the summative rating.

2. Background

In July 2011, the Tennessee State Board of Education approved four teacher evaluation models: Tennessee Educator Acceleration Model (TEAM), Project Coach (COACH), Teacher Effectiveness Measure (TEM), and Teacher Instructional Growth for Effectiveness and Results (TIGER). A fifth model, the Achievement Framework for Excellent Teaching (AFET), was first approved for use during the 2012-2013 school year. All of the models have the same goals – to monitor teacher performance and encourage teacher development – and include a post-observation conference to discuss teachers' strengths and weaknesses based on their classroom performance. Observers may

also directly recommend actions or resources to help teachers improve their instructional skills, and can follow up to see how teachers are responding to address their indicators of weakness.

We focus on teacher responses to ratings they received during the 2012-13 school year, the second year of the new evaluation system. Our focus on the second year is motivated in part by findings from Dee and Wyckoff (2015), who show that teachers became more responsive to the IMPACT evaluation system in Washington DC as it matured. Thus, we anticipate teachers' second-year evaluation ratings are more likely than their first-year ratings to affect professional improvement activities.[5]

The summative ratings for teachers during the 2012-2013 evaluation year are comprised of three components. For teachers with available growth measures based on student performance on standardized tests (i.e., teachers in tested grades and subjects), 35 percent of the final rating is based on the growth measure, 15 percent on additional measures of student achievement chosen through mutual agreement by the educator and evaluator, and the remaining 50 percent on qualitative measures including classroom observations, student perception surveys, personal conferences, and a review of prior evaluations and work. For teachers without an individual student growth measure, grade- or school-level growth is used as a substitute and the evaluation weights are changed so that 25, 15 and 60 percent of the final rating depends on growth, additional achievement measures, and classroom observations and other measures, respectively.[6] As a practical matter, teachers' scores on the latter component are primarily driven by classroom observations.

The overall effectiveness scores range from 0 to 500 in all evaluation models and are used to assign teachers to discrete performance categories. Denoting X as the teacher score, for all models

[5] We also perform a complementary analysis of teachers' responses to their initial ratings from 2011-12 and obtain similar – i.e., null – results (omitted for brevity). In addition to the "system maturity" issue mentioned in the text, another limitation of the analysis of initial ratings is that the first follow-up survey in Tennessee asked just two questions about professional improvement. The professional improvement portion of the 2012-13 survey consists of four questions (described below and in Appendix A).

[6] Approximately 41 percent of teachers in Tennessee have an individual growth measure (see Table 1).

teachers with $X < 200$ are categorized as "Significantly Below Expectation" (level 1), teachers with $200 \leq X < 275$ as "Below Expectation" (level 2), teachers with $275 \leq X < 350$ as "At Expectation" (level 3), teachers with $350 \leq X < 425$ as "Above Expectation" (level 4), and teachers with $X \geq 425$ as "Significantly Above Expectation" (level 5). Importantly, rating reports provided to teachers include the discrete rating but not the underlying score on the 0-500 scale. This is useful for interpreting our regression discontinuity results because it means that teachers with very similar underlying scores but different discrete ratings were not provided the information to determine their closeness to the threshold.

While Tennessee law indicates that teachers' evaluation ratings will be incorporated into compensation, promotion, retention, tenure, and certification decisions, only some of these policies had been drafted and implemented at the time of our study and they did not apply universally. An example of a policy that was in place is that teachers working in disadvantaged schools under the Tennessee Achievement School District (ASD) program who earn a rating of three or higher could earn salary increases and/or promotions not available to teachers with lower ratings. However, most teachers in Tennessee were not covered by any state policy explicitly linking ratings to rewards and sanctions at the time of our study. Thus, the primary drivers of any behavioral effects of teachers' ratings will be factors such as (1) the transmission of new information about relative performance to individual teachers; (2) psychological effects of different rating assignments; and (3) informal/local policies that reward and sanction teachers based on their ratings such as promotions and access to desirable teaching assignments (e.g., Bates, 2016; Cullen, Koedel and Parsons, 2016), and the use of improvement plans or targeted professional development to low performers. We provide a more detailed discussion of the policy context below.

3. Data

3.1 Ratings, Administrative and Survey Data

We combine three data sources for our analysis: teacher rating data, administrative data from the Tennessee Department of Education, and data from an annual survey administered to all teachers as part of the First to the Top (FTTT) initiative in Tennessee. As noted above, the rating data are based on teacher performance during the 2012-2013 school year. The administrative data include information on each teacher's gender, race, education level, years of experience, and school assignment (with corresponding school characteristics), also from the 2012-2013 school year. The survey data are from the annual Tennessee Alliance Educator Survey administered during the spring of the 2013-2014 school year, after teachers received their ratings from 2012-2013. The survey is designed to improve the state's understanding of how performance evaluations are implemented and how feedback is provided to and processed by teachers.

Table 1 provides descriptive statistics for all teachers in the Tennessee ratings database side-by-side with various subsamples. Column 2 limits the sample to individuals who participated in the survey. Column 3 further restricts the sample to individuals who answered at least one of the four professional improvement questions, which are the focus of the analytic work below. Column 4 shows our primary analytic sample, which is a subsample of column 3 selected based on a bandwidth requirement for our RD models (described below) and several other criteria.[7] Column 5 mirrors column 4 but includes only teachers evaluated under the TEAM model, which is the predominant evaluation model used in the state.

INSERT TABLE 1 HERE

[7] The largest reduction in the sample due to the bandwidth restriction is the loss of level-5 teachers with scores far above the 4/5 threshold, which is of limited consequence given the nature of identification in our RD specifications (we also lose level-2 teachers with scores far below the 2/3 cutoff, but there are fewer of these). We additionally exclude teachers evaluated by the COACH model because it produces a lumpy distribution of underlying scores that is not compatible with the RD models, and teachers without basic covariates.

The analytic work that follows is based primarily on the sample in column 4. A potential concern is that a large fraction of teachers did not submit a survey and/or did not answer the questions about professional improvement, which could generate sample-selection bias. We examine this issue below formally within our RD framework and find no evidence of differential survey-response behaviors around the discontinuities in teachers' ratings, which suggests a limited scope for bias. The other valuable aspect of Table 1 is that it provides an observational comparison of the teachers in our analytic sample and all Tennessee teachers. The analytic sample (column 4) is observationally similar to the full sample, which suggests that our findings will generalize, at least to some extent, to the broader teaching population.[8]

In addition to generally describing the teaching workforce, Table 1 also documents the distribution of teacher ratings in the system. Although ratings are more differentiated in Tennessee than has been typical of the education sector historically, very few teachers receive a score that puts them at level-1 (specifically, 2.2 percent of all Tennessee teachers received a level-1 rating during the 2012-2013 evaluation year). Because of the small sample size, we cannot formally evaluate the effects of ratings around the 1/2 threshold with reasonable precision. This is a limitation in the sense that a particularly low rating may trigger a teacher to be put on a formal improvement plan, which includes explicit improvement expectations. However, from a policy perspective, since so few teachers receive a level-1 rating, the other rating thresholds are more relevant for understanding rating effects on the teaching workforce more broadly.

3.2 Measuring Professional Improvement with the Survey

There are four questions on the survey that elicit feedback from teachers regarding their professional improvement activities. Teacher answers reflect self-reported, self-directed choices about

[8] While most of the differences between samples in Table 1 are statistically significant, this is the result of the very large sample sizes. For example, even differences between the samples in columns 1 and 4 that are clearly not different substantively, like across teacher education levels, are different statistically.

effort to improve practice and responsiveness to evaluation feedback. Table 2 summarizes the content of each question and Appendix A shows how the questions were presented to teachers on the survey. The table also shows how responses to the questions vary by rating level descriptively. For presentational convenience, we code each answer in Table 2 as either a positive or negative response (binary 0/1) by collapsing teachers' more detailed ordered responses (see Appendix A for more details). A positive response indicates more intense professional improvement, or professional improvement that is more responsive to evaluation feedback; and the opposite for a negative response. The share of positive responses to each question at each rating level is reported conditional on a non-missing response.[9]

INSERT TABLE 2 HERE

Table 2 shows that teachers who receive lower ratings report more-intense and more-responsive professional improvement behaviors. But while the descriptive statistics in Table 2 illustrate a clear association between teachers' ratings and professional improvement activities, attributing causality is not straightforward. We overcome the causal inference challenge using the RD design described in the next section.

4. Methodology

4.1 Specification

Our RD models compare teachers whose underlying performance scores are similar but who receive different ratings because of the discrete function that translates the underlying scores into summative ratings. The key identifying assumption is that teachers with similar underlying scores are similar in other respects, and thus conditional on underlying scores the discontinuous rating

[9] On average across the four questions, 9.3 percent of survey respondents did not answer by choice and 12.7 percent were directed to skip the professional improvement questions (along with other questions) due to a position change. Individuals who changed positions were directed to a different set of questions based on their new positions. Response patterns are similar if we include individuals who do not answer the question as "non-positive" in the denominator of each ratio – see Appendix B. We condition on non-missing responses in the regression discontinuity models below; this has no bearing on our findings qualitatively (as discussed in the next section).

assignments can be viewed as effectively random (Hahn, Todd and Van der Klaauw, 2001; Imbens and Lemieux, 2008).

The outcomes in our models are responses to the survey questions shown in Table 2. We use the four questions separately as dependent variables in reduced-form RD models. Our lead specification is an ordered logit with more positive responses coded as higher-ordered values. We also report results from linear RD models where answers to the survey questions are coded as binary positive/negative, and linearly on a 1-8 scale for question-1 and 1-4 scale for the other questions. Our findings are substantively similar regardless of how we code/model teachers' survey responses.

We considered the possibility of combining the information from the survey questions into an index of professional improvement via factor analysis, as in Koedel et al. (2017), but the factor analysis does not yield reliable latent factors.[10] This suggests weak informational overlap across the questions. Thus, we analyze outcomes on a question-by-question basis. We drop records with a missing response to each survey question when that question is used as the dependent variable. In results omitted for brevity we confirm that our findings are not qualitatively affected if we instead directly model missing responses (i.e., if we treat them as separate outcomes in multinomial models).

Our primary specification is a "stacked" model where we estimate the average treatment effect of the higher rating across the performance thresholds 2/3, 3/4 and 4/5 (recall that we do not study the 1/2 threshold because very few teachers receive a level-1 rating – see Table 1). We also estimate models that separately examine each threshold to investigate the potential for threshold-specific effect heterogeneity, which has been found in other recent studies (Koedel et al., 2017; Papay et al., 2016). The models all follow the general form:

[10] With one common factor, the average commonality among questions is 0.32, which is low for a typical factor analysis. That said, in results omitted for brevity we verified the robustness of our findings to a model that uses an "index of professional improvement intensity and responsiveness" as the outcome of interest. The index is a weighted average of answers to the four questions where the weights are determined by factor analysis, as in Koedel et al. (2017).

$$Y_i = \beta_0 + \mathbf{X_i}\boldsymbol{\beta_1} + [f(S_i)]\beta_2 + [f(S_i) * I(S_i \geq T)]\beta_3 + [I(S_i \geq T)]\beta_4 + \varepsilon_i \quad (1)$$

In equation (1), Y_i measures the answer to a professional improvement question on the survey, $\mathbf{X_i}$ is a vector of observable teacher and school characteristics, $f(S_i)$ is a function of the underlying score, or running variable, $I(S_i \geq T)$ is an indicator function equal to one if the score is above the threshold (i.e., the RD indicator), and ε_i is the error term, which we cluster at the school level.[11] The *X*-vector includes teacher gender, race, degree level, certification level, and experience, plus the model used to evaluate the teacher and whether the teacher had an individual growth score. It also includes controls for the shares of students at the teacher's school who are a disadvantaged minority, female and eligible for free or reduced-price lunch. Although we considered several functions for $f(S_i)$, we ultimately specify $f(S_i)$ as a simple linear function of the running variable on both sides of the discontinuity.[12] The parameter of interest is β_4, which under the RD assumptions is the causal effect of receiving a higher rating relative to a lower rating.

An issue that arises with the stacked model is that individual teachers can in principle be identified as both treatments and controls at different thresholds. For example, a teacher with an overall score of "4" is on the high side of the 3/4 cutoff and the low side of the 4/5 cutoff. We ensure that individual teachers are not double-counted in our main models by specifying the bandwidth around each RD cutoff at 37. Per the previous documentation of the score ranges, 37 is the highest

[11] The running variable is not perfectly continuous due to some discreteness in teachers' scores on the subcomponents. The end result is that the values of the running variable cluster around 0.5-unit intervals throughout the range of possible scores. Although the discreteness in the running variable is not egregious in our application by any means, in results omitted for brevity we investigate its implications by estimating variants of our models where we used two-dimensional clustering at the school level and the 0.5-unit-interval level, as suggested by Lee and Card (2008). The standard errors from the alternatively-clustered models are very similar to what we report below, and do not alter inference from our analysis in any way.

[12] We use the *Akaike information criterion* (AIC) test to determine the polynomial order for our primary specification. Adding higher order polynomial terms (up to quartic) of the running variable to the models does not influence our results qualitatively.

integer-value bandwidth that we can assign across all three rating thresholds and still ensure that no individual is double-counted. We also verify the robustness of our findings to narrower bandwidths.

4.2 Validation of the RD Design

The RD design offers a credible approach for identifying the causal effects of teacher ratings subject to several assumptions. In this section, we review and test these assumptions to provide evidence on the extent to which this approach can be useful for informing our research question.

We first examine whether the discontinuities are sharp or fuzzy. Figure 1 shows the probability of a teacher receiving treatment (i.e., the higher rating) as a function of his or her underlying score. In the figure, we aggregate the scores for individual teachers into 5-point bins, center them on the threshold value for the higher rating, and stack the data across the three thresholds. The figure shows that the discontinuities in converting the underlying performance measures into final ratings are fuzzy, albeit only slightly. Given the fuzziness, our reduced-form RD models are properly interpreted as identifying intent-to-treat (ITT) effects of higher ratings. But because the fuzziness is so mild, the ITT effects will be very similar to treatment effects.

INSERT FIGURE 1 HERE

We perform two common tests to look for potential violations of the RD assumptions. The first test examines whether there are other discontinuities in the data that align with the rating discontinuities. If other variables are discontinuous at the discontinuity thresholds, it would suggest that individuals with similar forcing-variable values near the cutoff are not otherwise similar.[13] To determine whether other discontinuities in the data are present and align with the discontinuities in teachers' evaluation scores, we estimate a stacked model of the following form:

$$X_i = \alpha_0 + [f(S_i)]\alpha_1 + [f(S_i) * I(S_i \geq T)]\alpha_2 + [I(S_i \geq T)]\alpha_3 + u_i \qquad (2)$$

[13] Although researchers can overcome the direct threat by controlling for violating covariates in a regression, if discontinuities in observables emerge then it raises the concern that there are other, unobserved discontinuities as well.

In equation (2) X_i is a teacher or school characteristic from equation (1), now used as a dependent variable, all other variables and functions are specified as in equation (1), and u_i is the error term.

Table 3 presents results from a series of linear RD regressions based on equation (2). Each cell shows results from a different regression. Estimates using our primary 37-point bandwidth are shown in Column 1, and for completeness, we show estimates for a range of other bandwidths down to 5 points. Focusing on the results in Column 1, two covariates are unbalanced – the Group-1 indicator and the percentage of minority students in the teacher's school.

With multiple hypothesis testing, some imbalance in Table 3 is likely to occur by chance. In order to determine the likelihood of the observed level of imbalance by chance, we use a randomized-inference test following Cullen, Jacob and Levitt (2005) and Fitzpatrick, Grissmer and Hastedt (2011). We first split the analytic dataset vertically, separately blocking off teachers' (a) covariates (dependent variables) and (b) underlying scores and ratings (independent variables). The key feature of the vertical blocking is that it maintains the covariance structure between the variables in the *X*-vector, which is important because the covariance structure will influence the probability of observing any given number of "statistically significant" relationships by chance with the real data. Next, we randomly sort the block of teacher scores, then re-connect it to the covariate block to assign each teacher a random score. We then estimate the model in equation (2) for each covariate and store the number of covariates that are unbalanced at the 5-percent level when ratings are assigned at random. We repeat this procedure 3,000 times to construct an empirical distribution of covariate imbalance under random assignment.

Based on the randomized-inference simulations, the bottom of Table 3 reports the likelihood of observing at least the number of unbalanced covariates by chance that we observe with the real data at the discontinuities, using the various bandwidths. With our preferred bandwidth in column 1, we observe two unbalanced covariates at the 5-percent level in the real data. Our procedure indicates

this is quite likely by chance, with a *p*-value of 0.44. Although the estimates using the narrower bandwidths bounce around some, which is expected, the remainder of Table 3 does not indicate covariate imbalance in our data at any bandwidth. Thus, we conclude that the degree of covariate imbalance in Table 3 is in line with what one would expect by chance.

Density tests are also commonly used to validate RD designs. These tests look for evidence of "bunching" of the running variable around the discontinuity and can be useful for detecting manipulating behavior. In instances where the running variable is not smoothly distributed around the discontinuity point, the concern is that the lack of smoothness could reflect unobserved differences between individuals near the threshold (i.e., the manipulation may be non-random). A textbook example is a test-score discontinuity where a continuous score is converted to pass-fail, but where students can re-take the test (e.g., see Jepsen, Mueser and Troske, 2016; Van Der Klaauw, 2002). We report results from density tests in Figure 2 and Table 4. We do not find evidence of bunching in the data at any threshold.[14]

INSERT FIGURE 2 AND TABLE 4 HERE

The other important context-specific issue we face in Tennessee is the large fraction of teachers who did not submit a survey and/or answer a professional-improvement question. The threat to identification is that if teachers' decisions to submit a survey are determined in part by their ratings, RD estimates conditional on submitting a survey will be biased by attrition from the dataset that is itself caused by treatment (a form of selection bias). To test whether teachers' decisions to submit a survey were influenced by the discontinuities that convert underlying scores into ratings, we estimate supplementary RD models analogous to Equation (1), but for all Tennessee teachers in the ratings database. We estimate models for two dependent variables: (1) a binary indicator for whether a survey

[14] Koedel et al. (2017) found some evidence of score rounding in the Tennessee system in the first year, but there is no evidence of the type of bunching one would expect from score rounding in our data from the second year.

was submitted at all, and (2) a binary indicator for whether a survey was submitted *and* at least one professional improvement question was answered.[15] The latter condition is the requirement for inclusion in our analytic sample.

Table 5 displays the estimated effects of treatment on these two outcomes, both by stacking the discontinuities and for each discontinuity separately. The estimates are substantively small and statistically insignificant for both outcomes, overall and at each threshold. We conclude that survey participation is not caused by the rating treatments, at least subject to the local interpretation of the RD estimates, which is most relevant for informing the credibility of the results to follow.

INSERT TABLE 5 HERE

5. Results

5.1 Graphical evidence

Figure 3 shows our primary results graphically for each professional improvement question. For ease of presentation in the figures, we show results from linear RD models where we code survey responses (a) as either positive or negative as in Table 2, where a positive response indicates more intense or responsive professional improvement, and (b) numerically – e.g., for a question with four possible answers, we use a 1-4 scale where the highest value is assigned to the answer that conveys the most intense or responsive behavior as indicated in Appendix A. In addition to the regression lines and 95-percent confidence intervals, we also report average values for teachers within 5-point bins of the underlying evaluation scores and center the stacked results around the threshold value for the

[15] This approach follows that of McCrary and Royer (2011), who encounter a related problem in their investigation of the effects of female education on fertility and infant health. We exclude teachers outside of the bandwidth, assessed under COACH, and with missing basic covariates from these regressions.

higher rating (similarly to Figure 1). The graphs provide no visual indication of significant differences in teachers' self-reported professional improvement activities at the rating thresholds.

INSERT FIGURE 3 HERE

5.2. Regression Results

Table 6 presents regression results that correspond to the results in Figure 3, but using our preferred ordered logit specification. The table reports ITT effects of higher ratings on professional improvement outcomes for all teachers, and for teachers working in TEAM-only school districts (who comprise 84 percent of our full analytic sample per Table 1). Recall that the discontinuities in ratings are nearly sharp and as such the ITT effects will be only slightly attenuated relative to treatment effects.

The ordered-logit estimates are presented as odds ratios. A value above 1.0 indicates that a higher performance rating causes an increase in the likelihood of a higher-ordered answer (i.e., more positive) to a particular survey question, and a value of less than 1.0 indicates the opposite. For example, taking the coefficient for question 4 from the full analytic sample at face value would imply that a higher rating causes a 1.013 times increase in the odds of a higher-ordered answer at the discontinuity. Standard errors are reported in parentheses and statistical significance is assessed relative to a value of 1.0, which would imply no difference between the groups above and below the threshold.[16]

None of the estimates in Table 6 are statistically significant at the 5 percent level for any question in any model using either the full or TEAM-only analytic samples. Moreover, the point estimates are inconsistent in sign and imply small effects nominally throughout. On the whole, we interpret the results as showing that the assignment of a higher rating itself does not impact teachers' self-directed professional improvement behaviors.

[16] The question-by-question sample sizes in Table 6 are slightly smaller than the sample sizes reported in columns 4 and 5 of Table 1 because all teachers who answered at least one survey question did not answer every question.

INSERT TABLE 6 HERE

5.3. Robustness and Sensitivity

We also estimate linear RD models where we code the dependent variable for each question using the positive/negative and numerical coding schemes described above and in Appendix A. These regression models align more closely with the graphs in Figure 3. The results are reported in Appendix Table C.1 and mirror the findings in Table 6. Specifically, across the two coding schemes, two samples (full and TEAM-only), and four survey questions, no coefficient is statistically different from zero at the 5-percent level (just one coefficient is statistically significant at the 10-percent level) and the point estimates are small and directionally inconsistent. We conclude that our null findings are robust to these adjustments to measurement and modeling structure.

We use the models of positive/negative coded responses to illustrate the bounds of our estimates. Our point estimates and standard errors in these models are small enough to rule out effects of higher ratings in either direction larger than 0.03-0.05 units across questions with 95 percent confidence. These are relatively small compared to the positive-response shares for the questions, which range from roughly 0.30 to 0.70.[17]

Appendix Table C.2 shows results from models that use the alternative bandwidths from Table 5. Although we lose precision as the bandwidth narrows, there is no indication that our findings are sensitive to bandwidth variation, further confirming the robustness of our null results.[18]

We also test for effect heterogeneity across rating thresholds in Appendix Table C.3. For this analysis we return to our preferred ordered-logit specification but run the model separately for each

[17] We reach a qualitatively similar conclusion with the numerically-coded models, where our point estimates and standard errors rule out effect sizes larger than 0.25 for question 1, and 0.05-0.07 for questions 2-4. The sample average value of the numerically-coded answers to question 1 is 3.6, and for questions 2-4 (with fewer choices, per Appendix A) the sample averages range from 2.2-2.8.

[18] In results omitted for brevity we further verify that our findings are robust to re-weighting the observations so that teachers with scores closer to the discontinuity thresholds receive higher weights than those farther away within the main bandwidth.

rating threshold. These tests are motivated by evidence from Koedel et al. (2017) and Papay et al. (2016) that rating-threshold effects can vary by the level of ratings distinguished, perhaps because of psychological factors associated with the performance labels assigned to different ratings. Effect heterogeneity in our application seems less likely given that our overall results are null – i.e., for a higher rating to cause a positive effect at some thresholds, it would need to cause a negative effect at others in order for the total effect to be zero. Consistent with this intuition, Appendix Table C.3 shows little evidence of effect heterogeneity. While four coefficients are significant at the 10 percent level with the proliferation of tests, it remains the case that no estimate at any threshold is significant at the 5-percent level or better. Moreover, unlike in Koedel et al. (2017) and Papay et al. (2016), there is not a consistent pattern in the estimates to suggest that teachers are systematically responding to signals related to specific rating thresholds.

Finally, in Appendix D we examine the robustness of our findings to dropping teachers from the sample who report not seeing their ratings. Although all teachers were provided access to their ratings online, approximately 7.5 percent of survey respondents indicate that they did not see them. We do not do anything differently for these teachers in the main analysis, but if these teachers truly did not receive their ratings then for the purpose of our study they can be viewed as not receiving treatment and their inclusion in the analytic sample will attenuate the results.[19] Our estimates in Appendix D remain substantively similar, further reinforcing our null findings.

6. Discussion and Interpretation

Perhaps the most important contextual factor for interpreting our results is the policy environment in which teachers' ratings are assigned. Some formal statewide policies were in place in

[19] This is a simplification in the following sense: seeing the rating is necessary for treatment if the treatment is viewed purely as informational to the individual teacher, but we cannot rule out that discrete ratings affect behaviors even if teachers never see the ratings. For example, school principals may respond to the ratings, which they also observe, and this could affect how they interact with teachers, and in turn affect teachers' behaviors related to professional improvement.

Tennessee to encourage teachers to value high ratings at the time of our study. Specifically, as noted above, teachers working in disadvantaged schools under the Tennessee Achievement School District program who earned a rating of three or higher were eligible for salary increases and/or promotions not available to teachers with lower ratings. Also, non-tenured teachers were required to receive ratings of four or five during the last two years of the pre-tenure probationary period to earn tenure. However, neither of these conditions apply to the vast majority of teachers in our sample. Thus, a potential explanation for our null findings is that insufficient incentives were in place to encourage teachers to respond to their ratings.

The lack of a widespread, formal incentive structure in Tennessee certainly distinguishes it from the high-profile IMPACT program in Washington DC studied by Dee and Wyckoff (2015). These authors find substantial changes in teachers' retention and improvement outcomes around rating thresholds using a similar research design, but the thresholds where they find large effects are associated with consequential labor outcomes such as large bonuses at the high end and threat of dismissal at the low end. While this obvious difference between Washington DC and Tennessee stands out, we offer the suggestion that our null results are owing to a lack of a formal statewide incentive structure cautiously. The lack of statewide incentives does not preclude local incentives, formal and informal, and these incentives can also be potent (Dixit, 2002; Prendergast, 1999). We are not aware of any district in Tennessee with rewards or sanctions nearly on par with Washington DC (to the best of our knowledge, the incentive structure in Washington DC is uniquely strong), but recent evidence from Bates (2016) and Cullen, Koedel and Parsons (2016) suggests that even with a weaker formal structure, teachers are able to leverage more informative, public (i.e., available to prospective employers) evaluations to facilitate moves to more desirable positions. Moreover, it is easy to imagine other ways for highly rated teachers to use their newly-available signals to achieve professional gains

(e.g., access to desirable teaching and extra-curricular assignments). We have no credible evidence to dismiss the possibility that these types of incentives matter.

It is also notable that at least two other studies of which we are aware identify rating effects in similar or less-incentivized environments. The most relevant is Koedel et al. (2017), who study the effects of ratings on teacher job satisfaction, also in Tennessee. In the same policy environment, these authors find statistically significant and substantively meaningful rating effects on teachers' perceptions of work. Specifically, a higher rating in the Tennessee system causes a roughly 0.10 standard deviation increase in job satisfaction, on average, as measured by an index constructed from 10 questions on the teacher survey. These authors also find effect heterogeneity across thresholds. A clear takeaway from our study is that the job-satisfaction effects of ratings documented by Koedel et al. (2017) are not accompanied by corresponding effects on teachers' self-directed improvement behaviors.

Another relevant study is Papay et al. (2016), who show that the discrete labels students receive summarizing their performance on standardized tests influence human capital investments. While the focus on students instead of teachers weakens the connection between their work and ours, their findings are of interest because of the low-stakes environment they study. Papay et al. (2016) offer psychological effects of the labels as one potential explanation for their results. The effect heterogeneity they identify at different rating thresholds is consistent with this explanation but their research design (like ours) is not well-suited to identify mechanisms. Together, the studies by Koedel et al. (2017) and Papay et al. (2016) make clear that formal, structured incentives are not necessary for rating treatments to be influential. While we could only speculate as to what our results would look like if the Tennessee system included a stronger and more systematic formal incentive structure, our null findings rule out some pathways by which teacher ratings could affect the behaviors we study.

Our finding that summative performance information alone is not sufficient to spur a response along the margin of self-directed improvement is consistent with the argument that workers require formative, actionable feedback about performance in order to improve (Loeb, 2013).[20] Building on this point, recall that the teachers who we leverage for identification are similar in other ways except the final rating, including the nature of the formative feedback received from classroom observers during the evaluation process.[21] If teachers respond to formative feedback from their classroom observers, but not the summative rating, this would produce null results in our regression-discontinuity models. That said, recall that teachers do not know how close they are to the threshold scores and the classroom observation is just one (noisy) component of the final rating for an individual teacher. Even conditional on similar formative feedback, the lack of responsiveness to differences in the summative rating is notable.

We also acknowledge that our regression-discontinuity estimates are limited in the standard way given their local interpretation. Our null results cannot rule out effects of summative ratings more broadly, such as between teachers further from the thresholds or even spanning multiple thresholds. Unfortunately we lack a credible causal design to push further on this dimension, but again note that similarly-designed studies have found rating effects elsewhere, and in Tennessee, with the same local-identification restriction.

Finally, we briefly turn to the issue that we use teachers' self-reported professional improvement behaviors as outcomes. It is not clear how accurate teachers' self-reported activities are relative to true behaviors, which are unobserved (and would be difficult to measure well objectively).

[20] Moreover, it suggests that the gains in teacher improvement observed by Taylor and Tyler (2012) in Cincinnati are unlikely to be the result of the revelation of summative performance information.

[21] In results omitted for brevity we have verified that the qualitative feedback that teachers receive from their observers, as measured by word counts in reinforcement (positive) and refinement (negative) areas, is not discontinuous at the rating thresholds used in our regression-discontinuity models. This is as expected under the RD assumptions. The analysis of qualitative feedback data was performed for TEAM teachers only because we do not have these data for teachers evaluated under the other evaluation models.

We make two points on this. First, one could argue that teachers' perceptions of their self-improvement activities are an independently important outcome, albeit of second-order importance. For example, even if we assume that perceptions do not translate to actual behaviors; at the least, a change in perceptions would indicate a frame-of-mind effect. Second, and focusing on the case where "true" professional improvement behavior is the desired outcome, a measurement threat is that teachers may be concerned that their individual surveys will not remain anonymous, and thus answer the questions how they think they should, regardless of actual behavior. Given the nature of the questions and their focus on responsiveness to evaluation feedback, a particular concern is that low-rated teachers might overstate their responsiveness because they feel that this is the "right" way to respond. Our null results suggest that bias of this nature is not a concern.[22,23]

7. Conclusion

Teachers receive high and compressed ratings in annual reviews that do little to differentiate them by performance (U.S. GAO, 2013; Kraft and Gilmour, 2017; Weisberg et al., 2009). Given that mounting empirical evidence on the distribution of teacher quality contradicts the compressed distribution of teachers' evaluation outcomes, there has been a push to develop more rigorous and informative review processes. More informative evaluations can be used to improve workforce quality in many ways (Koedel, Mihaly and Rockoff, 2015), including by providing new information to encourage and guide teachers' professional improvement activities.

The purpose of this study is to understand if the assignment of more differentiated ratings leads to differences in teachers' self-directed professional improvement, in terms of both total effort

[22] Unless it is offset by some sort of positive reporting bias among higher rated teachers at the thresholds. To be clear, this would not be positive improvement activity among higher rated teachers, in which case it would be part of a real effect, but positive bias in reporting relative to activity – while we have no means to formally rule out this possibility, it seems unlikely.

[23] There is also the technical concern that measurement error in the dependent variable will bias the estimates from our nonlinear models. However, the substantive importance of this concern is ruled out by the similarity of results from the linear RD models shown in Appendix C, in which dependent-variable measurement error does not cause bias.

and responsiveness to evaluation feedback. Our setting is Tennessee, where the depth of evaluations and dispersion of ratings are far from sectoral norms. Using regression discontinuity models that compare otherwise similar teachers who receive different discrete signals about performance from the system, we find no evidence of rating effects along these dimensions. The null results are robust to a variety of measurement and modeling adjustments. These findings indicate that the provision of differential summative performance information to teachers, in and of itself, does not alter behavior.

We conclude by again briefly addressing our focus on information conveyed by the summative rating. An alternative channel for conveying information in emerging teacher evaluation systems is via formative feedback from deeper classroom observations, which typically include observer-teacher conferences and written feedback. In results we omit for brevity, we find that teachers' self-reported professional improvement activities are more strongly associated with their scores on the classroom-observation component than the overall score. This descriptive result is consistent with the hypothesis that formative feedback is important, but we lack a causal identification strategy to pursue the line of inquiry further. We suggest formative assessment as a dimension of employee evaluation worthy of study in future research.

References

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

Bates, Michael. 2016. Public and Private Learning in the Market for Teachers: Evidence from the Adoption of Value-Added Measures. Unpublished manuscript.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633-79.

Condie, Scott, Lars Lefgren and David Sims. 2014. Teacher Heterogeneity, Value-Added and Education Policy. *Economics of Education Review* 40(1), 76-92.

Cullen, Julie Berry, Brian A. Jacob, and Steven D. Levitt. 2005. The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools. *Journal of Public Economics* 89(5/6), 729-760.

Cullen, Julie Berry, Cory Koedel and Eric Parsons. 2016. The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. NBER Working Paper No. 22805. National Bureau of Economic Research: Cambridge, MA.

Dee, Thomas and James Wyckoff. 2015. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34(2), 267-297.

Dixit, Avinash. 2002. Incentives and Organizations in the Public Sector. *Journal of Human Resources*, 37, 696-727.

Fitzpatrick, Maria D., David Grissmer and Sarah Hastedt. 2011. What a Difference a Day Makes: Estimating Daily Learning Gains During Kindergarten and First Grade Using a Natural Experiment. *Economics of Education Review* 30(2), 269-279.

Goldhaber, Dan, James Cowan and Joe Walch. 2013. Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects. *Economics of Education Review* 36(1), 216-228.

Grissom, Jason A. and Susanna Loeb. 2017. Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments. Education Finance and Policy 12(3), 369-395.

Hahn, Jinyong, Petra Todd and Wilbert Van der Klaauw. 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica* 69(1), 201-209.

Hanushek, Eric A. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3), 266-479.

Imbens, Guido W. and Thomas Lemieux. 2008. Regression discontinuity designs: A Guide to Practice. *Journal of Econometrics* 142(2), 615–635.

Jepsen, Christopher, Peter Mueser and Kenneth Troske. 2016. Labor-Market Returns to the GED Using Regression Discontinuity Analysis. *Journal of Political Economy* 124(3), 621-649.

Koedel, Cory, Jiaxi Li, Matthew Springer and Li Tan. 2017. The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. *American Education Research Journal* 54(2), 241-278.

Koedel, Cory, Kata Mihaly and Jonah E. Rockoff. 2015. Value-Added Modeling: A Review. *Economics of Education Review* 47, 180-195.

Kraft, Matthew A. and Gilmour, Allison F. 2017. Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher* 46(5), 234-249.

Lee, David S. and David Card. 2008. Regression Discontinuity Inference with Specification Error. *Journal of Econometrics* 142(2), 655-674.

Loeb, Susanna. 2013. How Can Value-Added Measures be Used for Teacher Improvement? Policy Report. Carnegie Foundation for the Advancement of Teaching.

McCrary, Justin and Heather Royer. 2011. The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth. *American Economic Review* 101(1), 158-195.

Murphy, Kevin R. and Jeannette Cleveland. 1991. *Performance Appraisal: An Organizational Perspective*. Boston, MA: Allyn and Bacon.

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

Papay, John P., Richard J. Murnane and John B. Willet. 2016. The Impact of Test-Score Labels on Human-Capital Investment Decisions. *Journal of Human Resources,* 51(2), 357-388.

Papay, John P., Eric S. Talyor, John H. Tyler and Mary Laski. 2016. Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data. NBER Working Paper No. 21986. National Bureau of Economic Research: Cambridge, MA.

Prendergast, Canice. 1999. The Provision of Incentives in Firms. *Journal of Economic Literature* 37, 7-63.

Springer, Matthew G., Swain, Walker A., and Rodriguez, Luis A. 2016. Effective Teacher Retention Bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis*, 38(2), 199-221.

Staiger, Douglas O. and Jonah E. Rockoff. 2010. Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.

Taylor, Eric S. and John H. Tyler. 2012. The Effect of Evaluation on Teacher Performance. *American Economic Review* 102(7), 3628-3651.

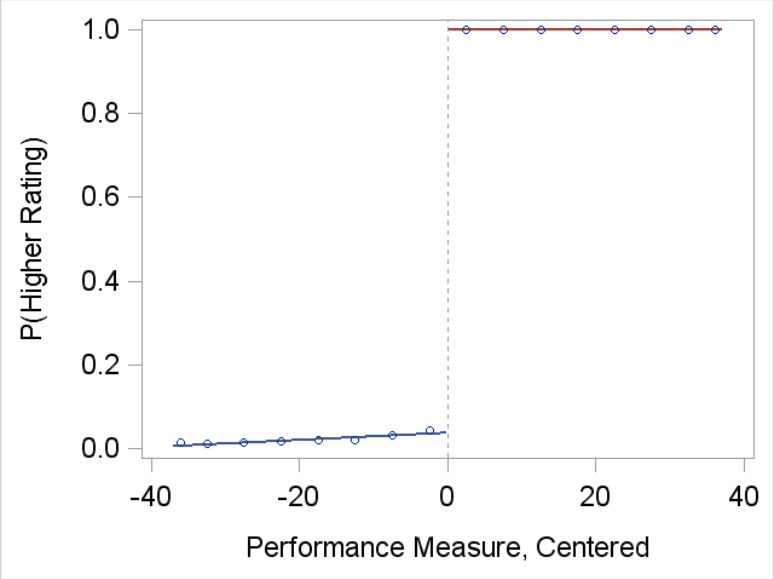
United States Government Accountability Office. 2013. Federal Workforce: Distribution of Performance Ratings Across the Federal Government, 2013. U.S. Government Accountability Office: Washington, DC.

Van Der Klaauw, Wilbert. 2002. Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach. *International Economic Review* 43 (4), 1249–1287.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern and David Keeling. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New York: The New Teacher Project.

Winters, Marcus A. and Joshua M. Cowen. 2013. Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management* 32(3), 634-654.

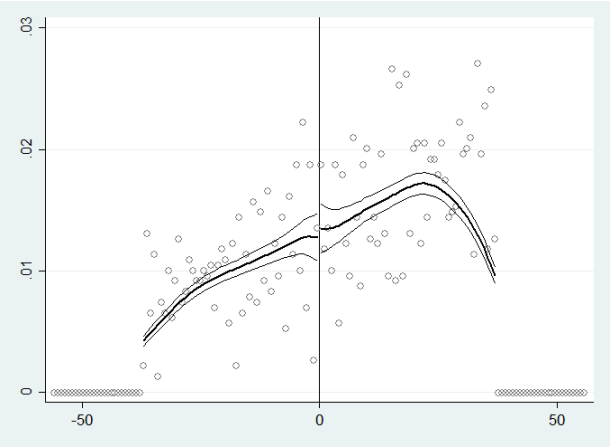
Figure 1. Illustration of the Stacked Rating Discontinuities at the Cut Scores between Levels 2/3, 3/4, and 4/5.



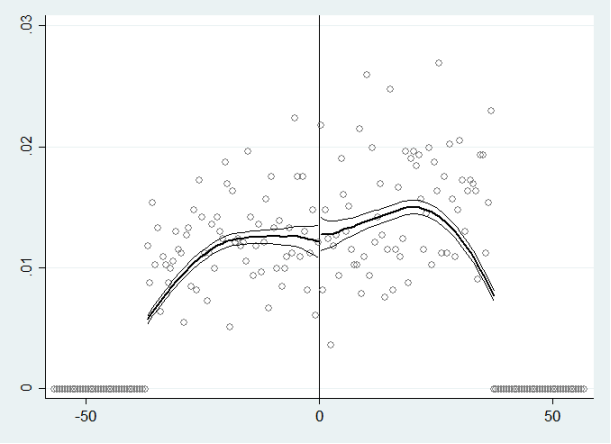
Notes: Teachers are aggregated into 5-point bins based on their underlying performance scores. Thus, each point on the graphs denotes the probability of being assigned to the higher rating for teachers with underlying performance at that point or less than 5 points above it.

Figure 2. Density Test Results at Each Rating Threshold.

Cutoff 2/3



Cut off 3/4



Cut off 4/5

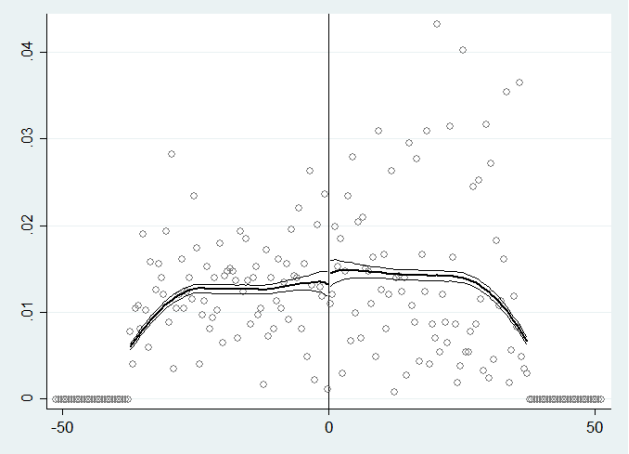
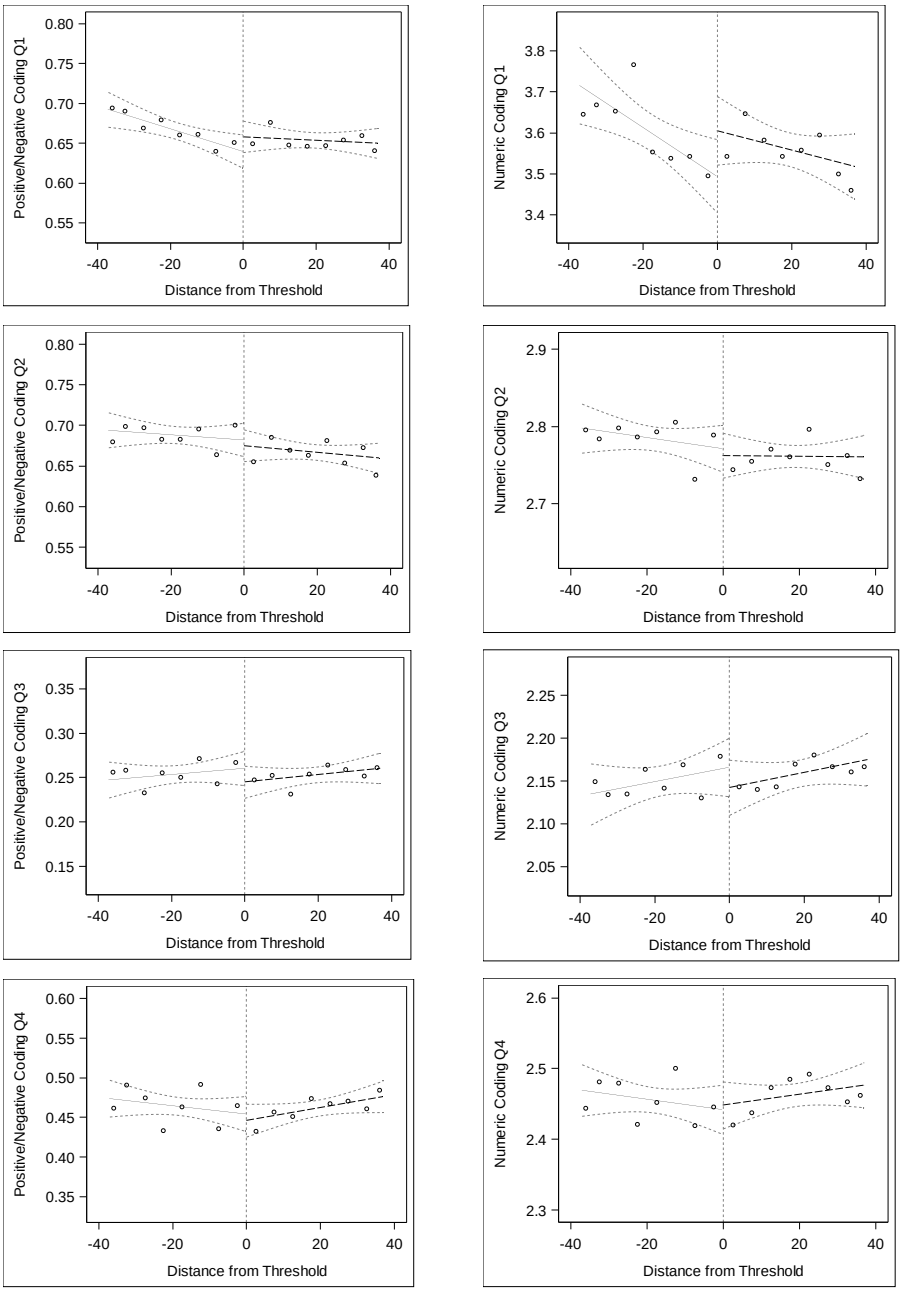


Figure 3. Illustration of Linear Regression-Discontinuity Results for Each Professional Improvement Question using the Positive/Negative (on the left) and Numerical (on the right) Coding Strategies.



Notes: Regression lines (solid lines) and 95% confidence intervals (dashed lines) are reported along with binned outcome data.

Table 1. Descriptive Statistics, Various Samples.

	All Tennessee Teachers	Survey Respondents	Responded to at Least One Professional Improv Question	RD Sample, Main Bandwidth	RD Sample, Main Bandwidth, TEAM Only
Female teacher	0.80	0.82	0.82	0.82	0.82
Black teacher	0.12	0.11	0.10	0.10	0.04
White teacher	0.87	0.89	0.90	0.90	0.96
Other race	0.01	0.00	0.00	0.00	0.00
Bachelor degree	0.41	0.38	0.41	0.42	0.43
Education specialist	0.07	0.09	0.08	0.08	0.08
Master degree	0.42	0.44	0.43	0.42	0.43
Other education	0.10	0.10	0.08	0.08	0.06
Teacher experience	13.36	14.13	13.90	13.72	13.63
Group 1 (with individual growth score)	0.45	0.47	0.51	0.48	0.50
AFET evaluation model	0.00	0.00	0.00	0.00	0.00
COACH evaluation model	0.06	0.07	0.07	0.00	0.00
TEAM evaluation model	0.77	0.76	0.77	0.84	1.00
TEM evaluation model	0.14	0.14	0.13	0.13	0.00
TIGER evaluation model	0.02	0.03	0.03	0.03	0.00
Level 1 (Sig. Below Expectations)	0.02	0.01	0.01	0.00	0.00
Level 2 (Below Expectations)	0.09	0.08	0.09	0.07	0.07
Level 3 (At Expectations)	0.22	0.22	0.23	0.27	0.27
Level 4 (Above Expectations)	0.33	0.33	0.33	0.41	0.41
Level 5 (Sig. Above Expectations)	0.33	0.35	0.34	0.25	0.25
School percentage minority students	30.80	28.86	28.17	27.94	19.65
School percentage female	48.50	48.46	48.48	48.45	48.37
School percentage FRL eligible	59.79	61.07	60.81	61.23	59.15
N	68171	25952	22099	16823	14053

Notes: Group-1 teachers are those with an individual growth score. The main analytic sample is shown in column 4. Some teachers are missing some of the information reported in this table. Teachers with missing information are omitted from the calculations on an item-by-item basis.

Table 2. Proportion of Positive Responses to Each Survey Question.

	Below Expectation	At Expectation	Above Expectation	Significantly Above Expectation
	(2)	(3)	(4)	(5)
Question 1: Spent time improving instructional skills	0.72	0.70	0.66	0.67
Question 2: Changed teaching based on evaluation results	0.81	0.77	0.74	0.69
Question 3: Changed non-teaching duties based on evaluation results	0.33	0.29	0.28	0.27
Question 4: Evaluation feedback influenced prof. development activities	0.54	0.52	0.52	0.51

Notes: The table reports the share of positive responses to each survey question conditional on a non-missing response. Non-responders are excluded from the ratios on a question-by-question basis. Appendix B presents analogous ratios where non-responders are included as "non-positive" responses in the denominator of each ratio.

Table 3. Regression Discontinuity Estimates of the "Effects" of a Higher Rating on Teacher Characteristics that Should Not be Affected, for the Purpose of Validating the Research Design.

	Bandwidth				
	37	30	20	10	5
Female teacher	0.009	0.005	0.012	0.003	-0.006
	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Black teacher	-0.004	-0.006	-0.001	0.016	0.035
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
White teacher	0.006	0.008	0.001	-0.02	-0.037
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Bachelor degree	-0.003	-0.005	-0.005	0.001	-0.031
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
Master degree	0.009	0.009	0.005	0.006	0.03
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
Education specialist	0.004	0.009	0.003	-0.006	0.022
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Teacher experience	0.284	0.447	0.635	0.427	-0.299
	(0.30)	(0.33)	(0.40)	(0.57)	(0.81)
Group 1 (with individual growth)	0.031	0.032	0.030	0.065	0.127
	(0.015)**	(0.017)*	(0.02)	(0.029)**	(0.041)***
School percentage minority	-2.014	-1.790	-0.913	1.970	2.987
	(0.94)**	(1.024)*	(1.20)	(1.76)	(2.50)
School percentage female	-0.639	-0.615	-0.201	0.639	2.455
	(0.67)	(0.74)	(0.87)	(1.28)	(1.86)
School percentage FRL eligible	0.023	0.051	0.082	0.174	0.15
	(0.10)	(0.11)	(0.13)	(0.20)	(0.31)
Randomized-inference *p* value	0.44	1.00	1.00	0.60	0.63
N	16823	13746	9087	4633	2339

***/**/* denotes significance level 0.01/0.05/0.10

Notes: Models are specified as linear probability models. Each estimate in each cell comes from a separate regression. Standard errors are clustered at the school level and reported in parentheses. The p-value at the bottom of each row indicates the likelihood of obtaining the observed number of statistically significant coefficients by chance at the 5-percent level based on 3,000 bootstrap repetitions. See Table 1 for details about the variables listed.

Table 4. Density Tests at Each Threshold Using McCrary's Method.

	Cutoff 2/3	Cutoff 3/4	Cutoff 4/5
Density Test Coefficient	0.053	0.047	0.090
	(0.114)	(0.081)	(0.080)
N	3068	5988	7767

***/**/* denotes significance level 0.01/0.05/0.10

Notes: Models are specified as linear probability models. Standard errors clustered at school level are reported in parentheses.

Table 5. Regression Discontinuity Estimates of the Effect of a Higher Rating on Survey Response Behaviors, Overall and for Each Threshold Separately.

	Stacked Model	Cutoff 2/3	Cutoff 3/4	Cutoff 4/5
Any survey submission	0.003	-0.023	0.001	0.008
	(0.009)	(0.021)	(0.016)	(0.013)
Survey submission and answered at least one professional improvement question	-0.001	-0.009	-0.017	0.008
	(0.009)	(0.021)	(0.015)	(0.012)
N	46857	8651	16442	21764

***/**/* denotes significance level 0.01/0.05/0.10

Notes: Models are specified as linear probability models. Standard errors clustered at school level are reported in parentheses. The sample used for these regressions includes all individuals with complete data, with the additional restriction that they are within 37 points of the discontinuity cutoffs at rating levels 2/3, 3/4, and 4/5.

Table 6. Effects of a Higher Ratings on Self-Reported Professional Improvement Activities.

	Ordered Logit Estimates		
	Higher Rating	*Higher Rating, TEAM Only*	*N*
Question 1: Spent time improving instructional skills	1.093 (0.061)	1.093 (0.066)	16724/13962
Question 2: Changed teaching based on evaluation results	0.995 (0.068)	0.942 (0.069)	15162/12725
Question 3: Changed non-teaching duties based on evaluation results	0.945 (0.060)	0.954 (0.065)	15026/12611
Question 4: Evaluation feedback influenced prof. development activities	1.013 (0.062)	1.044 (0.069)	15021/12600

***/**/* denotes significance level 0.01/0.05/0.10

Notes: Each estimate in each cell comes from a separate ordered logistic regression. The values in each cell are odds ratios. Standard errors are reported in parentheses and clustered at the school level. The question-by-question sample sizes differ slightly from what we report in columns 4 and 5 of Table 1 because all teachers who answered at least one survey question did not answer every question.

Appendix A
Coding and Other Details for the Professional Improvement Questions

Below we show each question listed in Table 2 in full form as it was presented to teachers.[24] The options in italics are the ones that we coded to indicate a positive response. Non-italicized options are coded to indicate a negative response. The coding choices are based in part on the underlying distribution of answers to each question. For the numerical coding scheme that we use in some parts of our analysis, answers are coded numerically in order where the highest-valued answer is the one that conveys the strongest professional-improvement response – e.g., for question 2, the answer "strongly agree" is assigned a value of 4, the answer "agree" a value of 3, etc. Of course, these are also the values assigned to these answers in the ordered logits. More information is available from the authors upon request.

Q1. Approximately how much time have you invested so far during the 2013-2014 school year in efforts to improve your instructional practices?

a. 0 hours.
b. 1-10 hours.
c. 11-20 hours.
d. *21-40 hours.*
e. *41-60 hours.*
f. *61-80 hours.*
g. *81-100 hours.*
h. *More than 100 hours.*

Q2. I made changes to my teaching based on my evaluation results.

1. Strongly Disagree.
2. Disagree.
3. *Agree.*
4. *Strongly Agree.*

[24] With the exception of Question 1, which as shown here is a combination of two questions on the survey. The first question asks whether the teacher worked on improving her instructional practice during the 2013-2014 school year at all. If yes, there is a follow-up question about the time investment. We combine the questions here for presentational brevity, using choice "a" to represent the first part of the two-part question.

Q3. I changed how I perform non-teaching duties based on my evaluation results.[25]

1. Strongly Disagree.
2. Disagree.
3. *Agree.*
4. *Strongly Agree.*

Q4. Feedback from my teacher evaluation influences the professional development activities in which I participate.

1. Strongly Disagree.
2. Disagree.
3. *Agree.*
4. *Strongly Agree.*

[25] There is no widely acknowledged definition of non-teaching duties. The 2015 Tennessee Code Annotated lists several duties not directly related to teaching which include duties such as keeping accurate attendance records and supervising educational assistants when they are working with pupils. Non-teaching duties may also include parent and teacher meetings, curriculum development, etc.

Appendix B
Unconditional Proportion of Positive Responses to Each Question

Table B.1. Proportion of Positive Responses to Each Question. Non-Responses Included in the Denominator.

	Below Expectations (2)	Meets Expectations (3)	Above Expectations (4)	Sig. Above Expectations (5)
Question 1: Spent time improving instructional skills	0.63	0.60	0.55	0.55
Question 2: Changed teaching based on evaluation results	0.65	0.61	0.57	0.52
Question 3: Changed non-teaching duties based on evaluation results	0.26	0.23	0.22	0.20
Question 4: Evaluation feedback influenced prof. development activities	0.43	0.41	0.40	0.38

Notes: This table is analogous to Table 2 in the text, but does not condition on non-missing responses (that is, missing responses are included in the denominator in each cell as "non-positive").

Appendix C
Robustness Tests

Table C.1. Effects of a Higher Ratings on Self-Reported Professional Improvement Activities from Linear RD Models.

	Panel A: Positive/Negative Question Coding		
	Higher Rating	*Higher Rating, TEAM Only*	*N*
Question 1: Spent time improving instructional skills	0.023 (0.015)	0.020 (0.016)	16724/13962
Question 2: Changed teaching based on evaluation results	-0.002 (0.014)	-0.009 (0.016)	15162/12725
Question 3: Changed non-teaching duties based on evaluation results	-0.014 (0.015)	-0.012 (0.016)	15026/12611
Question 4: Evaluation feedback influenced prof. development activities	-0.003 (0.016)	0.007 (0.017)	15021/12600
	Panel B: Numerical Question Coding		
	Higher Rating	*Higher Rating, TEAM Only*	*N*
Question 1: Spent time improving instructional skills	0.109 (0.063)*	0.106 (0.068)	16724/13962
Question 2: Changed teaching based on evaluation results	-0.005 (0.021)	-0.018 (0.023)	15162/12725
Question 3: Changed non-teaching duties based on evaluation results	-0.021 (0.024)	-0.016 (0.025)	15026/12611
Question 4: Evaluation feedback influenced prof. development activities	0.008 (0.024)	0.020 (0.026)	15021/12600

***/**/* denotes significance level 0.01/0.05/0.10.

Notes: Panel A and B are specified as linear models where the dependent variable is either a 0/1 indicator for a positive response or a numerically-coded value of the strength of the positive response based on its ordering in Appendix A. Each estimate in each cell comes from a separate regression. The values in each cell are regression coefficients. Standard errors are reported in parentheses and clustered at the school level.

Table C.2. Effects of a Higher Ratings on Self-Reported Professional Improvement Activities Using Different Bandwidths.

	Higher Rating					*Higher Rating, TEAM Only*				
	37 (Main Results)	30	20	10	5	37 (Main Results)	30	20	10	5
Question 1	1.093	1.081	1.054	0.973	1.033	1.093	1.078	1.068	1.027	1.075
	(0.061)	(0.066)	(0.078)	(0.104)	(0.158)	(0.066)	(0.072)	(0.087)	(0.120)	(0.177)
Question 2	0.995	0.987	0.970	0.881	1.394	0.942	0.929	0.954	0.838	1.383
	(0.068)	(0.074)	(0.090)	(0.115)	(0.265)*	(0.069)	(0.076)	(0.096)	(0.117)	(0.287)
Question 3	0.945	0.942	0.922	0.959	1.007	0.954	0.971	0.943	1.018	1.080
	(0.060)	(0.065)	(0.076)	(0.109)	(0.174)	(0.065)	(0.073)	(0.084)	(0.128)	(0.204)
Question 4	1.013	0.985	0.970	1.023	1.299	1.044	1.006	0.982	1.052	1.356
	(0.062)	(0.066)	(0.079)	(0.123)	(0.232)	(0.069)	(0.074)	(0.088)	(0.136)	(0.259)

***/**/* denotes significance level 0.01/0.05/0.10.

Notes: Each estimate in each cell comes from a separate ordered logistic regression. The values in each cell are odds ratios. Standard errors are reported in parentheses and clustered at the school level. The questions are in the same order as in Table C.1.

Table C.3. Effects of a Higher Ratings on Self-Reported Professional Improvement Activities at Different Thresholds.

	Higher Rating				Higher Rating, TEAM Only			
	Stacked	Separate			Stacked	Separate		
	(Main Results)	2/3	3/4	4/5	(Main Results)	2/3	3/4	4/5
Question 1	1.093	0.912	1.125	1.146	1.093	0.939	1.099	1.163
	(0.061)	(0.127)	(0.101)	(0.094)*	(0.066)	(0.141)	(0.107)	(0.105)*
Question 2	0.995	0.951	0.869	1.107	0.942	0.942	0.840	1.037
	(0.068)	(0.158)	(0.102)	(0.109)	(0.069)	(0.168)	(0.107)	(0.113)
Question 3	0.945	0.767	0.976	0.999	0.954	0.811	0.977	1.004
	(0.060)	(0.112)*	(0.105)	(0.090)	(0.065)	(0.127)	(0.114)	(0.100)
Question 4	1.013	0.992	1.153	0.918	1.044	1.016	1.224	0.931
	(0.062)	(0.143)	(0.122)	(0.081)	(0.069)	(0.152)	(0.139)*	(0.090)

***/**/* denotes significance level 0.01/0.05/0.10.

Notes: Each estimate in each cell comes from a separate regression. The values in each cell are odds ratios and column 1 repeats our main results from Table 6 for comparison. Standard errors are reported in parentheses and clustered at the school level. The questions are in the same order as Table C.1.

Appendix D
Robustness to Excluding Teachers Who Report Not Receiving a Rating

Column 1 of Table D.1 reports ordered-logit results analogous to Table 6 after restricting the sample to exclude teachers who indicated that they did not receive their evaluation ratings from the system. Columns 2 and 3 report results analogous to those presented in Appendix Table C.1. As reported in the text, approximately 7.5 percent of teachers indicated that they did not receive their ratings. Table D.1 shows that even if we exclude these teachers, our null results are retained.

Table D.1. Effects of a Higher Rating on Professional Improvement Outcomes for Teachers Who Report Receiving their Ratings.

	Panel A: Numerically Coding, Ordered Logit	Panel B: Positive/Negative Coding, Linear Model	Panel C: Numerical Coding, Linear Model	*N*
Question 1: Spent time improving instructional skills	1.099 (0.064)	0.021 (0.015)	0.120 (0.065)*	15187
Question 2: Changed teaching based on evaluation results	0.957 (0.070)	-0.005 (0.015)	-0.017 (0.022)	13672
Question 3: Changed non-teaching duties based on evaluation results	0.947 (0.064)	-0.012 (0.016)	-0.020 (0.025)	13546
Question 4: Evaluation feedback influenced prof. development activities	1.040 (0.066)	0.004 (0.016)	0.018 (0.025)	13535

***/**/* denotes significance level 0.01/0.05/0.10

Notes: Each estimate in each cell is from a separate regression. The values are odds ratios (column 1) and regression coefficients (columns 2/3). Standard errors are reported in parentheses and clustered at the school level.